INSTRUCTOR'S SOLUTIONS MANUAL

GEX PUBLISHING SERVICES

FUNDAMENTALS OF STATISTICS: INFORMED DECISIONS USING DATA

FIFTH EDITION

Michael Sullivan, III
Joliet Junior College

Pearson

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 330 Hudson Street, NY NY 10013

Pearson

ISBN-13: 978-0-13-450993-8
ISBN-10: 0-13-450993-5

Table of Contents

Preface

This solutions manual accompanies *Fundamentals of Statistics: Informed Decisions Using Data, 5/e* by Michael Sullivan, III. The Instructor's Solutions Manual contains detailed solutions to all exercises in the text. The Student's Solutions Manual contains detailed solutions to all odd exercises in the text and all solutions to chapter reviews and tests. A concerted effort has been made to make this manual as user-friendly and error free as possible.

Chapter 1
Data Collection

Section 1.1

1. Statistics is the science of collecting, organizing, summarizing, and analyzing information in order to draw conclusions and answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

2. The population is the group to be studied as defined by the research objective. A sample is any subset of the population.

3. Individual

4. Descriptive; Inferential

5. Statistic; Parameter

6. Variables

7. 18% is a parameter because it describes a population (all of the governors).

8. 72% is a parameter because it describes a population (the entire class).

9. 32% is a statistic because it describes a sample (the high school students surveyed).

10. 9.6% is a statistic because it describes a sample (the youths surveyed).

11. 0.366 is a parameter because it describes a population (all of Ty Cobb's at-bats).

12. 43.92 hours is a parameter because it describes a population (all the men who have walked on the moon).

13. 23% is a statistic because it describes a sample (the 6076 adults studied).

14. 44% is a statistic because it describes a sample (the 100 adults interviewed).

15. Qualitative

16. Quantitative

17. Quantitative

18. Qualitative

19. Quantitative

20. Quantitative

21. Qualitative

22. Qualitative

23. Discrete

24. Continuous

25. Continuous

26. Discrete

27. Continuous

28. Continuous

29. Discrete

30. Continuous

31. Nominal

32. Ordinal

33. Ratio

34. Interval

35. Ordinal

36. Nominal

37. Ratio

38. Interval

39. The population consists of all teenagers 13 to 17 years old who live in the United States. The sample consists of the 1028 teenagers 13 to 17 years old who were contacted by the Gallup Organization.

40. The population consists of all bottles of Coca-Cola filled by that particular machine on October 15. The sample consists of the 50 bottles of Coca-Cola that were selected by the quality control manager.

41. The population consists of all of the soybean plants in this farmer's crop. The sample consists of the 100 soybean plants that were selected by the farmer.

42. The population consists of all households within the United States. The sample consists of the 50,000 households that are surveyed by the U.S. Census Bureau.

43. The population consists of all women 27 to 44 years of age with hypertension. The sample consists of the 7373 women 27 to 44 years of age with hypertension who were included in the study.

44. The population consists of all full-time students enrolled at this large community college. The sample consists of the 128 full-time students who were surveyed by the administration.

45. Individuals: Alabama, Colorado, Indiana, North Carolina, Wisconsin.
Variables: Minimum age for driver's license (unrestricted); mandatory belt use seating positions, maximum allowable speed limit (rural interstate) in 2011.
Data for minimum age for driver's license: 17, 17, 18, 16, 18;
Data for mandatory belt use seating positions: front, front, all, all, all;
Data for maximum allowable speed limit (rural interstate) 2011: 70, 75, 70, 70, 65 (mph.)
The variable *minimum age for driver's license* is continuous; the variable *mandatory belt use seating positions* is qualitative; the variable *maximum allowable speed limit (rural interstate) 2011* is continuous (although only discrete values are typically chosen for speed limits.)

46. Individuals: 3 Series, 5 Series, 6 Series, 7 Series, X3, Z4 Roadster
Variables: Body Style, Weight (lb), Number of Seats
Data for body style: Coupe, Sedan, Convertible, Sedan, Sport utility, Roadster Coupe; Data for weight: 3362, 4056, 4277, 4564, 4012, 3505 (lb);
Data for number of seats: 4, 5, 4, 5, 5, 2. The variable *body style* is qualitative; the variable *weight* is continuous; the variable *number of seats* is discrete.

47. **(a)** The research objective is to determine if adolescents aged 18–21 who smoke have a lower IQ than nonsmokers.

(b) The population is all adolescents aged 18–21. The sample consisted of 20,211 18-year-old Israeli military recruits.

(c) Descriptive statistics: The average IQ of the smokers was 94, and the average IQ of nonsmokers was 101.

(d) The conclusion is that individuals with a lower IQ are more likely to choose to smoke.

48. **(a)** The research objective is to determine if the application of duct tape is as effective as cryotherapy in the treatment of common warts.

(b) The population is all people with warts. The sample consisted of 51 patients with warts.

(c) Descriptive statistics: 85% of patients in group 1 and 60% of patients in group 2 had complete resolution of their warts.

(d) The conclusion is that duct tape is significantly more effective in treating warts than cryotherapy.

49. **(a)** The research objective is to determine the proportion of adult Americans who believe the federal government wastes 51 cents or more of every dollar.

(b) The population is all adult Americans aged 18 years or older.

(c) The sample is the 1017 American adults aged 18 years or older that were surveyed.

(d) Descriptive statistics: Of the 1017 individuals surveyed, 35% indicated that 51 cents or more is wasted.

(e) From this study, one can infer that 31% to 39% of Americans believe the federal government wastes much of the money collected in taxes.

50. **(a)** The research objective is to determine what proportion of adults, aged 18 and over, believe it would be a bad idea to invest $1000 in the stock market.

(b) The population is all adults aged 18 and over living in the United States.

(c) The sample is the 1018 adults aged 18 and over living in the United States who completed the survey.

(d) Descriptive statistics: Of the 1018 adults surveyed, 46% believe it would be a bad idea to invest $1000 in the stock market.

(e) The conclusion is that a little fewer than half of the adults in the United States believe investing $1000 in the stock market is a bad idea.

51. *Jersey number* is nominal (the numbers generally indicate a type of position played). However, if the researcher feels that lower caliber players received higher numbers, then *jersey number* would be ordinal since players could be ranked by their number.

52. **(a)** Nominal; the ticket number is categorized as a winner or a loser.

(b) Ordinal; the ticket number gives an indication as to the order of arrival of guests.

(c) Ratio; the implication is that the ticket number gives an indication of the number of people attending the party.

53. **(a)** The research question is to determine if the season of birth affects mood later in life.

(b) The sample consisted of the 400 people the researchers studied.

(c) The season in which you were born (winter, spring, summer, or fall) is a qualitative variable.

(d) According to the article, individuals born in the summer are characterized by rapid, frequent swings between sad and cheerful moods, while those born in the winter are less likely to be irritable.

(e) The conclusion was that the season at birth plays a role in one's temperament.

54. Quantitative variables are numerical measures such that meaningful arithmetic operations can be performed on the values of the variable. Qualitative variables describe an attribute or characteristic of the individual that allows researchers to categorize the individual.

55. The values of a discrete random variable result from counting. The values of a continuous random variable result from a measurement.

56. The four levels of measurement of a variable are nominal, ordinal, interval, and ratio. Examples: Nominal—brand of clothing; Ordinal—size of a car (small, mid-size, large); Interval—temperature (in degrees Celsius); Ratio—number of students in a class (Examples will vary.)

57. We say data vary, because when we draw a random sample from a population, we do not know which individuals will be included. If we were to take another random sample, we would have different individuals and therefore different data. This variability affects the results of a statistical analysis because the results would differ if a study is repeated.

58. The process of statistics is to (1) identify the research objective, which means to determine what should be studied and what we hope to learn; (2) collect the data needed to answer the research question, which is typically done by taking a random sample from a population; (3) describe the data, which is done by presenting descriptive statistics; and (4) perform inference in which the results are generalized to a larger population.

59. Age could be considered a discrete random variable. A random variable can be discrete by allowing, for example, only whole numbers to be recorded.

Section 1.2

1. The response variable is the variable of interest in a research study. An explanatory variable is a variable that affects (or explains) the value of the response variable. In research, we want to see how changes in the value of the explanatory variable affect the value of the response variable.

2. An observational study uses data obtained by studying individuals in a sample without trying to manipulate or influence the variable(s) of interest. In a designed experiment, a treatment is applied to the individuals in a sample in order to isolate the effects of the treatment on a response variable. Only an experiment can establish causation between an explanatory variable and a response variable. Observational studies can indicate a relationship, but cannot establish causation.

3. Confounding exists in a study when the effects of two or more explanatory variables are not separated. So any relation that appears to exist between a certain explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study. A lurking variable is a variable not accounted for in a study, but one that affects the value of the response variable. A confounding variable is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

4. The choice between an observational study and an experiment depends on the circumstances involved. Sometimes there are ethical reasons why an experiment cannot be conducted. Other times the researcher may conduct an observational study first to validate a belief prior to investing a large amount of time and money into a designed experiment. A designed experiment is preferred if ethics, time, and money are not an issue.

5. Cross-sectional studies collect information at a specific point in time (or over a very short period of time). Case-control studies are retrospective (they look back in time). Also, individuals that have a certain characteristic (such as cancer) in a case-control study are matched with those that do not have the characteristic. Case-control studies are typically superior to cross-sectional studies. They are relatively inexpensive, provide individual level data, and give longitudinal information not available in a cross-sectional study.

6. A cohort study identifies the individuals to participate and then follows them over a period of time. During this period, information about the individuals is gathered, but there is no attempt to influence the individuals. Cohort studies are superior to case-control studies because cohort studies do not require recall to obtain the data.

7. There is a perceived benefit to obtaining a flu shot, so there are ethical issues in intentionally denying certain seniors access to the treatment.

8. A retrospective study looks at data from the past either through recall or existing records. A prospective study gathers data over time by following the individuals in the study and recording data as they occur.

9. This is an observational study because the researchers merely observed existing data. There was no attempt by the researchers to manipulate or influence the variable(s) of interest.

10. This is an experiment because the researchers intentionally changed the value of the explanatory variable (medication dose) to observe a potential effect on the response variable (cancer growth).

11. This is an experiment because the explanatory variable (teaching method) was intentionally varied to see how it affected the response variable (score on proficiency test).

12. This is an observational study because no attempt was made to influence the variable of interest. Voting choices were merely observed.

13. This is an observational study because the survey only observed preference of Coke or Pepsi. No attempt was made to manipulate or influence the variable of interest.

14. This is an experiment because the researcher intentionally imposed treatments on individuals in a controlled setting.

15. This is an experiment because the explanatory variable (carpal tunnel treatment regimen) was intentionally manipulated in order to observe potential effects on the response variable (level of pain).

16. This is an observational study because the conservation agents merely observed the fish to determine which were carrying parasites. No attempt was made to manipulate or influence any variable of interest.

17. **(a)** This is a cohort study because the researchers observed a group of people over a period of time.

(b) The response variable is whether the individual has heart disease or not. The explanatory variable is whether the individual is happy or not.

(c) There may be confounding due to lurking variables. For example, happy people may be more likely to exercise, which could affect whether they will have heart disease or not.

18. **(a)** This is a cross-sectional study because the researchers collected information about the individuals at a specific point in time.

(b) The response variable is whether the woman has nonmelanoma skin cancer or not. The explanatory variable is the daily amount of caffeinated coffee consumed.

(c) It was necessary to account for these variables to avoid confounding with other variables.

19. **(a)** This is an observational study because the researchers simply administered a questionnaire to obtain their data. No attempt was made to manipulate or influence the variable(s) of interest.
This is a cross-sectional study because the researchers are observing participants at a single point in time.

(b) The response variable is body mass index. The explanatory variable is whether a TV is in the bedroom or not.

(c) Answers will vary. Some lurking variables might be the amount of exercise per week and eating habits. Both of these variables can affect the body mass index of an individual.

(d) The researchers attempted to avoid confounding due to other variables by taking into account such variables as "socioeconomic status."

(e) No. Since this was an observational study, we can only say that a television in the bedroom is associated with a higher body mass index.

20. **(a)** This is an observational study because the researchers merely observed the individuals included in the study. No attempt was made to manipulate or influence any variable of interest.
This is a cohort study because the researchers identified the individuals to be included in the study, then followed them for a period of time (7 years).

(b) The response variable is weight gain. The explanatory variable is whether the individual is married/cohabitating or not.

(c) Answers will vary. Some potential lurking variables are eating habits, exercise routine, and whether the individual has children.

(d) No. Since this is an observational study, we can only say that being married or cohabitating is associated with weight gain.

21. **(a)** This is a cross-sectional study because information was collected at a specific point in time (or over a very short period of time).

(b) The explanatory variable is delivery scenario (caseload midwifery, standard hospital care, or private obstetric care).

(c) The two response variables are (1) cost of delivery, which is quantitative, and (2) type of delivery (vaginal or not), which is quantitative.

22. **(a)** The explanatory variable is web page design; qualitative

(b) The response variables are time on site and amount spent. Both are quantitative.

(c) Answers will vary. A confounding variable might be location. Any differences in spending may be due to location rather than to web page design.

23. Answers will vary. This is a prospective, cohort observational study. The response variable is whether the worker had cancer or not, and the explanatory variable is the amount of electromagnetic field exposure. Some possible lurking variables include eating habits, exercise habits, and other health-related variables such as smoking habits. Genetics (family history) could also be a lurking variable. This was an observational study, and not an experiment, so the study only concludes that high electromagnetic field exposure is associated with higher cancer rates.
The author reminds us that this is an observational study, so there is no direct control over the variables that may affect cancer rates. He also points out that while we should not simply dismiss such reports, we should consider the results in conjunction with results from future studies. The author concludes by mentioning known ways (based on extensive study) of reducing cancer risks that can currently be done in our lives.

24. **(a)** The research objective is to determine whether lung cancer is associated with exposure to tobacco smoke within the household.

(b) This is a case-controlled study because there is a group of individuals with a certain characteristic (lung cancer but never smoked) being compared to a similar group without the characteristic (no lung cancer and never smoked). The study is retrospective because lifetime residential histories were compiled and analyzed.

(c) The response variable is whether the individual has lung cancer or not. This is a qualitative variable.

(d) The explanatory variable is the number of "smoker years." This is a quantitative variable.

(e) Answers will vary. Some possible lurking variables are household income, exercise routine, and exposure to tobacco smoke outside the home.

(f) The conclusion of the study is that approximately 17% of lung cancer cases among nonsmokers can be attributed to high levels of exposure to tobacco smoke during childhood and adolescence. No, we cannot say that exposure to household tobacco smoke causes lung cancer since this is only an observational study. We can, however, conclude that lung cancer is associated with exposure to tobacco smoke in the home.

(g) An experiment involving human subjects is not possible for ethical reasons. Researchers would be able to conduct an experiment using laboratory animals, such as rats.

Section 1.3

1. The frame is a list of all the individuals in the population.

2. Simple random sampling occurs when every possible sample of size *n* has an equally likely chance of occurring.

3. Sampling without replacement means that no individual may be selected more than once as a member of the sample.

4. Random sampling is a technique that uses chance to select individuals from a population to be in a sample. It is used because it maximizes the likelihood that the individuals in the sample are representative of the individuals in the population. In convenience sampling, the individuals in the sample are selected in the quickest and easiest way possible (e.g. the first 20 people to enter a store). Convenience samples likely do not represent the population of interest because chance was not used to select the individuals.

5. Answers will vary. We will use one-digit labels and assign the labels across each row (i.e. *Pride and Prejudice* – 0, *The Sun Also Rises* – 1, and so on). In Table I of Appendix A, starting at row 5, column 11, and proceeding downward, we obtain the following labels: 8, 4, 3
In this case, the 3 books in the sample would be *As I Lay Dying*, *A Tale of Two Cities*, and *Crime and Punishment*. Different labeling order, different starting points in Table I in Appendix A, or use of technology will likely yield different samples.

6. Answers will vary. We will use one-digit labels and assign the labels across each row (i.e. *Mady* – 0, *Breanne* – 1, and so on). In Table I of Appendix A, starting at row 11, column 6, and then proceeding downward, we obtain the following labels: 1, 5
In this case, the two captains would be Breanne and Payton. Different labeling order, different starting points in Table I in Appendix A, or use of technology will likely yield different results.

7. **(a)** {616, 630}, {616, 631}, {616, 632}, {616, 645}, {616, 649}, {616, 650}, {630, 631}, {630, 632}, {630, 645}, {630, 649}, {630, 650}, {631, 632}, {631, 645}, {631, 649}, {631, 650}, {632, 645}, {632, 649}, {632, 650}, {645, 649}, {645, 650}, {649, 650}

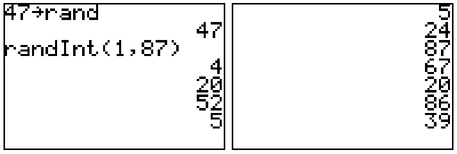
   (b) There is a 1 in 21 chance that the pair of courses will be EPR 630 and EPR 645.

8. **(a)** {1, 2}, {1, 3}, {1, 4}, {1, 5}, {1, 6}, {1, 7}, {2, 3}, {2, 4}, {2, 5}, {2, 6}, {2, 7}, {3, 4}, {3, 5}, {3, 6}, {3, 7}, {4, 5}, {4, 6}, {4, 7}, {5, 6}, {5, 7}, {6, 7}

   (b) There is a 1 in 21 chance that the pair *The United Nations* and *Amnesty International* will be selected.

9. **(a)** Starting at row 5, column 22, using two-digit numbers, and proceeding downward, we obtain the following values: 83, 94, 67, 84, 38, 22, 96, 24, 36, 36, 58, 34,.... We must disregard 94 and 96 because there are only 87 faculty members in the population. We must also disregard the second 36 because we are sampling without replacement. Thus, the 9 faculty members included in the sample are those numbered 83, 67, 84, 38, 22, 24, 36, 58, and 34.

(b) Answers will vary depending on the type of technology used. If using a TI-84 Plus, the sample will be: 4, 20, 52, 5, 24, 87, 67, 86, and 39.



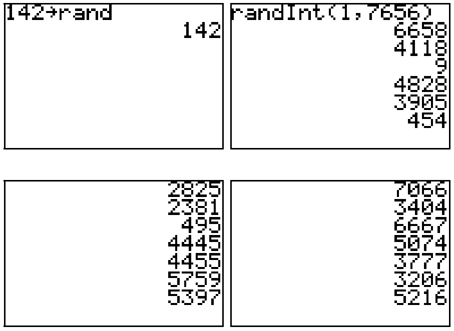
```
47→rand
             47
randInt(1,87)
              4
             20
             52
              5
```

```
              5
             24
             87
             67
             20
             86
             39
```

Note: We must disregard the second 20 because we are sampling without replacement.

10. **(a)** Starting at row 11, column 32, using four-digit numbers, and proceeding downward, we obtain the following values: 2869, 5518, 6635, 2182, 8906, 0603, 2654, 2686, 0135, 7783, 4080, 6621, 3774, 7887, 0826, 0916, 3188, 0876, 5418, 0037, 3130, 2882, 0662,…. We must disregard 8906, 7783, and 7887 because there are only 7656 students in the population.
Thus, the 20 students included in the sample are those numbered 2869, 5518, 6635, 2182, 0603, 2654, 2686, 0135, 4080, 6621, 3774, 0826, 0916, 3188, 0876, 5418, 0037, 3130, 2882, and 0662.

(b) Answers may vary depending on the type of technology used. If using a TI-84 Plus, the sample will be: 6658, 4118, 9, 4828, 3905, 454, 2825, 2381, 495, 4445, 4455, 5759, 5397, 7066, 3404, 6667, 5074, 3777, 3206, 5216.



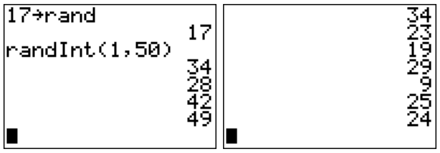
```
142→rand
            142
```

```
randInt(1,7656)
           6658
           4118
              9
           4828
           3905
            454
```

```
           2825
           2381
            495
           4445
           4455
           5759
           5397
```

```
           7066
           3404
           6667
           5074
           3777
           3206
           5216
```

11. **(a)** Answers will vary depending on the technology used (including a table of random digits). Using a TI-84 Plus graphing calculator with a seed of 17 and the labels provided, our sample would be North Dakota, Nevada, Tennessee, Wisconsin, Minnesota, Maine, New Hampshire, Florida, Missouri, and Mississippi.

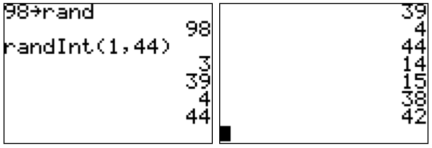


```
17→rand
             17
randInt(1,50)
             34
             28
             42
             49
```

```
             34
             23
             19
             29
              9
             25
             24
```

(b) Repeating part (a) with a seed of 18, our sample would be Michigan, Massachusetts, Arizona, Minnesota, Maine, Nebraska, Georgia, Iowa, Rhode Island, Indiana.

12. **(a)** Answers will vary depending on the technology used (including a table of random digits). Using a TI-84 Plus graphing calculator with a seed of 98 and the labels provided, our sample would be Jefferson, Carter, Madison, Obama, Pierce, Buchanan, Ford, Clinton.



```
98→rand
             98
randInt(1,44)
              3
             39
              4
             44
```

```
             39
              4
             44
             14
             15
             38
             42
```

(b) Repeating part (a) with a seed of 99, our sample would be L. B. Johnson, Truman, Pierce, Garfield, Obama, Grant, George H. Bush, T. Roosevelt.

13. **(a)** The list provided by the administration serves as the frame. Number each student in the list of registered students, from 1 to 19,935. Generate 25 random numbers, without repetition, between 1 and 19,935 using a random number generator or table. Select the 25 students with these numbers.

(b) Answers will vary.

14. **(a)** The list provided by the mayor serves as the frame. Number each resident in the list supplied by the mayor, from 1 to 5832. Generate 20 random numbers, without repetition, between 1 and 5832 using a random number generator or table. Select the 20 residents with these numbers.

(b) Answers will vary.

15. Answers will vary. Members should be numbered 1–32, though other numbering schemes are possible (e.g. 0–31). Using a table of random digits or a random-number generator, four different numbers (labels) should be selected. The names corresponding to these numbers form the sample.

16. Answers will vary. Employees should be numbered 1–29, though other numbering schemes are possible (e.g. 0–28). Using a table of random digits or a random-number generator, four different numbers (labels) should be selected. The names corresponding to these numbers form the sample.

Section 1.4

1. Stratified random sampling may be appropriate if the population of interest can be divided into groups (or strata) that are homogeneous and nonoverlapping.

2. Systematic sampling does not require a frame.

3. Convenience samples are typically selected in a nonrandom manner. This means the results are not likely to represent the population. Convenience samples may also be self-selected, which will frequently result in small portions of the population being overrepresented.

4. Cluster sample

5. Stratified sample

6. False. In a systematic random sample, every *k*th individual is selected from the population.

7. False. In many cases, other sampling techniques may provide equivalent or more information about the population with less “cost” than simple random sampling.

8. True. When the clusters are heterogeneous, the heterogeneity of each cluster likely resembles the heterogeneity of the population. In such cases, fewer clusters with more individuals from each cluster are preferred.

9. True. Because the individuals in a convenience sample are not selected using chance, it is likely that the sample is not representative of the population.

10. False. With stratified samples, the number of individuals sampled from each strata should be proportional to the size of the strata in the population.

11. Systematic sampling. The quality-control manager is sampling every 8th chip, starting with the 3rd chip.

12. Cluster sampling. The commission tests all members of the selected teams (clusters).

13. Cluster sampling. The airline surveys all passengers on selected flights (clusters).

14. Stratified sampling. The congresswoman samples some individuals from each of three different income brackets (strata).

15. Simple random sampling. Each known user of the product has the same chance of being included in the sample.

16. Convenience sampling. The radio station is relying on voluntary response to obtain the sample data.

17. Cluster sampling. The farmer samples all trees within the selected subsections (clusters).

18. Stratified sampling. The school official takes a sample of students from each of the five classes (strata).

19. Convenience sampling. The research firm is relying on voluntary response to obtain the sample data.

20. Systematic sampling. The presider is sampling every 5th person attending the lecture, starting with the 3rd person.

21. Stratified sampling. Shawn takes a sample of measurements during each of the four time intervals (strata).

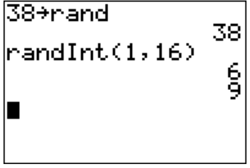
22. Simple random sampling. Each club member has the same chance of being selected for the survey.

23. The numbers corresponding to the 20 clients selected are 16, $16+25=41$, $41+25=66$, $66+25=91$, $91+25=116$, 141, 166, 191, 216, 241, 266, 291, 316, 341, 366, 391, 416, 441, 466, 491.

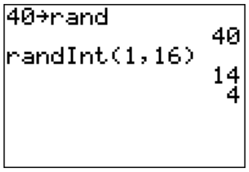
24. Since the number of clusters is more than 100, but less than 1000, we assign each cluster a three-digit label between 001 and 795. Starting at row 8, column 38 in Table I of Appendix A, and proceeding downward, the 10 clusters selected are numbered 763, 185, 377, 304, 626, 392, 315, 084, 565, and 508. Note that we discard 822 and 955 in reading the table because we have no clusters with these labels. We also discard the second occurrence of 377 because we cannot select the same cluster twice.

25. Answers will vary. To obtain the sample, number the Democrats 1 to 16 and obtain a simple random sample of size 2. Then number the Republicans 1 to 16 and obtain a simple random sample of size 2. Be sure to use a different starting point in Table I or a different seed for each stratum.

For example, using a TI-84 Plus graphing calculator with a seed of 38 for the Democrats and 40 for the Republicans, the numbers selected would be 6, 9 for the Democrats and 14, 4 for the Republicans. If we had numbered the individuals down each column, the sample would consist of Haydra, Motola, Thompson, and Engler.



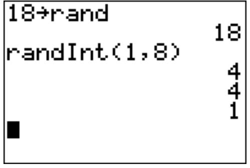
```
38→rand
                38
randInt(1,16)
                 6
                 9
```



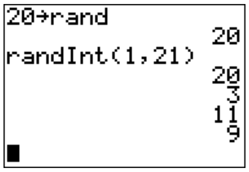
```
40→rand
                40
randInt(1,16)
                14
                 4
```

26. Answers will vary. To obtain the sample, number the managers 1 to 8 and obtain a simple random sample of size 2. Then number the employees 1 to 21 and obtain a simple random sample of size 4. Be sure to use a different starting point in Table I or a different seed for each stratum.

For example, using a TI-84 Plus graphing calculator with a seed of 18 for the managers and 20 for the employees, the numbers selected would be 4, 1 for the managers and 20, 3, 11, 9 for the employees. If we had numbered the individuals down each column, the sample would consist of Lindsey, Carlisle, Weber, Bryant, Hall, and Gow.



```
18→rand
                18
randInt(1,8)
                 4
                 4
                 1
```



```
20→rand
                20
randInt(1,21)
                20
                 3
                11
                 9
```

27. **(a)** $\frac{N}{n}=\frac{4502}{50}=90.04\rightarrow 90$; Thus, $k=90$.

(b) Randomly select a number between 1 and 90. Suppose that we select 15. Then the individuals to be surveyed will be the 15th, 105th, 195th, 285th, and so on up to the 4425th employee on the company list.

28. **(a)** $\frac{N}{n}=\frac{945035}{130}=7269.5\rightarrow 7269$; Thus, $k=7269$.

(b) Randomly select a number between 1 and 7269. Suppose that we randomly select 2000. Then we will survey the individuals numbered 2000, 9269, 16,538, and so on up to the individual numbered 939,701.

29. Simple Random Sample:

Number the students from 1 to 1280. Use a table of random digits or a random-number generator to randomly select 128 students to survey.

Stratified Sample:

Since class sizes are similar, we would want to randomly select $\frac{128}{32}=4$ students from each class to be included in the sample.

Cluster Sample:

Since classes are similar in size and makeup, we would want to randomly select $\frac{128}{32}=4$ classes and include all the students from those classes in the sample.

30. No. The clusters were not randomly selected. This would be considered convenience sampling.

31. Answers will vary. One design would be a stratified random sample, with two strata being commuters and noncommuters, as these two groups each might be fairly homogeneous in their reactions to the proposal.

32. Answers will vary. One design would be a cluster sample, with classes as the clusters. Randomly select clusters and then survey all the students in the selected classes. However, care would need to be taken to make sure that no one was polled twice. Since this would negate some of the ease of cluster sampling, a simple random sample might be the more suitable design.

33. Answers will vary. One design would be a cluster sample, with the clusters being city blocks. Randomly select city blocks and survey every household in the selected blocks.

34. Answers will vary. One appropriate design would be a systematic sample, after doing a random start, clocking the speed of every tenth car, for example.

35. Answers will vary. Since the company already has a list (frame) of 6600 individuals with high cholesterol, a simple random sample would be an appropriate design.

36. Answers will vary. Since a list of all the households in the population exists, a simple random sample is possible. Number the households from 1 to *N*, then use a table of random digits or a random-number generator to select the sample.

37. **(a)** For a political poll, a good frame would be all registered voters who have voted in the past few elections since they are more likely to vote in upcoming elections.

(b) Because each individual from the frame has the same chance of being selected, there is a possibility that one group may be over- or underrepresented.

(c) By using a stratified sample, the strategist can obtain a simple random sample within each strata (political party) so that the number of individuals in the sample is proportionate to the number of individuals in the population.

38. Random sampling means that the individuals chosen to be in the sample are selected by chance. Random sampling minimizes the chance that one part of the population is over- or underrepresented in the sample. However, it cannot guarantee that the sample will accurately represent the population.

39. Answers will vary.

40. Answers will vary.

Section 1.5

1. A closed question is one in which the respondent must choose from a list of prescribed responses. An open question is one in which the respondent is free to choose his or her own response. Closed questions are easier to analyze, but limit the responses. Open questions allow respondents to state exactly how they feel, but are harder to analyze due to the variety of answers and possible misinterpretation of answers.

2. A certain segment of the population is underrepresented if it is represented in the sample in a lower proportion than its size in the population.

3. Bias means that the results of the sample are not representative of the population. There are three types of bias: sampling bias, response bias, and nonresponse bias. Sampling bias is due to the use of a sample to describe a population. This includes bias due to convenience sampling. Response bias involves intentional or unintentional misinformation. This would include lying to a surveyor or entering responses incorrectly. Nonresponse bias results when individuals choose not to respond to questions or are unable to be reached. A census can suffer from response bias and nonresponse bias, but would not suffer from sampling bias.

4. Nonsampling error is the error that results from undercoverage, nonresponse bias, response bias, or data-entry errors. Essentially, it is the error that results from the process of obtaining and recording data. Sampling error is the error that results because a sample is being used to estimate information about a population. Any error that could also occur in a census is considered a nonsampling error.

5. **(a)** Sampling bias. The survey suffers from undercoverage because the first 60 customers are likely not representative of the entire customer population.

(b) Since a complete frame is not possible, systematic random sampling could be used to make the sample more representative of the customer population.

6. **(a)** Sampling bias. The survey suffers from undercoverage because only homes in the southwest corner have a chance to be interviewed. These homes may have different demographics than those in other parts of the village.

(b) Assuming that households within any given neighborhood have similar household incomes, stratified sampling might be appropriate, with neighborhoods as the strata.

7. **(a)** Response bias. The survey suffers from response bias because the question is poorly worded.

(b) The survey should inform the respondent of the current penalty for selling a gun illegally and the question should be worded as "Do you approve or disapprove of harsher penalties for individuals who sell guns illegally?" The order of "approve" and "disapprove" should be switched from one individual to the next.

8. **(a)** Response bias. The survey suffers from response bias because the wording of the question is ambiguous.

(b) The question might be worded more specifically as "How many hours per night do you sleep, on average?"

9. **(a)** Nonresponse bias. Assuming the survey is written in English, non-English speaking homes will be unable to read the survey. This is likely the reason for the very low response rate.

(b) The survey can be improved by using face-to-face or phone interviews, particularly if the interviewers are multi-lingual.

10. **(a)** Nonresponse bias

(b) The survey can be improved by using face-to-face or phone interviews, or possibly through the use of incentives.

11. **(a)** The survey suffers from sampling bias due to undercoverage and interviewer error. The readers of the magazine may not be representative of all Australian women, and advertisements and images in the magazine could affect the women's view of themselves.

(b) A well-designed sampling plan not in a magazine, such as a cluster sample, could make the sample more representative of the population.

12. **(a)** The survey suffers from sampling bias due to a bad sampling plan (convenience sampling) and possible response bias due to misreported weights on driver's licenses.

(b) The teacher could use cluster sampling or stratified sampling using classes throughout the day. Each student should be weighed to get a current and accurate weight measurement.

13. **(a)** Response bias due to a poorly worded question

(b) The question should be reworded in a more neutral manner. One possible phrasing might be "Do you believe that a marriage can be maintained after an extramarital relation?"

14. **(a)** Sampling bias. The frame is not necessarily representative of all college professors.

(b) To remedy this problem, the publisher could use cluster sampling and obtain a list of faculty from the human resources departments at selected colleges.

15. **(a)** Response bias. Students are unlikely to give honest answers if their teacher is administering the survey.

(b) An impartial party should administer the survey in order to increase the rate of truthful responses.

16. **(a)** Response bias. Residents are unlikely to give honest answers to uniformed police officers if their answer would be seen as negative by the police.

(b) An impartial party should administer the survey in order to increase the rate of truthful responses.

17. No. The survey still suffers from sampling bias due to undercoverage, nonresponse bias, and potentially response bias.

18. The General Social Survey uses random sampling to obtain individuals who take the survey, so the results of their survey are more likely to be representative of the population. However, it may suffer from response bias since the survey is conducted by personal interview rather than anonymously on the Internet. The online survey, while potentially obtaining more honest answers, is basically self-selected so may not be representative of the population, particularly if most respondents are clients of the family and wellness center seeking help with health or relationship problems.

19. It is very likely that the order of these two questions will affect the survey results. To alleviate the response bias, either question B could be asked first, or the order of the two questions could be rotated randomly.

20. It is very likely that the order of these two questions will affect the survey results. To alleviate the response bias, the order of the two questions could be rotated randomly. Prohibit is a strong word. People generally do not like to be prohibited from doing things. If the word must be used, it should be offset by the word "allow." The use of the words "prohibit" and "allow" should be rotated within the question.

21. The company is using a reward in the form of the $5.00 payment and an incentive by telling the reader that his or her input will make a difference.

22. The two choices need to be rotated so that any response bias due to the ordering of the questions is minimized.

23. For random digit dialing, the frame is anyone with a phone (whose number is not on a do-not-call registry). Even those with unlisted numbers can still be reached through this method.

Any household without a phone, households on the do-not-call registry, and homeless individuals are excluded. This could result in sampling bias due to undercoverage if the excluded individuals differ in some way than those included in the frame.

24. Answers will vary. The use of caller ID has likely increased nonresponse bias of phone surveys since individuals may not answer calls from numbers they do not recognize. If individuals with caller ID differ in some way from individuals without caller ID, then phone surveys could also suffer from sampling bias due to undercoverage.

25. It is extremely likely, particularly if households on the do-not-call registry have a trait that is not part of those households that are not on the registry.

26. There is a higher chance that an individual at least 70 years of age will be at home when an interviewer makes contact.

27. Some nonsampling errors presented in the article as leading to incorrect exit polls were poorly trained interviewers, interviewer bias, and over representation of female voters.

28. – 32. Answers will vary.

33. The *Literary Digest* made an incorrect prediction due to sampling bias (an incorrect frame led to undercoverage) and nonresponse bias (due to the low response rate).

34. Answers will vary. (Gallup incorrectly predicted the outcome of the 1948 election because he quit polling weeks before the election and missed a large number of changing opinions.)

35. **(a)** Answers will vary. Stratified sampling by political affiliation (Democrat, Republican, etc.) could be used to ensure that all affiliations are represented. One question that could be asked is whether or not the person plans to vote in the next election. This would help determine which registered voters are likely to vote.

(b) Answers will vary. Possible explanations are that presidential election cycles get more news coverage or perhaps people are more interested in voting when they can vote for a president as well as a senator. During non-presidential cycles it is very informative to poll likely registered voters.

(c) Answers will vary. A higher percentage of Democrats in polls versus turnout will lead to overstating the predicted Democrat percentage of Democratic votes.

36. It is difficult for a frame to be completely accurate since populations tend to change over time and there can be a delay in identifying individuals who have joined or left the population.

37. Nonresponse can be addressed by conducting callbacks or offering rewards.

38. Trained, skillful interviewers can elicit responses from individuals and help them give truthful responses.

39. Conducting a presurvey with open questions allows the researchers to use the most popular answers as choices on closed-question surveys.

40. Answers will vary. Phone surveys conducted in the evening may result in reaching more potential respondents; however some of these individuals could be upset by the intrusion.

41. Provided the survey was conducted properly and randomly, a high response rate will provide more representative results. When a survey has a low response rate, only those who are most willing to participate give responses. Their answers may not be representative of the whole population.

42. The order of questions on a survey should be carefully considered, so the responses are not affected by previous questions.

43. There is more than one type of CD. This can be interpreted as a medium used to store music or information electronically: a compact disk. It could also be understood as a special type of savings account: a certificate of deposit. The question can be improved by asking, "Do you own any certificates of deposit, which are a special type of savings account at a bank?"

44. Higher response rates typically suggest that the sample represents the population well. Using rewards can help increase response rates, allowing researchers to better understand the population. There can be disadvantages to offering rewards as incentives. Some people may hurry through the survey, giving superficial answers, just to obtain the reward.

Section 1.6

1. (a) An experimental unit is a person, object, or some other well-defined item upon which a treatment is applied.

(b) A treatment is a condition applied to an experimental unit. It can be any combination of the levels of the explanatory variables.

(c) A response variable is a quantitative or qualitative variable that measures a response of interest to the experimenter.

(d) A factor is a variable whose effect on the response variable is of interest to the experimenter. Factors are also called explanatory variables.

(e) A placebo is an innocuous treatment, such as a sugar pill, administered to a subject in a manner indistinguishable from an actual treatment.

(f) Confounding occurs when the effect of two explanatory variables on a response variable cannot be distinguished.

2. Replication occurs when each treatment is applied to more than one experimental unit.

3. In a single-blind experiment, subjects do not know which treatment they are receiving. In a double-blind experiment, neither the subject nor the researcher(s) in contact with the subjects knows which treatment is received.

4. Completely randomized; matched-pair

5. False

6. True

7. (a) The research objective of the study was to determine the association between number of times one chews food and food consumption.

(b) The response variable is food consumption; quantitative.

(c) The explanatory variable is chew level (100%, 150%, 200%); qualitative.

(d) The experimental units are the 45 individuals aged 18 to 45 who participated in the study.

(e) Control is used by determining a baseline number of chews before swallowing; same type of food is used in the baseline as in the experiment; same time of day (lunch); age (18 to 45).

(f) Randomization reduces the effect of the order in which the treatments are administered. For example, perhaps the first time through the subjects are more diligent about their chewing than the last time through the study.

8. (a) The researchers used an innocuous treatment to account for effects that would result from any treatment being given (i.e. the placebo effect). The placebo is a drug that looks and tastes like topiramate and serves as the baseline against which to compare the results when topiramate is administered.

(b) Being double-blind means that neither the subject nor the researcher in contact with the subjects knows whether the placebo or topiramate is being administered. Using a double-blind procedure is necessary to avoid any intentional or unintentional bias due to knowing which treatment is being given.

(c) The subjects were randomly assigned to the treatment groups (either the placebo or topiramate).

(d) The population is all men and women aged 18 to 65 years diagnosed with alcohol dependence. The sample is the 371 men and women aged 18 to 65 years diagnosed with alcohol dependence who participated in the 14-week trial.

(e) There are two treatments in the study: 300 mg of topiramate or a placebo daily.

(f) The response variable is the percentage of heavy drinking days.

---

9. **(a)** The response variable is the achievement test scores.

(b) Answers may vary. Some factors are teaching methods, grade level, intelligence, school district, and teacher.
Fixed: grade level, school district, teacher
Set at predetermined levels: teaching method

(c) The treatments are the new teaching method and the traditional method. There are 2 levels of treatment.

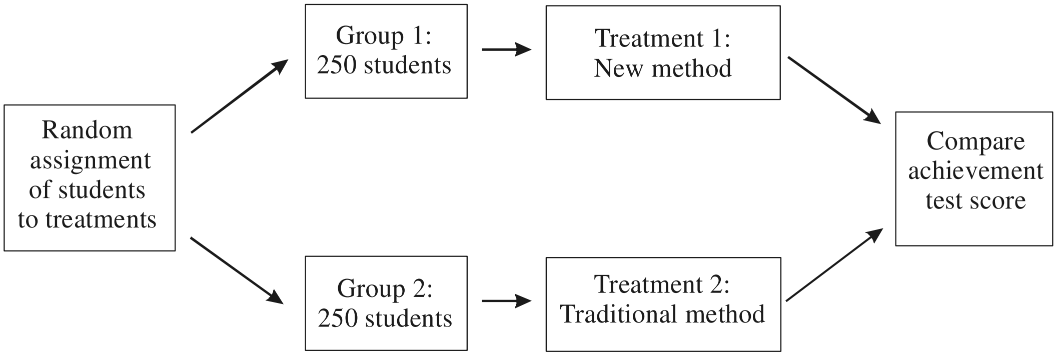
(d) The factors that are not controlled are dealt with by random assignment into the two treatment groups.

(e) Group 2, using the traditional teaching method, serves as the control group.

(f) This experiment has a completely randomized design.

(g) The subjects are the 500 first-grade students from District 203 recruited for the study.

(h)


Random assignment of students to treatments
Group 1: 250 students
Treatment 1: New method
Group 2: 250 students
Treatment 2: Traditional method
Compare achievement test score

10. **(a)** The response variable is the proportion of subjects with a cold.

(b) Answers may vary. Some factors are gender, age, geographic location, overall health, and drug intervention.
Fixed: gender, age, location
Set at predetermined levels: drug intervention

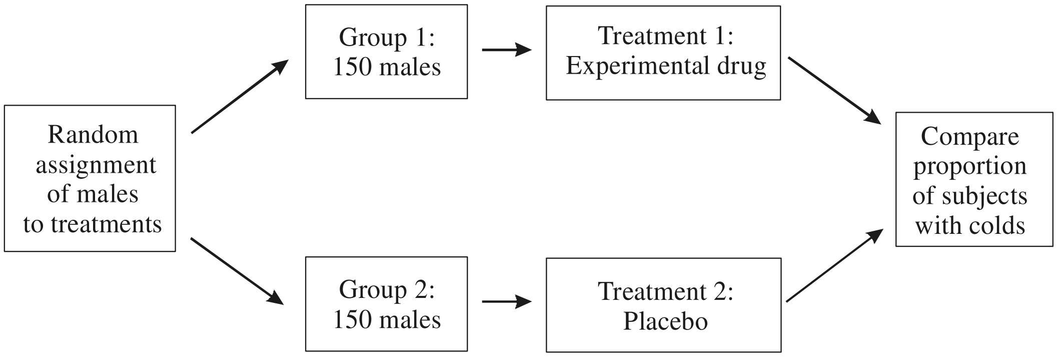
(c) The treatments are the experimental drug and the placebo. There are 2 levels of treatment.

(d) The factors that are not controlled are dealt with by random assignment into the two groups.

(e) This experiment has a completely randomized design.

(f) The subjects are the 300 adult males aged 25 to 29 who have the common cold.

(g)


Random assignment of males to treatments
Group 1: 150 males
Treatment 1: Experimental drug
Group 2: 150 males
Treatment 2: Placebo
Compare proportion of subjects with colds

11. **(a)** This experiment has a matched-pairs design.

(b) The response variable is the level of whiteness.

(c) The explanatory variable or factor is the whitening method. The treatments are Crest Whitestrips Premium in addition to brushing and flossing, and just brushing and flossing alone.

(d) Answers will vary. One other possible factor is diet. Certain foods and tobacco products are more likely to stain teeth. This could impact the level of whiteness.

(e) Answers will vary. One possibility is that using twins helps control for genetic factors such as weak teeth that may affect the results of the study.

12. **(a)** This experiment has a matched-pairs design.

(b) The response variable is the difference in test scores.

(c) The treatment is the mathematics course.

13. **(a)** This experiment has a completely randomized design.

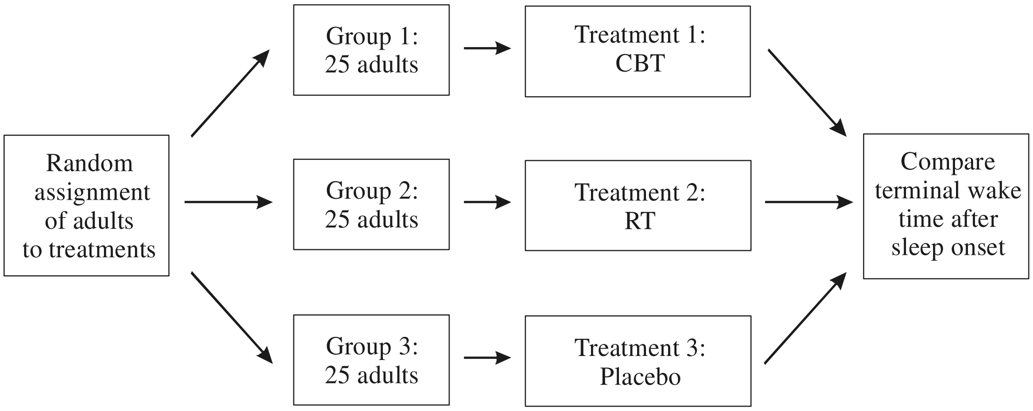
(b) The population being studied is adults with insomnia.

(c) The response variable is the terminal wake time after sleep onset (WASO).

(d) The explanatory variable or factor is the type of intervention. The treatments are cognitive behavioral therapy (CBT), muscle relaxation training (RT), and the placebo.

(e) The experimental units are the 75 adults with insomnia.

(f)



14. **(a)** This experiment has a completely randomized design.

(b) The population being studied is adult outpatients diagnosed as having major depression and having a baseline Hamilton Rating Scale for Depression (HAM-D) score of at least 20.

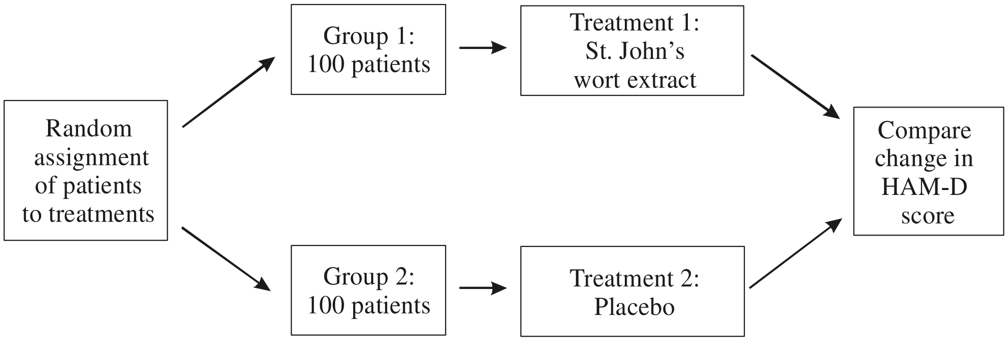
(c) The response variable is the change in the HAM-D over the treatment period.

(d) The explanatory variable or factor is the type of drug. The treatments are St. John's wort extract and the placebo.

(e) The experimental units are the 200 adult outpatients diagnosed with depression.

(f) The control group is the placebo group.

(g)



Random assignment of patients to treatments
Group 1: 100 patients
Treatment 1: St. John's wort extract
Group 2: 100 patients
Treatment 2: Placebo
Compare change in HAM-D score

15. **(a)** This experiment has a completely randomized design.

(b) The population being studied is adults over 60 years old and in good health.

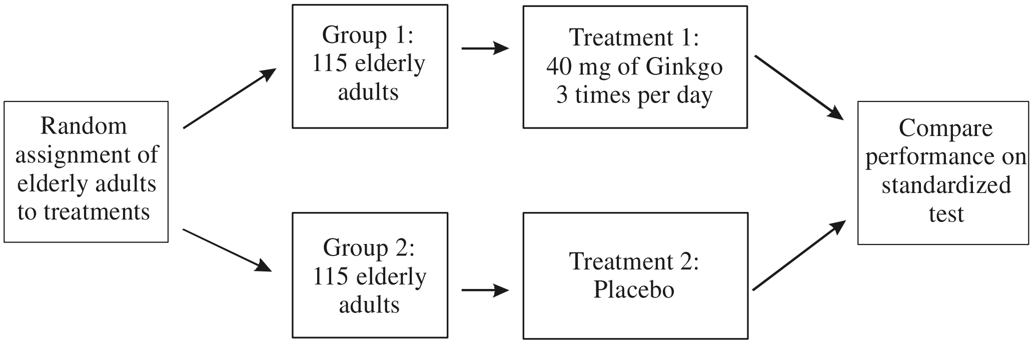
(c) The response variable is the standardized test of learning and memory.

(d) The factor set to predetermined levels (explanatory variable) is the drug. The treatments are 40 milligrams of ginkgo 3 times per day and the matching placebo.

(e) The experimental units are the 98 men and 132 women over 60 years old and in good health.

(f) The control group is the placebo group.

(g)



Random assignment of elderly adults to treatments
Group 1: 115 elderly adults
Treatment 1: 40 mg of Ginkgo 3 times per day
Group 2: 115 elderly adults
Treatment 2: Placebo
Compare performance on standardized test

16. **(a)** This experiment has a completely randomized design.

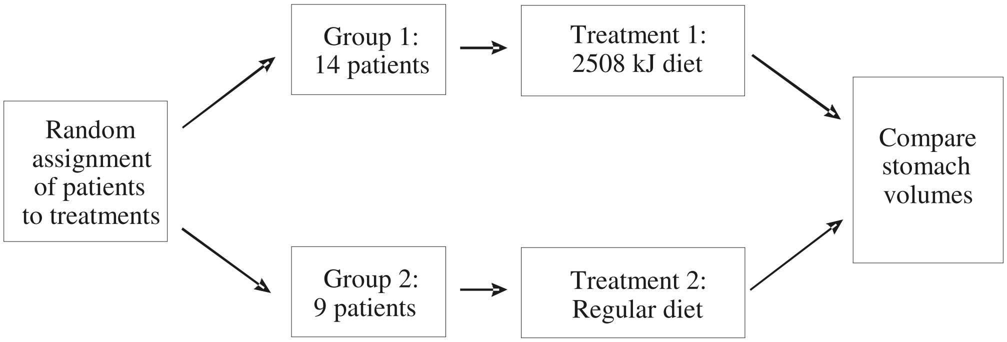
(b) The population being studied is obese patients.

(c) The response variable is the volume of the stomach. This is a quantitative variable.

(d) The treatments are the 2508 kJ diet versus the regular diet.

(e) The experimental units are the 23 obese patients.

(f)


Random assignment of patients to treatments
Group 1: 14 patients
Treatment 1: 2508 kJ diet
Group 2: 9 patients
Treatment 2: Regular diet
Compare stomach volumes

17. **(a)** This experiment has a matched-pairs design.

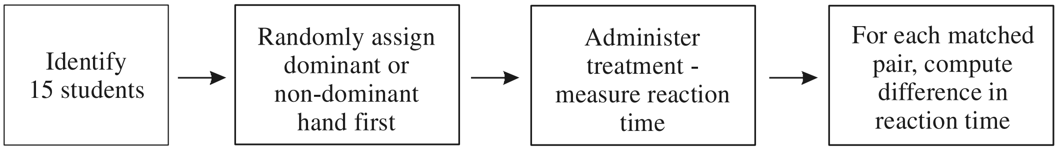
(b) The response variable is the distance the yardstick falls.

(c) The explanatory variable or factor is hand dominance. The treatment is dominant versus non-dominant hand.

(d) The experimental units are the 15 students.

(e) Professor Neil used a coin flip to eliminate bias due to starting on the dominant or non-dominant hand first on each trial.

(f)


Identify 15 students
Randomly assign dominant or non-dominant hand first
Administer treatment - measure reaction time
For each matched pair, compute difference in reaction time

18. **(a)** This experiment has a matched-pairs design.

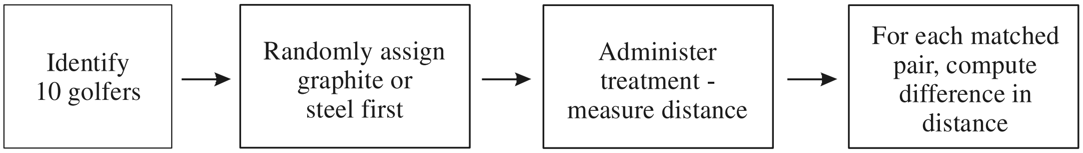
(b) The response variable is the distance the ball is hit.

(c) The explanatory variable or factor is the shaft type. The treatment is graphite shaft versus steel shaft.

(d) The experimental units are the 10 golfers.

(e) The golf pro used a coin flip to eliminate bias due to the type of shaft used first.

(f)


Identify 10 golfers
Randomly assign graphite or steel first
Administer treatment - measure distance
For each matched pair, compute difference in distance

19. Answers will vary. Using a TI-84 Plus graphing calculator with a seed of 195, we would pick the volunteers numbered 8, 19, 10, 12, 13, 6, 17, 1, 4, and 7 to go into the experimental group. The rest would go into the control group. If the volunteers were numbered in the order listed, the experimental group would consist of Ann, Kevin, Christina, Eddie, Shannon, Randy, Tom, Wanda, Kim, and Colleen.

20. **(a)** This experiment has a completely randomized design.

(b) Answers will vary. Using a TI-84 Plus graphing calculator with a seed of 223, we would pick the volunteers numbered 6, 18, 13, 3, 19, 14, 8, 1, 17, and 5 to go into group 1.

21. **(a)** This is an observational study because there is no intent to manipulate an explanatory variable or factor. The explanatory variable or factor is whether the individual is a green tea drinker or not, which is qualitative.

(b) Some lurking variables include diet, exercise, genetics, age, gender, and socioeconomic status.

(c) The experiment is a completely randomized design.

(d) To make this a double-blind experiment, we would need the placebo to look, taste, and smell like green tea. Subjects would not know which treatment is being delivered. In addition, the individuals administering the treatment and measuring the changes in LDL cholesterol would not know the treatment either.

(e) The factor that is manipulated is the tea, which is set at three levels; qualitative.

(f) Answers will vary. Other factors you might want to control in this experiment include age, exercise, and diet of the participants.

(g) Randomization could be used by numbering the subjects from 1 to 120. Randomly select 40 subjects and assign them to the placebo group. Then randomly select 40 from the remaining 80 subjects and assign to the one cup of green tea group. The remaining subjects will be assigned to the two cups of green tea group. By randomly assigning the subjects to the treatments, the expectation is that uncontrolled variables (such as genetic history, diet, exercise, etc.) are neutralized (even out).

(h) Exercise is a confounding variable because any change in the LDL cholesterol cannot be attributed to the tea. It may be the exercise that caused the change in LDL cholesterol.

22. (a) The research objective is to determine if alerting shoppers about the healthiness of energy-dense snack foods changes the shopping habits of overweight individuals.

(b) The subjects were 42 overweight shoppers.

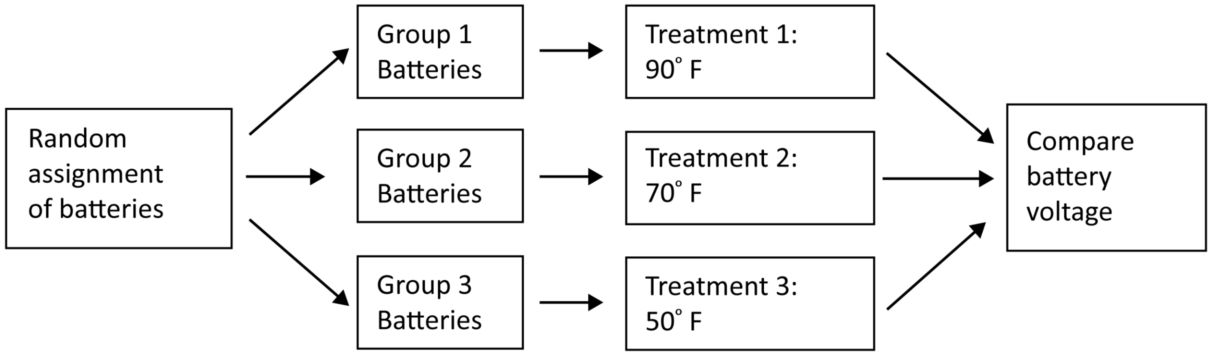
(c) Blinding is not possible because health information is visible.

(d) The explanatory variable is health information or not.

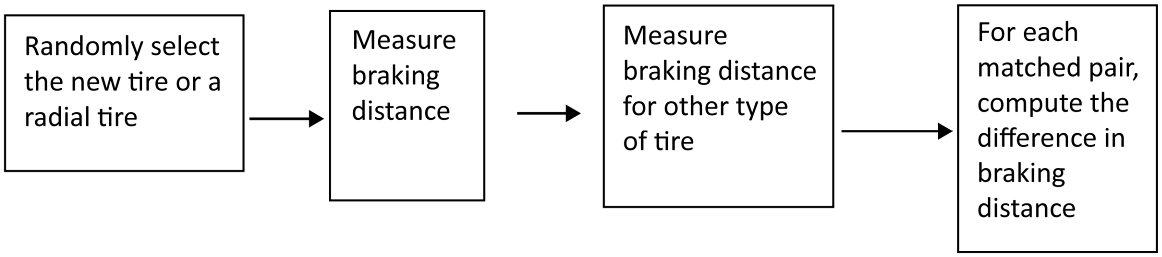
(e) The number of unhealthy snacks purchased is quantitative.

(f) The researchers would not be able to distinguish whether it was the priming or the weight status that played a role in purchase decisions.

23. Answers will vary. A completely randomized design is probably best.



24. Answers will vary. A matched-pairs design matched by car model is likely the best.


Randomly select the new tire or a radial tire
Measure braking distance
Measure braking distance for other type of tire
For each matched pair, compute the difference in braking distance

25. **(a)** The response variable is blood pressure.

(b) Three factors that have been identified are daily consumption of salt, daily consumption of fruits and vegetables, and the body's ability to process salt.

(c) The daily consumption of salt and the daily consumption of fruits and vegetables can be controlled. The body's ability to process salt cannot be controlled. To deal with variability of the body's ability to process salt, randomize experimental units to each treatment group.

(d) Answers will vary. Three levels of treatment might be a good choice – one level below the recommended daily allowance, one equal to the recommended daily allowance, and one above the recommended daily allowance.

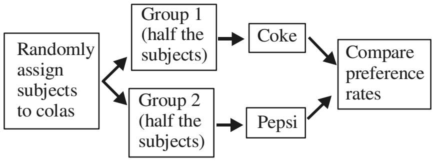
26. Answers will vary.

27. Answers will vary.

28. Answers will vary for the design preference.

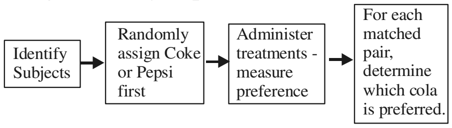
Completely Randomized Design

The researcher would randomly assign each subject to either drink Coke or Pepsi. The response variable would be whether the subject likes the soda or not. Preference rates would be compared at the end of the experiment. The subject would be blinded, but the researcher would not. Therefore, this would be a single-blind experiment.


Randomly assign subjects to colas
Group 1 (half the subjects)
Group 2 (half the subjects)
Coke
Pepsi
Compare preference rates

Matched-Pairs Design

The researcher would randomly determine whether each subject drinks Coke first or Pepsi first. To avoid confounding, subjects should eat something bland between drinks to remove any residual taste. The response variable would be either the proportion of subjects who prefer Coke or the proportion of subjects who prefer Pepsi. This would also be a single-blind experiment since the subject would not know which drink was first but the researcher would. The matched-pairs design is likely superior.


Identify Subjects
Randomly assign Coke or Pepsi first
Administer treatments - measure preference
For each matched pair, determine which cola is preferred.

29. Answers will vary. Control groups are needed in a designed experiment to serve as a baseline against which other treatments can be compared.

30. **(a)** Answers will vary.

(b) Answers will vary.

31. The purpose of randomization is to minimize the effect of factors whose levels cannot be controlled. (Answers will vary.) One way to assign the experimental units to the three groups is to write the numbers 1, 2, and 3 on identical pieces of paper and to draw them out of a "hat" at random for each experimental unit.

Chapter 1 Review Exercises

1. Statistics is the science of collecting, organizing, summarizing, and analyzing information in order to draw conclusions.

2. The population is the group of individuals that is to be studied.

3. A sample is a subset of the population.

4. An observational study uses data obtained by studying individuals in a sample without trying to manipulate or influence the variable(s) of interest. Observational studies are often called *ex post facto* studies because the value of the response variable has already been determined.

5. In a designed experiment, a treatment is applied to the individuals in a sample in order to isolate the effects of the treatment on the response variable.

6. The three major types of observational studies are (1) cross-sectional studies, (2) case-control studies, and (3) cohort studies.

Cross-sectional studies collect data at a specific point in time or over a short period of time. Cohort studies are prospective and collect data over a period of time, sometimes over a long period of time. Case-controlled studies are retrospective, looking back in time to collect data either from historical records or from recollection by subjects in the study.

Individuals possessing a certain characteristic are matched with those that do not.

7. The process of statistics refers to the approach used to collect, organize, analyze, and interpret data. The steps are to
(1) identify the research objective,
(2) collect the data needed to answer the research question,
(3) describe the data, and
(4) perform inference.

8. The three types of bias are sampling bias, nonresponse bias, and response bias. Sampling bias occurs when the techniques used to select individuals to be in the sample favor one part of the population over another. Bias in sampling is reduced when a random process is to select the sample. Nonresponse bias occurs when the individuals selected to be in the sample that do not respond to the survey have different opinions from those that do respond. This can be minimized by using callbacks and follow-up visits to increase the response rate. Response bias occurs when the answers on a survey do not reflect the true feelings of the respondent. This can be minimized by using trained interviewers, using carefully worded questions, and rotating question and answer selections.

9. Nonsampling errors are errors that result from undercoverage, nonresponse bias, response bias, and data-entry errors. These errors can occur even in a census. Sampling errors are errors that result from the use of a sample to estimate information about a population. These include random error and errors due to poor sampling plans, and result because samples contain incomplete information regarding a population.

10. The following are steps in conducting an experiment:

(1) *Identify the problem to be solved.*
Give direction and indicates the variables of interest (referred to as the claim).

(2) *Determine the factors that affect the response variable.*
List all variables that may affect the response, both controllable and uncontrollable.

(3) *Determine the number of experimental units.*
Determine the sample size. Use as many as time and money allow.

(4) *Determine the level of each factor.*
Factors can be controlled by fixing their level (e.g. only using men) or setting them at predetermined levels (e.g. different dosages of a new medicine). For factors that cannot be controlled, random assignment of units to treatments helps average out the effects of the uncontrolled factor over all treatments.

(5) *Conduct the experiment.*
Carry out the experiment using an equal number of units for each treatment. Collect and organize the data produced.

(6) *Test the claim.*
Analyze the collected data and draw conclusions.

11. “Number of new automobiles sold at a dealership on a given day” is quantitative because its values are numerical measures on which addition and subtraction can be performed with meaningful results. The variable is discrete because its values result from a count.

12. “Weight in carats of an uncut diamond” is quantitative because its values are numerical measures on which addition and subtraction can be performed with meaningful results. The variable is continuous because its values result from a measurement rather than a count.

13. “Brand name of a pair of running shoes” is qualitative because its values serve only to classify individuals based on a certain characteristic.

14. 73% is a statistic because it describes a sample (the 1011 people age 50 or older who were surveyed).

15. 70% is a parameter because it describes a population (all the passes completed by Cardale Jones in the 2015 Championship Game).

16. Birth year has the *interval* level of measurement since differences between values have meaning, but it lacks a true zero.

17. Marital status has the *nominal* level of measurement since its values merely categorize individuals based on a certain characteristic.

18. Stock rating has the *ordinal* level of measurement because its values can be placed in rank order, but differences between values have no meaning.

19. Number of siblings has the *ratio* level of measurement because differences between values have meaning and there is a true zero.

20. This is an observational study because no attempt was made to influence the variable of interest. Sexual innuendos and curse words were merely observed.

21. This is an experiment because the researcher intentionally imposed treatments (experimental drug vs. placebo) on individuals in a controlled setting.

22. This was a cohort study because participants were identified to be included in the study and then followed over a period of time with data being collected at regular intervals (every 2 years).

23. This is convenience sampling since the pollster simply asked the first 50 individuals she encountered.

24. This is a cluster sample since the ISP included all the households in the 15 randomly selected city blocks.

25. This is a stratified sample since individuals were randomly selected from each of the three grades.

26. This is a systematic sample since every 40th tractor trailer was tested using a random start with the 12th tractor trailer.

27. **(a)** Sampling bias; undercoverage or nonrepresentative sample due to a poor sampling frame. Cluster sampling or stratified sampling are better alternatives.

(b) Response bias due to interviewer error. A multilingual interviewer could reduce the bias.

(c) Data-entry error due to the incorrect entries. Entries should be checked by a second reader.

28. Answers will vary. Using a TI-84 Plus graphing calculator with a seed of 1990, and numbering the individuals from 1 to 21, we would select individuals numbered 14, 6, 10, 17, and 11. If we numbered the businesses down each column, the businesses selected would be Jiffy Lube, Nancy's Flowers, Norm's Jewelry, Risky Business Security, and Solus, Maria, DDS.

29. Answers will vary. The first step is to select a random starting point among the first 9 bolts produced. Using row 9, column 17 from Table I in Appendix A, he will sample the 3rd bolt produced, then every 9th bolt after that until a sample size of 32 is obtained. In this case, he would sample bolts 3, 12, 21, 30, and so on, until bolt 282.

30. Answers will vary. The goggles could be numbered 00 to 99, then a table of random digits could be used to select the numbers of the goggles to be inspected. Starting with row 12, column 1 of Table 1 in Appendix A and reading down, the selected labels would be 55, 96, 38, 85, 10, 67, 23, 39, 45, 57, 82, 90, and 76.

31. **(a)** To determine the ability of chewing gum to remove stains from teeth

(b) This is an experimental design because the teeth were separated into groups that were assigned different treatments.

(c) Completely randomized design

(d) Percentage of stain removed

(e) Type of stain remover (gum or saliva); Qualitative

(f) The 64 stained bovine incisors

(g) The chewing simulator could impact the percentage of the stain removed.

(h) Gum A and B remove significantly more stain.

32. **(a)** Matched-pairs

(b) Reaction time; Quantitative

(c) Alcohol consumption

(d) Food consumption; caffeine intake

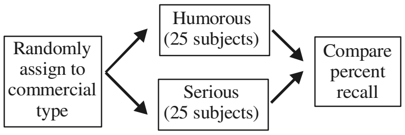
(e) Weight, gender, etc.

(f) To act as a placebo to control for the psychosomatic effects of alcohol

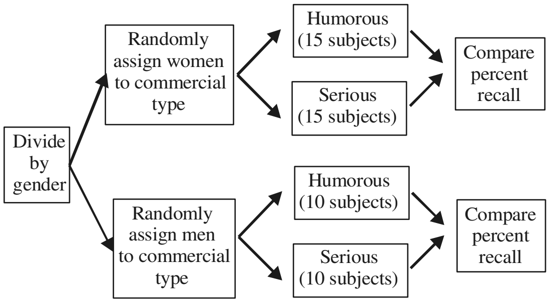
(g) Alcohol delays the reaction time significantly in seniors for low levels of alcohol consumption; healthy seniors that are not regular drinkers.

33. Answers will vary. Since there are ten digits (0 – 9), we will let a 0 or 1 indicate that (a) is to be the correct answer, 2 or 3 indicate that (b) is to be the correct answer, and so on. Beginning with row 1, column 8 of Table 1 in Appendix A, and reading downward, we obtain the following:
2, 6, 1, 4, 1, 4, 2, 9, 4, 3, 9, 0, 6, 4, 4,
8, 6, 5, 8, 5
Therefore, the sequence of correct answers would be:
b, d, a, c, a, c, b, e, c, b, e, a, d, c, c, e, d, c,
e, c

34. **(a)** Answers will vary. One possible diagram is shown below.



(b) Answers will vary. One possible diagram is shown below.



35. A matched-pairs design is an experimental design where experimental units are matched up so they are related in some way.

In a completely randomized design, the experimental units are randomly assigned to one of the treatments. The value of the response variable is compared for each treatment. In a matched-pairs design, experimental units are matched up on the basis of some common characteristic (such as husband-wife or twins). The differences between the matched units are analyzed.

36. Answers will vary.

37. Answers will vary.

38. Randomization is meant to even out the effect of those variables that are not controlled for in a designed experiment. Answers to the randomization question may vary; however, each experimental unit must be randomly assigned. For example, a researcher might randomly select 25 experimental units from the 100 units and assign them to treatment #1. Then the researcher could randomly select 25 from the remaining 75 units and assign them to treatment #2, and so on.

Chapter 1 Test

1. Collect information, organize and summarize the information, analyze the information to draw conclusions, provide a measure of confidence in the conclusions drawn from the information collected.

2. The process of statistics refers to the approach used to collect, organize, analyze, and interpret data. The steps are to
(1) identify the research objective,
(2) collect the data needed to answer the research question,
(3) describe the data, and
(4) perform inference.

3. The time to complete the 500-meter race in speed skating is quantitative because its values are numerical measurements on which addition and subtraction have meaningful results. The variable is continuous because its values result from a measurement rather than a count. The variable is at the *ratio* level of measurement because differences between values have meaning and there is a true zero.

4. Video game rating is qualitative because its values classify games based on certain characteristics but arithmetic operations have no meaningful results. The variable is at the *ordinal* level of measurement because its values can be placed in rank order, but differences between values have no meaning.

5. The number of surface imperfections is quantitative because its values are numerical measurements on which addition and subtraction have meaningful results. The variable is discrete because its values result from a count. The variable is at the *ratio* level of measurement because differences between values have meaning and there is a true zero.

6. This is an experiment because the researcher intentionally imposed treatments (brand-name battery versus plain-label battery) on individuals (cameras) in a controlled setting. The response variable is the battery life.

7. This is an observational study because no attempt was made to influence the variable of interest. Fan opinions about the asterisk were merely observed. The response variable is whether or not an asterisk should be placed on Barry Bonds' 756th homerun ball.

8. A *cross-sectional study* collects data at a specific point in time or over a short period of time; a *cohort study* collects data over a period of time, sometimes over a long period of time (prospective); a *case-controlled study* is retrospective, looking back in time to collect data.

9. An experiment involves the researcher actively imposing treatments on experimental units in order to observe any difference between the treatments in terms of effect on the response variable. In an observational study, the researcher observes the individuals in the study without attempting to influence the response variable in any way. Only an experiment will allow a researcher to establish causality.

10. A control group is necessary for a baseline comparison. This accounts for the placebo effect that says that some individuals will respond to any treatment. Comparing other treatments to the control group allows the researcher to identify which, if any, of the other treatments are superior to the current treatment (or no treatment at all). Blinding is important to eliminate bias due to the individual or experimenter knowing which treatment is being applied.

11. The steps in conducting an experiment are to (1) identify the problem to be solved, (2) determine the factors that affect the response variable, (3) determine the number of experimental units, (4) determine the level of each factor, (5) conduct the experiment, and (6) test the claim.

12. Answers will vary. The franchise locations could be numbered 01 to 15 going across. Starting at row 7, column 14 of Table I in Appendix, and working downward, the selected numbers would be 08, 11, 03, and 02. The corresponding locations would be Ballwin, Chesterfield, Fenton, and O'Fallon.

13. Answers will vary. Using the available lists, obtain a simple random sample from each stratum and combine the results to form the stratified sample. Start at different points in Table I or use different seeds in a random number generator. Using a TI-84 Plus graphing calculator with a seed of 14 for Democrats, 28 for Republicans, and 42 for Independents, the selected numbers would be
Democrats: 3946, 8856, 1398, 5130, 5531, 1703, 1090, and 6369
Republicans: 7271, 8014, 2575, 1150, 1888, 3138, and 2008
Independents: 945, 2855, and 1401

14. Answers will vary. Number the blocks from 1 to 2500 and obtain a simple random sample of size 10. The blocks corresponding to these numbers represent the blocks analyzed. All trees in the selected blocks are included in the sample. Using a TI-84 Plus graphing calculator with a seed of 12, the selected blocks would be numbered 2367, 678, 1761, 1577, 601, 48, 2402, 1158, 1317, and 440.

15. Answers will vary. $\frac{600}{14} \approx 42.86$, so we let $k = 42$. Select a random number between 1 and 42 that represents the first slot machine inspected. Using a TI-84 Plus graphing calculator with a seed of 132, we select machine 18 as the first machine inspected. Starting with machine 18, every 42nd machine thereafter would also be inspected (60, 102, 144, 186, …, 564).

16. In a completely randomized design, the experimental units are randomly assigned to one of the treatments. The value of the response variable is compared for each treatment.

17. **(a)** Sampling bias due to voluntary response

(b) Nonresponse bias due to the low response rate

(c) Response bias due to poorly worded questions.

(d) Sampling bias due to poor sampling plan (undercoverage)

18. **(a)** This experiment has a matched-pairs design.

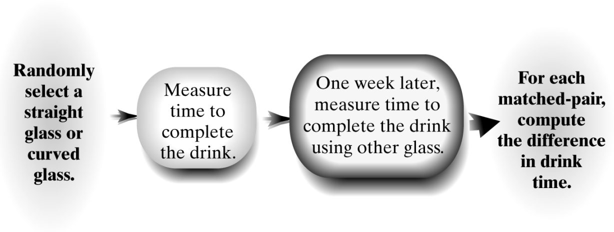
(b) The subjects are the 159 social drinkers who participated in the study.

(c) Treatments are the types of beer glasses (straight glass or curved glass).

(d) The response variable is the time to complete the drink; quantitative.

(e) The type of glass used in the first week is randomly determined. This is to neutralize the effect of drinking out of a specific glass first.

(f)


Randomly select a straight glass or curved glass.
Measure time to complete the drink.
One week later, measure time to complete the drink using other glass.
For each matched-pair, compute the difference in drink time.

19. **(a)** This experiment has a completely randomized design.

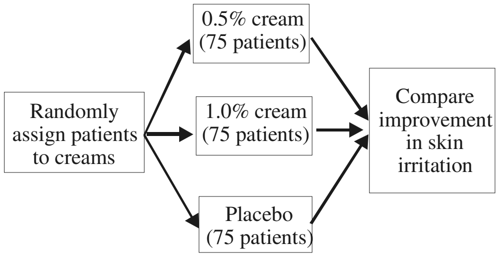
(b) The factor set to predetermined levels is the topical cream concentration. The treatments are 0.5% cream, 1.0% cream, and a placebo (0% cream).

(c) The study is double-blind if neither the subjects, nor the person administering the treatments, are aware of which topical cream is being applied.

(d) The control group is the placebo (0% topical cream).

(e) The experimental units are the 225 patients with skin irritations.

(f)


Randomly assign patients to creams
0.5% cream (75 patients)
1.0% cream (75 patients)
Placebo (75 patients)
Compare improvement in skin irritation

20. **(a)** This was a cohort study because participants were identified to be included in the study and then followed over a long period of time with data being collected at regular intervals (every 4 years).

(b) The response variable is bone mineral density. The explanatory variable is weekly cola consumption.

(c) The response variable is quantitative because its values are numerical measures on which addition and subtraction can be performed with meaningful results.

(d) The researchers observed values of variables that could potentially impact bone mineral density (besides cola consumption), so their effect could be isolated from the variable of interest.

(e) Answers will vary. Some possible lurking variables that should be accounted for are smoking status, alcohol consumption, physical activity, and calcium intake (form and quantity).

(f) The study concluded that women who consumed at least one cola per day (on average) had a bone mineral density that was significantly lower at the femoral neck than those who consumed less than one cola per day. The study cannot claim that increased cola consumption *causes* lower bone mineral density because it is only an observational study. The researchers can only say that increased cola consumption is *associated* with lower bone mineral density for women.

21. A confounding variable is an explanatory variable that cannot be separated from another explanatory variable. A lurking variable is an explanatory variable that was not considered in the study but affects the response variable in the study.

Chapter 2
Summarizing Data in Tables and Graphs

Section 2.1

1. Raw data are the data as originally collected, before they have been organized or coded.

2. Number (or count); proportion (or percent)

3. The relative frequencies should add to 1, although rounding may cause the answers to vary slightly.

4. A bar graph is used to illustrate qualitative data. It is a chart in which rectangles are used to illustrate the frequency or relative frequency with which a category appears. A Pareto chart is a bar chart with bars drawn in order of decreasing frequency or relative frequency.

5. **(a)** The largest segment in the pie chart is for "Washing your hands" so the most commonly used approach to beat the flu bug is washing your hands. 61% of respondents selected this as their primary method for beating the flu.

(b) The smallest segment in the pie chart is for "Drinking Orange Juice" so the least used method is drinking orange juice. 2% of respondents selected this as their primary method for beating the flu.

(c) 25% of respondents felt that flu shots were the best way to beat the flu.

6. **(a)** $\frac{128{,}000}{1{,}350{,}000} \approx 0.0948$; approximately 9.48% of cosmetic surgeries in 2009 were for tummy tucks.

(b) $\frac{138{,}000}{1{,}350{,}000} \approx 0.102$; approximately 10.2% of cosmetic surgeries in 2009 were for nose reshaping.

(c) The graph accounts for 312,000 + 284,000 + 150,000 + 138,000 + 128,000 = 1,012,000 surgeries. Thus, 1,350,000 − 1,012,000 = 338,000 surgeries are not accounted for in the graph.

7. **(a)** The highest bar corresponds to the position OF (outfield), so OF is the position with the most MVPs.

(b) The bar for first base (1B) reaches the line for 15. Thus, there were 15 MVPs who played first base.

(c) The bar for outfield (OF) is 30 on the vertical axis. The bar for first base (1B) reaches 15. Since 30 – 15 = 15, there were 15 more MVPs who played outfield than first base.

(d) Each of the three outfield positions should be reported as MVPs, rather than treating the three positions as one position.

8. **(a)** 29,936,000 whites were living in poverty.

(b) 12745/(29936 + 11041 + 12745 + 1974) = 0.229 = 22.9%
In 2013, about 22.9% of the impoverished in the United States were Hispanic.

(c) This graph should use relative frequencies, rather than frequencies. The graph does not account for the different population size of each ethnic group. Without knowing the population sizes, we cannot determine whether a group is disproportionally impoverished.

9. **(a)** 69% of the respondents believe divorce is morally acceptable.

(b) 23% believe divorce is morally wrong. So, 240 million · 0.23 = 55.2 million adult Americans believe divorce is morally wrong.

(c) This statement is inferential, since it is a generalization based on the observed data.

10. **(a)** 5% of identity theft was loan fraud.

(b) 26% of the identity fraud cases in a recent year involved credit card fraud. So, 10 million · 0.26 = 2.6 million cases of credit card fraud occurred in a recent year.

11. **(a)** The proportion of 18–34 year old respondents who are more likely to buy when made in America is 0.42. For 34–44 year olds, the proportion is 0.61.

(b) The 55+ age group has the greatest proportion of respondents who are more likely to buy when made in America.

(c) The 18–34 age group has a majority of respondents who are less likely to buy when made in America.

(d) As age increases, so does the likelihood that a respondent will be more likely to buy a product that is made in America.

12. **(a)** The proportion of males who would like to be richer is 0.46. The proportion of females who would like to be richer is 0.41.

(b) The attribute that females desire more than males is to be thinner.

(c) The attribute that males prefer over females two-to-one is to be younger.

(d) Equal proportions of males and females desire to be smarter.

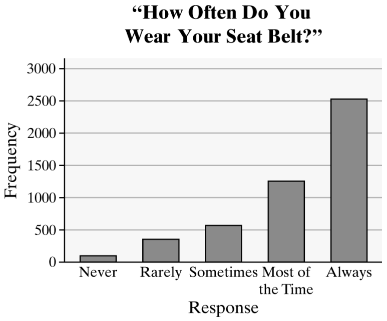
13. **(a)** Total students surveyed = 125 + 324 + 552 + 1257 + 2518 = 4776
Relative frequency of "Never"
$= 125/4776 \approx 0.0262,$ and so on.

Response	Relative Frequency
Never	0.0262
Rarely	0.0678
Sometimes	0.1156
Most of the time	0.2632
Always	0.5272

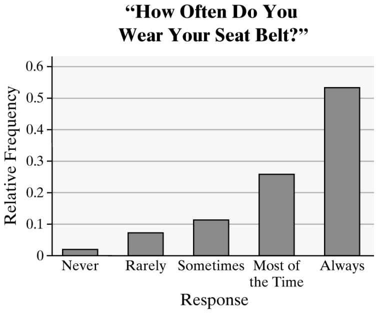
(b) 52.72%

(c) $0.0262 + 0.0678 = 0.0940$ or 9.40%

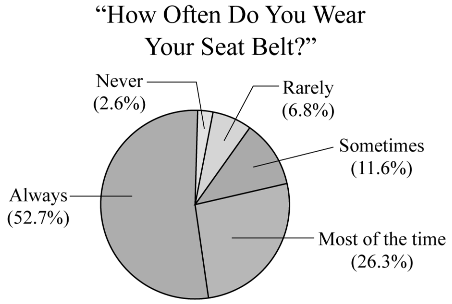
(d)



(e)



(f)



(g) This is a descriptive statement because it is reporting a result of the sample.

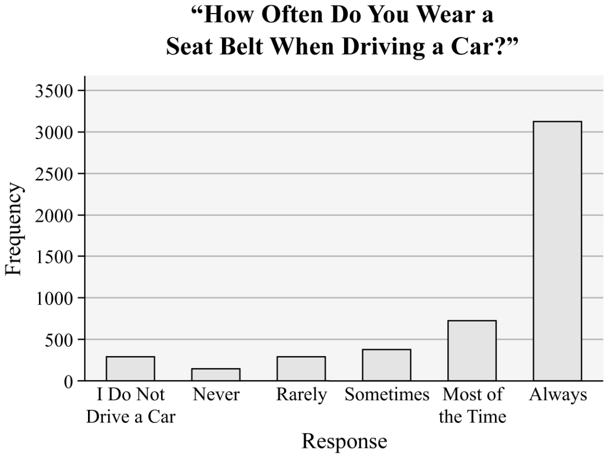
14. **(a)** Total students surveyed = 249 + 118 + 249 + 345 + 716 + 3093 = 4770
Relative frequency of " I do not drive"
$= \dfrac{249}{4770} \approx 0.0522,$ and so on.

Response	Relative Frequency
I do not drive	0.0522
Never	0.0247
Rarely	0.0522
Sometimes	0.0723
Most of the time	0.1501
Always	0.6484

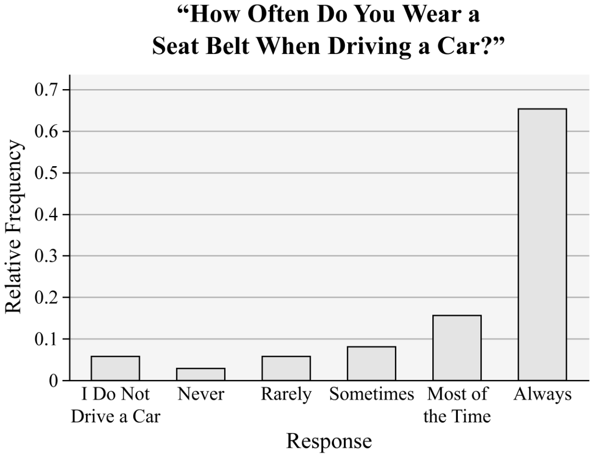
(b) 64.84%

(c) $0.0247 + 0.0522 = 0.0769$ or 7.69%

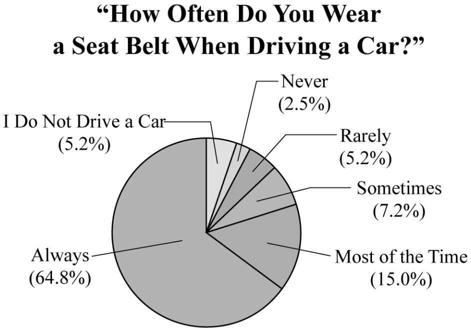
(d)



(e)



(f)



(g) Total students = 118 + 249 + 345 + 716 + 3093 = 4521
Relative frequency of "Never"
$= \dfrac{118}{4521} \approx 0.0261,$ and so on.

Response	Relative Frequency
Never	0.0261
Rarely	0.0551
Sometimes	0.0763
Most of the time	0.1584
Always	0.6841

The relative frequencies of all categories are very similar except that students are more likely to wear their seatbelt 'Always' when driving.

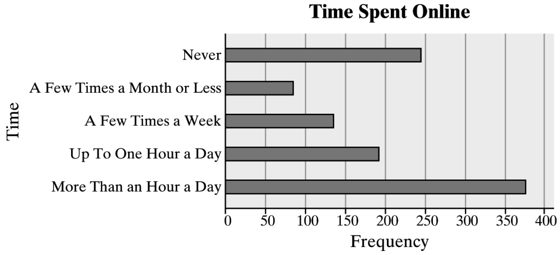
(h) The statement is descriptive because it is describing the particular sample.

15. (a) Total adults surveyed = 377 + 192 + 132 + 81 + 243 = 1025
Relative frequency of "More than 1 hour a day" = $377/1025 \approx 0.3678,$ and so on.

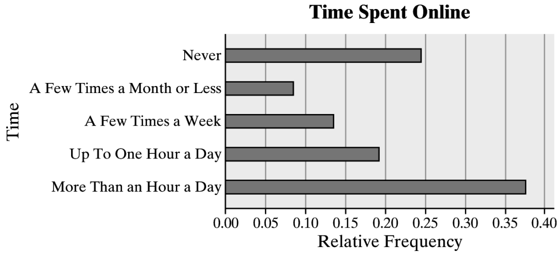
Response	Relative Frequency
More than 1 hr a day	0.3678
Up to 1 hr a day	0.1873
A few times a week	0.1288
A few times a month or less	0.0790
Never	0.2371

(b) 0.2371 (about 24%)

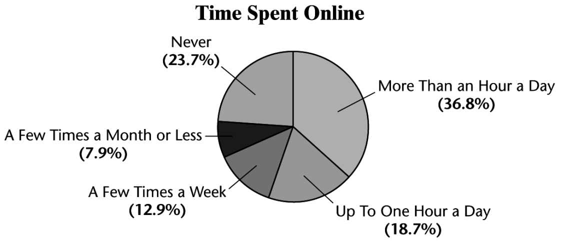
(c)



(d)



(e)



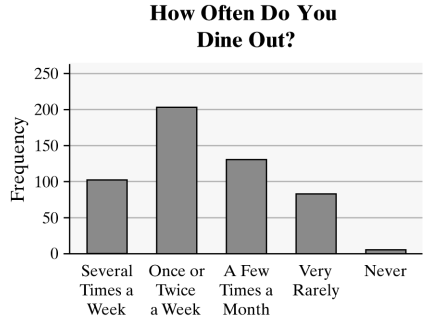
(f) The statement provides an estimate, but no level of confidence is given.

16. **(a)** Total adults surveyed = 103 + 204 + 130 + 79 + 5 = 521
Relative frequency of "Several times a week" = $\dfrac{103}{521} \approx 0.197$, and so on.

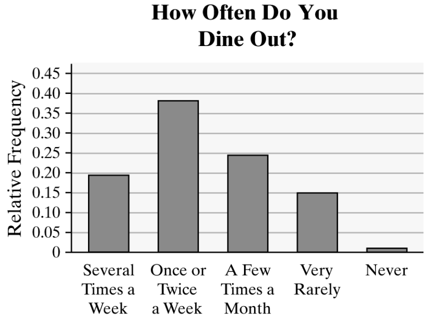
Response	Relative Frequency
Several times a week	0.1977
Once or twice a week	0.3916
A few times a month	0.2495
Vary rarely	0.1516
Never	0.0096

(b) The proportion surveyed who dine out once or twice a week is 204/(103 + 204 + 130 + 79 + 5) = 0.392

(c)



(d)



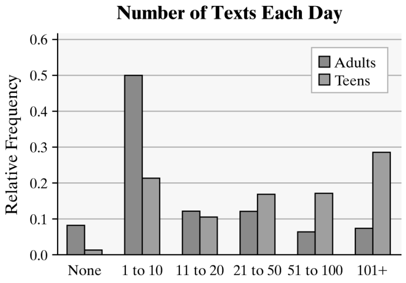
17. **(a)** Total adults = 1936
Relative frequency for "none" is:
173/1936 = 0.09, and so on.

Number of Texts	Rel. Freq. (Adults)
None	0.0894
1 to 10	0.5052
11 to 20	0.1286
21 to 50	0.1286
51 to 100	0.0692
101+	0.0790

(b) Total teens = 627
Relative frequency for "none" is:
13/627 = 0.021, and so on.

Number of Texts	Rel. Freq. (Teens)
None	0.0207
1 to 10	0.2201
11 to 20	0.1100
21 to 50	0.1802
51 to 100	0.1802
101+	0.2887

(c)



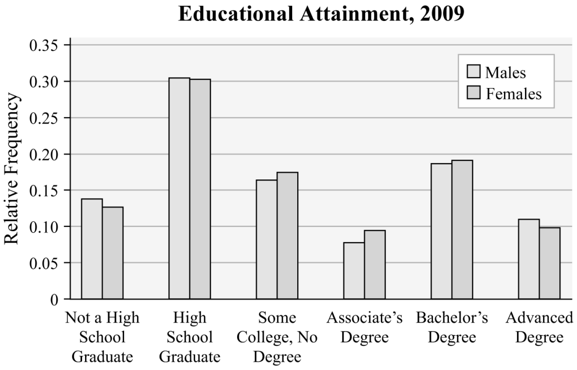
(d) Answers will vary. Adults are much more likely to send fewer texts per day, while teens are much more likely to do more texting.

18. **(a), (b)**
Total males = 99.4 million
Relative frequency for "Not HS graduate" is 12.3/99.4 = 0.124, and so on.

Total females = 107.6 million
Relative frequency for "Not HS graduate" is 12.2/107.6 = 0.113, and so on.

Educational Attainment	Males	Females
Not a HS graduate	0.1237	0.1134
High school graduate	0.3018	0.2946
Some college, no degree	0.1660	0.1701
Associate's degree	0.0885	0.1078
Bachelor's degree	0.2002	0.2017
Advanced degree	0.1197	0.1125

(c)



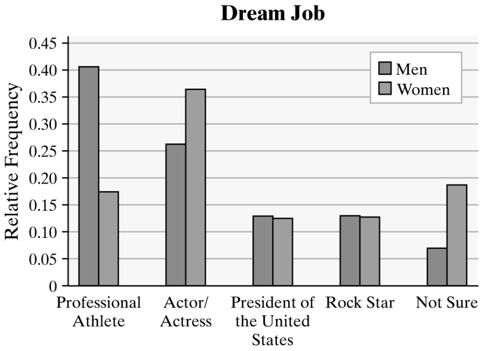
(d) Answers will vary. It appears that females are slightly more likely to start, but not finish college. Males appear to be slightly more likely to attain an advanced degree.

19. **(a)** Total males = 99; Relative frequency for "Professional Athlete" is 40/99 = 0.404, and so on.

Total number of females = 100; Relative frequency for "Professional Athlete" is 18/100 = 0.18, and so on.

Dream Job	Men	Women
Professional Athlete	0.4040	0.180
Actor/Actress	0.2626	0.370
President of the United States	0.1313	0.130
Rock Star	0.1313	0.130
Not Sure	0.0707	0.190

(b)

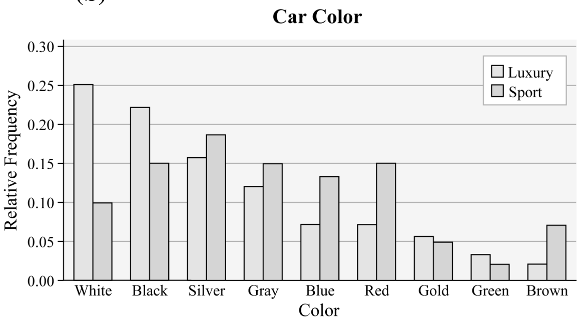


(c) Answers will vary. Males are much more likely to want to be a professional athlete. Women are more likely to aspire to a career in acting than men. Men's desire to become athletes may be influenced by the prominence of male sporting figures in popular culture. Women may aspire to careers in acting due to the perceived glamour of famous female actresses.

20. **(a)** Relative frequency for "White" luxury cars $= \dfrac{25}{100} = 0.25,$ and so on.
Relative frequency for "White" sport cars $= \dfrac{10}{100} = 0.10,$ and so on.

	Relative Frequencies	
Color	Luxury Cars	Sport Cars
White	0.25	0.10
Black	0.22	0.15
Silver	0.16	0.18
Gray	0.12	0.15
Blue	0.07	0.13
Red	0.07	0.15
Gold	0.06	0.05
Green	0.03	0.02
Brown	0.02	0.07

(b)

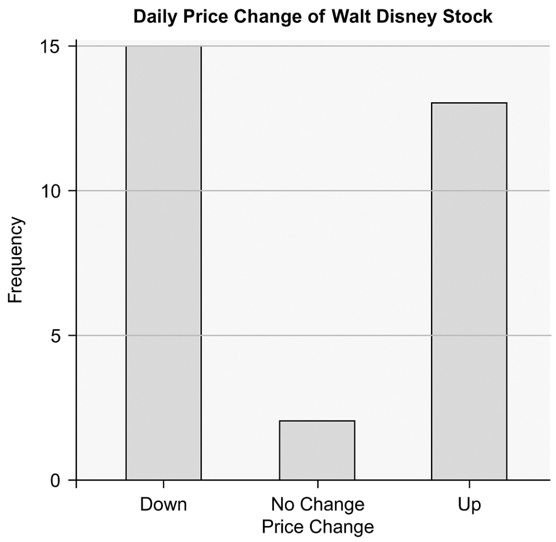


(c) Answers will vary. White is the most popular color for luxury cars, while silver is the most popular for sports cars. People who drive luxury cars may enjoy the clean look of a white vehicle. People who drive sports cars may prefer the flashier look of silver.

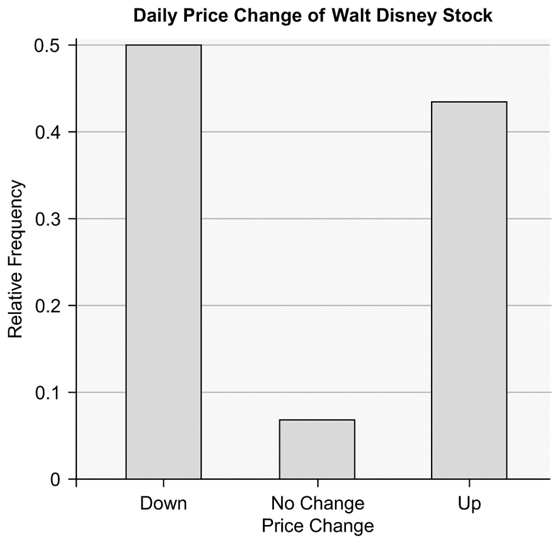
21. **(a)**, **(b)**
Total number of Trading Days = 30; relative frequency for Down is 15/30 = 0.5, and so on.

Price Change	Freq.	Rel. Freq.
Down	15	0.500
No Change	2	0.067
Up	13	0.433

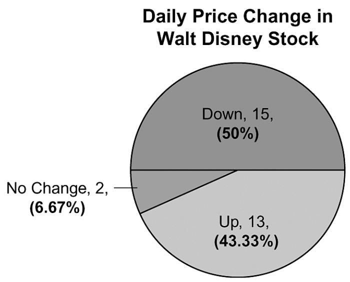
(c)



(d)



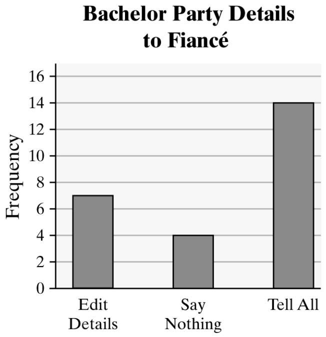
(e)



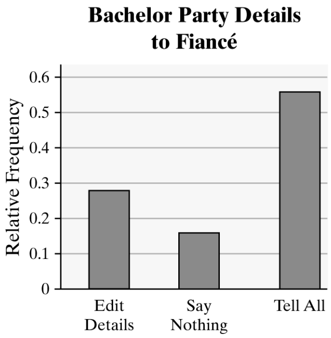
22. **(a)**, **(b)** Total number of responses = 25; relative frequency for "edit details" is 7/25 = 0.28.

Response	Freq.	Rel. Freq.
Edit details	7	0.28
Say nothing	4	0.16
Tell all	14	0.56

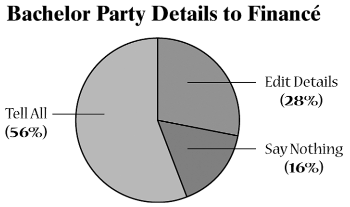
(c)



(d)



(e)



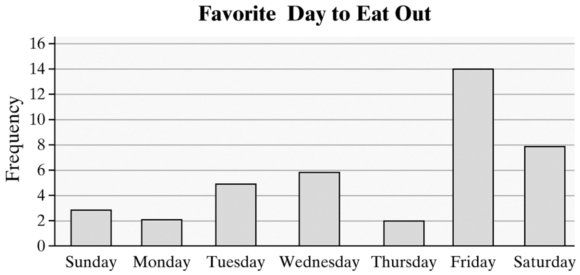
23. **(a)**, **(b)**

Total number of responses = 40; relative frequency for "Sunday" is 3/40 = 0.075.

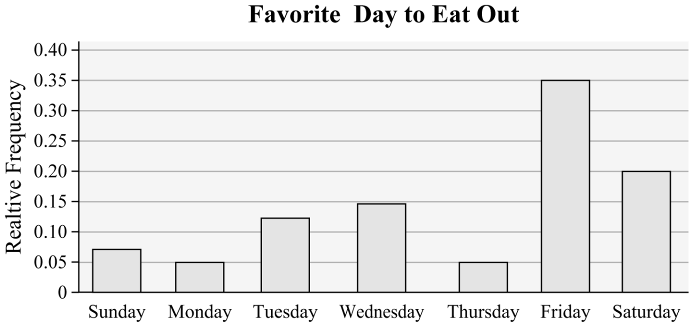
Day	Freq.	Rel. Freq.
Sunday	3	0.075
Monday	2	0.050
Tuesday	5	0.125
Wednesday	6	0.150
Thursday	2	0.050
Friday	14	0.350
Saturday	8	0.200

(c) Answers will vary. If you own a restaurant, you will probably want to advertize on the days when people will be most likely to order takeout: Friday. You might consider avoiding placing an ad on Monday and Thursday, since the readers are least likely to choose to order takeout on these days.

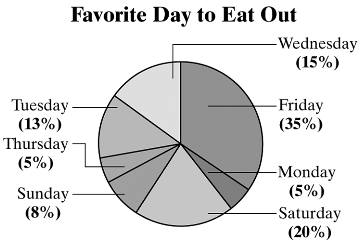
(d)



(e)



(f)



24. **(a)**, **(b)**

Total number of patients = 50
Relative frequency for "Type A"

$$= \frac{18}{50} = 0.36, \text{ and so on.}$$

Blood Type	Freq.	Rel. Freq.
A	18	0.36
AB	4	0.08
B	6	0.12
O	22	0.44

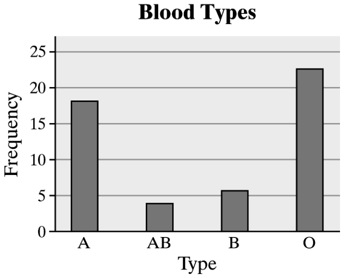
(c) Type O is the most common.

(d) Type AB is the least common.

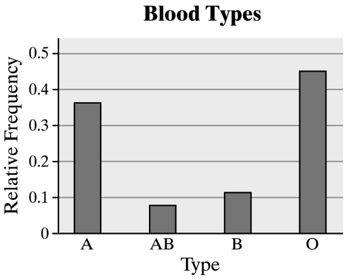
(e) We estimate that 44% of the population has type O blood. This is considered inferential statistics because a conclusion about the population is being drawn based on sample data.

(f) Answers will vary; in 2008 the Red Cross reported that 45% of the population had type O blood (either + or –). Results will differ because of sampling variability.

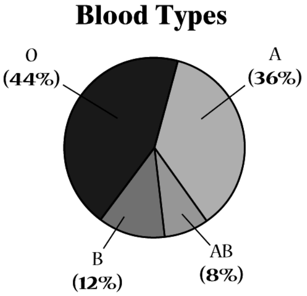
(g)



(h)



(i)



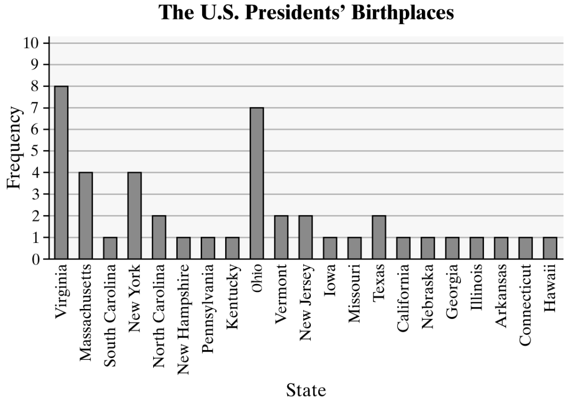
25. **(a)**

State	AR	CA	CT	GA	HI	IL
Freq.	1	1	1	1	1	1

State	IA	KY	MA	MO	NE
Freq.	1	1	4	1	1

State	NH	NJ	NY	NC	OH
Freq.	1	2	4	2	7

State	PA	SC	TX	VT	VA
Freq.	1	1	2	2	8

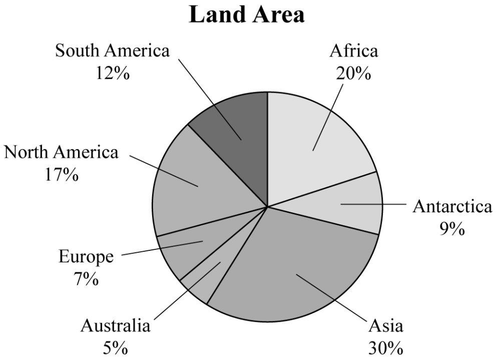


(b) More presidents were born in Virginia than in any other state.

(c) Answers will vary. The data do not take the year of statehood into account. For example, Virginia has been a state for roughly 62 years more than California. The population of the United States was more concentrated in the east in the early years, so it was more likely that the president would be from that part of the country.

26. **(a)** It would make sense to draw a pie chart for land area since the 7 continents contain all the land area on Earth.
Total land area is 11,608,000 + 5,100,000 + … + 9,449,000 + 6,879,000 = 57,217,000 square miles.
The relative frequency (percentage) for Africa is $\frac{11,608,000}{57,217,000} = 0.2029$.

Continent	Land Area (mi^2)	Rel. Freq.
Africa	11,608,000	0.2029
Antarctica	5,100,000	0.0891
Asia	17,212,000	0.3008
Australia	3,132,000	0.0547
Europe	3,837,000	0.0671
North America	9,449,000	0.1651
South America	6,879,000	0.1202



(b) It would not make sense to draw a pie chart for the highest elevation because there is no whole to which to compare the parts.

27. Answers will vary.

28. Answers will vary.

29. **(a)** The researcher wants to determine if online homework improves student learning over traditional pencil-and-paper homework.

(b) This study is an experiment because the researcher is actively imposing treatments (the homework style) on subjects.

(c) Answers will vary. Some examples are same teacher, same semester, and same course.

(d) Assigning different homework methods to entire classes could confound the results because there may be differences between the classes. The instructor may give more instruction to one class than the other. The instructor is not blinded, so he or she may treat one group differently from the other.

(e) *Number of students*: quantitative, discrete
Average age: quantitative, continuous
Average exam score: quantitative, continuous
Type of homework: qualitative
College experience: qualitative

(f) Letter grade is a qualitative variable at the ordinal level of measurement.
Answers will vary. It is possible that ordering the data from A to F is better because it might give more "weight" to the higher grade and the researcher wants to show that a higher percent of students passed using the online homework.

(g) The graph being displayed is a side-by-side relative frequency bar graph.

(h) Yes; the "whole" is the set of students who received a grade for the course for each homework method.

(i) The table shows that the two groups with no prior college experience had roughly the same average exam grade. From the bar graph, we see that the students using online homework had a lower percent for As, but had a higher percent who passed with a C or better.

30. Relative frequencies should be used when the size of two samples or populations differ.

31. Answers will vary. If the goal is to illustrate the levels of importance, then arranging the bars in a bar chart in decreasing order makes sense. Sometimes it is useful to arrange the categorical data in a bar chart in alphabetical order. A pie chart does not readily allow for arranging the data in order.

32. A bar graph is preferred when trying to compare two specific values. Pie charts are helpful for comparing parts of a whole. A pie chart cannot be drawn if the data do not include all possible values of the qualitative variable.

33. No, the percentages do not sum to 100%.

Section 2.2

1. classes

2. lower; upper

3. class width

4. Skewed left means that the left tail is longer than the right tail.

5. True

6. False

7. False. The distribution shape shown is skewed right.

8. False. The distribution shape is bell-shaped.

9. **(a)** The value with the highest frequency is 8.

(b) The value with the lowest frequency is 2.

(c) The value of 7 was observed 15 times.

(d) The value of 5 was observed 11 times and the value of 4 was observed 7 times. Therefore, the value of 5 was observed 4 more times than the value of 4 (e.g. $11-7=4$).

(e) $\frac{15}{100}=0.15$ or 15% of the time a 7 was observed.

(f) The distribution is approximately bell-shaped.

10. **(a)** The most frequent number of cars sold in a week was 4 cars.

(b) There were 9 weeks in which 2 cars sold.

(c) Total frequency = 4 + 2 + 9 + 8 + 12 + 8 + 5 + 2 + 1 + 1 = 52 (as required)
Percentage of time two cars are sold
$= \frac{9}{52}\cdot 100 = 17.3\%$

(d) Slightly skewed to the right

11. **(a)** Total frequency = 2 + 3 + 13 + 42 + 58 + 40 + 31 + 8 +2 + 1 = 200

(b) 10 (e.g. 70 – 60 = 10)

(c)

IQ Score (class)	Frequency
60–69	2
70–79	3
80–89	13
90–99	42
100–109	58
110–119	40
120–129	31
130–139	8
140–149	2
150–159	1

(d) The class "100 – 109" has the highest frequency.

(e) The class "150 – 159" has the lowest frequency.

(f) $\frac{8+2+1}{200} = 0.055 = 5.5\%$

(g) No, there were no IQs above 159.

12. **(a)** The class width is 200 (e.g. 200 – 0 = 200).

(b) 0–199, 200–399, 400–599, 600–799, 800–999, 1000–1199, 1200–1399

(c) The highest frequency is in class 0–199.

(d) The distribution is skewed right.

(e) Answers will vary. The statement is incorrect because they are comparing counts from populations of different size. To make a fair comparison, the reporter should use rates of fatalities such as the number of fatalities per 1000 residents.

13. **(a)** Likely skewed right. Most household incomes will be to the left (perhaps in the $50,000 to $150,000 range), with fewer higher incomes to the right (in the millions).

(b) Likely bell-shaped. Most scores will occur near the middle range, with scores tapering off equally in both directions.

(c) Likely skewed right. Most households will have, say, 1 to 4 occupants, with fewer households having a higher number of occupants.

(d) Likely skewed left. Most Alzheimer's patients will fall in older-aged categories, with fewer patients being younger.

14. **(a)** Likely skewed right. More individuals would consume fewer alcoholic drinks per week, while less individuals would consume more alcoholic drinks per week.

(b) Likely uniform. There will be approximately an equal number of students in each age category.

(c) Likely skewed left. Most hearing-aid patients will fall in older-aged categories, with fewer patients being younger.

(d) Likely bell-shaped. Most heights will occur, say, in the 66- to 70-inch range, with heights tapering off equally in both directions.

15. **(a)** From the graph, it appears the unemployment rate in 2011 was about 9%.

(b) The highest unemployment rate was about 9.8%. This occurred in 2010.

(c) The highest inflation rate was about 4.3%. This occurred in 2008.

(d) The unemployment rate and inflation rate were closest in 2001. The unemployment rate and inflation rate were furthest in 2009.

(e) The misery index for 1999 was approximately 4.2 + 1.8 = 6. The misery index for 2014 was approximately 6.5 + 1.5 = 8. According to the misery index, the year 2014 was more "miserable" than the year 1999.

(f) Since 2010, the misery index has been declining due to the decreases in unemployment each year.

16. **(a)** To the nearest year, the average age of a man who first married in 1980 was 25.

(b) To the nearest year, the average age of a woman who first married in 1960 was 21.

(c) The largest difference in the average age of men and women at which they first married occurred in 1950. The approximate age difference was 24 – 20.5 = 3.5 years.

(d) The least amount of difference in the average age of men and women at which they first married occurred in 2000. The approximate age difference was 26.5 – 25.2 = 1.3 years.

17. (a) Total number of households = $16+18+12+3+1=50$

Relative frequency of 0 children = 16/50 = 0.32, and so on.

Number of Children Under Five	**Relative Frequency**
0	0.32
1	0.36
2	0.24
3	0.06
4	0.02

(b) $\frac{12}{50} = 0.24$ or 24% of households have two children under the age of 5.

(c) $\frac{18+12}{50} = \frac{30}{50} = 0.6$ or 60% of households have one or two children under the age of 5.

18. (a) Total number of free throws = $16+11+9+7+2+3+0+1+0+1=50$. Relative frequency of 1 throw until a miss = 16/50 = 0.32, and so on.

Number of Free Throws Until a Miss	**Relative Frequency**
1	0.32
2	0.22
3	0.18
4	0.14
5	0.04
6	0.06
7	0.00
8	0.02
9	0.00
10	0.02

(b) $\frac{7}{50} = 0.14$; 14% of the time she first missed on the fourth try.

(c) $\frac{1}{50} = 0.02$; 2% of the time she first missed on the tenth try.

(d) "At least 5" means that the basketball player misses on the 6th shot or 7th shot or 8th, etc. $\frac{3+0+1+0+1}{50} = \frac{5}{50} = 0.10$ or 10% of the time.

19. From the legend, 1|0 represents 10, so the original data set is 10, 11, 14, 21, 24, 24, 27, 29, 33, 35, 35, 35, 37, 37, 38, 40, 40, 41, 42, 46, 46, 48, 49, 49, 53, 53, 55, 58, 61, 62.

20. From the legend, 24|0 represents 240, so the original data set is 240, 244, 247, 252, 252, 253, 259, 259, 263, 264, 265, 268, 268, 269, 270, 271, 271, 273, 276, 276, 282, 283, 288.

21. From the legend, 1|2 represents 1.2, so the original data set is 1.2, 1.4, 1.6, 2.1, 2.4, 2.7, 2.7, 2.9, 3.3, 3.3, 3.3, 3.5, 3.7, 3.7, 3.8, 4.0, 4.1, 4.1, 4.3, 4.6, 4.6, 4.8, 4.8, 4.9, 5.3, 5.4, 5.5, 5.8, 6.2, 6.4.

22. From the legend, 12|3 represents 12.3, so the original data set is 12.3, 12.7, 12.9, 12.9, 13.0, 13.4, 13.5, 13.7, 13.8, 13.9, 13.9, 14.2, 14.4, 14.4, 14.7, 14.7, 14.8, 14.9, 15.1, 15.2, 15.2, 15.5, 15.6, 16.0, 16.3.

23. (a) There are six classes.

(b) Lower class limits: 10, 14, 18, 22, 26, 30
Upper class limits: 13.9, 17.9, 21.9, 25.9, 29.9, 33.9

(c) The class width can be found by subtracting consecutive lower class limits. For example, 14 – 10 = 4. Therefore, the class width is 4 (players).

24. (a) There are eight classes.

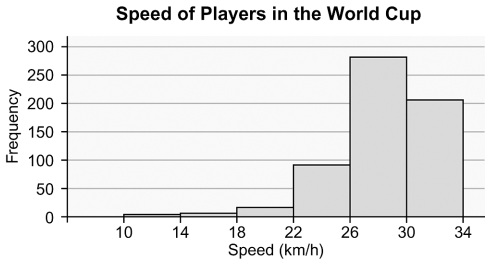
(b) Lower class limits: 0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0
Upper class limits: 0.9, 1.9, 2.9, 3.9, 4.9, 5.9, 6.9, 7.9

(c) The class width can be found by subtracting consecutive lower class limits. For example, $2.0-1.0=1.0$. Therefore, the class width is 1.0.

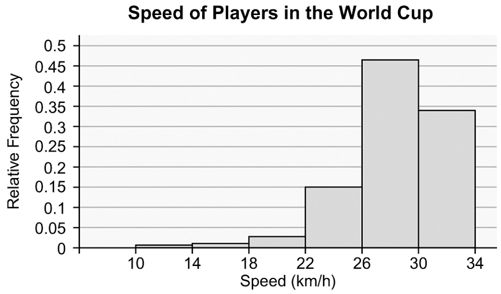
25. **(a)** Total frequency =
4 + 7 + 17 + 91+ 282 + 206 = 607
Relative frequency for 10–13.9 is
4/607 = 0.0066, and so on.

Speed (Km/hr)	Relative Frequency
10–13.9	0.0066
14–17.9	0.0115
18–21.9	0.0280
22–25.9	0.1499
26–29.9	0.4646
30–33.9	0.3394

(b)


Speed of Players in the World Cup

(c)

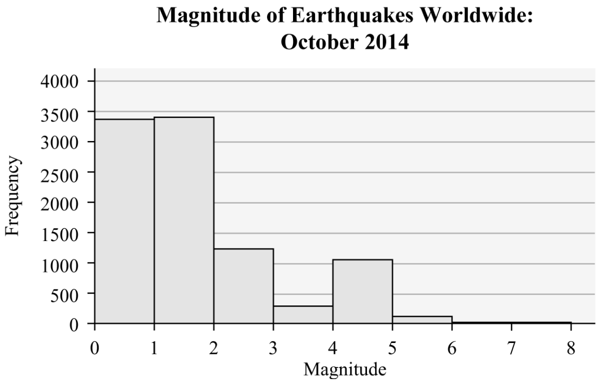

Speed of Players in the World Cup

The percentage of players who had a top speed between 30 and 33.9 km/h is 33.94%. The percent of players who had a top speed less than 13.9 km/h is 0.66%.

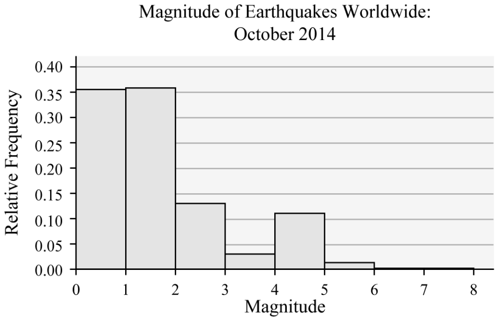
26. **(a)** Total frequency =
3371 + 3400 + 1237 + 286 + 1045 + 121
+ 7 + 2 = 9469
Relative frequency for 0–0.9 is
3371/9469 = 0.3561, and so on.

Magnitude	Relative Frequency
0–0.9	0.3560
1.0–1.9	0.3591
2.0–2.9	0.1306
3.0–3.9	0.0302
4.0–4.9	0.1104
5.0–5.9	0.0128
6.0–6.9	0.0007
.0–7.9	0.0002

(b)


Magnitude of Earthquakes Worldwide: October 2014

(c)


Magnitude of Earthquakes Worldwide: October 2014

The percentage of earthquakes that registered between 4.0 and 4.9 km/h is 11.04%. The percent of earthquakes that registered 4.9 or less is 98.63%.

27. **(a)** The data are discrete. The possible values for the number of color televisions in a household are countable.

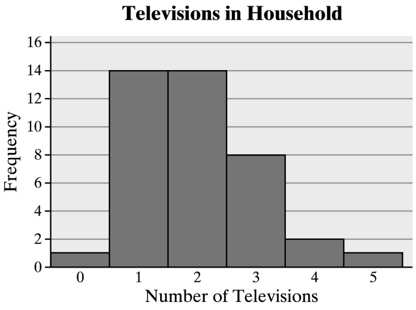
(b), **(c)**
The relative frequency for 0 color televisions is 1/40 = 0.025, and so on.

Number of TVs	Frequency	Relative Frequency
0	1	0.025
1	14	0.350
2	14	0.350
3	8	0.200
4	2	0.050
5	1	0.025

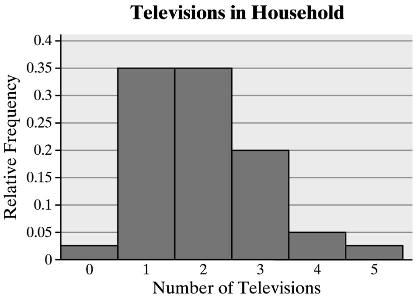
(d) The relative frequency is 0.2, so 20% of the households surveyed had 3 televisions.

(e) $0.05 + 0.025 = 0.075$
7.5% of the households in the survey had 4 or more televisions.

(f)


Televisions in Household

(g)


Televisions in Household

(h) The distribution is skewed right.

28. **(a)** The data are discrete. The possible values for the number of customers waiting for a table are countable.

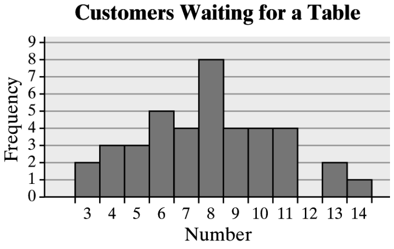
(b), **(c)**
Relative frequency of 3 customers waiting = 2/40 = 0.05, and so on.

Number of Customers	Freq.	Rel. Freq.
3	2	0.050
4	3	0.075
5	3	0.075
6	5	0.125
7	4	0.100
8	8	0.200
9	4	0.100
10	4	0.100
11	4	0.100
12	0	0.000
13	2	0.050
14	1	0.025

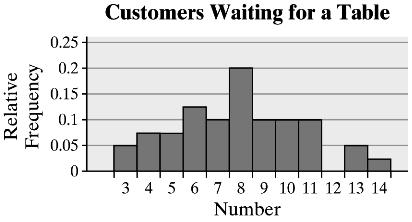
(d) $10.0 + 10.0 + 0.0 + 5.0 + 2.5 = 27.5\%$ of the Saturdays had 10 or more customers waiting for a table at 6 p.m.

(e) $5.0 + 7.5 + 7.5 = 20.0\%$ of the Saturdays had 5 or fewer customers waiting for a table at 6 p.m.

(f)


Customers Waiting for a Table

(g)

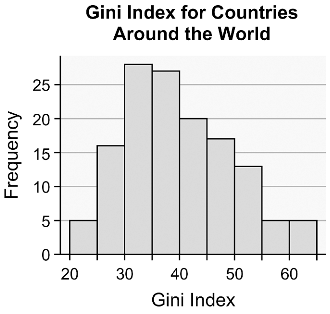

Customers Waiting for a Table

(h) The distribution is more or less symmetric.

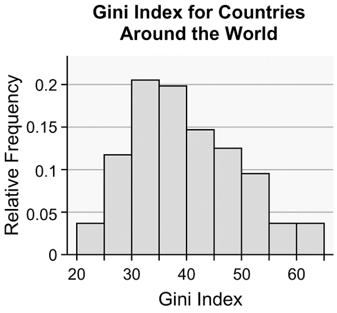
29. **(a), (b)** Relative frequency of a Gini Index of 20–24.9 = 5/136 = 0.037, and so on.

Gini Index	Freq.	Rel. Freq.
20–24.9	5	0.037
25–29.9	16	0.118
30–34.9	28	0.206
35–39.9	27	0.199
40–44.9	20	0.147
45–49.9	17	0.125
50–54.9	13	0.096
55–59.9	5	0.037
60–64.9	5	0.037

(c)



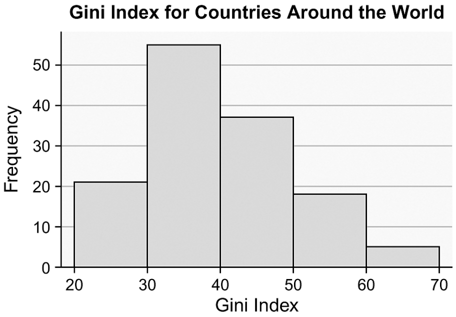
(d)

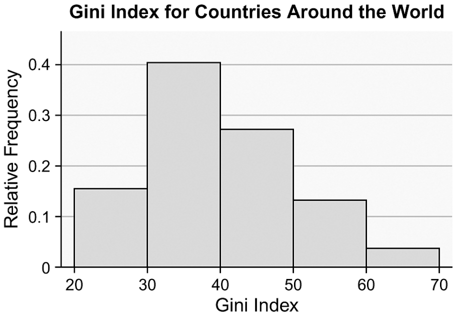


(e) The shape of the distribution is skewed right.

(f) Relative frequency of a Gini Index of 20–29.9 = 21/136 = 0.154, and so on.

Gini Index	Freq.	Rel. Freq.
20–29.9	21	0.154
30–39.9	55	0.404
40–49.9	37	0.272
50–59.9	18	0.132
60–69.9	5	0.037





The shape of the distribution is skewed right.

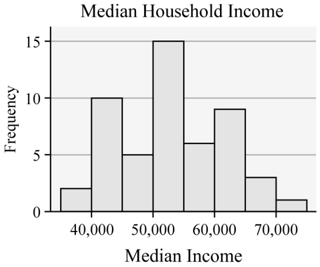
(g) Answers will vary. The graph with a class width of 5 provides more detail, so it seems to be a superior graph.

30. **(a), (b)**

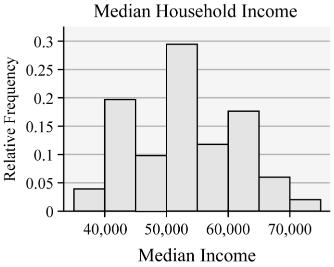
Relative frequency for the median income 35,000–39,999 is 2/51 = 0.0392, and so on.

Income	Freq.	Rel. Freq.
35,000–39,999	2	0.0392
40,000–44,999	10	0.1961
45,000–49,999	5	0.0980
50,000–54,999	15	0.2941
55,000–59,999	6	0.1176
60,000–64,999	9	0.1765
65,000–69,999	3	0.0588
70,000–74,999	1	0.0196

(c)



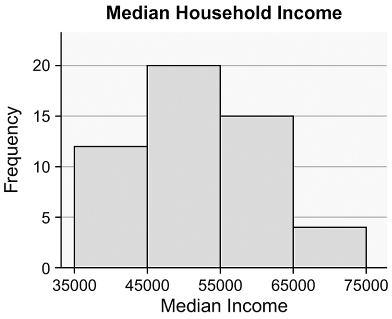
(d)

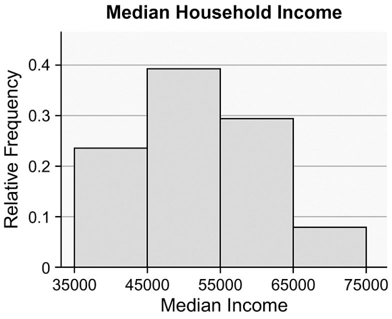


(e) The shape of the distribution is approximately symmetric.

(f) Relative frequency for the median income 35,000–44,999 is
12/51 = 0.2353, and so on.

Income	Freq.	Rel. Freq.
35,000–44,999	12	0.2353
45,000–54,999	20	0.3922
55,000–64,999	15	0.2941
65,000–74,999	4	0.0784





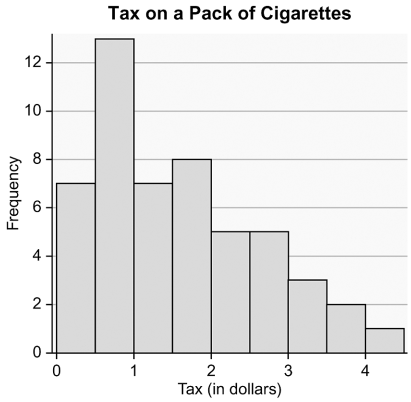
Distribution appears to be approximately symmetric, however, one could argue it is slightly skewed to the right (since the tail on the right is longer than the tail on the left).

(g) Answers will vary, but the graph with a class width of $5000 seems to show more details about the data so it seems better.

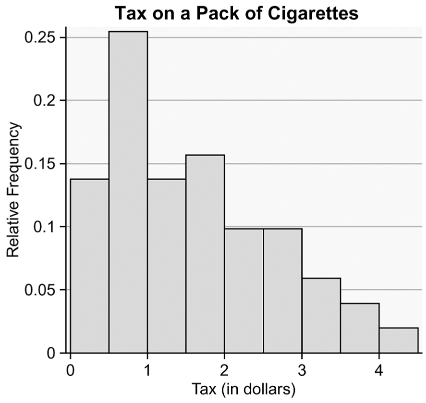
31. (a), (b)
Total number of data points = 51
Relative frequency of 0–0.499 is
7/51 = 0.1373, and so on.

Cigarette Tax	Frequency	Relative Frequency
0.00–0.499	7	0.1373
0.50–0.999	13	0.2549
1.00–1.499	7	0.1373
1.50–1.999	8	0.1569
2.00–2.499	5	0.0980
2.50–2.999	5	0.0980
3.00–3.499	3	0.0588
3.50–3.999	2	0.0392
4.00–4.499	1	0.0196

(c)



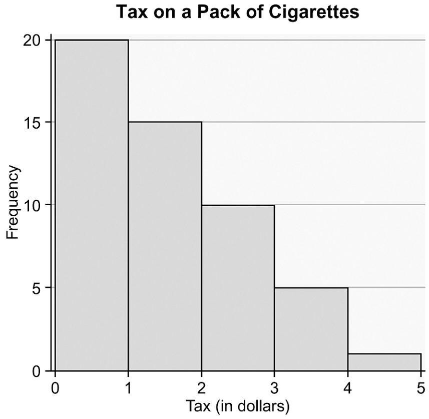
(d)

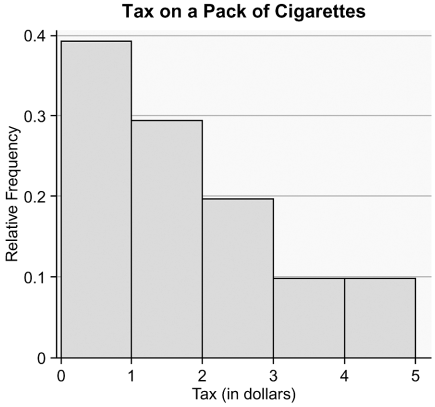


(e) The distribution appears to be right skewed.

(f) Relative frequency of 0–0.999 is: 20/51 = 0.3922, and so on.

Cigarette Tax	Frequency	Relative Frequency
0.00–0.999	20	0.3922
1.00–1.999	15	0.2941
2.00–2.999	10	0.1961
3.00–3.999	5	0.0980
4.00–4.999	1	0.0196





The distribution is right skewed.

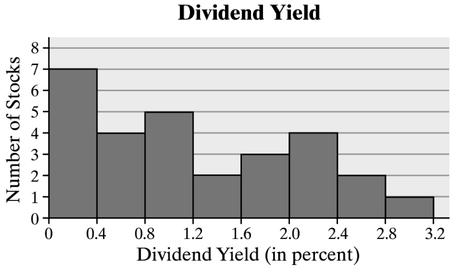
(g) Answers will vary. The first distribution gives a more detailed pattern and does a nice job summarizing the data.

32. (a), (b)

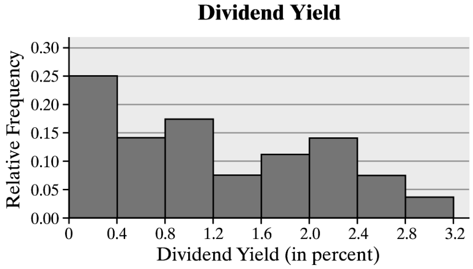
Relative frequency for 0.00–0.39 = 7/28 = 0.2500, and so on.

Class (Dividend)	Freq.	Rel. Freq.
0.00 – 0.39	7	0.2500
0.40 – 0.79	4	0.1429
0.80 – 1.19	5	0.1786
1.20 – 1.59	2	0.0714
1.60 – 1.99	3	0.1071
2.00 – 2.39	4	0.1429
2.40 – 2.79	2	0.0714
2.80 – 3.19	1	0.0357

(c)



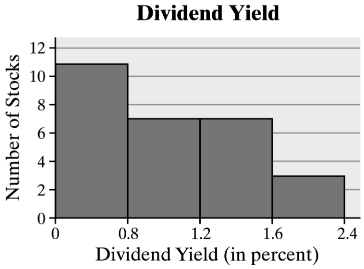
(d)

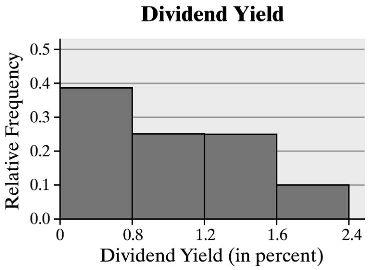


(e) The distribution is skewed right.

(f) Relative frequency for 0.00–0.79 = 11/28 = 0.3929, and so on.

Class (Dividend)	Freq.	Rel. Freq.
0.00 – 0.79	11	0.3929
0.80 – 1.59	7	0.2500
1.60 – 2.39	7	0.2500
2.40 – 3.19	3	0.1071





The distribution is skewed right.

(g) Answers will vary. Both distributions indicate the data are skewed right. The first graph is preferred because it gives more detailed information. The second graph is a little too compressed to get a complete view of what is happening with the data.

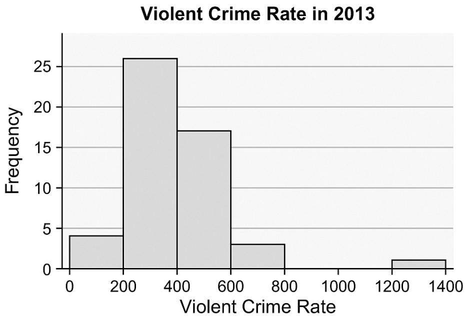
33. Answers will vary. One possibility follows.

(a) Choose a lower class limit of first class of 0 with a class width of 200.

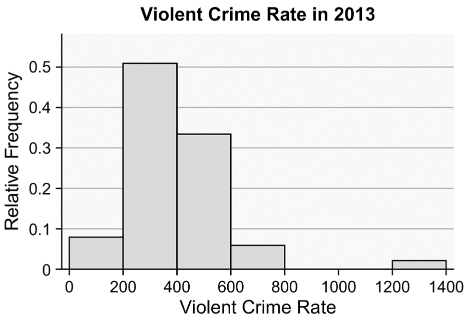
(b), (c)
Relative frequency for 0–199 is 4/51 = 0.0784, and so on.

Violent Crime Rate	Frequency	Relative Frequency
0–199.9	4	0.0784
200–399.9	26	0.5098
400–599.9	17	0.3333
600–799.9	3	0.0588
800–999.9	0	0.0000
1000–1199.9	0	0.0000
1200–1399.9	1	0.0196

(d)



(e)



(f) The distribution is skewed right.

34. Answers will vary. One possibility follows.

(a) We can determine a class width by subtracting the smallest value from the largest, dividing by the desired number of classes, then rounding up. For example,

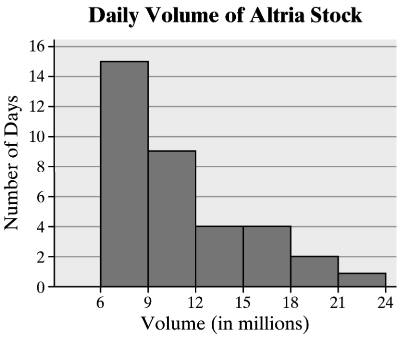
$$\frac{23.59-6.37}{6} = 2.87 \to 3$$

Our first lower class limit should be a nice number below the smallest data value. In this case, 6 is a good first lower limit since it is the nearest whole number below the smallest data value. Thus, we will have a class width of 3, and the first class will have a lower limit of 6.

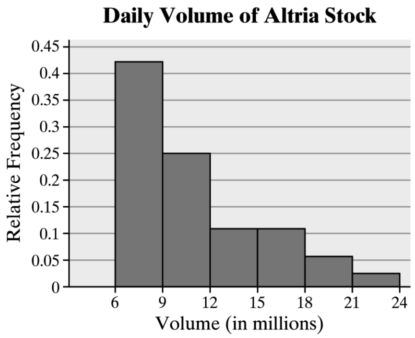
(b), (c)
Relative frequency for 6–8.99 = 15/35 = 0.4286, and so on.

Class (Volume)	Freq.	Rel. Freq.
6 – 8.99	15	0.4286
9 – 11.99	9	0.2571
12 – 14.99	4	0.1143
15 – 17.99	4	0.1143
18 – 20.99	2	0.0571
21 – 23.99	1	0.0286

(d)



(e)



(f) The distribution is skewed right.

35. **(a)** **President Ages at Inauguration**

4	23
4	6677899
5	0011112244444
5	555566677778
6	0111244
6	589

Legend: 4 | 2 represents 42 years.

(b) The distribution appears to be roughly symmetric and bell-shaped.

36. **(a)** Divorce Rates in 2011

2	4
2	677899999
3	12223344
3	56677888999
4	00123344
4	5889
5	223
5	6

Legend: 2|4 represents 2.4 per 1000 population

(b) The distribution appears to be roughly symmetric and bell-shaped. One could argue that the distribution is slightly skewed right.

37. **(a)** **Fat in McDonald's Breakfast**

0	39
1	1266
2	1224577
3	0012267
4	6
5	159

Legend: 5 | 1 represents 51 grams of fat.

(b) The distribution appears to be roughly symmetric and bell-shaped.

38. **(a)** **Gasoline Mileages**

2	233
2	55567889999
3	000001111111112222223333333333334444444444444
3	5555555555666666666666778
4	0223

Legend: 2 | 2 represents 22 miles per gallon.

(b) The distribution appears to be symmetric and bell-shaped.

39. **(a)** Five Year Rate of Return Rounded to the nearest tenth:

10.9	14.2	12.4	13.6	13.0
10.5	10.3	13.1	15.7	14.9
14.1	12.8	13.3	9.9	15.6
12.3	13.9	13.4	19.4	13.4
12.2	14.8	11.9	10.1	13.6
14.6	14.8	13.5	13.9	13.2
14.0	15.2	8.3	9.0	8.7
14.9	16.0	13.7	13.9	12.8

(b) Five Year Rate of Return

8	37
9	09
10	1359
11	9
12	23488
13	0123445667999
14	01268899
15	267
16	0
17	
18	
19	4

Legend: 8|3 represents 8.3%

(c) The distribution is bell-shaped.

40. **(a)** Home appreciation values rounded to the nearest whole number:

69	149	94	118	87
113	130	65	113	109
350	122	104	94	101
185	150	225	117	107
136	135	113	87	113
115	197	71	96	91
85	105	210	109	125
220	136	127	110	87
75	105	104	97	133
207	93	80	145	67
80				

Home Appreciation

6	579
7	15
8	005777
9	134467
10	14455799
11	03333578
12	257
13	03566
14	59
15	0
16	
17	
18	5
19	7
20	7
21	0
22	05
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	0

Legend: 6|5 represents 65

(b) The shape of the distribution is skewed to the right.

(c) Answers will vary. However, a histogram is probably a better choice because of the wide range of possible values.

41. (a) Violent crime rates rounded to the nearest tens:

450	1240	350	260	410	560	320
600	490	220	450	350	320	280
430	380	500	270	240	640	200
470	240	120	260	300	410	
420	210	480	610	470	210	
310	410	410	190	250	140	
280	350	450	290	350	190	
550	260	230	560	250	300	

(b) Violent Crime Rates by State, 2013

1	2499
2	0112344556667889
3	0012255558
4	1111235557789
5	0566
6	014
7	
8	
9	
10	
11	
12	4

Legend: 1|2 represents 120 violent crimes per 100,000 population

(c) Violent Crime Rates by State, 2013

1	24
1	99
2	0112344
2	556667889
3	00122
3	55558
4	111123
4	5557789
5	0
5	566
6	014
6	
7	
7	
8	
8	
9	
9	
10	
10	
11	
11	
12	4

Legend: 1|2 represents 120 violent crimes per 100,000 population

(d) Answers will vary. The first display is decent. It clearly shows that the distribution is skewed right and has an outlier. The second display is not as good as the first. Splitting the stems did not reveal any additional information and has made the display more cluttered and cumbersome.

42. (a) **Ages of Academy Award Winners**

Best Actor		Best Actress
9	2	125668999
998877766220	3	012233333455689
8765554332200	4	11245599
5432100	5	
200	6	112
6	7	4
	8	0

Legend: 6|7|4 represents 76 years old for Best Actor and 74 years old for Best Actress.

(b) Answers will vary. It appears that Academy Award winners for best actor tend to be older on the whole than winners for best actress.

43. **(a)**

Home Run Distances

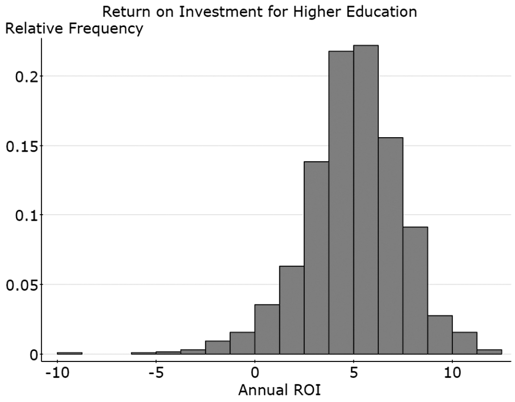
McGwire		Bonds
	32	0 0
	33	
1 0	34	7
0 0	35	0
9 0 0 0	36	0 0 0 1 5
7 0 0 0 0	37	0 0 5 5 5 5
8 5 5 0 0 0 0 0	38	0 0 0 0 0 5
8 0 0 0 0	39	0 0 1 4 6
9 0 0	40	0 0 0 0 4 5
0 0 0 0 0	41	0 0 0 0 0 0 0 0 0 0 1 5 5 6 7 7
5 3 0 0 0 0 0	42	0 0 0 0 0 0 0 0 9
0 0 0 0 0 0 0	43	0 0 0 0 0 5 5 6
0 0 0	44	0 0 0 0 2
8 2 0 0 0 0	45	0 4
1 0 0	46	
8 0 0 0	47	
0	48	8
	49	
0	50	
0 0	51	
7	52	
	53	
	54	
0	55	

Legend: 0 | 34 | 7 represents 340 feet for McGwire and 347 feet for Bonds.

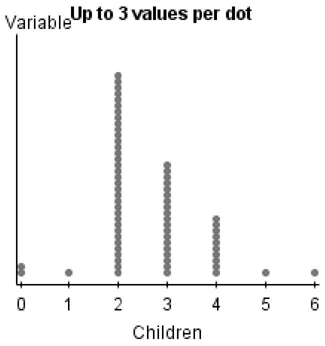
(b) Answers will vary. For both players, the distances of home runs mainly fall from 360 to 450 feet. McGwire has quite a few extremely long distances.

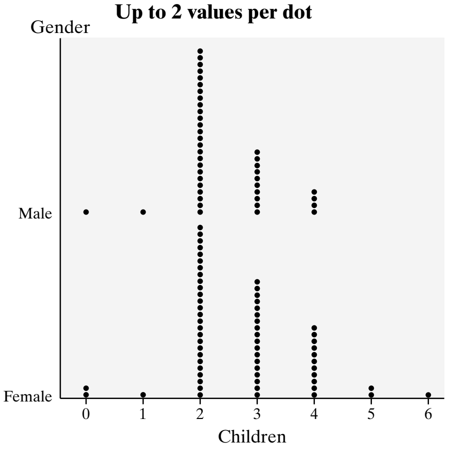
44. Answers will vary.

45. Answers will vary. It is disconcerting that some schools have a negative ROI.



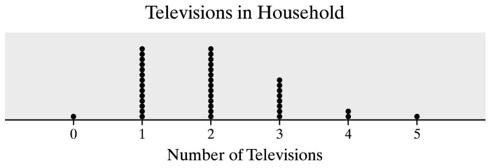
46.



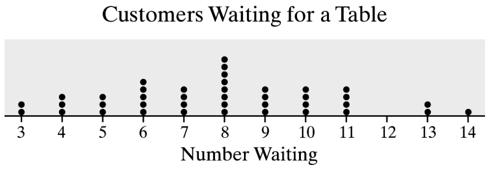


There are several similarities in the distribution of the ideal number of children, as reported by males and females. However, females seem more likely to deem larger families as ideal.
A histogram would better serve us in comparing the preferences between males and females.

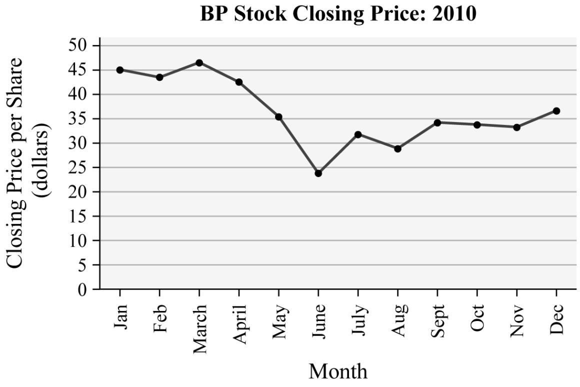
47.



48.

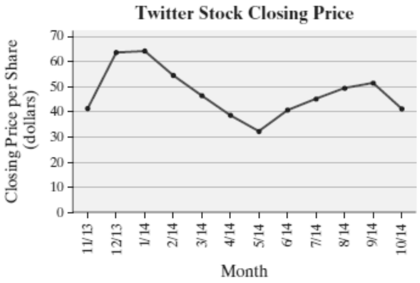


49. **(a)**



(b) The value of the BP stock at the end of May 2010 was 35.72 and was only 24.02 at the end of June 2010. The percentage change in the BP stock price from May to June 2010 was (24.02–35.72)/35.72 = –0.328, which is a decrease of 32.8%.

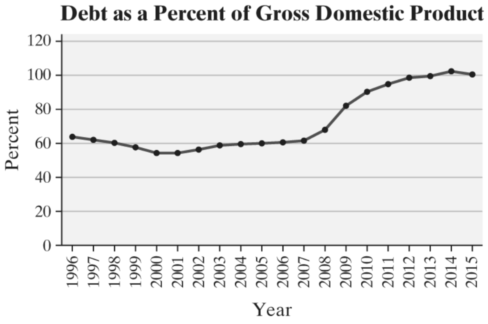
50. **(a)**



(b) The closing price of Twitter stock in November 2013 was 41.57 and was 63.65 in December 2013. The percentage change from November to December 2013 was (63.65–41.57)/41.57 = 0.531, which is an increase of 53.1%.

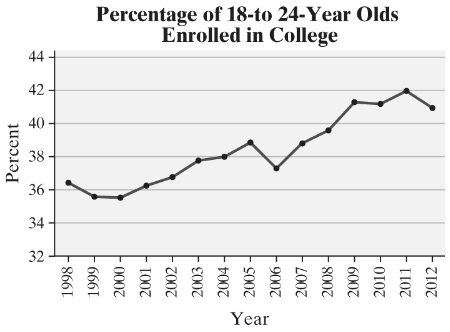
The closing price of Twitter stock in November 2013 was 41.57 and was 41.47 in October 2014. The percentage change from November 2013 to October 2014 was (41.47–41.57)/41.57 = –0.002, which is a decrease of 0.2%. The closing price of Twitter stock increased 53.1% between November and December 2013, so an investor would have been wise to sell in December 2013.

51.



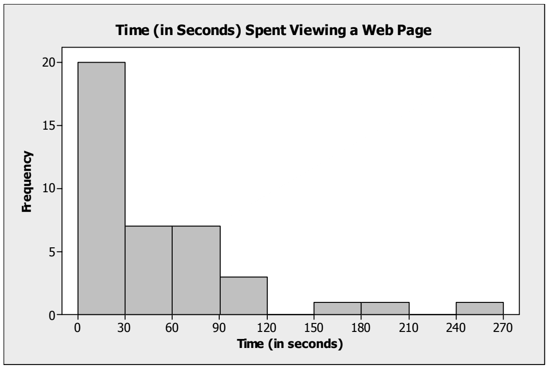
Answers will vary. The time-series plot shows that debt as a percent of gross domestic product remained relatively stable around 60% from 1996 to 2007 and then began to increase steadily from 2008 to 2015.

52.



Answers will vary. The time-series plot shows that the percentage of high school graduates enrolling in college seems to have increased over the given time period amid a variety of fluctuations with a slight downturn in 2012.

53. Because the data are continuous, either a stem-and-leaf plot or a histogram would be appropriate. There were 20 people who spent less than 30 seconds, 7 people spent at least 30 seconds but less than 60 seconds, etc. One possible histogram is:



The data appear to be skewed right with a gap and one potential outlier. It seems as if the majority of surfers spent less than one minute viewing the page, while a few surfers spent several minutes viewing the page.

54. Age: histogram, stem-and-leaf plot, or dot plot; Income: histogram or stem-and-leaf plot; Marital status: bar graph or pie chart; Number of vehicles: histogram, stem-and-leaf plot, or dot plot

55. Answers will vary. Reports should address the fact that the number of people going to the beach and participating in underwater activities (e.g. scuba diving, snorkeling) has also increased, so an increase in shark attacks is not unexpected. A better comparison would be the rate of attacks per 100,000 beach visitors. The number of fatalities could decrease due to better safety equipment (e.g. bite resistant suits) and better medical care.

56. Classes should not overlap to avoid any confusion as to which class an observation belongs to.

57. Histograms are useful for large data sets or data sets with a large amount of spread. Stem-and-leaf plots are nice because the raw data can easily be retrieved. A disadvantage of stem-and-leaf plots is that sometimes the data must be rounded, truncated, or adjusted in some way that requires extra work. Furthermore, if these steps are taken, the original data is lost and a primary advantage of stem-and-leaf plots is lost.

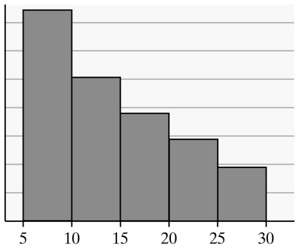
58. There is no such thing as the correct choice for a class width, however some choices are better than others. For example, if the class width is too small, the histogram will show many gaps between the bars. If the class width is too large, the histogram may not provide enough detail.

59. Relative frequencies should be used when comparing two data sets with different sample sizes.

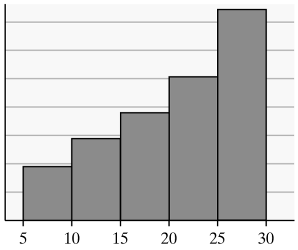
60. Answers will vary. The exercise illustrates the fact that there is no such thing as the "correct" histogram. However, some histograms are better than others and class width can affect the shape of a graph.

61. Answers will vary. Sample histograms are given below.

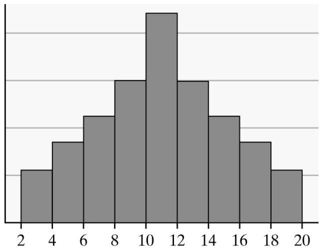
Skewed Right



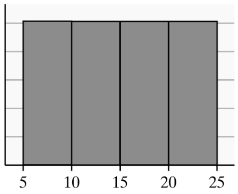
Skewed Left



Bell-Shaped



Uniform



A histogram is skewed left if it has a long tail on the left side. A histogram is skewed right if it has a long tail on the right side. A histogram is symmetric if the left and right sides of the graph are roughly mirror images of each other.

62. Time-series plots are drawn with quantitative variables. They are drawn to see trends in the data.

Section 2.3

1. The lengths of the bars are not proportional. For example, the bar representing the cost of Clinton's inauguration should be slightly more than 9 times as long as the one for Carter's cost, and twice as long as the bar representing Reagan's cost.

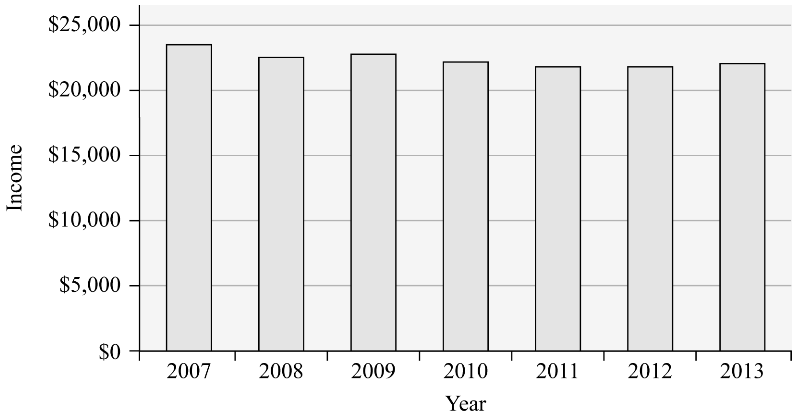
2. **(a)** Answers will vary. The lengths of the bars are not proportional. For example, the bar for soda is 1/3 the size of the bar for a cheeseburger, but the number of steps for a cheeseburger is just over twice that for the soda. In addition, it is unclear where the graph begins: at the base of each figure or the bottom of the platform.

(b) Answers will vary. The pictures could be replaced by simple bars (of the same width) that are proportional in area.

3. **(a)** The vertical axis starts at $21,500 instead of $0. This tends to indicate that the median earnings for females changed at a faster rate than actually occurred.

(b) This graph indicates that the median earnings for females has decreased slightly over the given time period.

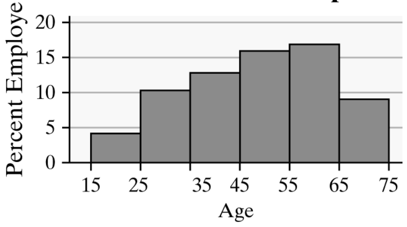
Median Income for Females in Constant 2013 Dollars



4. **(a)** The vertical axis starts at 4 instead of 0. This may lead the reader to conclude, for example, the percentage of employed people aged 55–64 who are members of a union is more than double the percentage of those aged 25–34 years.

(b)

Union Membership



5. The bar for 12p–6p covers twice as many hours as the other bars. By combining two 3-hour periods, this bar looks larger compared to the others, making afternoon hours look more dangerous. If this bar were split into two periods, the graph may give a different impression. For example, the graph may show that daylight hours are safer.

6. The article is basing its conclusion on a comparison of categories that do not cover the same number of years. A better comparison is the incidence rate (number of accidents per 100,000 licensed drivers). [Note: only about 14% of licensed drivers in 2005 were aged 24 years or younger.]

7. Answers will vary. This graph is misleading because it does not take into account the size of the population of each state. For example, Vermont is going to pay less in total taxes than California simply because its population is so much lower. There are many variables that should be considered on per capita (per person) basis. For example, this graph would be less misleading if it was drawn to represent taxes paid per capita (per person).

8. **(a)** The oil reserves in 2014 were 691.0 million barrels, whereas the oil reserves in 1977 were 7.5 million barrels. The oil reserves in 2014 were 92 times as large as in 1977 (e.g. 691/7.5=92.1). Thus, the graphic for 2014 should be roughly 92 times larger than the graphic for 1977.

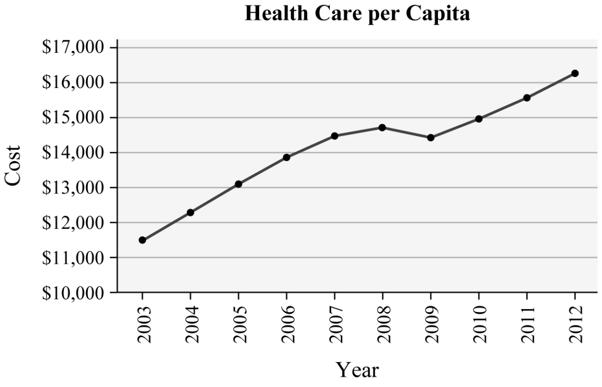
(b) Assuming no change in U.S. oil production, the U.S. strategic oil reserves would last approximately 90 days (e.g. 691/7.7 = 89.74 days).

9. **(a)** The graphic is misleading because the bars are not proportional. The bar for housing should be a little more than twice the length of the bar for transportation, but it is not.

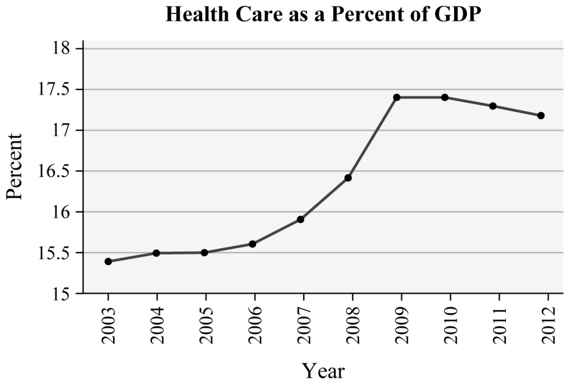
(b) The graphic could be improved by adjusting the bars so that their lengths are proportional.

10. The graph does not support the safety manager's claim. The vertical scale starts at 0.17 instead of 0, so the difference between the bars is distorted. While there was a decrease, it appears that the decrease is roughly 10% of the 1992 rate.

11. **(a)** Answers will vary. Here is a time-series plot that a politician might use to support the position that health care is increasing.

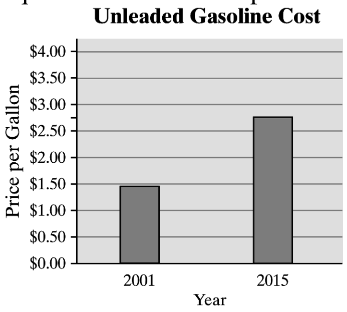


(b) Answers will vary. Here is a time-series plot that the health care industry might use to refute the opinion of the politician.

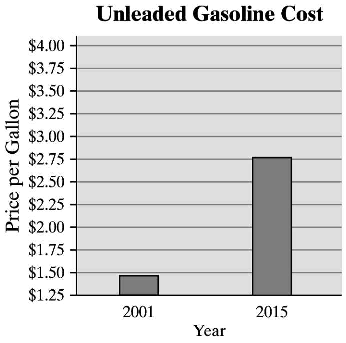


(c) Answers will vary. Changing the scale on the graph will affect the message. The message is also affected by using the variable "Health Care per Capita" rather than "Health Care as a Percent of GDP."

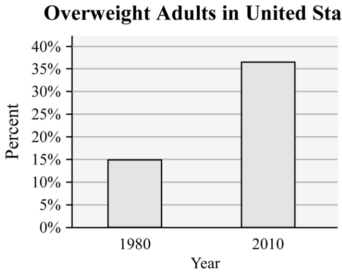
12. **(a)** A graph that is not misleading will use a vertical scale starting at $0 and bars of equal width. One example:



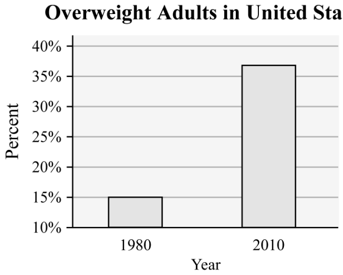
(b) A graph that is misleading might use bars of unequal width or will use a vertical scale that does not start at $0. One example, as follows, is misleading because it starts at $1.25 instead of 0 without indicating a gap. This might cause the reader to conclude that cost of unleaded gasoline has risen more sharply than actually occurred.



13. **(a)** A graph that is not misleading will use a vertical scale starting at 0% and bars of equal width. One example:



(b)

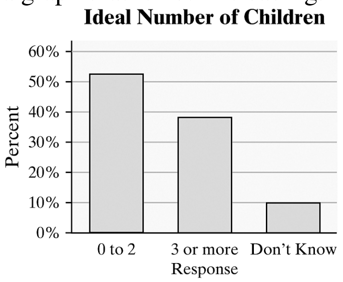


This graphic is misleading because the vertical scale starts at 10% instead of 0% without indicating a gap. This might cause the reader to think that the proportion of overweight adults in the United States is increasing more quickly than they really are.

14. **(a)** A bar graph

(b) A reader cannot tell whether the graph ends at the top of the nipple on the baby bottle, or at the end of the milk.

(c) Answers will vary. Here is an example of a graph that is not misleading.



15. Answers will vary. Three-dimensional graphs are deceptive because the pieces are not proportional. For example, the area for P (pitcher) looks substantially larger than the area for 3B (third base), even though both are the same percentage. Graphs should not be drawn using three dimensions. Instead, use two dimensions.

16. Answers will vary. This is a histogram so the bars should touch. In addition, there are no labels and no title.

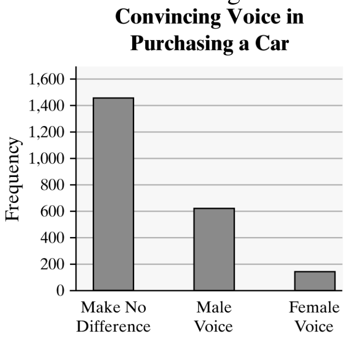
Chapter 2 Review Exercises

1. **(a)** There are 614 + 154 + 1448 = 2216 participants.

(b) The relative frequency of the respondents indicating that it makes no difference is

$$\frac{1448}{2216} \approx 0.653$$

(c) A Pareto chart is a bar chart where the bars are in descending order.



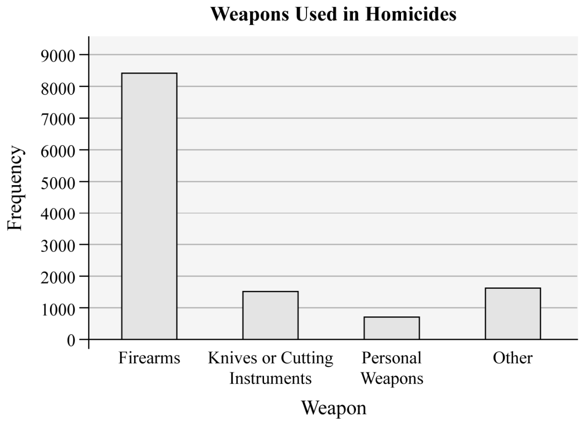
(d) Answers will vary.

2. **(a)** Total homicides = 8438 + 1486 + 685 + 1621 = 12230
Relative frequency for firearms is 8438/12230 = 0.6899, and so on.

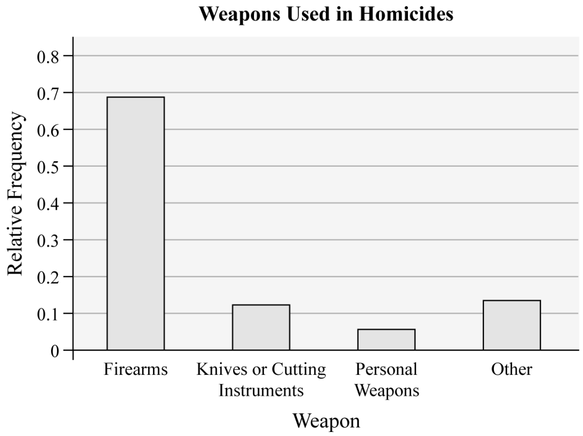
Type of Weapon	Relative Frequency
Firearms	0.6899
Knives or cutting instruments	0.1215
Personal weapons	0.0560
Other weapon	0.1325

(b) The relative frequency is 0.6899, so 68.99% of the homicides were committed using a firearm.

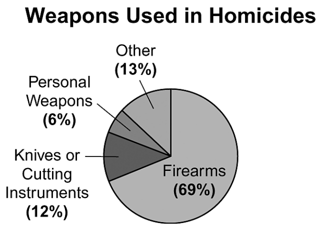
(c)



(d)



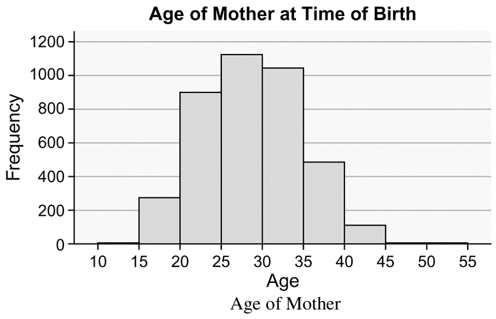
(e)



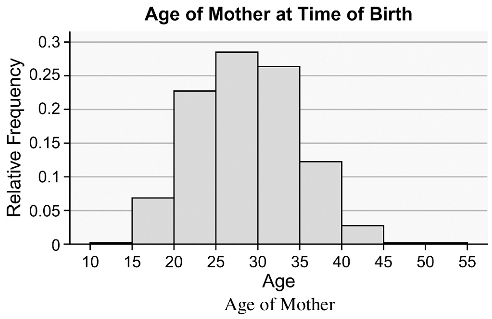
3. **(a)**
Total births (in thousands) = 3 + 275 + 902 + 1128 + 1044 + 487 + 110 + 7 + 1 = 3957
Relative frequency for 10–14 year old mothers = $3/3957 \approx 0.0008$, and so on.
Cumulative frequency for 15–19 year old mothers = 3 + 275 = 278, and so on.
Cumulative relative frequency for 15–19 year old mothers = $278/3957 \approx 0.0703$, and so on.

Age of Mother	Rel. Freq.
10–14	0.0008
15–19	0.0695
20–24	0.2280
25–29	0.2851
30–34	0.2638
35–39	0.1231
40–44	0.0278
45–49	0.0018
50–54	0.0003

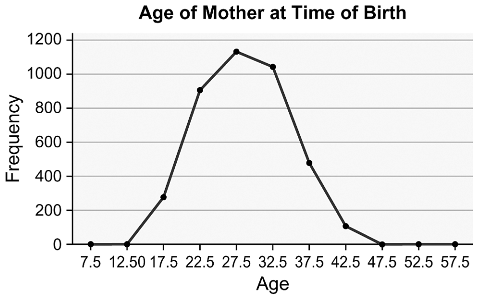
(b) The distribution is roughly symmetric and bell-shaped.



(c)



(d)



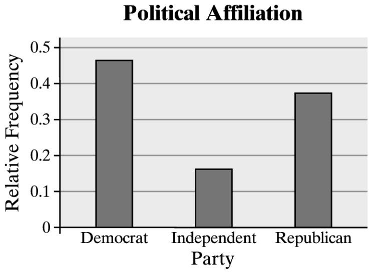
(e) From the relative frequency table, the relative frequency of 20–24 is 0.2280, so the percentage is 22.80%.

(f) $\dfrac{1044+487+110+7+1}{3957}=\dfrac{1649}{3957}\approx 0.4167$
41.67% of live births were to mothers aged 30 years or older.

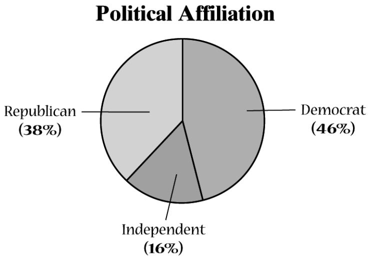
4. **(a), (b)**

Affiliation	Frequency	Relative Frequency
Democrat	46	0.46
Independent	16	0.16
Republican	38	0.38

(c)



(d)

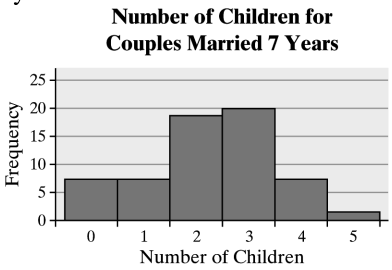


(e) Democrat appears to be the most common affiliation in Naperville.

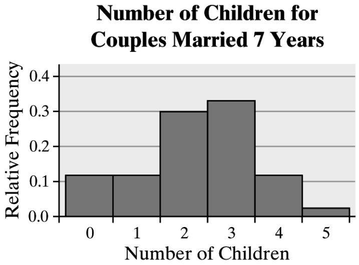
5. **(a), (b)**

Number of Children	Freq.	Rel. Freq.
0	7	0.1167
1	7	0.1167
2	18	0.3000
3	20	0.3333
4	7	0.1167
5	1	0.0167

(c) The distribution is more or less symmetric.



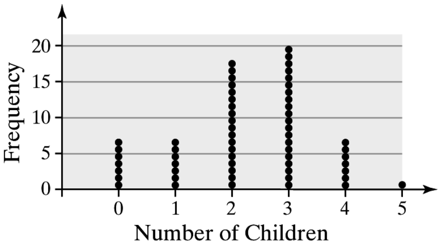
(d)



(e) From the relative frequency table, the relative frequency of two children is 0.3000, so 30% of the couples have two children.

(f) From the frequency table, the relative frequency of at least two children (i.e. two or more) is $0.3000+0.3333+0.1167+0.0167=0.7667$ or 76.67%. So, 76.67% of the couples have at least two children.

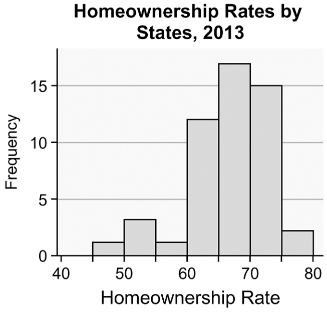
(g)



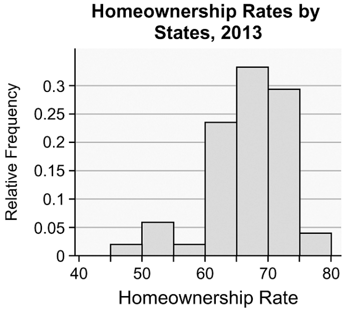
6. (a), **(b)**

Homeownership Rate	Frequency	Relative Frequency
45–49.9	1	0.0196
50–54.9	3	0.0588
55–59.9	1	0.0196
60–64.9	12	0.2353
65–69.9	17	0.3333
70–74.9	15	0.2941
75–75.9	2	0.0392

(c)



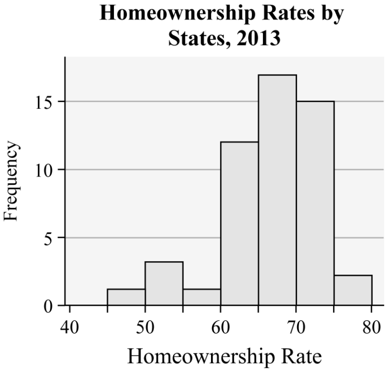
(d)

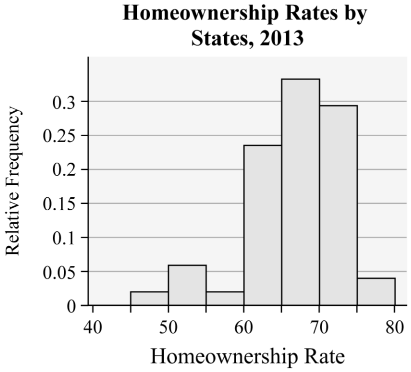


(e) The distribution is slightly skewed left.

(f)

Homeownership Rate	Frequency	Relative Frequency
40–49.9	1	0.0196
50–59.9	4	0.0784
60–69.9	29	0.5686
70–79.9	17	0.3333





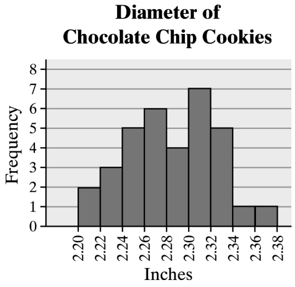
(g) Answers will vary. Both class widths give a good overall picture of the distribution. The first class width provides a little more detail to the graph, but not necessarily enough to be worth the trouble. An intermediate value, say a width of 8, might be a reasonable compromise.

7. (a), **(b)**

Answers will vary. Using 2.2000 as the lower class limit of the first class and 0.0200 as the class width, we obtain the following.

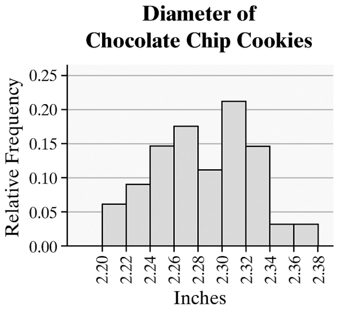
Class	Freq.	Rel. Freq.
2.2000 – 2.2199	2	0.0588
2.2200 – 2.2399	3	0.0882
2.2400 – 2.2599	5	0.1471
2.2600 – 2.2799	6	0.1765
2.2800 – 2.2999	4	0.1176
2.3000 – 2.3199	7	0.2059
2.3200 – 2.3399	5	0.1471
2.3400 – 2.3599	1	0.0294
2.3600 – 2.3799	1	0.0294

(c)



The distribution is roughly symmetric.

(d)



8 **Hours Spent Online**

12	
13	467
14	05578
15	1236
16	456
17	113449
18	066889
19	2
20	168
21	119
22	29
23	48
24	4
25	7
26	

Legend: 13 | 4 = average 13.4 hours per week.

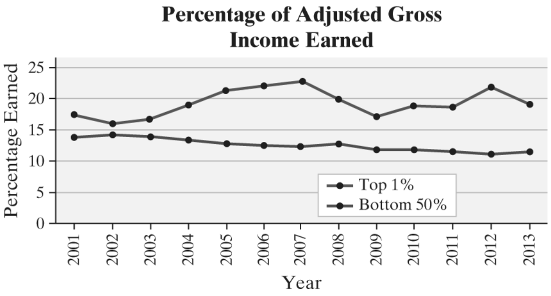
The distribution is slightly skewed right.

9. **(a)** Grade inflation seems to be happening in colleges. GPAs have increased every time period for all schools.

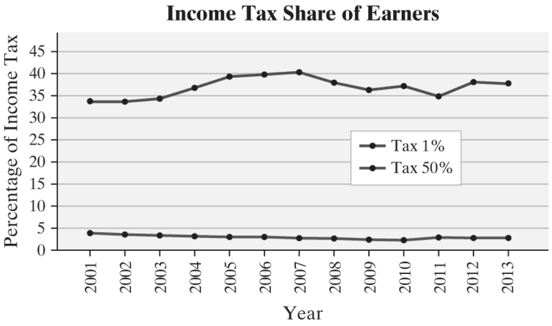
(b) GPAs increased about 5.6% for public schools. GPAs increased about 6.8% for private schools. Private schools have higher grade inflation because the GPAs are higher and they are increasing faster.

(c) The graph is misleading because it starts at 2.6 on the vertical axis.

10. **(a)** Answers will vary. The adjusted gross income share of the top 1% of earners shows steady increases overall, with a few minor exceptions. The adjusted gross income share of the bottom 50% of earners shows steady decreases overall, with a few minor exceptions.

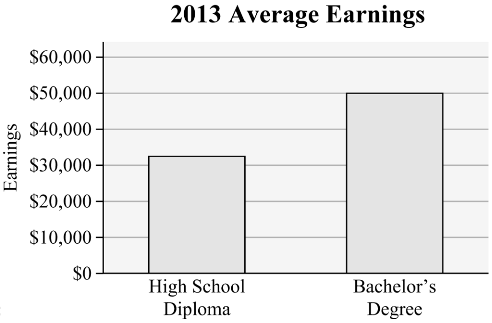


(b) Answers will vary. The income tax share of the top 1% of earners shows steady increases overall, with few exceptions, including a notable decrease from 2007 to 2008. The income tax share of the bottom 50% of earners shows steady decreases over time.



11. **(a)** Graphs will vary. One way to mislead would be to start the vertical scale at a value other than 0. For example, starting the vertical scale at $30,000 might make the reader believe that college graduates earn more than three times what a high school graduate earns (on average).

(b) A graph that does not mislead would use equal widths for the bars and would start the vertical scale at $0. Here is an example of a graph that is not misleading:



=

12. **(a)** Flats are preferred the most (40%) and extra-high heels are preferred the least (1%).

(b) The graph is misleading because the bar heights and areas for each category are not proportional.

Chapter 2 Test

1. **(a)** A 5 Star rating was the most popular rating with 1675 votes.

(b) 35 + 67 +246 + 724 + 1675 = 2747 postings were posted on Yelp for Hot Doug's restaurant.

(c) $1675 - 724 = 951$
There were 951 more 5 Star ratings than 4 Star ratings.

(d) There were 1675 5 Star ratings out of a total of 2747 ratings. $\frac{1675}{2747} \approx 0.6098$
Approximately 61% of all ratings were 5 Star ratings.

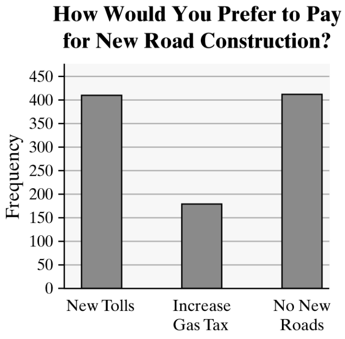
(e) No, it is not appropriate to describe the shape of the distribution as skewed right. The data represented by the graph are qualitative, so the bars in the graph could be placed in any order.

2. **(a)** There were 1005 responses. The relative frequency who indicated they preferred new tolls was $\frac{412}{1005} = 0.4100,$ and so on.

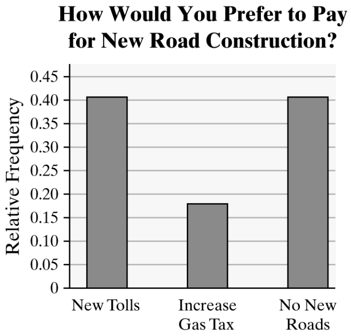
Response	Freq.	Rel. Freq.
New Tolls	412	0.4100
Inc. Gas Tax	181	0.1801
No New Roads	412	0.4100

(b) The relative frequency is 0.1801, so the percentage of respondents who would like to see an increase in gas taxes is 18.01%.

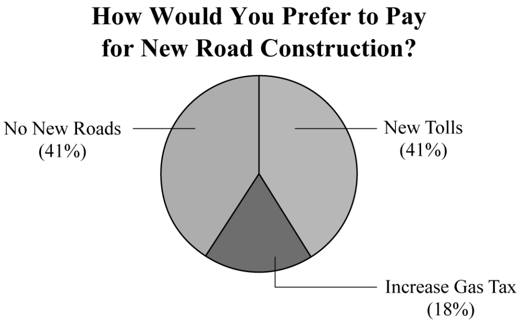
(c)



(d)



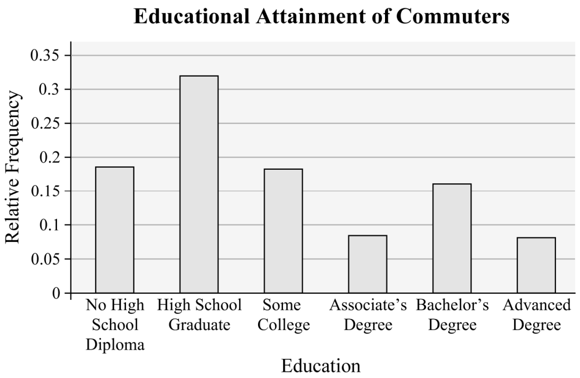
(e)



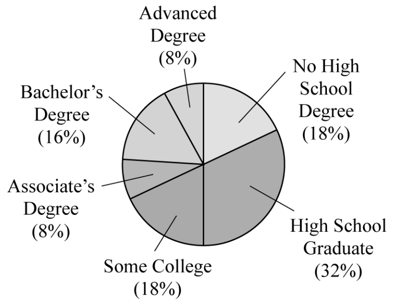
3. **(a), (b)**

Education	Freq.	Rel. Freq.
No high school diploma	9	0.18
High school graduate	16	0.32
Some college	9	0.18
Associate's degree	4	0.08
Bachelor's degree	8	0.16
Advanced degree	4	0.08

(c)



(d) **Educational Attainment of Commuters**

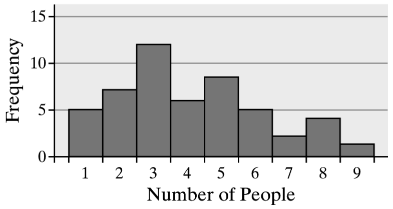


(e) The largest bar (and largest pie segment) corresponds to "High School Graduate," so high school graduate is the most common educational level of a commuter.

4. **(a)**, **(b)**

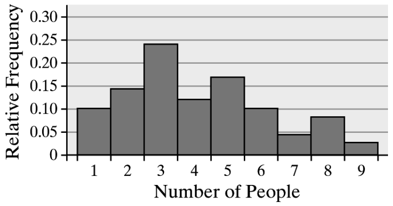
No. of Cars	Freq.	Rel. Freq.
1	5	0.10
2	7	0.14
3	12	0.24
4	6	0.12
5	8	0.16
6	5	0.10
7	2	0.04
8	4	0.08
9	1	0.02

(c) **Number of Cars Arriving at McDonald's**



The distribution is skewed right.

(d) **Number of Cars Arriving at McDonald's**



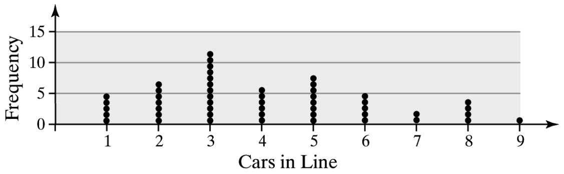
(e) The relative frequency of exactly 3 cars is 0.24. So, for 24% of the weeks, exactly three cars arrived between 11:50 am and 12:00 noon.

(f) The relative frequency of 3 or more cars

$$= 0.24 + 0.12 + 0.16 + 0.10$$
$$+ 0.04 + 0.08 + 0.02 = 0.76$$

So, for 76% of the weeks, three or more cars arrived between 11:50 am and 12:00 noon.

(g)

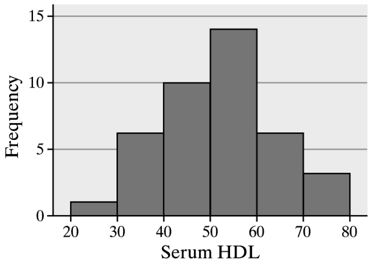


5. Answers may vary. One possibility follows:

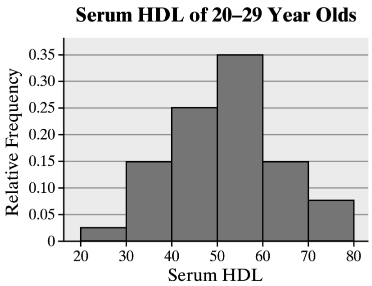
(a), (b)
Using a lower class limit of the first class of 20 and a class width of 10:
Total number of data points = 40
Relative frequency of 20 – 29 = 1/40 = 0.025, and so on.

HDL Cholesterol	Frequency	Relative Frequency
20–29	1	0.025
30–39	6	0.150
40–49	10	0.250
50–59	14	0.350
60–69	6	0.150
70–79	3	0.075

(c) **Serum HDL of 20–29 Year Olds**



(d)



(e) The distribution appears to be roughly bell-shaped.

6. The stem-and-leaf diagram below shows an approximately uniform distribution.

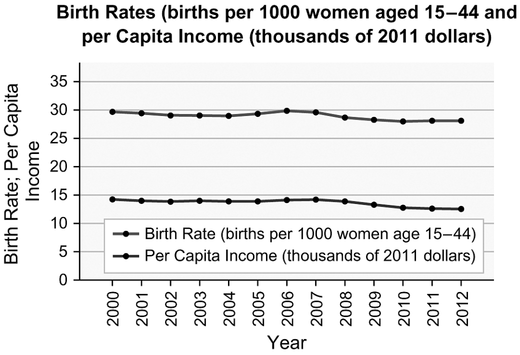
Time Spent on Homework

4	0567
5	26
6	13
7	01338
8	59
9	1369
10	3899
11	0018
12	556

Legend: 4 | 0 represents 40 minutes.

The distribution is symmetric (uniform).

7. The curves in the figure below appear to follow the same trend. Birth rate increases as per capita income increases.



8. Answers may vary. It is difficult to interpret this graph because it is not clear whether the scale is represented by the height of the steps, the width of the steps, or by the graphics above the steps. The graphics are misleading because they must be increased in size both vertically and horizontally to avoid distorting the image. Thus, the resulting areas are not proportionally correct. The graph could be redrawn using bars whose widths are the same and whose heights are proportional based on the given percentages. The use of graphics should be avoided, or a standard size graphic representing a fixed value could be used and repeated as necessary to illustrate the given percentages.

Chapter 3
Numerically Summarizing Data

Section 3.1

1. A statistic is resistant if its value is not sensitive to extreme data values.

2. Since the distribution of household incomes in the United States is skewed to the right, the mean is greater than the median. Thus, the mean household income is $75,738 and the median is $53,657.

3. HUD uses the median because the data are skewed. Explanations will vary. One possibility is that the price of homes has a distribution that is skewed to the right, so the median is more representative of the typical price of a home.

4. The mean will be larger because it will be influenced by the extreme data values that are to the right end (or high end) of the distribution.

5. $\frac{10,000+1}{2}=5000.5$. The median is between the 5000th and the 5001st ordered values.

6. False. A data set may have multiple modes, or it may have no mode at all.

7. $\bar{x}=\frac{20+13+4+8+10}{5}=\frac{55}{5}=11$

8. $\bar{x}=\frac{83+65+91+87+84}{5}=\frac{410}{5}=82$

9. $\mu=\frac{3+6+10+12+14}{5}=\frac{45}{5}=9$

10. $\mu=\frac{1+19+25+15+12+16+28+13+6}{9}=\frac{135}{9}=15$

11. $\mu=\frac{218,469,636}{82,566}=\$2,646$

The mean price per ticket was $2,646.

12. Let x represent the missing value. Since there are 6 data values in the list, the median 26.5 is between the 3rd and 4th ordered values, which are 21 and x, respectively. Thus,

$$\frac{21+x}{2}=26.5$$
$$21+x=53$$
$$x=32$$ The missing value is 32.

13. $\sum x_i=34.0+33.2+37.0+29.4+23.6+25.9=183.1$

Mean $=\bar{x}=\frac{\sum x_i}{n}=\frac{183.1}{6}=30.52$ mpg

Data in order: 23.6, 25.9, 29.4, 33.2, 34.0, 37.0

Median $=\frac{29.4+33.2}{2}=\frac{62.6}{2}=31.3$ mpg

No data value occurs more than once, so there is no mode.

14. $\sum x_i=60.5+128.0+84.6+122.3+78.9+94.7+85.9+89.9=744.8$

Mean $=\bar{x}=\frac{\sum x_i}{n}=\frac{744.8}{8}=93.1$ minutes

Data in order:
60.5, 78.9, 84.6, 85.9, 89.9, 94.7, 122.3, 128.0

Median $=\frac{85.9+89.9}{2}=\frac{175.8}{2}=87.9$ minutes

No data value occurs more than once, so there is no mode.

15. $\sum x_i=3960+4090+3200+3100+2940+3830+4090+4040+3780=33,030$ psi

Mean $=\bar{x}=\frac{\sum x_i}{n}=\frac{33,030}{9}=3670$ psi

Data in order: 2940, 3100, 3200, 3780, 3830, 3960, 4040, 4090, 4090
Median = the 5th ordered data value = 3830 psi
Mode = 4090 psi (because it is the only data value to occur twice)