INSTRUCTOR'S MANUAL

to accompany

STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fourth Edition

Alan Agresti and Barbara Finlay

published by Pearson Education

Manual prepared by:

Jackie Miller
404 Cockins Hall
Department of Statistics
The Ohio State University
Columbus, OH 43210

Instructors: Please notify Alan Agresti of any errors in this manual or the text so they can be corrected for future printings. Please send e-mail to AA@STAT.UFL.EDU.

Table of Contents

Chapter 1

1.1. (a) An individual Prius (automobile). (b) All Prius automobiles used in the EPA tests. (c) All Prius automobiles that are or may be manufactured.

1.2. (a) All 7 million voters. (b) A statistic is the 56.5% who voted for Schwarzenegger from the exit poll sample of size 2705; a parameter is the 55.9% who actually voted for Schwarzenegger.

1.3. (a) All students at the University of Wisconsin. (b) A statistic, since it's calculated only for the 100 sampled students.

1.4. A statistic, since it is based on the approximately 1200 Floridians in the sample.

1.5. (a) All adult Americans. (b) Proportion of all adult Americans who would answer definitely or probably true. (c) The sample proportion 0.523 estimates the population proportion. (d) No, it is a *prediction* of the population value but will not equal it exactly, because the sample is only a very small subset of the population.

1.6. (a) The most common response was 2 hours per day. (b) This is a descriptive statistic because it describes the results of a sample.

1.7. (a) A total of 85.7% said "yes, definitely" or "yes, probably." (b) In 1998, a total of 85.8% said "yes, definitely" or "yes, probably." (c) A total of 74.4% said "yes, definitely" or "yes, probably." The percentages of *yes* responses were higher for HEAVEN than for HELL.

1.8. (a) Statistics, since they're based on a sample of 60,000 households, rather than *all* households. (b) Inferential, predicting for a population using sample information.

1.9. (a)

1.10.

Race	Age	Sentence	Felony?	Prior Arrests	Prior Convictions
white	19	2	no	2	1
black	23	1	no	0	0
white	38	10	yes	8	3
Hispanic	20	2	no	1	1
white	41	5	yes	5	4

1.14. (a) A *statistic* is a numerical summary of the sample data, while a *parameter* is a numerical summary of the population. For example, consider an exit poll of voters on election day. The proportion voting for a particular candidate is a statistic. Once all of the votes have been counted, the proportion of voters who voted for that candidate would be known (and is the parameter). (b) *Description* deals with describing the available data (sample or population), whereas *inference* deals with making predictions about a population using information in the sample. For example,

consider a sample of voters on election day. One could use descriptive statistics to describe the voters in terms of gender, race, party, etc., and inferential statistics to predict the winner of the election.

1.15. If you have a census, you do not need to use the information from a sample to describe the population since you have information from the population as a whole.

1.16. (a) The descriptive part of this example is that the average age in the sample is 24.1 years. (b) The inferential part of this example is that the sociologist estimates the average age of brides at marriage for the population to between 23.5 and 24.7 years. (b) The population of interest is women in New England in the early eighteenth century.

1.17. (a) A statistic is the 45% of the sample of subjects interviewed in the UK who said *yes*. (b) A parameter is the true percent of the 48 million adults in the UK who would say *yes*. (c) A descriptive analysis is that the percentage of *yes* responses in the survey varied from 10% (in Bulgaria) to 60% in Luxembourg). (d) An inferential analysis is that the percentage of adults in the UK who would say *yes* falls between 41% and 49%.

Chapter 2

2.1. (a) Discrete variables take a finite set of values (or possible all nonnegative integers), and we can enumerate them all. Continuous variables take an infinite continuum of values. (b) Categorical variables have a scale that is a set of categories; for quantitative variables, the measurement scale has numerical values that represent different magnitudes of the variable. (c) Nominal variables have a scale of *unordered* categories, whereas ordinal variables have a scale of *ordered* categories. The distinctions among types of variables are important in determining the appropriate descriptive and inferential procedures for a statistical analysis.

2.2. (a) Quantitative (b) Categorical (c) Categorical (d) Quantitative (e) Categorical (f) Quantitative (g) Categorical (h) Quantitative (i) Categorical

2.3. (a) Nominal (b) Nominal (c) Interval (d) Nominal (e) Nominal (f) Ordinal (g) Interval (h) Ordinal (i) Nominal (j) Interval (k) Nominal

2.4. (a) Nominal (b) Nominal (c) Ordinal (d) Interval (e) Interval (f) Interval (g) Ordinal (h) Interval (i) Nominal (j) Interval

2.5. (a) Interval (b) Ordinal (c) Nominal

2.6. (a) State of residence. (b) Number of siblings. (c) Social class (high, medium, low). (d) Student status (full time, part time). (e) Number of cars owned. (f) Time (in minutes) needed to complete an exam. (g) Number of siblings.

2.7. (a) Ordinal, since there is a sense of order to the categories. (b) Discrete. (c) These values are statistics since them come from a sample.

2.8. Ordinal.

2.9. (b), (c), (d)

2.10. (a), (c), (e), (f)

2.11. Students numbered 10, 22, 24.

2.12. Number names 00001 to 52000. First five that are selected are 15011, 46573, 48360, 39975, 06907.

2.13. Observational study (b) Experiment (c) Observational study (d) Experiment

2.14. (a) Experimental study, since the researchers are assigning subjects to treatments. (b) An observational study could look those who grew up in nonsmoking or smoking environments and examine incidence of lung cancer.

2.15. (a) Sample-to-sample variability causes the results to vary. (b) The sampling error for the Gallup poll is –2.4% for Gore, 0.1% for Bush, and 1.3% for Nader.

2.16. (a) This is a volunteer sample because viewers chose whether to call in. (b) Randomly sample the population.

2.17. The first question is confusing in its wording. The second question has clearer wording.

2.18. (a) Skip number is $k = 52{,}000/5 = 10{,}400$. Randomly select one of the first 10,400 names and then skip 10,400 names to get each of the next names. For example, if the first name picked is 01536, the other four names are 01536 + 10400 = 11936, 11936 + 10400 = 22336, 22336 + 10400 = 32736, 32736 + 10400 = 43136. (b) We could treat the pages as clusters. We would select a random sample of pages, and then sample every name on the pages selected. Its advantage is that it is much easier to select the sample than it is with random sampling. A disadvantage is as follows: Suppose there are 100 “Martinez” listings in the directory, all falling on the same page. Then with cluster sampling, either all or none of the Martinez families would end up in the sample. If they are all sampled, certain traits which they might have in common (perhaps, e.g., religious affiliation) might be over-represented in the sample.

2.19. Draw a systematic sample form the student directory, using skip number $k = 5000/100 = 50$.

2.20. (a) This is not a simple random sample since the sample with necessarily have 40 women and 40 men. A simple random sample may or may not have exactly 40 men and 40 women. (b) This is stratified random sampling. You ensure that neither men nor women are over-sampled.

2.21. (a) The clusters. (b) The subjects within every stratum. (c) The main difference is that a stratified random sample uses *every* stratum, and we want to compare the strata. By contrast, we have a *sample* of clusters, and not all clusters are represented—the goal is not to compare the clusters but to use them to obtain a sample.

2.22. (a) Categorical are GE, VE, AB, PI, PA, RE, LD, AA; quantitative are AG, HI, CO, DH, DR, NE, TV, SP, AH. (b) Nominal are GE, VE, AB, PA, LD, AA; ordinal are PI and RE; interval are AG, HI, CO, DH, DR, NE, TV, SP, AH.

2.24. (a) Draw a systematic sample from the student directory, using skip number $k = N/100$, where N = number of students on the campus. (b) High school GPA on a 4-point scale, treated as quantitative, interval, continuous; math and verbal SAT on a 200 to 800 scale, treated as quantitative, interval, continuous; whether work to support study (yes, no), treated as categorical, nominal, discrete; time spent studying in average day, on scale (none, less than 2 hours, 2-4 hours, more than 4 hours), treated as quantitative, ordinal, discrete.

2.25. This is nonprobability sampling; certain segments may be over- or under-represented, depending on where the interviewer stands, time of day, etc. Quota sampling fails to incorporate randomization into the selection method.

2.26. Responses can be highly dependent on nonsampling errors such as question wording.

2.27. (a) This is a volunteer sample, so results are unreliable; e.g., there is no way of judging how close 93% is to the actual population who believe that benefits should be reduced. (b) This is a volunteer sample; perhaps an organization opposing gun control laws has encouraged members to send letters, resulting in a distorted picture for the congresswoman. The results are completely unreliable as a guide to views of the overall population. She should take a probability sample of her constituents to get a less biased reaction to the issue. (c) The physical science majors who take the course might tend to be different from the entire population of physical science majors (perhaps more liberal minded on sexual attitudes, for example). Thus, it would be better to take random samples of students of the two majors from the population of all social science majors and all physical science majors at the college. (d) There would probably be a tendency for students within a given class to be more similar than students in the school as a whole. For example, if the chosen first period class consists of college-bound seniors, the members of the class will probably tend to be less opposed to the test than would be a class of lower achievement students planning to terminate their studies with high school. The design could be improved by taking a simple random sample of students, or a larger random sample of classes with a random sample of students then being selected from each of those classes (a two-stage random sample).

2.28. A systematic sample with a skip number of 7 (or a multiple of 7) would be problematic since the sampled editions would all be from the same day of the week (e.g., Friday). The day of the week may be related to the percentage of newspaper space devoted to news about entertainment.

2.29. Because of skipping names, two subjects listed next to each other on the list cannot both be in the sample, so not all samples are equally likely.

2.30. If we do not take a disproportional stratified random sample, we might not have enough Native Americans in our sample to compare their views to those of other Americans.

2.31. If a subject is in one of the clusters that is not chosen, then this subject can never be in the sample. Not all samples are equally likely.

2.33. The nursing homes can be regarded as *clusters*. A *systematic random sample* is taken of the clusters, and then a *simple random sample* is taken of residents from within the selected clusters.

2.34. (b)

2.35. (c)

2.36. (c)

2.37. (a)

2.38. False. This is a convenience sample.

2.39. False. This is a voluntary response sample.

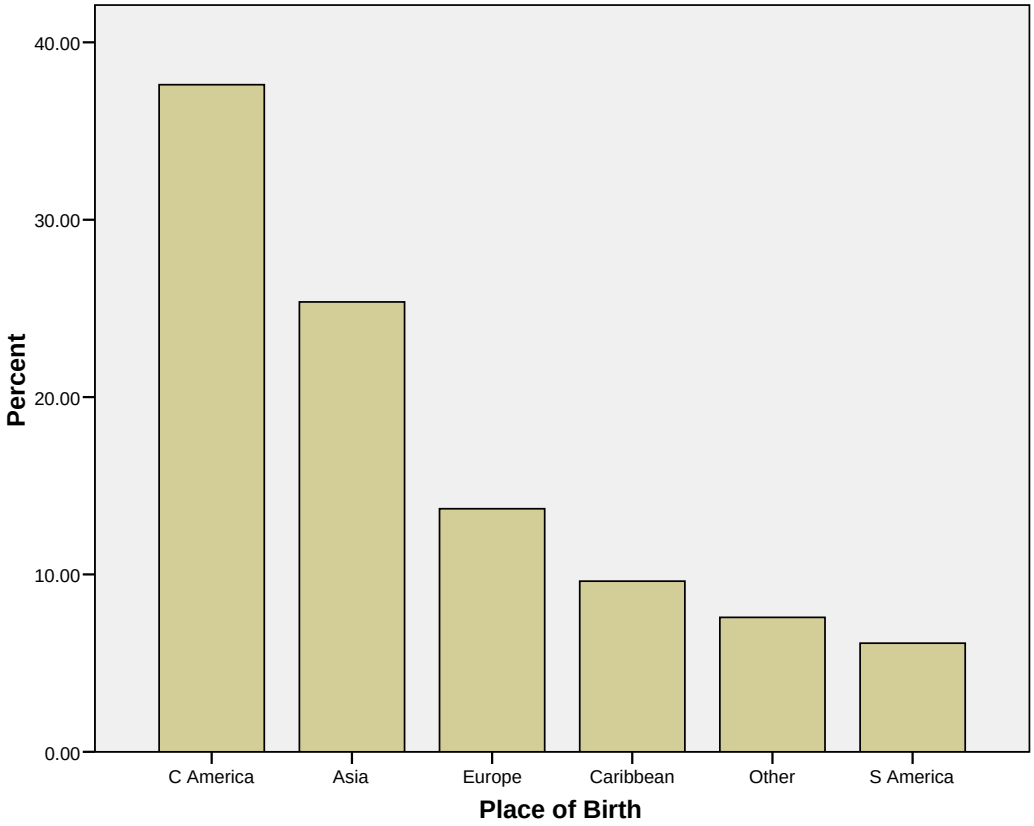
2.40. An annual income of $40,000 is twice the annual income of $20,000. However, 70 degrees Fahrenheit is not twice as hot as 35 degrees Fahrenheit. (Note that income has a meaningful zero and temperature does not.) IQ is not a ratio-scale variable.

Chapter 3

3.1. (a)

Place of Birth	Relative Frequency
Europe	13.7%
Asia	25.4%
Caribbean	9.6%
Central America	37.6%
South America	6.1%
Other	7.6%

(b)

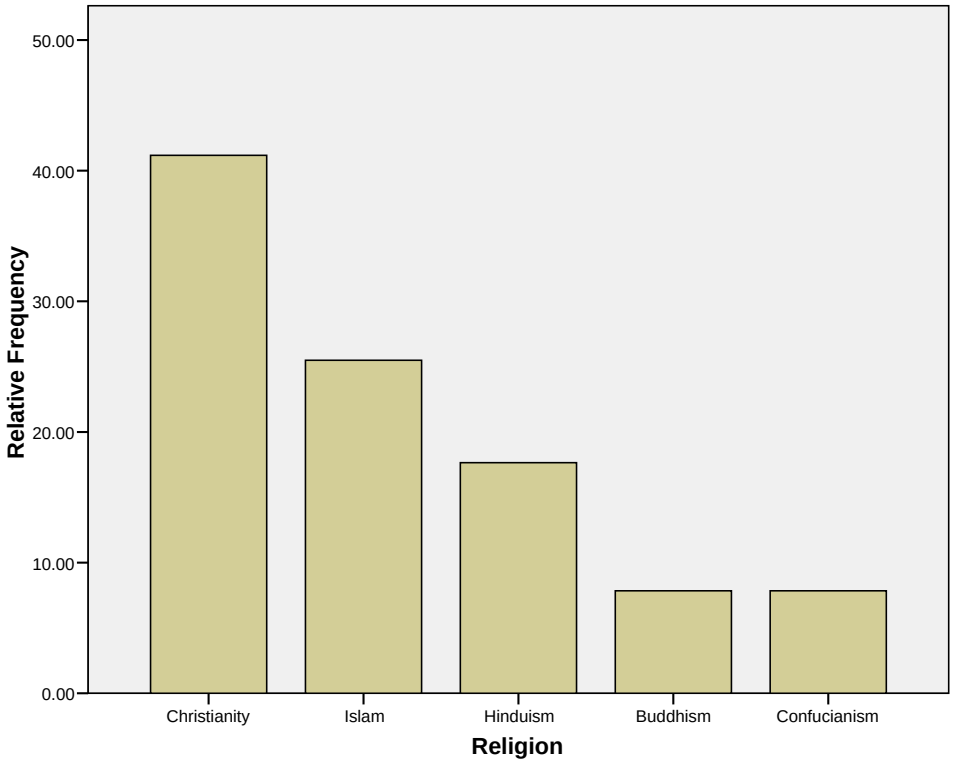


(c) "Place of birth" is categorical. (d) The mode is Central America.

3.2. (a)

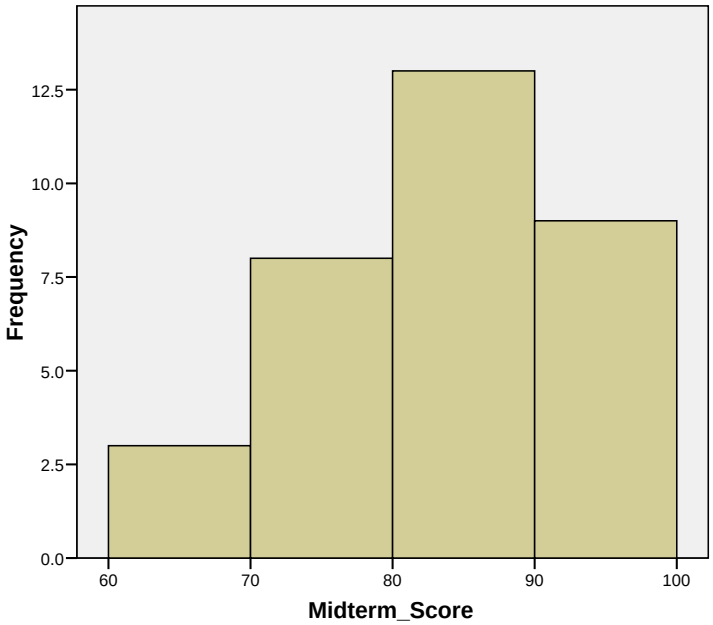
Religion	Relative Frequency
Christianity	41.2%
Islam	25.5%
Hinduism	17.6%
Confucianism	7.8%
Buddhism	7.8%

(b)



(c) The mode of these five religions is Christianity. Christianity is also the mode of all religions.

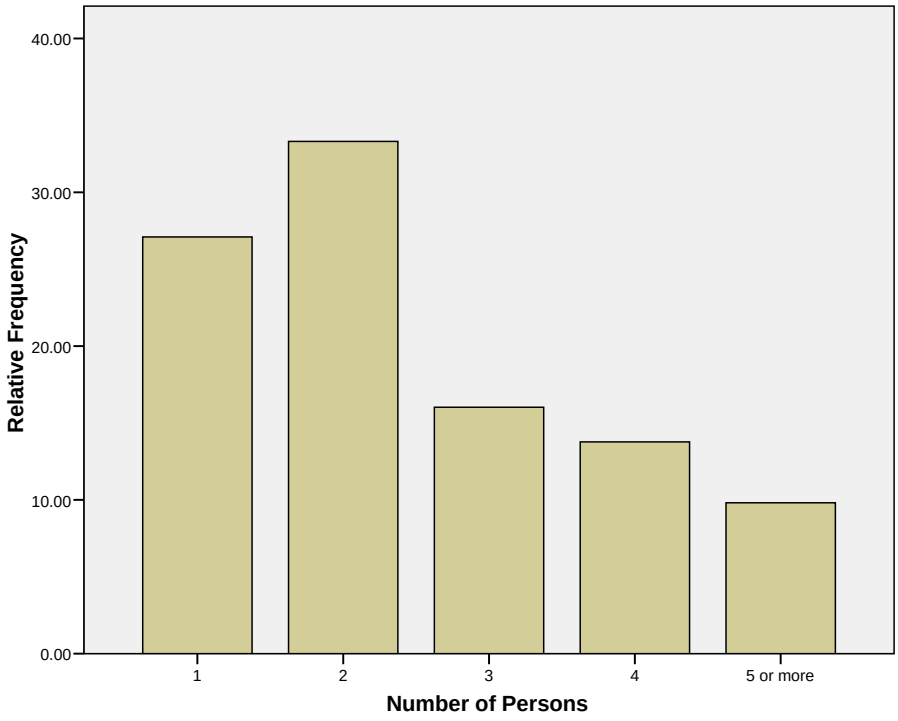
3.3. (a) There are 33 students. The minimum score is 65, and the maximum score is 98.
(b)



3.4. (a)

Number Persons	Relative Frequency
1	27.1%
2	33.3%
3	16.0%
4	13.8%
5 or more	9.8%

(b)

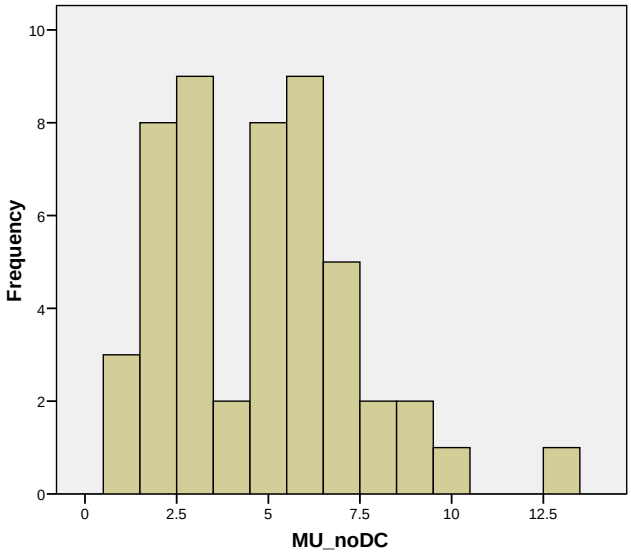


(c) The median household size is 2 persons, and the mode is also 2 persons.

3.5. (a)

		Frequency	Percent	Cumulative Percent
Valid	1	3	6.0	6.0
	2	8	16.0	22.0
	3	9	18.0	40.0
	4	2	4.0	44.0
	5	8	16.0	60.0
	6	9	18.0	78.0
	7	5	10.0	88.0
	8	2	4.0	92.0
	9	2	4.0	96.0
	10	1	2.0	98.0
	13	1	2.0	100.0
	Total	50	100.0	

(b)



The distribution appears to be bimodal and skewed to the right.

(c)

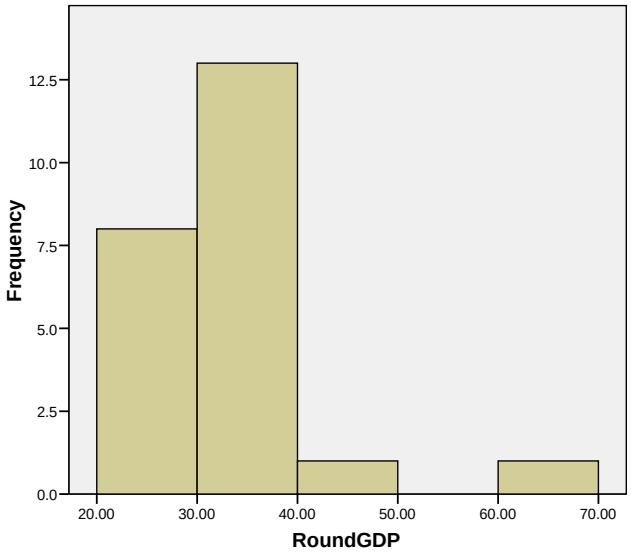
Stem	Leaves
1	000
2	00000000
3	000000000
4	00
5	00000000
6	000000000
7	00000
8	00
9	00
10	0
11	
12	
13	0

The stem-and-leaf plot shows the same bimodality and right skew that the histogram does.

3.6. (a) GDP is rounded to the nearest thousand

Stem (10 thousands)	Leaves (thousands
2	023
2	58899
3	00011122233
3	8
4	
4	
5	
5	
6	
6	
7	0

(b)



(c) The outlier in each plot is Luxembourg.

3.7. (a) The mean is (26 + 17 + 236 + 2 + 6)/5 = 287/5 = 57.4 abortions per 1000 women 15 to 41 years of age. (b) The median is 17 abortions per 1000 women 15 to 41 years of age. The mean and median are so different because California is an extreme outlier in this small data set.

3.8. (a) The mean is (0.3 + 1.8 + 2.3 + 1.2 + 1.4 + 0.7 + 9.9 + 20.1)/8 = 37.7/8 = 4.7 metric tons per person. The median is 1.6 metric tons per person. (b) The United States appears to be an outlier, since it is far greater than any other data value. (Without the United States, the mean is 2.5 and the median is 1.4.)

3.9. (a) The response "not far enough" is the mode. (b) We cannot compute and mean or median with these data since they are categorical.

3.10. (a)

Stem	Leaves
0	4679
1	133
2	0
3	9
4	4

(b) The mean is 16.6 days, and the median is 12 days.

(c)

Leaves 25 years ago	Stem	Leaves
	0	4679
875	1	133
440	2	0
21	3	9
0	4	4
5	5	

For the data from 25 years ago, the mean was 27.6 days, and the median was 24 days. The mean has decreased by 11 days, and the median has decreased by 12 days since 25 years ago. (d) Of the

11 observations, the median is 13 days. We cannot calculate the mean, but substituting 40 for the censored observation gives a mean of 18.7 days.

3.11. (a)

TV Hours	Frequency	Relative Frequency
0	79	4.0
1	422	21.2
2	577	29.0
3	337	17.0
4	226	11.4
5	136	6.8
6	99	5.0
7	23	1.2
8	34	1.7
9	4	0.2
10	23	1.2
12	14	0.7
13	1	0.1
14	7	0.4
15	2	0.1
18	2	0.1
24	1	0.1
Total	1987	100.0

(b) The distribution is unimodal and right skewed. (c) The median is the 994th data value, which is 2. (d) The mean is larger than 2 because the data is skew right by a few high values.

3.12.

Central America	Stem	Western Europe
8540	4	
85210	5	488
82	6	003678
	7	1268
	8	567
	9	0

Female economic activity seems greater, on average, in Western Europe than in Central America. Most of the values in Western Europe exceed the highest value in Central America. There appear to be more women in the labor force (per 100 men) in Western Europe than in Central America.

3.13. Since the mean is much greater than the median, the distribution of 2000 household income in Canada is most likely skewed to the right.

3.14. (a) The median is "2 or 3 times a month." The mode is "not at all." The data are centered around the respondents having sex about 2 or 3 months in the past 12 months. The most frequent answer to the question is "not at all." (b) The sample mean is 4.1, which means that, on average, the respondents had sex about 4 times a month in the past 12 months.

3.15. (a) The mode is "every day." The median is "a few times a week." (b) The mean is 3.7 times per week, which is lower than the 4.4 times a week in 1994.

3.16. (a) For each gender, the distribution of earnings is skewed to the right, since each mean is greater than its respective median. (b) The overall mean income is ($39,890×73.8 + $56,724×83.4)/(73.8 + 83.4) = $7674663.6/157.2 = $48,821.

3.17. (a) The response variable is median family income, and the explanatory variable is race. (b) We cannot find the median income for the combined groups since we do not know how many families are in each group. (c) We would need to know how many families were in each group.

3.18. (a) The distribution is skewed to the right. (b) The Empirical Rule only applies to bell-shaped distributions, so it does not apply here. (c) The median is 0. If the 500 observations were to shift from 0 to 6, the median would remain zero, since half of the data values fall below 0 and half fall above 0. This illustrates the resistance of the median to skewness and extreme values.

3.19. (a) Median: $10.13; mean: $10.18; range: $0.46; standard deviation: $0.22. (b) Median: $10.01; mean: $9.17; range: $5.31; standard deviation: $2.26. The median is resistant to outliers, but the mean, range, and standard deviation are highly impacted by outliers.

3.20. (a) Mean: 30; standard deviation: 9.0. (b) Minimum: 13; lower quartile: 25.5; median: 31; upper quartile: 36; maximum: 42.

3.21. The mean is 28.7, and the standard deviation is 12.5. The 2006 HDI ratings for the top 10 nations vary greatly.

3.22. (a) The life expectancies in Africa vary more than the life expectancies in Western Europe, because the life expectancies for the African countries are more spread out than those for the Western European countries. (b) The standard deviation is 1.1 for the Western European nations and 7.1 for the African nations.

3.23. (a) (i) $40,000 to $60,000; (ii) $30,000 to $70,000; (iii) $20,000 to $80,000. (b) A salary of $100,000 would be unusual because it is 5 standard deviations above the mean.

3.24. (a) Approximately 68% of the values are contained in the interval 32 to 38 days; approximately 95% of the values are contained in the interval 29 to 41 days; all or nearly all of the values are contained in the interval 26 to 44 days. (b) (i) The mean would decrease if the observation for the U.S. was included. (ii) The standard deviation would increase if the observation for the U.S. was included. (c) The U.S. observation is 5.3 standard deviations below the mean.

3.25. (a) 88.8% of the observations fall within one standard deviation of the mean. (b) The Empirical Rule is not appropriate for this variable, since the data are highly skewed to the right.

3.26. 10 is realistic; –20 is impossible since the standard deviation cannot be negative; 0 implies that every student scored 76 on the exam, which is highly improbable; 50 is too large (it is half of the possible range of scores).

3.27. (a) The most realistic value is 0.4, because the range is 5 times the length of this value. (b) The value of –10.0 is impossible since the standard deviation cannot be negative.

3.28. (d)