**CHAPTER 2 SOLUTIONS**

1. 180

****

1. It is impossible to know the exact number. However, we can estimate the value by assuming that equal numbers of students were absent 3, 4, 5, and 6 times. Accordingly, the total number absent exactly 3 times in twelfth grade is estimated to be 180/4 = 45.
2. 36 percent.
3. 86 percent.
4. Positively skewed. The bulk of students have six or fewer absences, and there are only a few students with an unusually high number of absences.
5. That the student’s parents were separated at the time the student was in eighth grade.

****

1. 8.
2. Of all students that reported their parents’ marital status in eighth grade, 1.7 percent reported that their parents were separated . Of all students in the *NELS* data set, 1.6 percent reported that their parents were separated.
3. Because PARMARL8 is nominal there is no intrinsic ordering to the categories; the numbers are assigned arbitrarily to them. As such, it does not make sense to consider the percent of people “below separated”, for example.
4. Married.
5. 9.

****

1. 52.
2. 9 + 52 = 61.
3. 10.4 percent.
4. 12.2 percent.
5. *Q2* is approximately 4.
6. GENDER is a nominal variable.



1. There are more females. There are 273 females and 227 males.
2. percent.
3. Yes. Every person in the data set is classified as either male or female, so the categories are mutually exclusive and exhaustive.
4. The effect of the change in starting point on the vertical axis is that it appears that there is a greater difference in the number of people of each gender in the data set.



* 1. For this solution, the graph was edited so that the bars are labeled with exact counts. If you do not do that, your answer will be an approximation of the answers given.
1. For males the answer is: (66/227) x 100 or 29.1 For females the answer is: (42/273) x 100 or 15.4 percent.

 

1. To accommodate the larger number of females who had never binged on alcohol.



1. The values are not equally-spaced, making it difficult to judge the shape of the distribution, in general, and the distance of the values in the lower tail from the bulk of the distribution, in particular.
2. A stem-and-leaf, histogram, or interactive line graph.
3. The shape of the distribution is negatively skewed, even more negatively skewed than it appears to be given the unequal spacing of values along the x-axis. Most schools have attendance rates above 90 percent; however, there are two schools with attendance rates below 50 percent.

****

1. The ordering of categories along the X-axis corresponds to way in which the data were input into the data file.
2. 5.



1. It is useful for ranking statisticians by their respective number of recent citations as recorded in *The American Statistician*.
2. He is 6th from the top.
3. The variable, number of recent citations/references, is represented by the X-axis in this graph. Since this variable is discrete, with a relatively small range of values, a bar graph would continue to be appropriate.



1. The distribution appears to be skewed positively, with a center at approximately 5 references (there are five values above 5 and six values below 5), and a spread from 0 to 16 references.



1. 123 students, or 24.6 percent, come from an urban area.



1. There is no “correct” answer to this question. It is a matter of preference.
2. Most students are from suburban areas, followed by rural, and the fewest are from urban areas.

Math Comprehension Stem-and-Leaf Plot

 Frequency Stem & Leaf

 2.00 6 . 12

 7.00 6 . 6678999

 9.00 7 . 002223334

 19.00 7 . 5566777777888888889

 12.00 8 . 000122444444

 8.00 8 . 56667789

 12.00 9 . 011222333333

 10.00 9 . 5566777788

 6.00 10 . 001333

 1.00 10 . 8

 3.00 11 . 223

 4.00 11 . 5678

 1.00 12 . 1

 Stem width: 10

 Each leaf: 1 case(s)

1. The variable is interval, so a line graph, histogram, or boxplot, could also be used.
2. 2 + 7 + 9 + 19 + 12 + 8 + 12 + 10 + 6 + 1 + 3 + 4 + 1 = 94.
3. 9 + 19 = 28.
4. No.
5. It is reasonably symmetric. The two tails are about equal in size and there are no outliers.
6. Because there are an even number of scores (94), the 50th percentile is the average of the two middle scores, which are in the 47th and 48th positions. Both of these values are 84, so the 50th percentile is 84.
7. 78; it occurs 8 times.
8. 121.
9. Yes.

Year of Birth Stem-and-Leaf Plot

 Frequency Stem & Leaf

 1.00 Extremes (=<1822)

 1.00 185 . 7

 .00 186 .

 1.00 187 . 6

 .00 188 .

 5.00 189 . 03345

 3.00 190 . 069

 1.00 191 . 5

 Stem width: 10

 Each leaf: 1 case(s)

1. 1.
2. Yes. One statistician was born on or before 1822.
3. The distribution is negatively skewed because of the low outlier.
4. Two statisticians were born in 1893. No other pair of statisticians was born in the same year.
5. 1915.
6. No. The exact value of the outlier cannot be determined from the display.

Number of References in American Statistician 1995-2005 Stem-and-Leaf Plot

 Frequency Stem & Leaf

 6.00 0 . 000011

 4.00 0 . 5677

 1.00 1 . 2

 1.00 1 . 6

 Stem width: 10

 Each leaf: 1 case(s)

1. 2.
2. The distribution is positively skewed despite the fact that there are no outliers. Most statisticians have fewer than 8 citations, but two have more than 10 citations.
3. The most frequently occurring number of citations is 0.
4. The 50th percentile is (1+5)/2 = 3.

Serum Cholesterol mg/dL (1) Stem-and-Leaf Plot

 Frequency Stem & Leaf

 1.00 1 . 3

 4.00 1 . 5555

 24.00 1 . 666666666666677777777777

 52.00 1 . 8888888888888888888899999999999999999999999999999999

 62.00 2 . 00000000000000000000000000011111111111111111111111111111111111

 68.00 2 . 22222222222222222222222222222222222233333333333333333333333333333333

 74.00 2 . 44444444444444444444444444444444444455555555555555555555555555555555555555

 46.00 2 . 6666666666666666666666666666777777777777777777

 31.00 2 . 8888888888888899999999999999999

 16.00 3 . 0000000011111111

 9.00 3 . 222223333

 3.00 3 . 444

 2.00 Extremes (>=353)

 Stem width: 100

 Each leaf: 1 case(s)

1. The exact value of the lowest score cannot be determined because it has been truncated to 130. All we know is that the exact value is between 130 and 140, inclusive.
2. Yes. There are two people with initial cholesterol levels that are greater or equal to 353.
3. Positively skewed due to the two high outliers.
4. Approximately 230. This stem-and-leaf plot reports values rounded to two significant digits, so that a value of 233, for example, is reported as 230. Given that there are 392 scores in this stem-and-leaf, the 50th percentile for these data, based on this stem-and-leaf, is estimated to be 230, which is the average of the 196th and 197th scores in this distribution. The 196th and 197th score each has a recorded value of 230.
5. The exact value of the highest score cannot be determined because it is recorded simply as an extreme value, noted to be greater than or equal to 353.
6. No, for two reasons. First, the scores are rounded to two significant digits and so we would not know the actual units digit of each value (e.g., an original score of 253 would be represented in this stem-and-leaf as 250). Second, we do not know the values of the two outliers.

Math Achievement in Eighth Grade Stem-and-Leaf Plot

 Frequency Stem & Leaf

 12.00 3 . 78899&

 47.00 4 . 0000112222233333444444

 73.00 4 . 555555666666667777777788888889999999

 101.00 5 . 0000000001111111122222222222333333344444444444444

 81.00 5 . 555555566666666677777777788888888889999

 77.00 6 . 0000000001111111222222223333334444444

 70.00 6 . 5555555666666667777777788888899999

 28.00 7 . 000123334444

 11.00 7 . 55777

 Stem width: 10.00

 Each leaf: 2 case(s)

 & denotes fractional leaves.

* 1. Approximately 37. The scores in the data set are given to two places after the decimal, but the stem and leaf plot reports only the two leading digits.
	2. Approximately two, because there is a single leaf and each leaf is noted to represent two cases.
	3. No.
	4. Approximately symmetric.
	5. 54.
	6. For general questions about the shape, level, and spread of a distribution, a stem-and-leaf plot provides a better visual summary of the distribution. However, when exact frequencies and variable values are desired, the frequency distribution table often provides a greater level of detail.

Systolic BP mmHg (1) Stem-and-Leaf Plot for

SEX= Men

 Frequency Stem & Leaf

 2.00 9 . 89

 3.00 10 . 012

 13.00 10 . 5555777788889

 24.00 11 . 000001111122222233344444

 22.00 11 . 5555556666777889999999

 22.00 12 . 0001111112222223444444

 20.00 12 . 55566677777777888999

 26.00 13 . 00000000011112222222333344

 16.00 13 . 5566677778888999

 14.00 14 . 00001111123444

 6.00 14 . 566789

 4.00 15 . 0023

 9.00 15 . 567899999

 4.00 16 . 0222

 2.00 16 . 78

 4.00 17 . 1234

 1.00 17 . 5

 8.00 Extremes (>=179)

 Stem width: 10

 Each leaf: 1 case(s)

Systolic BP mmHg (1) Stem-and-Leaf Plot for

SEX= Women

 Frequency Stem & Leaf

 1.00 9 . 2

 5.00 9 . 56778

 13.00 10 . 0000002233334

 18.00 10 . 555567777778888999

 17.00 11 . 00001122233333444

 19.00 11 . 5555666666778888999

 20.00 12 . 00000012223333334444

 27.00 12 . 556666677777778888889999999

 21.00 13 . 000111222222222333344

 13.00 13 . 5566778899999

 10.00 14 . 0000111223

 4.00 14 . 6788

 5.00 15 . 02234

 4.00 15 . 5589

 3.00 16 . 001

 4.00 16 . 5558

 3.00 17 . 033

 2.00 17 . 66

 11.00 Extremes (>=178)

 Stem width: 10

 Each leaf: 1 case(s)

1. Female. For males, the lowest systolic blood pressure value is 98, while for females it is 92.
2. Female. In the female distribution, the most frequently occurring systolic blood pressure is 132, while in the male distribution, it is 130.
3. The distributions appear to be quite similar in spread.
4. The shapes of the distributions appear to be quite similar; they are both similarly positively skewed.
	1.



1. Estimating from the histogram, the lowest score is approximately 34 and it appears 2 times.
2. Between 54 and 55.
3. It is slightly negatively skewed. There is a longer tail on the left than on the right.

* 1.
1. The group that owned a computer in eighth grade.
2. Yes, it appears to be higher for the group that owns a computer as the distribution for this group begins and ends at a higher point along the ses scale and the bulk of the scores in that distribution is located higher up on the ses scale than for the other group.
3. No. The spread of the two distributions appears to be quite similar.



1. The level of self-concept is higher for the twelfth grade – the bulk of the scores in the histogram for twelfth grade is further to the right on the x-axis than for eighth grade.
2. In eighth grade the scores are more closely clustered.
3. The scores are more negatively skewed in twelfth grade – the tail is longer in the negative direction or to the left.



1. In the interactive line graph.
2. The interactive line graph. The regular line graph gives a distorted view due the way in which values are formatted along the x-axis. The distance between values on the x-axis shrinks toward the lower end of the tail.



1. Approximately 75.
2. Negatively skewed.



1. The compressed scale reduces the appearance of a change in the percentages of students who use marijuana over time. Said differently, student behavior with respect to smoking marijuana appears to be more consistent over time with the compressed scale.
2. The percentage of seniors who have tried marijuana increases rather dramatically from 1992 through 1999.



1. $30,000. This means that 15 percent of the students are expecting to be earning less than $30,000 at age 30.
2. $40,000.
3. 73.9.
	1. Because PARMARL8 is only nominal. Variables must be at least ordinal-leveled for percentiles to be meaningfully calculated.
4. 102.
5. 25.
6. According to the frequency distribution table, the percentile rank of his eighth grade self-concept score is 74.8 inclusive. While higher than almost three-fourths of the individuals in the data set, one might have expected this person’s self-concept to be even higher.
7. He was least self-confident in twelfth grade. In eighth grade, his self-concept score of 25, gave him a percentile rank of 74.8 percent. In tenth grade, his self-concept score of 32, gave him a percentile rank of 88 percent. In twelfth grade, his self-concept score of 22, gave him a percentile rank of 7.2 percent.



* 1. The value of the 50th percentile is 1. Since the value 1 corresponds to being late one or two times, we know that half the students in the NELS study were late to school no more than only one or two times.
	2. The value of the IQR is the box height, or 2 – 0 = 2.
	3. The graph is positively skewed because there is a long whisker above the box and no whisker at all below the box.
	4. The minimum score is the same as the 25th percentile score. That is, at least 25% of students in the *NELS* data set were never late in twelfth grade.



1. Negatively skewed. All of the outliers are on the more negative end of the scale.
2. Eighth grade.
3. Twelfth grade as measured by the 50th percentile.
4. The twelfth grade distribution contains the highest self-concept score, which is approximately 43.
5. Twelfth grade.



1. West.
2. West.
3. There is a student in the Northeast that took only one year of math in high school.
4. The distributions of all regions appear to be similar in spread according to the interquartile range.
5. Because 25 percent of the scores in any distribution fall between *Q1* and *Q2,* the number ofstudents in the Northeast whose scores fall between these two quartiles will be .25x106 or approximately 27 for those in the Northeast regardless of the variable in question.



1. Schools that offer only rigorous high school academic programs are likely to be smaller in scale (e.g., they may be non-public schools) than those that offer academic programs, which could explain the relatively smaller number of AP course offerings overall.
2. By using the percentage of all course offerings per school.



1. Yes.
2. The large presence of extremely low self-concept scores for females as compared to males.
3. Yes. Box and whiskers are themselves comparable for males and females.









* 1.
1. Positively skewed. .
2. Negatively skewed.
3. Slightly positively skewed.
4. Positively skewed.
5. Positively skewed.
6. Positively skewed.

1. South
2. West
3. Midwest
4. Northeast
5. (1) The variable is nominal.
6. (1) The variable is ordinal with only three categories.
7. (1) The variable is nominal.
8. (1) or (2) The variable is interval, so (2) is appropriate, but it takes on only four values, so (1) is also appropriate.
9. (1) or (2) The variable is ratio, so (2) is appropriate, but there are only four categories, so (1) is also appropriate.
10. (2) The variable is interval and takes on many values.
11. (2) The variable is ratio and takes on many values.
12. (1) or (2) The variable is ordinal, so (2) is appropriate, but it takes on only six values, so (1) is also appropriate.
13. 72.89.
14. 13 students scored 72.89.
15. 72.89.
16. 55.7.
17. 59.9.
18. 94.9 percent.
19. According to the histogram, it is reasonably symmetric.



1. 1 – 4 hours was spent weekly on extracurricular activities in twelfth grade.



1. 125.
2. 1 – 4 hours.
3. Slightly positively skewed. The bulk of the students spend less than 20 hours per week on extracurricular activities, but a few students spend quite a bit more than that.
4. May 2003. It is highlighted because it is when President Bush signed the Jobs and Growth Act.
5. 130,000,000 and 5.7 percent.
6. August 2005.
7. It is not fair to conclude that. Among the most serious flaws with the graph is the restriction of the time range. It depicts fewer than two of the eight years President Bush was in office. In a graph created by Brendan Nyhan, we see that a longer time frame shows that President Bush’s term began with an unemployment rate that was lower than it was even after the Jobs and Growth Act took effect.



Even though there is only a relatively small difference between the velocities, the graphic accentuates the difference because the vertical axis does not start at 0, one bar is about twice the height of the other, making it appear that Dickey’s knuckleball was twice as fast in 2012 than it was in 2013 When the graph is edited so that the vertical axis starts at 0, the comparison is more accurate.



1. 31% of the students in one of the top six most in-demand college majors majored in business.
2. The pie graph implies that the percentages are out of a meaningful whole. Instead, a bar graph of counts could give us a sense of the relative sizes of these majors without implying that there are not any others.
3. 10.5%.
4. Assuming that recent graduates and experienced graduates can be combined to represent all graduates in a certain time period it makes sense for this bar graph to be stacked.
	1. *C4 D1  Q1  D3  C50  Q3*
	2. The statement is wrong. For example, if everyone scores 10, then a raw score of 75 is not surpassed by any of the scores in the distribution. A raw score does not, in general, give information about percentages.
	3. The statement is wrong. For example, if 100 people take the test, then the highest percentile rank possible is 99, that is, the highest score will have 99 percent of the scores falling below it, unless there is a tie.
	4. The statement is wrong. This depends on the percentage of people below Alice’s and Ellen’s scores. For example, if there were 10 people taking the exam and eight people scored 40, Alice scored 50, and Ellen scored 100, then the percentile rank of Alice’s score is 80, while the percentile rank of Ellen’s score is 90, not twice that of Alice’s score. Of course, a percentile rank can never be 100 or above.
	5. The statement is wrong. Percentiles are not necessarily interval.
	6. The statement is wrong. If the students at this student’s school are exceptionally good in math, having a percentile rank of 85 in math relative to this school could be better than having a percentile rank of 95 in science relative to the entire city.
	7. The statement is wrong. The validity of the conclusion depends on the number of people questioned. Whereas 5 out of 5 is 100 percent, 5 out of 100 is only 5 percent. A description in terms of relative frequency or percentage would probably be more appropriate in this situation than a description in terms of frequency.
	8. They are all possible values for percentiles. Only 50 is a possible value for a percentile rank.
	9. A distribution with outliers that are balanced in both the positive and negative directions can be symmetric.
	10. A boxplot would have only one whisker, for example, if the 75th percentile was the same as the maximum score, that is, if the top 25 percent of the scores in the distribution were all the same value.



