Chapter 1 Looking at Data—Distributions

1.1. The regular price for the Smokey Grill Ribs coupon is 20; the discount price is 11.

1.2. The value of the coupon is computed by subtracting the DiscPrice from the RegPrice. It is quantitative because arithmetic operations such as the average value would make sense.

	A	B	C	D	E	F	G
1	ID	Type	Name	Item	RegPrice	DiscPrice	Value
2	1	Italian	Domo's	Pizza	20	10	10
3	2	Italian	Mama Rita's	Pizza	20	12	8
4	3	BBQ	Smokey McSween's	Barbecue	30	17	13
5	4	BBQ	Smokey Grill	Ribs	20	11	9
6	5	Mexican	Dos Amigos	Tacos	16	8	8
7	6	Mexican	Holy Guacamole	Steak fajitas	13	8	5
8	7	Seafood	Sea Grille	Shrimp platter	20	11	9

1.3. Who: The cases are coupons; there are seven cases. What: There are six variables: ID, Type, Name, Item, RegPrice, and DiscPrice. Only RegPrice and DiscPrice have units in dollars. The data might be used to compare coupons with one another to see which is better. We would not want to draw conclusions about other coupons not listed.

1.4. Cases: apartments. Five variables: rent (quantitative), cable (categorical), pets (categorical), bedrooms (quantitative), distance to campus (quantitative).

1.5. Answers will vary. For example, giving the actual exam scores will allow the student to see the absolute change in grade.

1.6. Answers will vary. **(a)** For example, number of graduates could be used for similar-sized colleges. **(b)**One possibility might be to compare graduation rates between private and public colleges.

1.7. (a) The cases are students. **(b)** Four variables: Favorite choice for online research ("Google or Google Scholar," "Library database or website," "Wikipedia or online encyclopedia," "Other"), Age (reasonable age range for first-year college students—17 to 30, etc.), Sex (M or F), and Major (could be a big list—statistics, math, engineering, English, etc.). **(c)**Age is quantitative; the rest are categorical. **(d)** The label is the number1 to 552. **(e)** Who: part **(a)** answer. What: part **(b)** and **(c)**. Why: We could look at the distribution of favorite choice across different age groups, majors, and sex.

1.8. (a) The cases are summer jobs. **(b)** Variables might include position, company, hourly wage, whether the job is on or off campus, and hours per week (other answers are possible). **(c)** Position (categorical), company (categorical), hourly wage (quantitative), on or off campus (categorical), hours per week (quantitative) (other answers are possible). **(d)** We could use a number as a label. The reason for doing so is that there could be several jobs with the same company or position that you would need to differentiate from one another. **(e)** Who: part **(a)** answer. What: part **(b)** and **(c)**. Why: To compile a list of available summer jobs and possibly compare them. We would not want to draw conclusions about other jobs not listed.

1.9. (a) The cases are employees. **(b)** Employee identification number (label), last name (label), first name (label), middle initial (label), department (categorical), number of years (quantitative), salary (quantitative), education (categorical), age (quantitative). **(c)** Sample data would vary.

	A	B	C	D	E	F	G	H	I
1	EIN	Last	First	Middle	Department	Years	Salary	Education	Age
2	001	Marley	Bob	M	Sales	4	45000	some college	34
3	002	Fisher	Margeret	A	Sales	8	54000	college degree	37
4	003	Marin	Jane	E	Admin	2	39000	high school	25

1.10. Answers will vary. For example, variables might include crime rates (quantitative): cities with high crime rates are less favorable, income (quantitative); cities with higher mean income might mean better opportunities for the residents, cost of living (quantitative); high cost of living means less income available for social factors, entertainment, and cultural activities (categorical); more activities make a city more appealing, or taxes (quantitative); lower tax rates lead to more disposable income. However, higher tax rates could mean a city has more free services.

1.11. Age: quantitative, possible values 16to ? (what would the oldest student's age be?). Sing: categorical, yes/no. Can you play: categorical, no, a little, pretty well. Food: quantitative, possible values $0 to ? (what would be the most a person might spend in a week?). Height: quantitative, possible values 2feet to 9 feet (check the *Guinness Book of World Records*).

1.12. Answers will vary. For example, "Do you like math?" Answer choices could be yes/no (categorical).This could also be rated on a 1 (I hate math) to 5 (I love math) scale.

1.13. Answers will vary. A few possibilities would be graduation rate, student–professor ratio, and job placement rate.

1.14. (a) The cases are states. **(b)** The label would be the name of the state. **(c)** The number of students from the state who attend college and the number of students who attend college in their home state are both quantitative. **(d)** Answers will vary. One possibility is that the number of students who attend college would address the education level of the state. **(e)** The ratio would determine how popular staying in-state is and, possibly, the availability of colleges.

1.15. Answers will vary. For example, each state could be divided as a percentage of the total of the nation's fatalities to show state differences; the disadvantage is that states with more population would have a higher number of fatalities. Instead, each state's fatalities could be divided by the state population to get a percentage for each state; this would be a better way to compare state-to-state rates of drunk-driving fatalities.

1.16. Answers will vary. The bar graph is easier to read, but the pie chart does a better job showing the dominance of Google as a source, filling almost three-quarters of the pie.

1.17. Both regular and split stemplots are shown. The first exam scores are left-skewed; the middle is around 80.

5	79
6	158
7	00335558
8	000223557
9	00122448

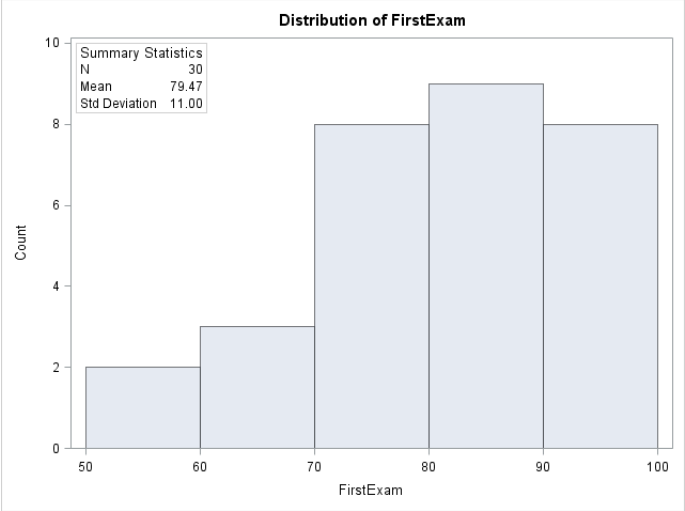
5	79
6	1
6	58
7	0033
7	5558
8	000223
8	557
9	0012244
9	8

1.18. Answers will vary. However, the stemplot in Figure 1.8 shows a bit more detail, which is useful for comparing the two distributions.

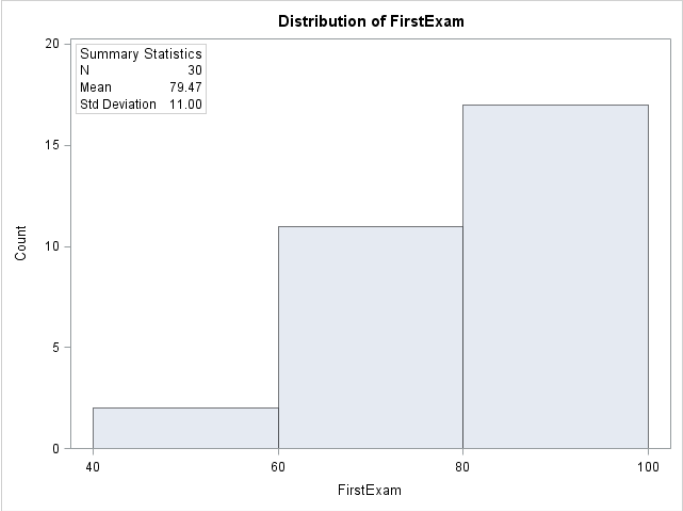
1.19. (a) The stemplot is below.**(b)** Use two stems, even though one is blank. Seeing the gap is useful.

3	3
3	58
4	1223
4	6777899
5	1135
5	9
6	344
6	
7	2

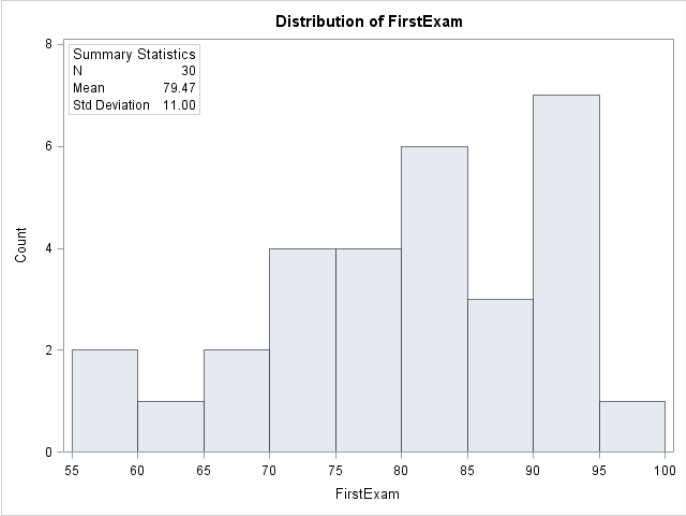
1.20. (a) Shown below. **(b)** Student preferences will vary. The stemplot has the advantage of showing each individual score. Note that this histogram has the exact same shape as the first stemplot in **Exercise 1.17.**



1.21. The larger classes hide a lot of detail; there are now only three bars in the histogram.



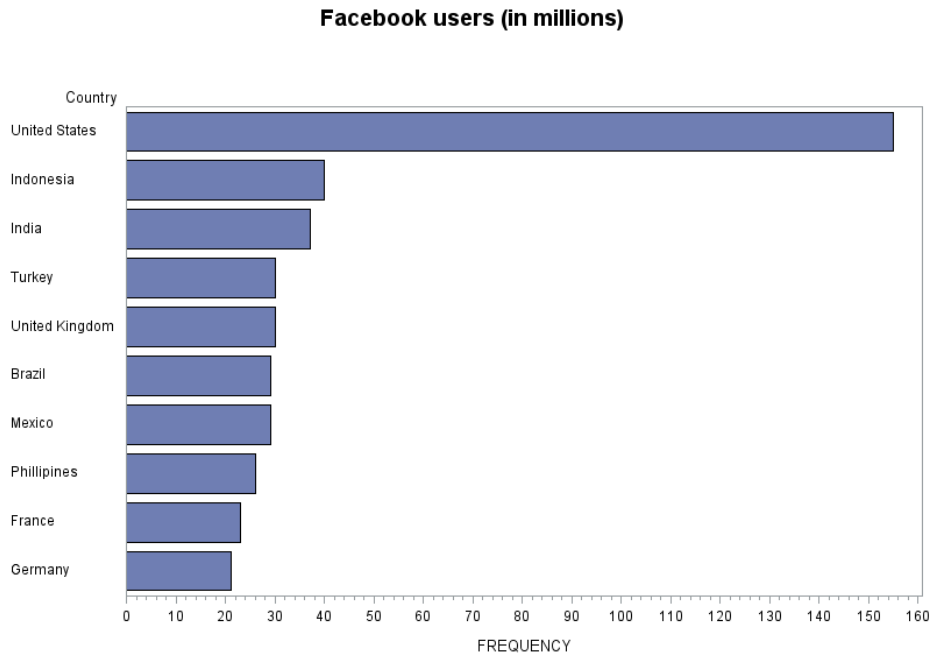
1.22. This histogram shows more detail about the distribution (perhaps more detail than is useful). Note that this histogram has the same shape as the stemplot with split stems shown in the solution to **Exercise 1.17.**



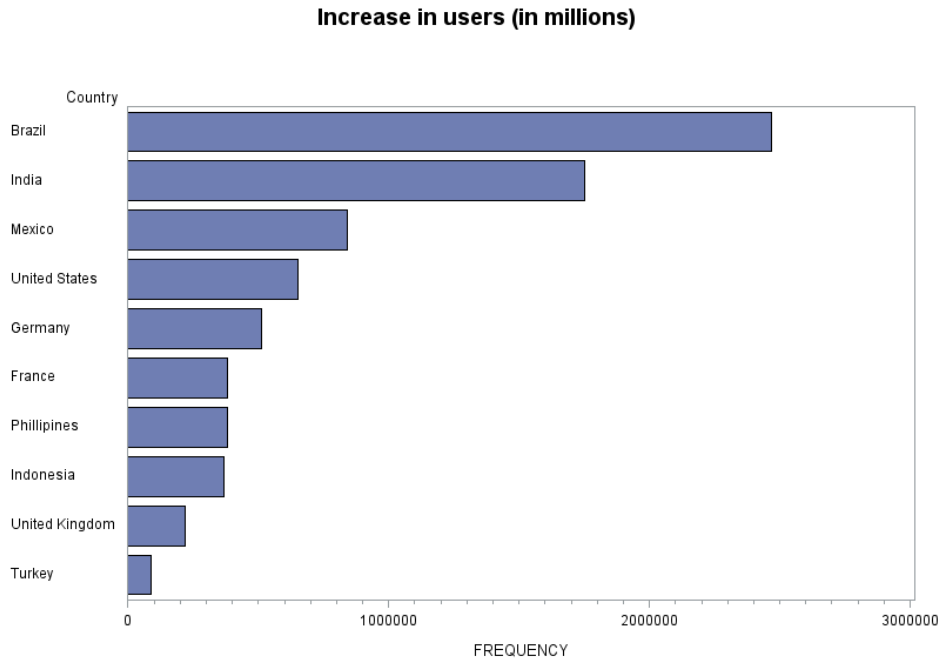
1.23. A stemplot or histogram can be used (possible stemplots are shown in the solution to **Exercise 1.17** and histograms in **Exercises 1.20** and **1.22**); the distribution is unimodal and left-skewed, centered near 80, and range from 55 to 98. There are no apparent outliers.

1.24. (a) Florida, California, Texas, and New York have more than 15,000,000 people. **(b)** None is particularly influential. Florida, Texas, and New York are all in the 40s (toward the lower end of the distribution). California has about 60.14 students per 1000 people (toward the upper end but not particularly high). The states with the largest number of undergraduates are Arizona, Iowa, and Utah (relatively "small" states).

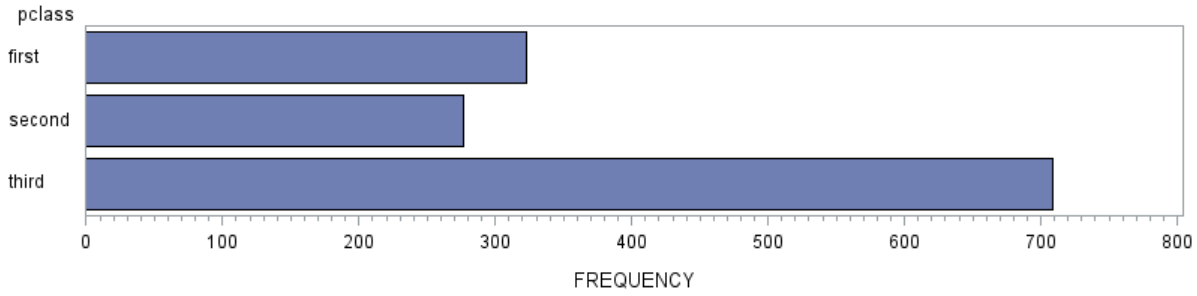
1.25. (a) Shown below. **(b)** The United States is a clear outlier. It has four or five times as many Facebook users as the other countries, despite having a population smaller than some of the other countries. **(c)** The United States dominates; many other countries shown have similar amounts of Facebook users.



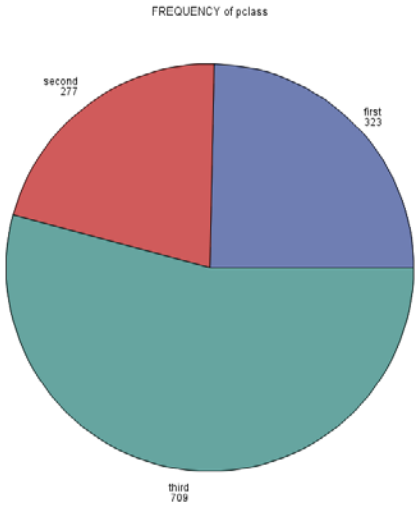
1.26. (a) Shown below. **(b)** Brazil is the leading country in Facebook user growth, followed by India, then Mexico. **(c)** A stemplot would not be better because the data are categorical and better represent the different countries. **(d)** Countries with higher Facebook user growth show more online presence and would have potential for growth among online marketing and other online business ventures.



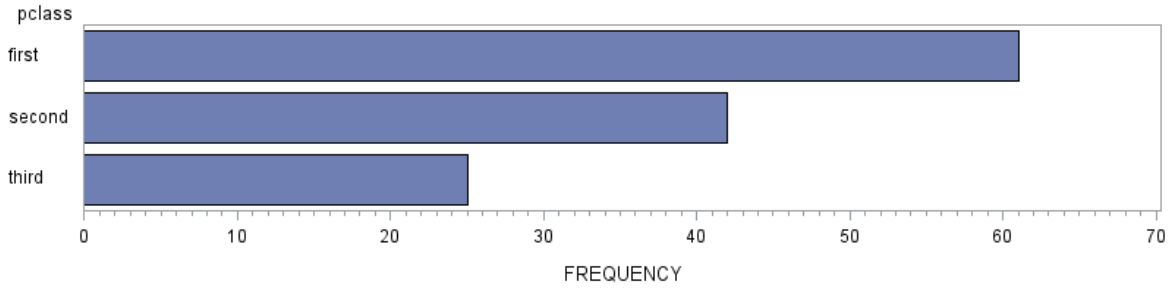
1.27. (a) Bar graph below. **(b)** Second class had the fewest passengers. Third class had by far the most, over twice as many as in first class. **(c)** A bar graph of the percents (relative frequency) would have the same features.



1.28. (a) Pie chart below. **(b)** What is clear from the pie chart is that third-class passengers were over half of the passengers. First class was about 25%.The relative sizes are not quite as easy to see in the bar chart.



1.29. We divided the number of survivors by the number of passengers in each class and then multiplied by 100 to get a percent. A bar graph is appropriate because we now have three "wholes" to consider



1.30. (a) Stemplot shown below; each number is rounded to the nearest 10, so 266 is 2660. **(b)** The distribution is somewhat symmetric. **(c)** There appears to be one large outlier at 4210. **(d)** The shape is roughly symmetric, the center is around 3010, the range is from 2660 to 4210.

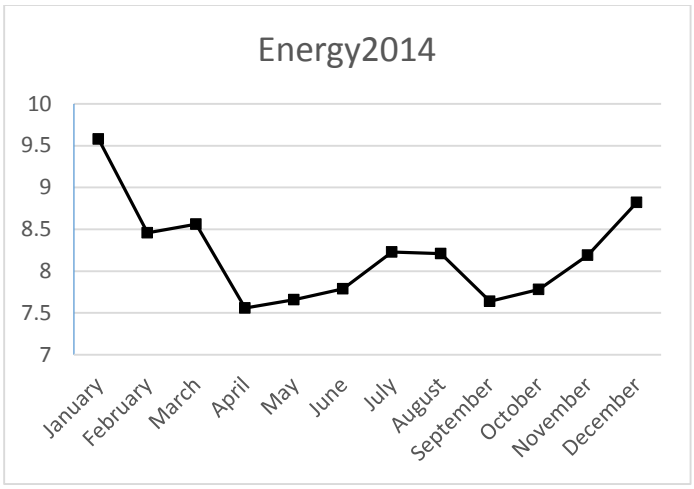
26	6
27	9
28	
29	5688

30	3577
31	02235
32	336689
33	
34	9
35	148
36	1
37	
38	
39	
40	
41	
42	1

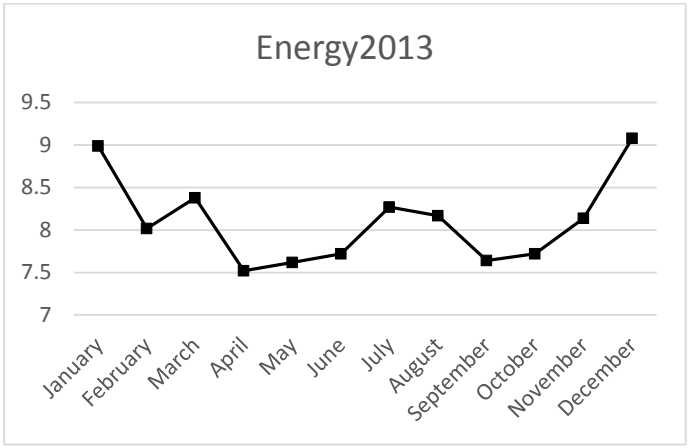
1.31. (a) Stemplot shown below; each number is rounded to the nearest 10, so 268 is 2680. **(b)** The distribution is somewhat right-skewed. **(c)** There appears to be one small outlier at 2680. **(d)** The shape is right-skewed, the center is around 3200, and the range is from 2680 to 3950.

26	8
27	
28	
29	5
30	37
31	02347779
32	0358
33	2333
34	68
35	29
36	28
37	
38	02
39	5

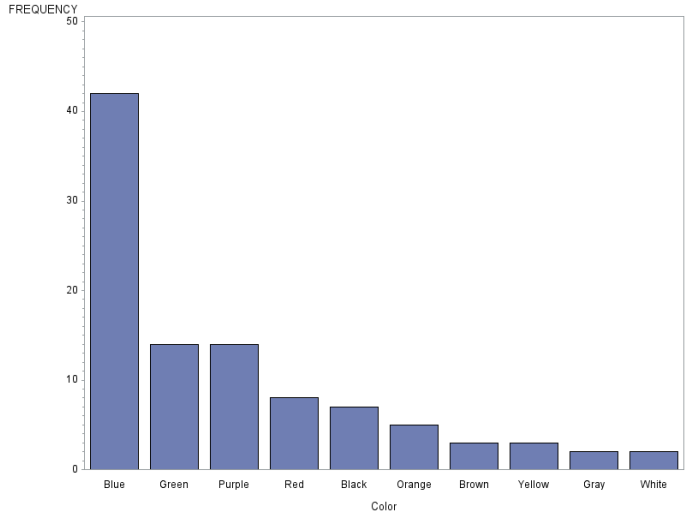
1.32. (a) Energy usage peaks in winter (December–January) and is again high in midsummer (July and August). **(b)** See time plot below. **(c)** Answers will vary. One possibility is that the time plot makes the variability easier to distinguish.



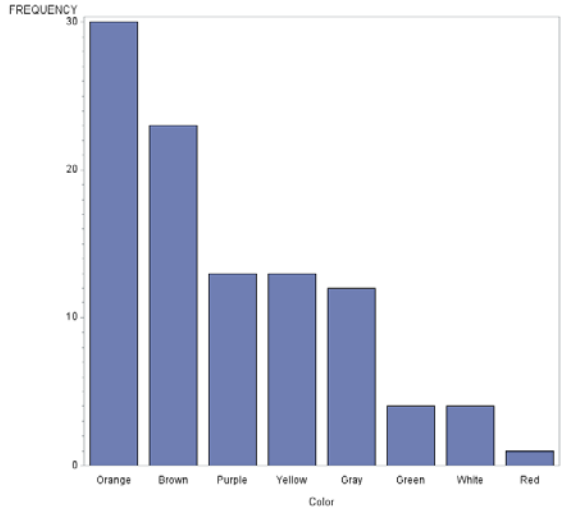
1.33. (a) 2013 still has the highest usage in December and January. See time plot below. **(b)** The patterns are very similar, but the values for the winter months in 2014 are somewhat higher than those in the 2013 winter months. These differences are most likely due to weather.



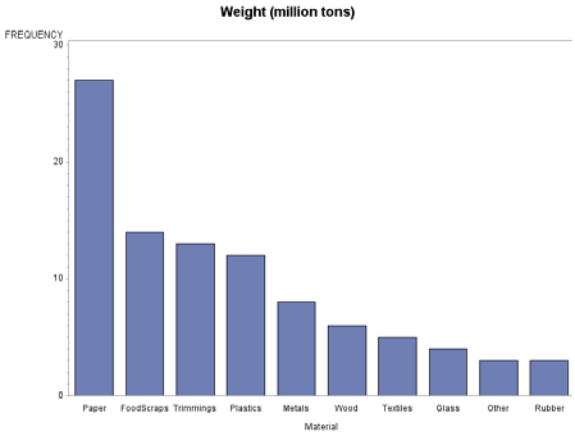
1.34. Bar graph below left. For example, blue is by far the most popular choice; 70% of respondents chose three of the 10 options (blue, green, and purple).

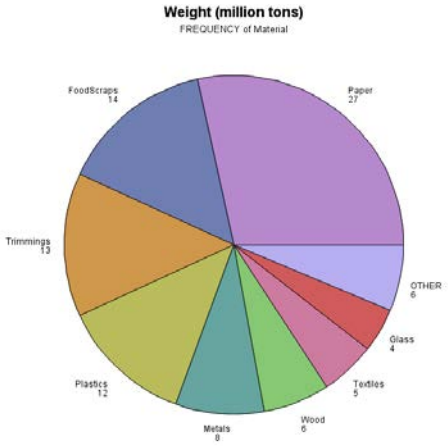


1.35. Bar graph below right. For example, opinions about least-favorite color are somewhat more varied than favorite colors. Interestingly, purple is liked and disliked by about the same percentage of people.

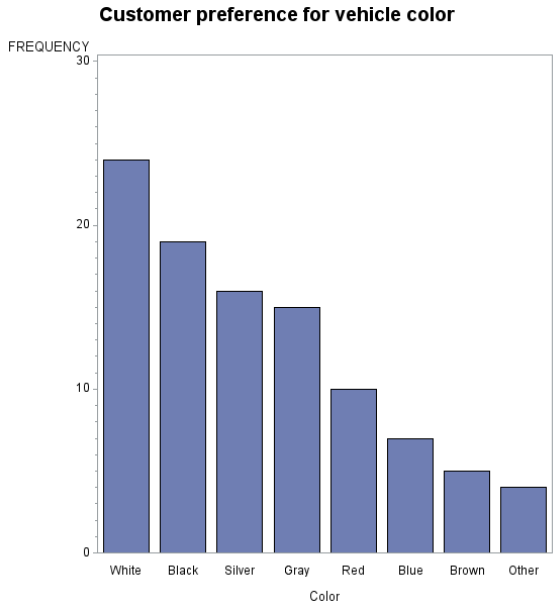


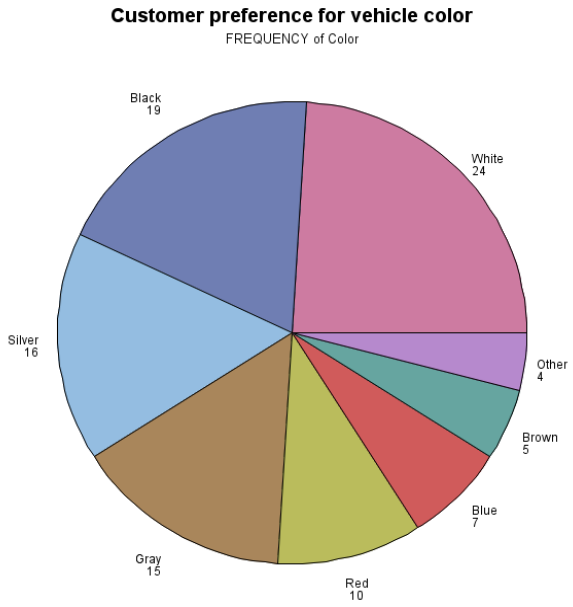
1.36. (a) Values are rounded. **(b)** Shown below. **(c)** Shown below.





1.37. White is the most popular color in 2012 for North America, followed by Black, Silver, and Gray. For marketing techniques, answers will vary.





1.38. Sketches will vary. The distribution of coin years would be left-skewed because newer coins are more common than older coins.

1.39. (a) Four variables: GPA, IQ, and self-concept are quantitative; gender is categorical. **(b)** See below, left. **(c)** Unimodal and skewed to the left, centered near 7.8, spread from 0.5 to 10.8. **(d)** There is more variability among the boys; in fact, there seems to be two groups of boys: those with GPAs below 5 and those with GPAs above 5.

0	5
1	8
2	4
3	4689
4	0679
5	1259
6	0112249
7	22333556666666788899
8	0000222223347899
9	002223344556668
10	01678

Female		Male
	0	5
	1	8
	2	4
4	3	689
7	4	069
952	5	1
4210	6	129
98866533	7	223566666789
997320	8	0002222348
65300	9	2223445668
710	10	68

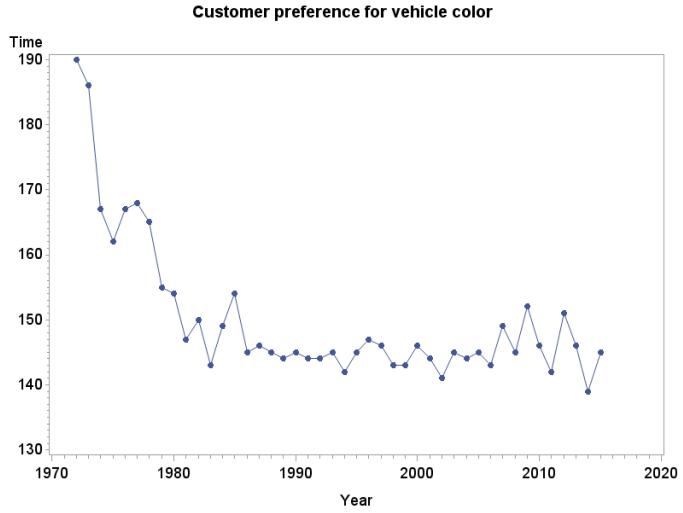
1.40. The stemplot is at right, with split stems. The distribution is unimodal and fairly symmetric—perhaps slightly left-skewed—with center around 110 (clearly above 100). IQs range from the low 70s to the high 130s, with a "gap" in the low 80s.

7	24
7	79
8	
8	69
9	0133
9	6778
10	0022333344
10	555666777789
11	0000111122223334444
11	55688999
12	003344
12	677888
13	02
13	6

1.41. The stemplot is at right, with split stems. The distribution is unimodal and skewed to the left, with center around 59.5. Most self-concept scores are between 35 and 73, with a few below that, and one high score of 80 (but not really high enough to be an outlier).

2	01
2	8
3	0
3	5679
4	02344
4	6799
5	1111223344444
5	556668899
6	00001233344444
6	55666677777899
7	0000111223
7	
8	0

1.42. The time plot below shows that women's times decreased quite rapidly from 1972 until the mid-1980s. Since that time, they have been fairly consistent, with slightly more scatter in recent years.



1.43. Without Suriname: $\overline{X}$ = 16.29. With Suriname: $\overline{X}$ = 23.96.

1.44. $\overline{x}$ = 82.8.

1.45. The ordered list is: 2 4 5 5 5 5 6 6 7 8 10 11 12 13 16 17 19 19 24 25 32 38 49 53 208.

M = 12. Without the outlier, the median is 11.5; with the outlier, the median is 12. The outlier does not greatly influence the median.

1.46. *M* = 103.5.

1.47. *M* = 84.

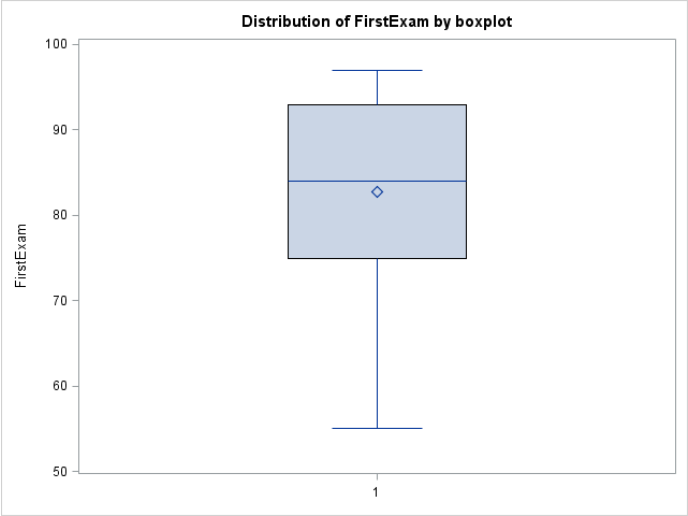
1.48. The median was in position 5.5, so there are five observations on either side. (5 + 1)/2 = 3. Counting in 3 from either end, we find Q_1 = 75 and Q_3 = 93.

1.49. The mean is $\overline{x} = \frac{77 + 289 + 128 + \ldots + 25}{80} = 196.575$ min (the value 196 in the text was rounded). The quartiles and median are in positions 20.5, 40.5, and 60.5. The sorted data are given below. Based on this:Q_1 = 54.5, *M* = 103.5, Q_3 = 200.

1	2	2	3	4	9	9	9	11	19
19	25	30	35	40	44	48	51	52	54
55	56	57	59	64	67	68	73	73	75
75	76	76	77	80	88	89	90	102	103
104	106	115	116	118	121	126	128	137	138
140	141	143	148	148	157	178	179	182	199
201	203	211	225	274	277	289	290	325	367
372	386	438	465	479	700	700	951	1148	2631

1.50. Min = 55, Q_1 = 75, *M* = 84, Q_3 = 93, Max = 97.

1.51. Shown below.



1.52. $IQR = Q_3 - Q_1 = 93 - 75 = 18$. $Q_1 - 1.5IQR = 75 - 1.5(18) = 48$. $Q_3 + 1.5IQR = 93 + 1.5(18) = 120$. The lowest score would need to be lower than 48 to be considered an outlier according to this rule.

1.53. From software: $s^2 = 157.07$, $s = 12.53$.

1.54. Picking the same number for all six cases results in a standard deviation of 0.

1.55. Without Suriname: Min = 55, $Q_1 = 75$, $M = 84$, $Q_3 = 93$, Max = 97. With Suriname: Min = 2, $Q_1 = 5.5$, $M = 11.5$, $Q_3 = 21.5$, Max = 53. Of course, the max changes drastically with the outlier removed; otherwise, the other numbers in the five number summary do not change drastically. This shows that, generally, the five number summary is robust.

1.56. 950(300/1200) = 237.5.

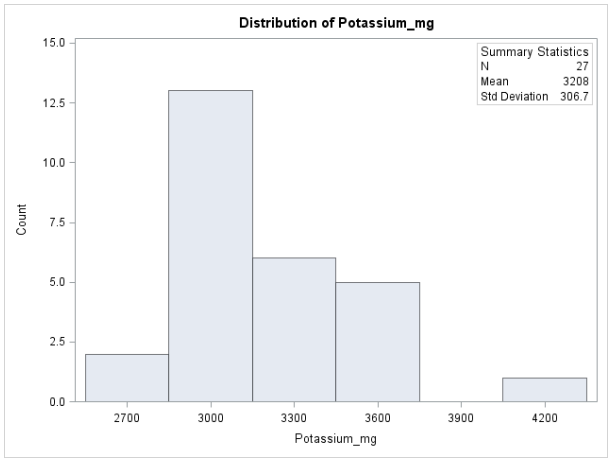
1.57. (a) $\bar{x} = 3208.44$. **(b)** $M = 3130.37$. **(c)** Because the distribution is right-skewed with a potential outlier, the median is a better measure of center. See histogram in **Exercise 1.61**.

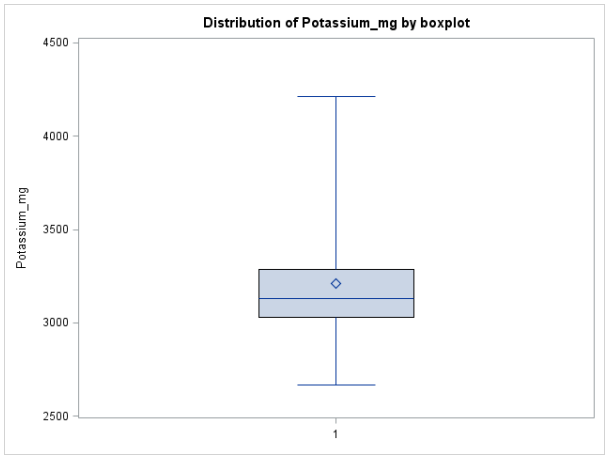
1.58. (a) $\bar{x} = 3313.38$. **(b)** $M = 3245.62$. **(c)** Because the distribution is symmetric with no outliers, the mean is a better measure of center. See histogram in **Exercise 1.62**.

1.59. (a) $s = 306.68$. **(b)** $Q_1 = 3027.64$, $Q_3 = 3286.95$. **(c)** Min = 2664.38 (this is the smallest value), $Q_1 = 3027.64$ (this value has 25% of the observations below it), $M = 3130.37$ (this is the middle observation or has 50% of the observations below or above it), $Q_3 = 3286.95$ (this value has 75% of the observations below it), Max = 4213.49 (this is the largest value). **(d)** The five number summary would be better for this distribution because it is right-skewed with a potential outlier. See histogram in **1.57**.

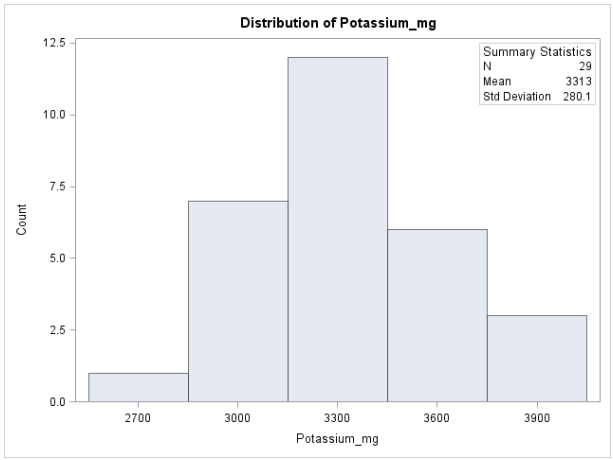
1.60. (a) $s = 280.12$. **(b)** $Q_1 = 3136.48$, $Q_3 = 3481.39$. **(c)** Min = 2680.07 (this is the smallest value), $Q_1 = 3136.48$ (this value has 25% of the observations below it), $M = 3245.62$ (this is the middle observation, or has 50% of the observations below or above it), $Q_3 = 3481.39$ (this value has 75% of the observations below it), Max = 3946.31 (this is the largest value). **(d)** The mean and standard deviation would be better for this distribution because it is symmetric without outliers. See histogram in **1.58**.

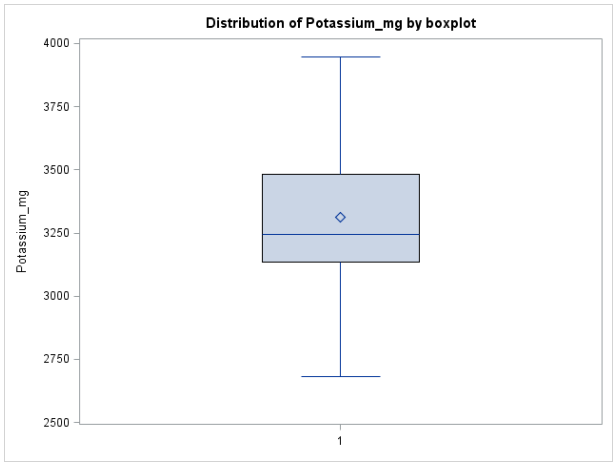
1.61. (a) Shown below. The distribution is right-skewed with a potential outlier **(b)** Shown below. **(c)** Preference will vary. The only advantage of the stemplot is that it preserves the data; otherwise, the histogram is likely better. The boxplot is also fine but hides some of the details that the histogram shows.





1.62. (a) Shown below. The distribution is symmetric with no outliers. **(b)** Shown below. **(c)** Preference will vary. The only advantage of the stemplot is that it preserves the data; otherwise, the histogram is likely better. The boxplot is also fine but hides some of the details that the histogram shows.

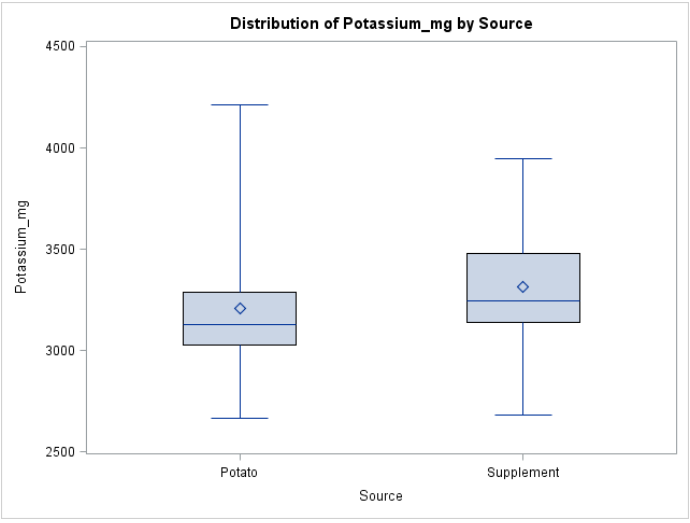




1.63. The KPOT values (shown on the left) are right-skewed, whereas the KSUP (shown on the right) values are fairly symmetric. The center for KSUP (right) is higher than the center for the KPOT (left). Also, the KPOT values (left) are more spread out than the KSUP values (right).

6	26	
9	27	
	28	8
8865	29	
7753	30	
53220	31	5
986633	32	37
	33	02347779
9	34	0358
841	35	2333
1	36	68
	37	29
	38	28
	39	
	40	02
	41	5
1	42	

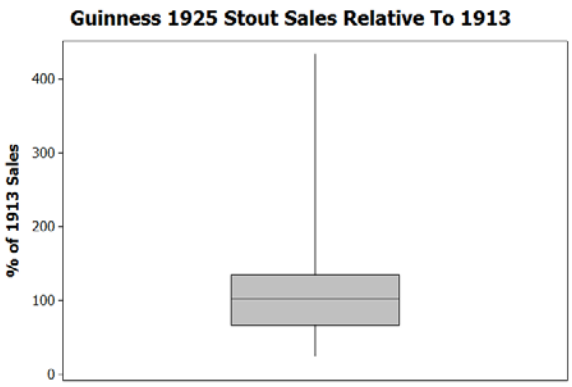
1.64. **(a)** The information in the boxplots is nearly identical to what we saw in the stemplots: The Potato values are right-skewed, whereas the Supplement values are fairly symmetric. The center for the Supplement values is higher than the center for the Potato values. Also, the Potato values are more spread out than the supplement values. **(b)** Preference will vary. The only advantage of the stemplots is that they preserve the data; otherwise, the boxplots are likely better.

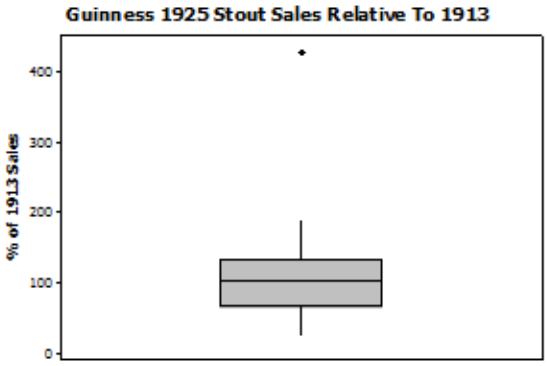


1.65. (a) $\bar{x}$ = 122.9. **(b)** M = 102.5. **(c)** The data set is right-skewed with an outlier (London), so the median is a better center.

1.66. (a) s= 105.7. **(b)** Q_1 = 67, Q_3 = 129. **(c)** The data set is right-skewed with an outlier (London), so the quartiles are a better way to describe the spread.

1.67. (a) *IQR* = 129 – 67 = 62. **(b)** Outliers are below 67 – 1.5*62 = – 26 or above 129 + 1.5*62 = 222. London is confirmed as an outlier. **(c)** The boxplots for **(c)** and **(d)** are shown on the following page. The first three-quarters are about equal in length, and the last (upper quarter) is extremely long. **(d)** The main part of the distribution is relatively symmetric; there is one extreme high outlier. The minimum is about 25, the first quartile is about 70, the median is about 100, and the third quartile is about 125. There is a gap in the data from roughly 200 to about 425.





(e) The stemplot is at right. **(f)** Answers will vary. For example, the stemplot and the boxplot both indicate the same shape distribution: relatively symmetric with an extremely high outlier. One can simply approximate locations of the median and quartiles from the boxplot but get more exact (at least to within rounding) with the stemplot.

```
0 | 24
0 | 6679
1 | 1114
1 | 9
2 |
2 |
3 |
3 |
4 | 2
```

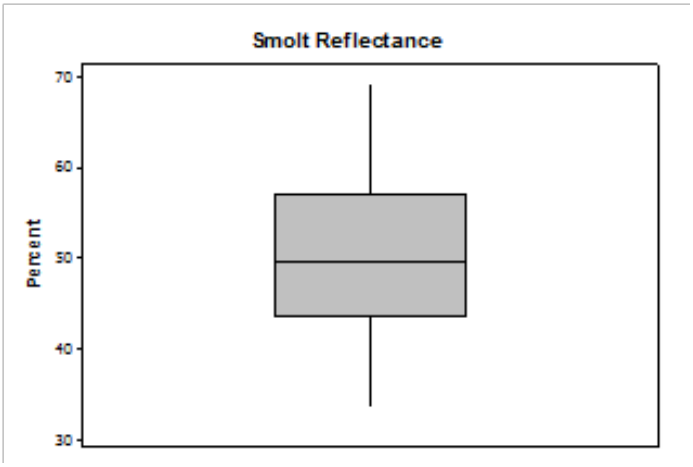
1.68. (a) $\bar{x}$ = 50.89. **(b)** *M* = 49.51. **(c)**Answers will vary. In the stemplot below, we see no outliers. The shape could possibly be called somewhat right-skewed. If the distribution is seen as right-skewed, the median would be a better measure of center, but the two statistics are close to one another.

```
3 | 3
3 | 788
4 | 222222233344
4 | 5566777799
5 | 0013344
5 | 556677889
6 | 03334
6 | 899
```

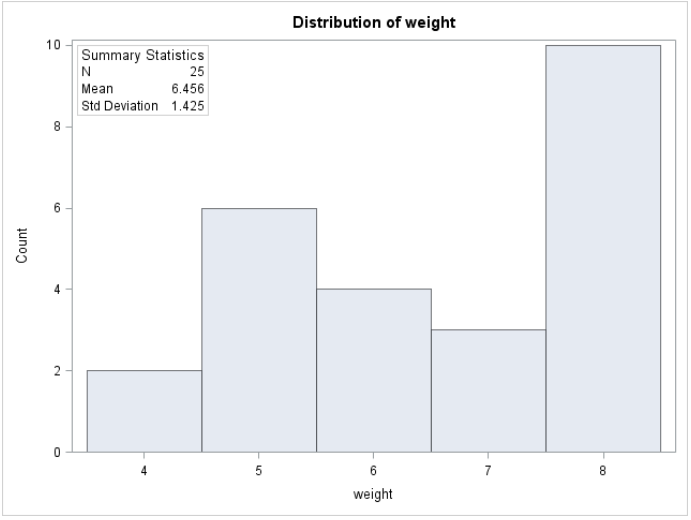
1.69. (a) *s* = 8.80. **(b)** With *n* = 50, the positions of Q_1 and Q_3 will be at 13 and 38 (51 – 13, or 13 in from the upper end of the sorted distribution). We find Q_1 = 43.79 and Q_3 = 57.02. **(c)** Answers will vary.

However, if students said the median was the better center in **Exercise 1.64**, they should pair that with the quartiles and not the standard deviation.

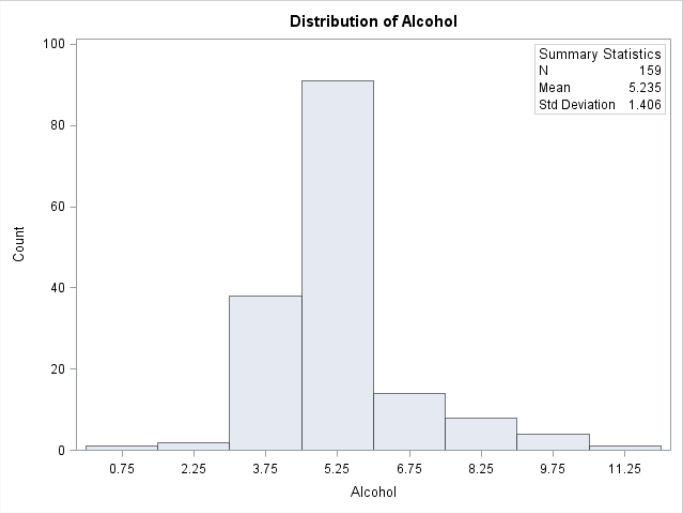
1.70. (a) From **Exercise 1.65**, Q_1 = 43.79 and Q_3 = 57.02. IQR = 57.02 – 43.79 = 13.23. **(b)** Outliers are below 43.79 – 1.5*13.23 = 23.95 and above 57.02 + 1.5*13.23 = 76.87. There are no outliers. **(c – d)** See the boxplot (with no outliers, there is no need for a modified boxplot). The distribution is roughly symmetric with median about 50. The minimum is about 33 and the maximum about 68. The quartiles are about 43 and 57. This distribution has no outliers. **(e)** The stemplot (see **Exercise 1.64** above) seems a bit right-skewed, but it also shows no outliers (there are no gaps). **(f)** Answers will vary. For these data, the two graphs give essentially the same information. Locating more "exact" (to within rounding) values for positions such as the median and quartiles might be termed easier with the stemplot.



1.71. (a) Answers will vary. Because weight is quantitative and has a decent amount of observations (n = 25), a histogram is a good choice. Mean and standard deviation are a good starting point for numerical summaries. **(b)** Answers will vary. Now that we see the distribution is left-skewed, the choice of using the mean and standard deviation was not a good choice. Median and quartiles would have been a better choice. **(c)** Answers will vary. One possible break is between 5.3 and 6.0. The summaries for the groups should provide better summary statistics than the grouped data.

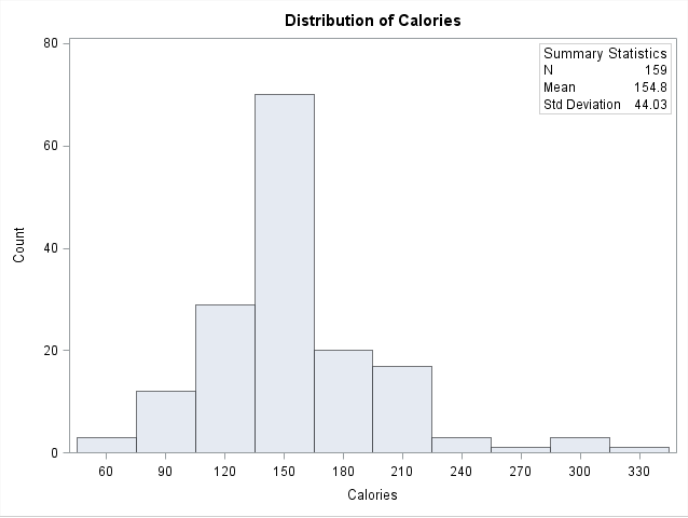


1.72. (a) Answers will vary. $\bar{X}$ = 5.17, s = 1.337, M = 4.9, $Q1$ = 4.415, $Q3$ = 5.600. A histogram is shown below. **(b)** O'Doul's is the outlier with only 0.4% alcohol, it is unique because it is considered a form of nonalcoholic beer. **(c)** Answers will vary.



1.73. (a) With the outlier: $\bar{x}$ = 5.235, M = 4.90. Without the outlier: $\bar{x}$ = 5.265, M = 4.905. The values are nearly identical with and without the outlier. **(b)** With the outlier: s = 1.406, Q_1= 4.40, Q_3= 5.60. Without the outlier: s = 1.356, Q_1= 4.430, Q_3= 5.600. The values are nearly identical with and without the outlier. **(c)** Even though there is one outlier, its removal does not change the numerical summaries at all. This is partly due to the large sample and partly due to the fact that this outlier is not too far from the other observations, so that removing it doesn't have a huge effect on the analysis.

1.74. (a) The distribution of calories is fairly symmetric (with maybe a small right-skew); the mean calories is 154.8. **(b)** O'Doul's has one of the smallest amount of calories per 12 oz, 70, but is not an outlier. **(c)** Answers will vary.



1.75. Celebrities and business executives (Bill Gates of Microsoft, Warren Buffett, Oprah, etc.) typically have a very large net net worth, which will pull the mean worth, making it much larger than the median.

1.76. Answers will vary. With $n = 7$ observations, the median is the middle number. After deleting the lowest observation, the median will be the average of that middle number and the next number after it; if that latter number is much larger, the median will change substantially. For example, start with 0, 1, 2, 3,997, 998, 999; after removing 0, the median changes from 3 to 500.

1.77. The mean is $\frac{7*55000+3*80000+650000}{11} = \$115,909.09.$ Ten of the employees make less than the mean. $M = \$55{,}000$.

1.78. If three individuals earn \$0, \$0, and \$20,000, the reported median is \$20,000. If the two individuals with no income take jobs at \$14,000 each, the median decreases to \$14,000. The same thing can happen to the mean: In this example, the mean drops from \$20,000 to \$16,000.

1.79. The median doesn't change, but the mean increases to \$138,636.36.

1.80. See details below. The mean is $\bar{x} = \frac{1792+1666+\cdots+1439}{7} = 1600.$

x	x − x bar	(x − x bar)^2
1792	192	36,864
1666	66	4356
1362	−238	56,644
1614	14	196
1460	−140	19,600
1867	267	71,289
1439	−161	25,921

$s^2 = \frac{36,864+4,356+\cdots+25,921}{6} = \frac{214,870}{6} = 35,811.667.$

$s = \sqrt{35,811.667} = 189.24.$

1.81. The average would be 2.5 or less (an earthquake that isn't usually felt). These do little or no damage.

1.82. (a) The mean of this distribution appears to be higher than 100. (There is no substantial difference between the actual standard deviation 13.17 and the stated standard deviation 15.) **(b)** $M = 110$, which is close to 108.92. **(c)** In addition to the mean and median, the standard deviation is given for reference (the exercise did not ask for it). The grade distribution is left-skewed, so we expect the mean to be somewhat less than the median.

	X bar	*s*	*M*
IQ	108.92	13.17	110
GPA	7.447	2.1	7.829

Note: *Students may be somewhat puzzled by the statement in **(b)** that the median is "close to the mean" (when they differ by 1.1), followed by **(c)**, where they "differ a bit" (when $M - \bar{x} = 0.382$). It may be useful to emphasize that we judge the size of such differences relative to the spread of the distribution. For example, we can note that (1.1/13.17) = 0.08 for **(b)** and (0.382/2.1) = 0.18 for **(c)**.*

1.83. For $n = 2$ the median is also the average of the two values.

1.84. (a) The mean (green arrow) moves along with the moving point (in fact, it moves in the same direction as the moving point, at one third the speed). At the same time, as long as the moving point remains to the right of the other two, the median (red arrow) points to the middle point (the rightmost

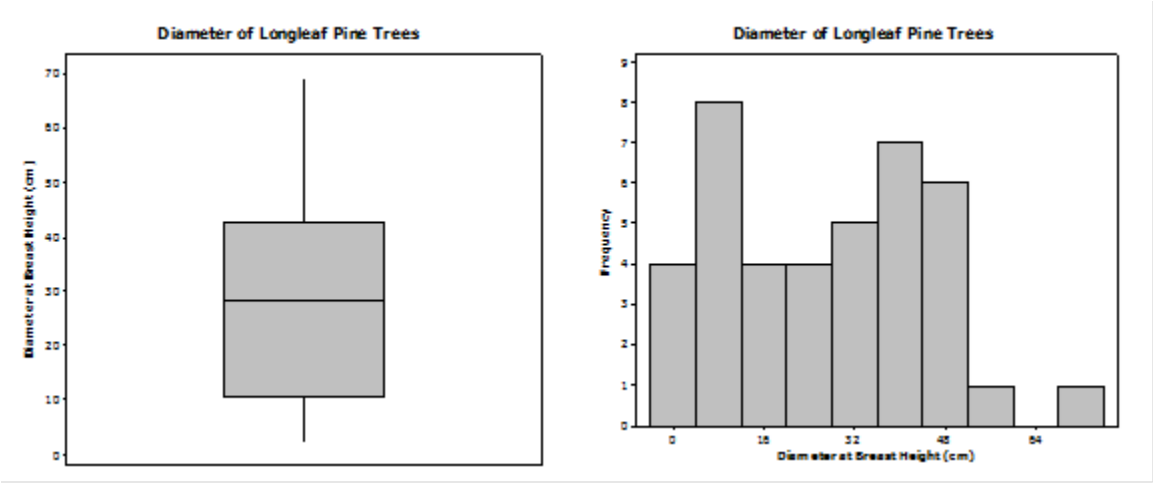
nonmoving point) and does not move. **(b)** The mean follows the moving point as before. When the moving point passes the rightmost fixed point, the median slides along with it until the moving point passes the leftmost fixed point, then the median stays there.

1.85. (a) The median of seven (sorted) points is the fourth, while the median of eight points is the average of the fourth and fifth. If these are to be the same, added points must be equal to the fourth point of the original seven, so that the fourth and fifth points are now the same. **(b)** Regardless of the configuration of the first seven points, if the eighth point is added so as to leave the median unchanged, then in that (sorted) set of eight, the fourth and fifth points must the same. Once we add a ninth point, one of these two points will be the new middle (fifth) point, so the median will not change.

1.86. (a) The mean is $\bar{x} = 15$, and the standard deviation is $s = 5.4365$. **(b)** The mean is still 15; the new standard deviation is 3.7417. **(c)** Using the mean as a substitute for missing data will not change the mean, but it will decrease the standard deviation.

1.87. (a) Picking the same number for all four observations results in a standard deviation of 0. **(b)** Picking 10, 10, 20, and 20 results in the largest standard deviation = 5.77. **(c)** For part **(a)**, you may pick any number as long as all observations are the same. For part **(b)**, only one choice provides the largest standard deviation.

1.88. (a) Min = 2.2, $Q_1 = 10.73$, $M = 28.50$, $Q_3 = 42.60$, Max = 69.30. **(b – c)** Shown below. **(d)** Answers will vary. Both graphs show the right-skew. The histogram shows a possible outlier not identified in the modified boxplot.



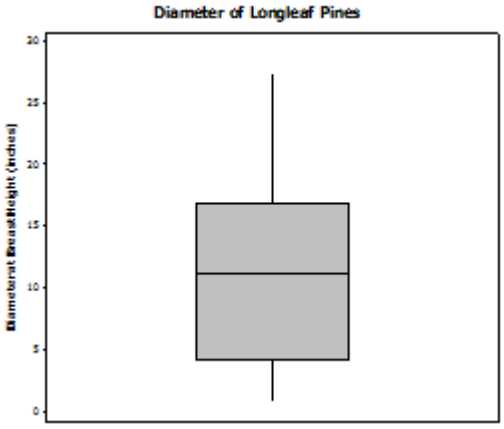
1.89. $\bar{x} = 2.41(2.2) = 5.302$ pounds and $s = 1.25(2.2) = 2.75$ pounds.

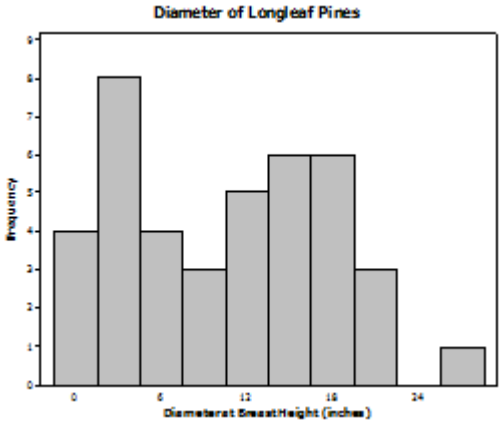
1.90. Because the variance has squared units, the variance will be multiplied by $2.54^2 = 6.4516$.

1.91. Full data set: $\bar{x} = 196.575$ and $M = 103.5$ minutes. The 10% and 20% trimmed means are $\bar{x} = 127.734$ and $\bar{x} = 111.917$ minutes. While still larger than the median of the original data set, they are much closer to the median than the ordinary untrimmed mean.

1.92. After changing the scale from centimeters to inches, the five-number summary values change by the same ratio (that is, they are multiplied by 0.39). The shape of the histogram might change slightly because of the change in class intervals. **(a)** The five-number summary (in inches) is Min = 0.858, $Q_1 = 4.2705$, M

= 11.115, Q_3= 16.341, Max = 27.027. **(b – c)** Shown below. **(d)** As in **Exercise 1.88**, the histogram reveals more detail.





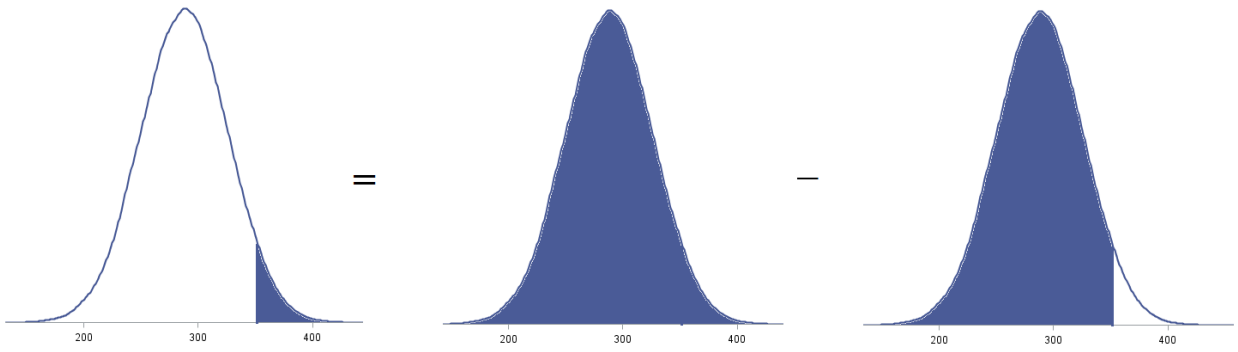
1.93. According to the rule, 95% of scores will fall between $\mu \pm 2\sigma$. Therefore, 95% of scores are between 288 – 2*38 = 212 and 288 + 2*38 = 364.

1.94. According to the rule, 99.7% of scores will fall between $\mu \pm 3\sigma$. Therefore, 99.7% of scores are between 288 – 3*38 = 174 to 288 + 2*38 = 402.

1.95. $z = \dfrac{350 - 288}{38} = 1.63.$

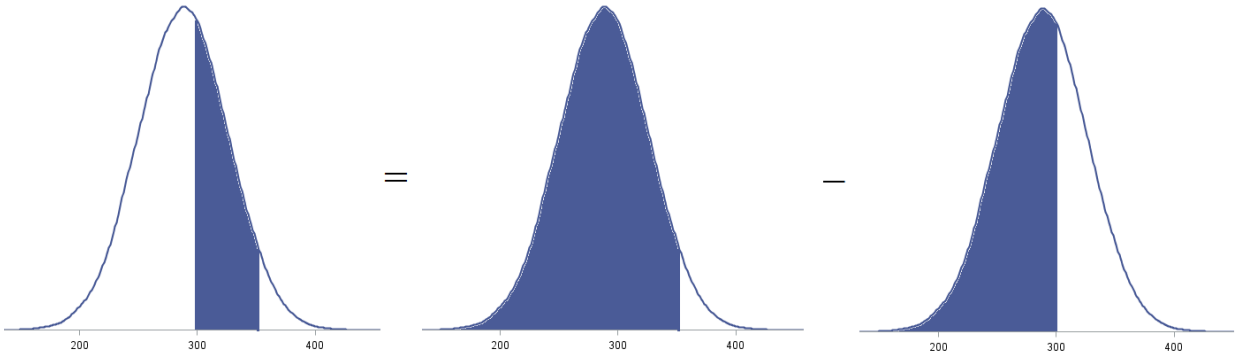
1.96. $z = \dfrac{240 - 288}{38} = -1.26.$ The *z*-score is negative because a score of 240 is below the mean.

1.97. For X = 350, $z = \dfrac{350 - 288}{38} = 1.63$, the proportion less than 350 is the area to the left, which is 0.9484. For the proportion greater than or equal to 350, we calculate 1 – 0.9484 = 0.0516.



Area greater than 350 = Entire area under the Normal – Area less than 350

1.98. For $X = 350$, $z = \dfrac{350 - 288}{38} = 1.63$; the area to the left of this is 0.9484. For $X = 300$, $z = \dfrac{300 - 288}{38} = 0.32$; the area to the left of this is 0.6255. Subtracting gives $0.9484 - 0.6255 = 0.3229$. So the proportion between 300 and 350 is 0.3229.

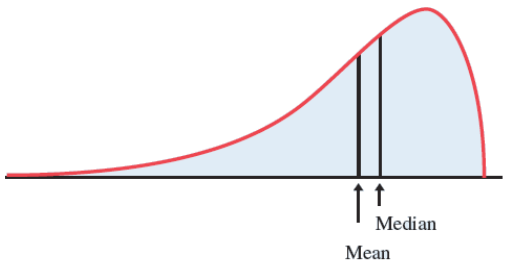


Area between 300 and 350 = Area to the left of 350 – Area to the left of 300

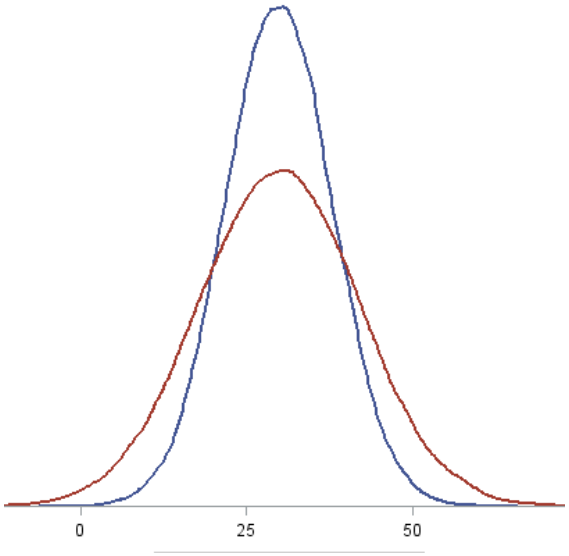
1.99. To get the top 20% of students, we need to solve for the 80th percentile. The corresponding z is 0.84. So $x = 288 + 38(0.84) = 319.92$.

1.100. If 75% score above x, x is the 25th percentile. The corresponding z is -0.67. So $x = 288 + 38(-0.67) = 262.54$.

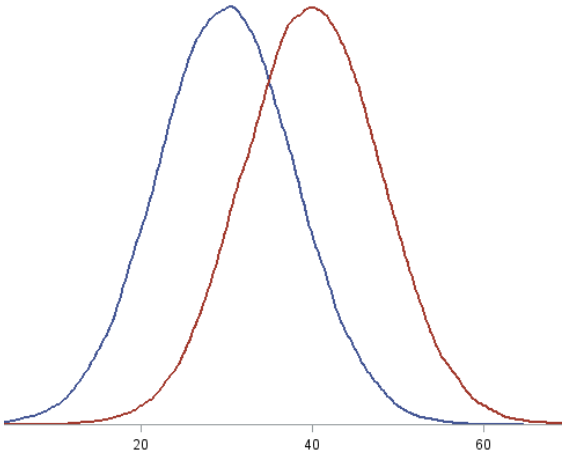
1.101. (a) Answers will vary; the graph should be a mirror image. The mean and median should be equal. **(b)** Answers will vary; example below. The mean should be farther left than the median.



1.102. (a – b) Shown below. **(c)** The curve narrows or widens but remains centered at the same point (at the mean).

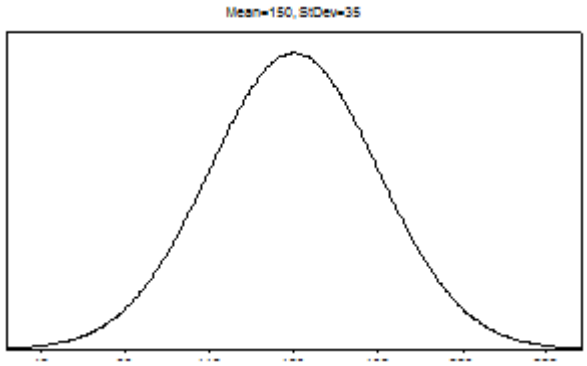


1.103. (a – b) Shown below. **(c)** The curve shifts to the left or right, but the spread remains the same.



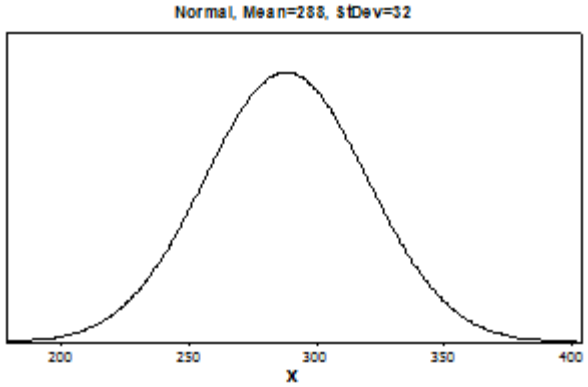
1.104. (a) The distribution is shown at right. **(b –c)** The table below indicates the desired ranges. These values are shown on the *x* axis of the distribution.

	Low	High
68%	115	185
95%	80	220
99.7%	45	255



1.105. (a) The distribution is shown at right. **(b – c)** The table below indicates the desired ranges. These values are shown on the *x* axis of the distribution.

	Low	High
68%	256	320
95%	224	352
99.7%	192	384



1.106.

Value	Standardized Score
150	0
140	–0.29
100	−1.43
180	0.86
230	2.29

Note: *The standardized score is computed by (Value – 150)/35.*

1.107.

Value	Percentile (Table A)	Percentile (Software)
150	50	50
140	38.6	38.8
100	7.6	7.7
180	80.5	80.4
230	98.9	98.9

1.108. Using the $N(288,32)$ distribution, we find the values corresponding to the given percentiles as given below (using Table A). The actual scores are close to the percentiles of the Normal distribution. We can conclude these scores are at least approximately Normal.

Percentile	Score	Score with $N(288, 32)$
10%	246	247
25%	276	266
50%	290	288
75%	311	310
90%	328	330

1.109. Using the $N(153,34)$ distribution, we find the values corresponding to the given percentiles as given below (using Table A). The actual scores are very close to the percentiles of the Normal distribution. We can conclude these scores are at least approximately Normal.

Percentile	Score	Score with $N(153, 34)$
10%	110	109
25%	130	130
50%	154	153
75%	177	176
90%	197	197

1.110. (a) 68% of women would speak between 7856 and 20,738 words per day. 95% of women would speak between 1415 and 27,179 words per day. 99.7% of women would speak between − 5026 and 33,620 words per day. **(b)** Because it is impossible to speak a negative number of words in a day, this distribution cannot be truly Normal. **(c)** 68% of men would speak between 5004 and 23,116 words per day. 95% of men would speak between − 4052 and 32,172 words per day. 99.7% of men would speak between − 13,108 and 41,228 words per day. This distribution must be even more non-Normal because

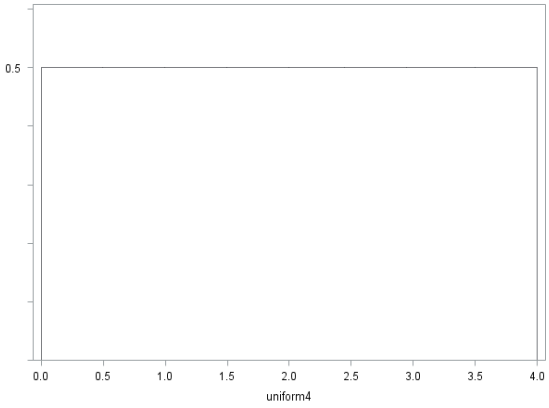
the lower end of the 95% region is quite a bit less than 0. **(d)** Men have a larger standard deviation, so the 68% region for women fits inside the 68% region for men. It is possible that men speak less than women but also possible they speak more words in a day.

1.111. (a) Ranges are shown in the table below. In both cases, some of the lower limits are negative, which does not make sense; this happens because the women's distribution is skewed, and the men's distribution has an outlier. Contrary to the conventional wisdom, the men's mean is slightly higher, although the outlier is at least partly responsible for that. **(b)** The means suggest that Mexican men and women tend to speak more than people of the same gender from the United States.

	Women	Men
68%	8489 to 20,919	7158 to 22,886
95%	2274 to 27,134	−706 to 30,750
99.7%	−3941 to 33,349	−8570 to 38,614

1.112. (a) The curve forms a 1 × 1 square (i.e., the length is 1 and the width is 1). The area of a square is L × W which equals 1. **(b)** 1 – 0.44 = 0.56 or 56%. **(c)** 0.70 – 0.44 = 0.26.

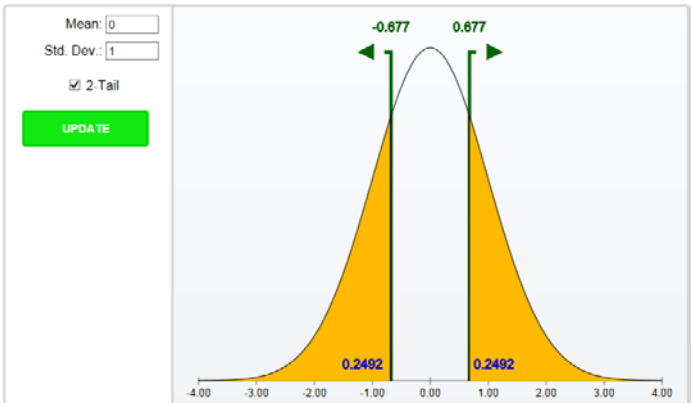
1.113. (a) 0.25 or 1/4. **(b)** Graph shown below. **(c)** (4 – 3)/4 = 0.75 or 75%. **(d)** (2.5 – 1.5)/4 = 0.25 or 25%.



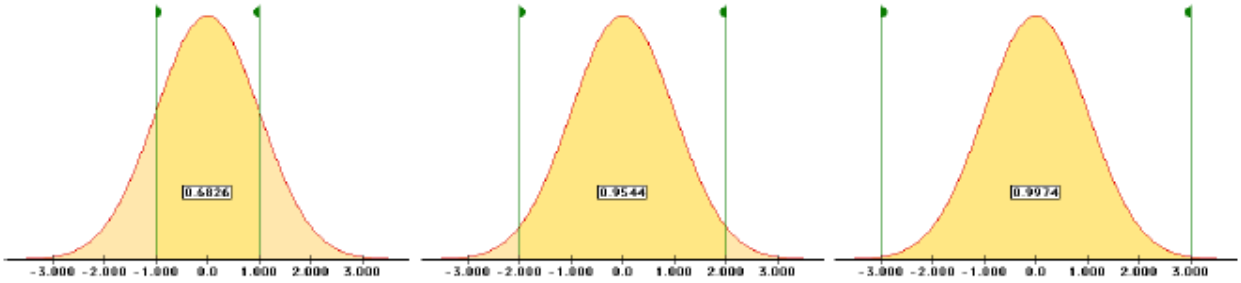
1.114. $\mu = M = 0.5$. $Q_1 = 0.25$, $Q_3 = 0.75$.

1.115. (a) The mean is at point C; the median is at point B. **(b)** The mean and median are both at point A. **(c)** The mean is at point A; the median is at point B.

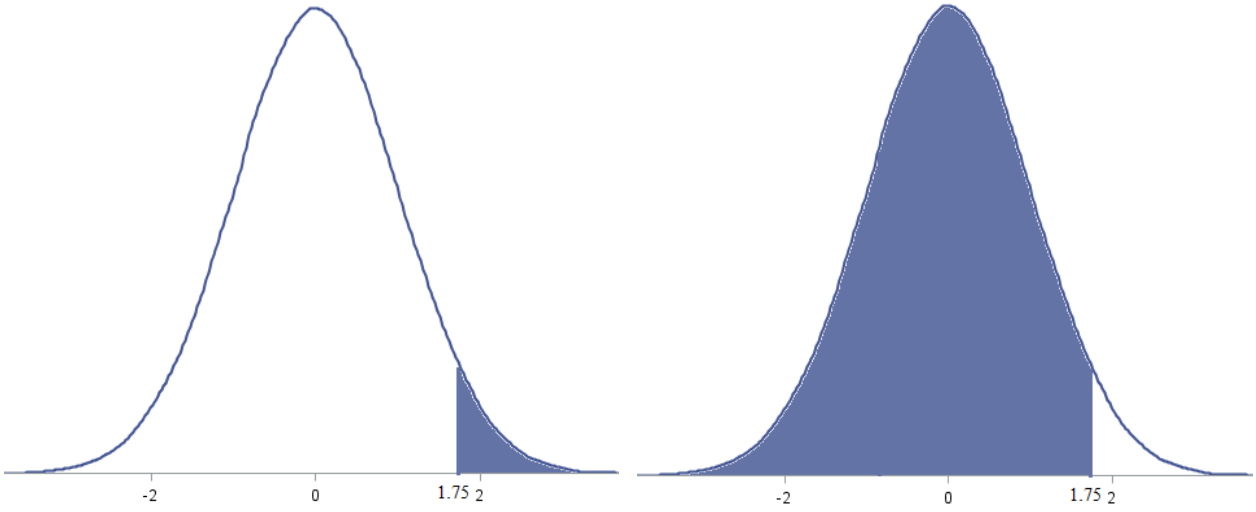
1.116. Because the quartiles of any distribution have 25% on either end, we seek to place the flags so that the reported areas are each 0.25. The closest the applet gets is an area of 0.2492, between − 0.677 and 0.677. Thus, the quartiles of any Normal distribution are about 0.68 standard deviations above and below the mean.

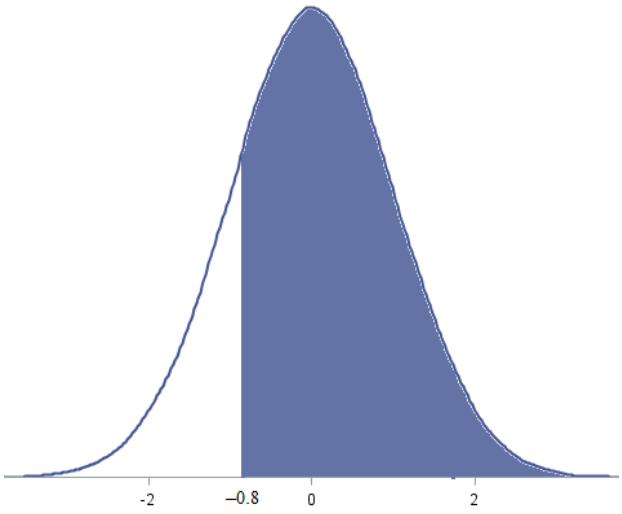


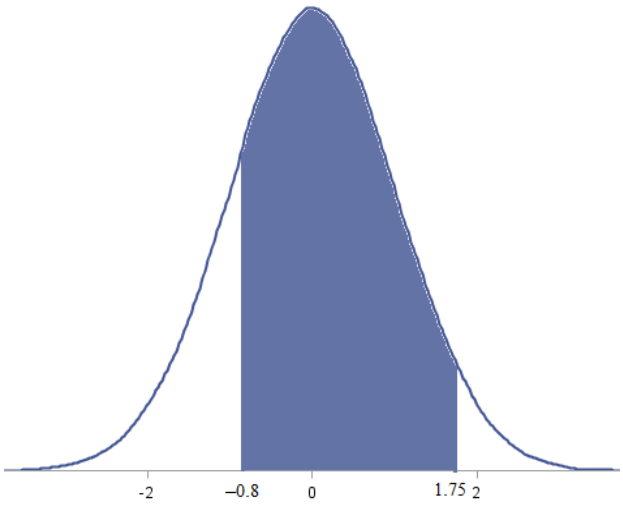
1.117. (a) The applet shows an area of 0.6826 between − 1.000 and 1.000, while the 68–95–99.7 rule rounds this to 0.68. **(b)** Between − 2.000 and 2.000, the applet reports 0.9544 (compared with the rounded 0.95 from the 68–95–99.7 rule). Between −3.000 and 3.000, the applet reports 0.9974 (compared with the rounded 0.997).



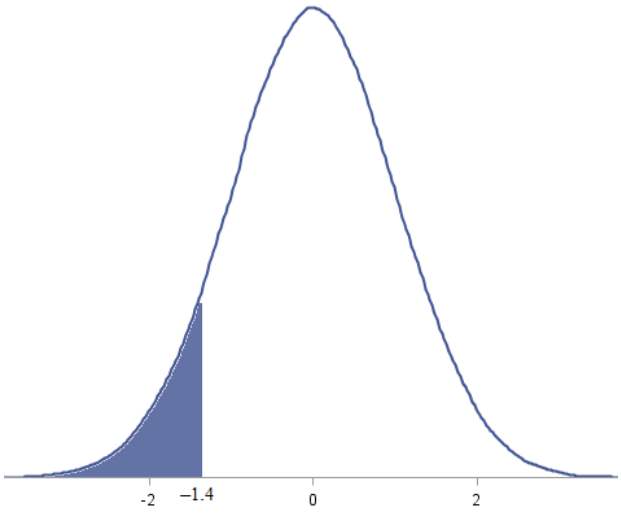
1.118. Graphs shown below from top left to bottom right. **(a)** 0.0401. **(b)** 0.9599. **(c)** 0.7881. **(d)** 0.9599 − 0.2119 = 0.7480.

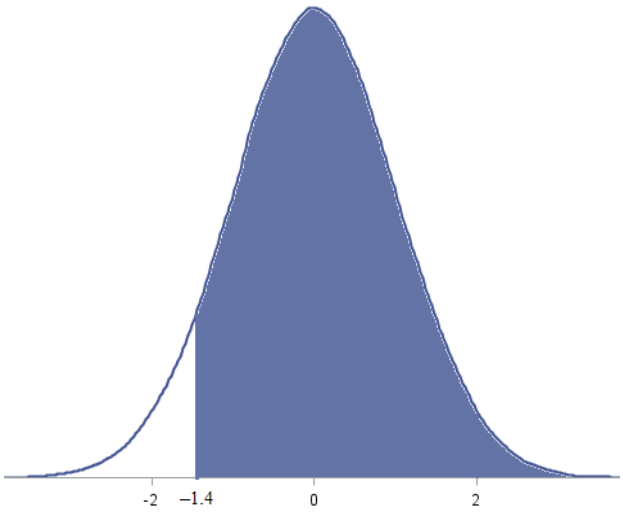


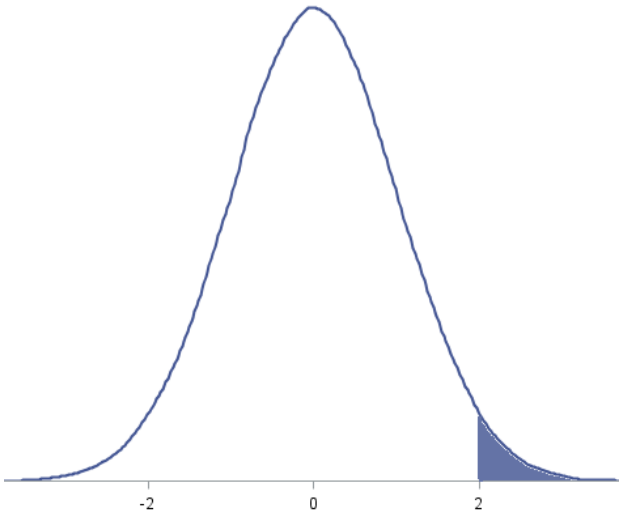


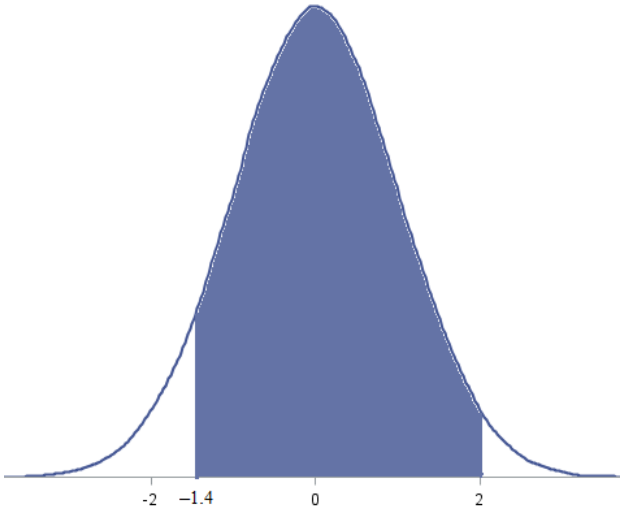


1.119. Graphs shown below from top left to bottom right. **(a)** 0.0808. **(b)** 0.9192. **(c)** 0.0228. **(d)** 0.9772 – 0.0808 = 0.8964.

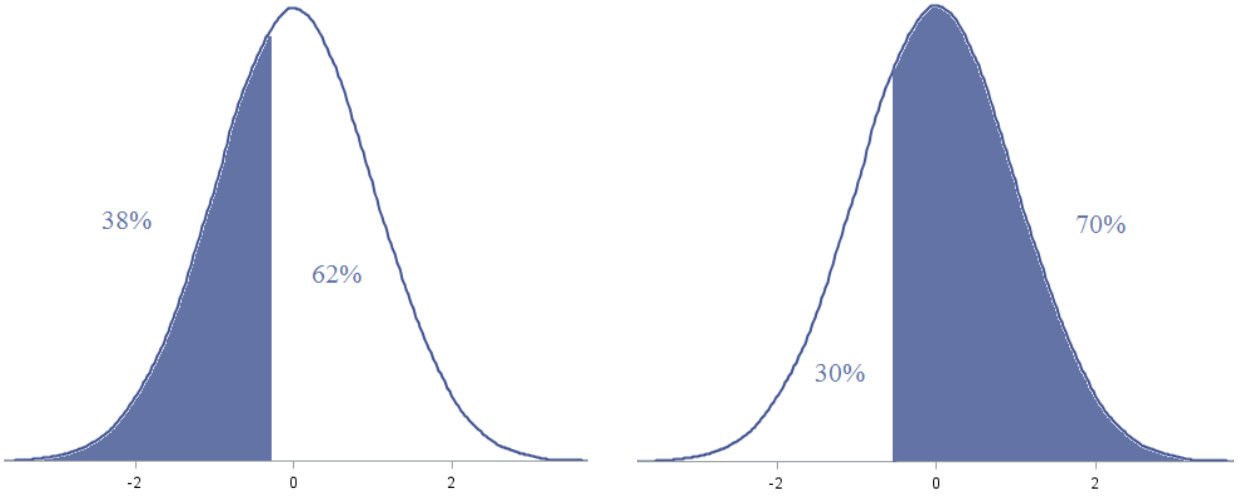








1.120. Graphs shown below from left to right. **(a)** This is the 38th percentile; using Table A gives $z = -0.31$. **(b)** This is the 30th percentile; using Table A gives $z = -0.52$.



1.121. (a) $z = 1.17$ or 1.18. **(b)** $z = 1.17$ or 1.18.

1.122. $z = \dfrac{70-100}{15} = -2.$ From Table A, 0.0228 or 2.28% are developmentally disabled by this criterion.

1.123. $z = \dfrac{130-100}{15} = 2.$ From Table A, 0.0228 or 2.28% qualify for membership.

1.124. For Jessica, $z = \dfrac{1830-1498}{316} = 1.05.$ For Ashley, $z = \dfrac{27-21.5}{5.4} = 1.02.$ Jessica has the higher standardized score.

1.125. For Joshua, $z = \dfrac{16-21.5}{5.4} = -1.02.$ For Anthony, $z = \dfrac{1050-1498}{316} = -1.42.$ Joshua has the higher standardized score.

1.126. $z = \dfrac{2090-1498}{316} = 1.87.$ Using $x = \mu + z\sigma$. The equivalent ACT score is 21.5 + 5.4(1.87) = 31.598.

1.127. $z = \dfrac{30-21.5}{5.4} = 1.57.$ Using $x = \mu + z\sigma$. The equivalent SAT score is 1498 + 316(1.57) = 1994.12.

1.128. $z = \dfrac{2050-1498}{316} = 1.75.$ From Table A, we get 0.9599, so about the 96th percentile.

1.129. $z = \dfrac{19-21.5}{5.4} = -0.46.$ From Table A, we get 0.3228, so about the 32nd percentile.

1.130. The top 12% corresponds to the 88th percentile; using Table A gives $z = 1.17$ or 1.18. Using 1.18, the SAT score needed is 1498 + 316(1.18) = 1870.88. So scores 1870.88 and higher make up the top 12% of all scores.

1.131. The bottom 12% corresponds to the 12th percentile; using Table A gives $z = -1.17$ or -1.18. Using -1.18, the SAT score needed is $1498 + 316(-1.18) = 1125.12$. So scores 1125.12 and lower make up the bottom 12% of all scores.

1.132. From Table A, the quintiles have *z*-scores of -0.84, -0.25, 0.25, and 0.84. Using $21.5 + 5.4(z)$ yields scores of 17, 20, 23, and 26 (rounded to the nearest integer).

1.133. From Table A, the quartiles have *z*-scores of -0.675, 0, and 0.675. Using $1498 + 316(z)$ yields scores of 1285, 1498, and 1711 (rounded to the nearest integer).

1.134. (a) $z = (40 - 55)/15.5 = -0.97$. From Table A, 16.6% of women have low values of HDL. (Software gives 16.66%.) **(b)** $z = (60 - 55)/15.5 = 0.32$. From Table A, 37.45% of women have protective levels of HDL. (Software gives 37.35%.) **(c)**$(1 - 0.3745) - 0.1660 = 0.4595$. 45.95% of women are in the intermediate range for HDL. (Software gives 0.4599.)

1.135. (a) $z = (40 - 46)/13.6 = -0.44$. From Table A, 33% of men have low values of HDL. (Software gives 32.95%.) **(b)** $z = (60 - 46)/13.6 = 1.03$. From Table A, 15.15% of men have protective levels of HDL. (Software gives 15.16%.) **(c)** $(1 - 0.1515) - 0.33 = 0.5185$. 51.85% of men are in the intermediate range for HDL. (Software gives 0.5188.)

1.136. (a) About 0.6% of healthy young adults have osteoporosis (the cumulative probability below a standard score of -2.5 is 0.0062). **(b)** About 31% of this population of older women has osteoporosis: The BMD level, which is 2.5 standard deviations below the young adult mean, would standardize to -0.5 for these older women, and the cumulative probability for this standard score is 0.3085.

1.137. (a) The first and last deciles for a standard Normal distribution are ±1.2816. **(b)** Fora $N(9.12, 0.15)$ distribution, the first and last deciles are $\mu - 1.2816\sigma = 8.93$ and $\mu + 1.2816\sigma = 9.31$ oz.

1.138. (a) The quartiles for a standard Normal distribution are ±0.6745. **(b)** For a $N(\mu, \sigma)$ distribution, $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$.

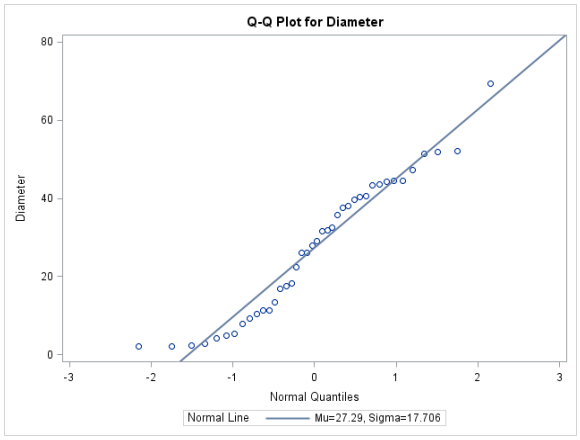
1.139. (a) As the quartiles for a standard Normal distribution are ±0.6745, we have $IQR = 1.3490$. **(b)** $c = 1.3490$: For a $N(\mu, \sigma)$ distribution, the quartiles are $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$, so $IQR = (\mu + 0.6745\sigma) - (\mu - 0.6745\sigma) = 1.3490\sigma$.

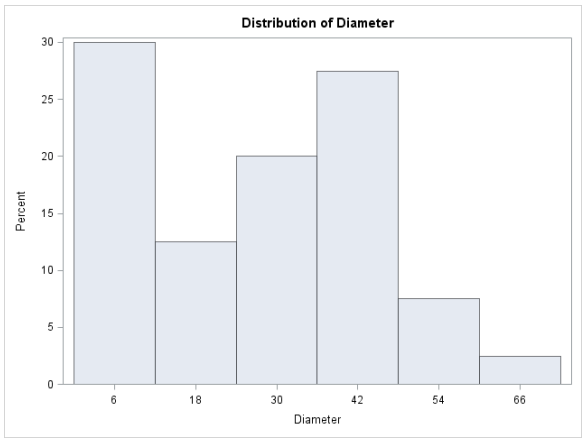
1.140. In the previous two exercises, we found that, for a $N(\mu, \sigma)$ distribution, $Q_1 = \mu - 0.6745\sigma$, $Q_3 = \mu + 0.6745\sigma$, and $IQR = 1.3490\sigma$. Therefore, $1.5 \times IQR = 2.0235\sigma$, and the suspected outliers are below $Q_1 - 1.5 \times IQR = \mu - 2.698\sigma$, and above $Q_3 + 1.5 \times IQR = \mu + 2.698\sigma$. The percent outside of this range is $2 \times 0.0035 = 0.70\%$.

1.141. To find these levels, use $55 + z^*15.5$, where *z* from Table A has the given percent as the area to the left of *z*. With symmetry, $z_{10\%} = -z_{90\%}$.

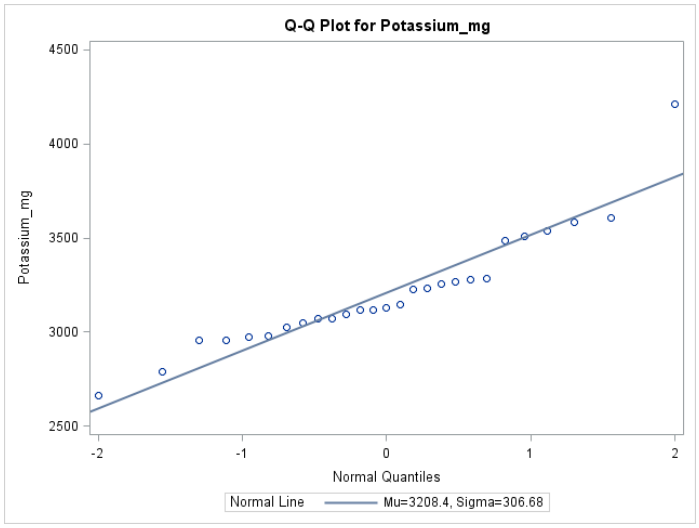
Percentile	10%	20%	30%	40%	50%	60%	70%	80%	90%
HDL level	35.2	42.0	46.9	51.1	55	58.9	63.1	68.0	74.8

1.142. The data are not normal but have a right-skew. The included histogram shows the right skew.

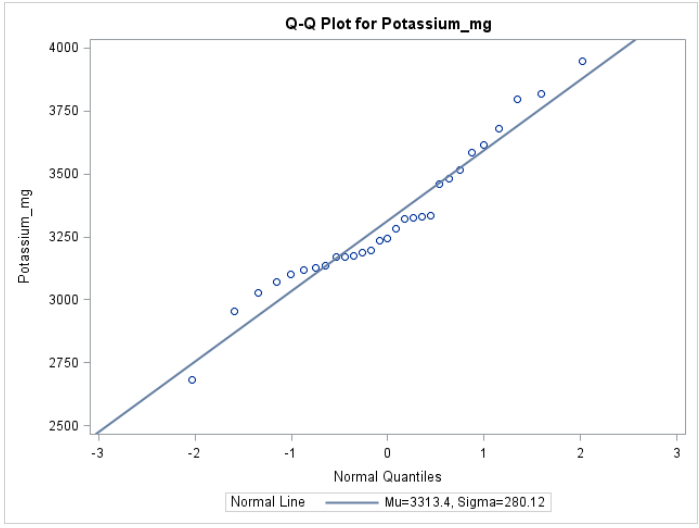




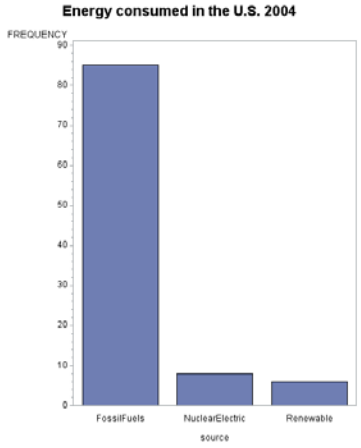
1.143. (a) See solutions for **Exercises 1.30** and **1.61**. **(b)** The Normal quantile plot is shown below. The data are roughly normal, but there is one potential high outlier.

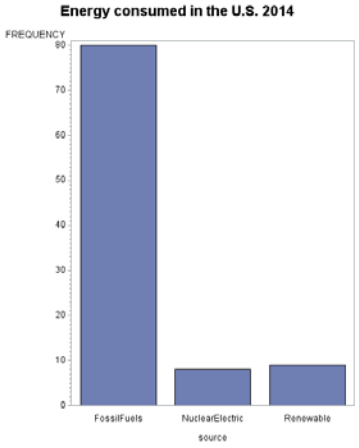


1.144. (a) See solutions for **Exercises 1.31** and **1.62**. **(b)** The Normal quantile plot is shown below. The data are normally distributed.

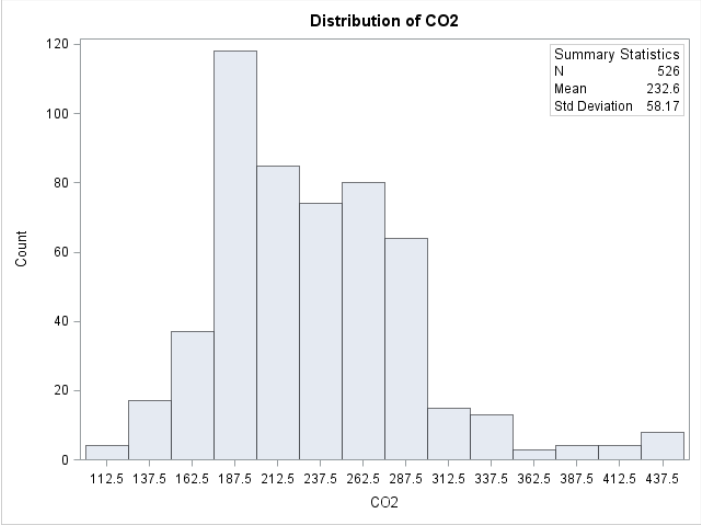


1.145. There is less energy used in 2014 from fossil fuels and more used from renewable sources. There was almost no change in nuclear and electric power usage. Graphs shown below.

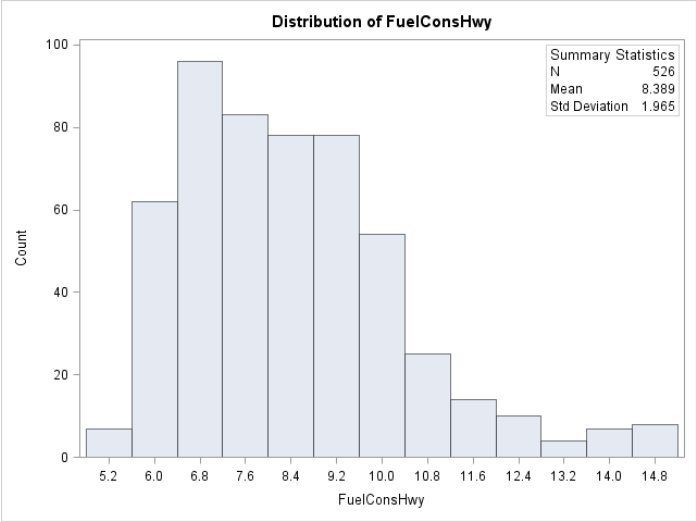




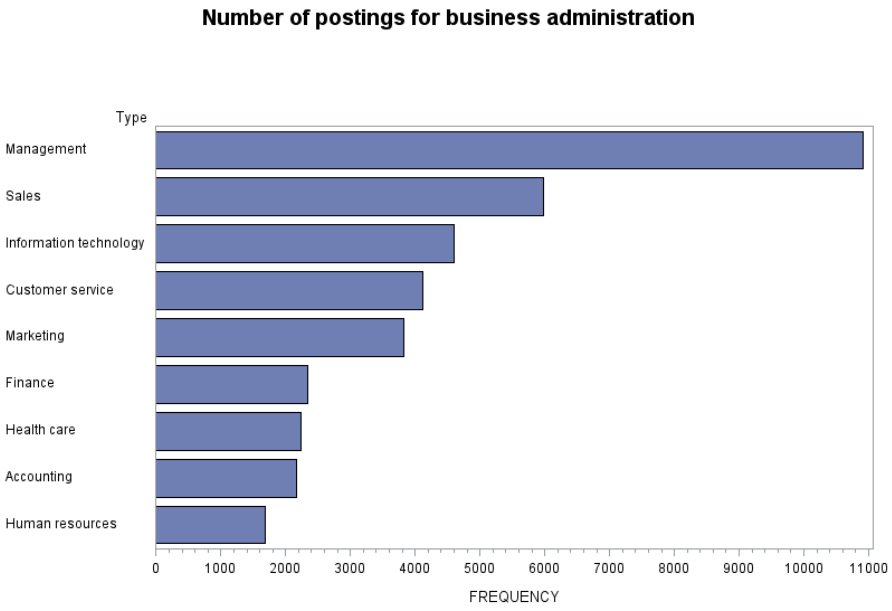
1.146. $\overline{x}$ = 232.64, s = 58.17. Min = 108, Q_1 = 191, M = 225, Q_3 = 267, Max = 437. As shown in the histogram, the distribution is somewhat right-skewed, so the five-number summary is a better description for CO2 emissions.



1.147. $\overline{x}$ = 8.389, s = 1.965. Min = 4.9, Q_1 = 6.9, M = 8.2, Q_3 = 9.5, Max = 15.1. As shown in the histogram, the distribution is somewhat right-skewed, so the five-number summary is a better description for highway fuel consumption.

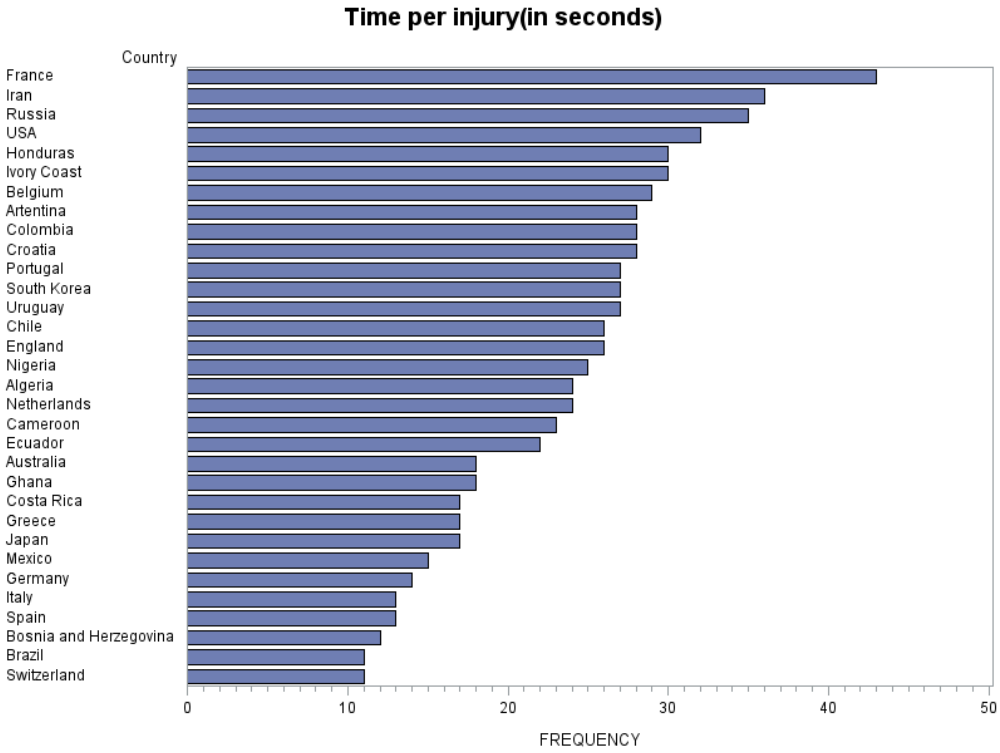


1.148. The most prevalent job for those who have a business degree is in Management, with that category having nearly twice as much as the next closest category. Sales is the second most prevalent job available, followed by Information Technology, Customer Service, and Marketing. The limitation on using these data is that it is likely to change over time, potentially even day to day. It is also restricted to the classification specifications of the particular website, which may classify jobs differently than others.

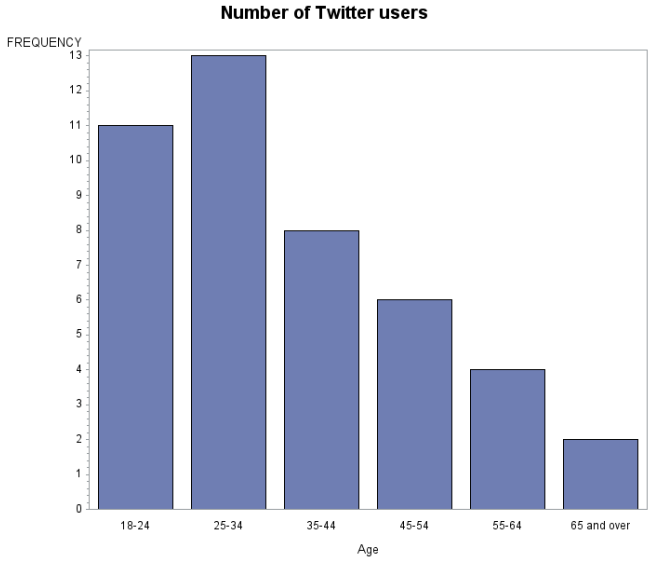


1.149. Answers may vary. One useful statistic might be the amount of time per injury for each country. To get this, we take the Time divided by the number of injuries and then multiply by 60 to convert it to seconds. The graph below shows the Time per injury (in seconds). It does look like some countries may be flopping as several spend between two to three times as long per injury as other countries.

Note: *One possible confounding factor is the severity of each injury.*



1.150. Users of Twitter are much more likely to be from younger age groups, particularly the 25–34 age group has the highest number of users at 13.3 million followed by the 18–24 age group with 11.7; after age 35, the number of users diminishes with age.



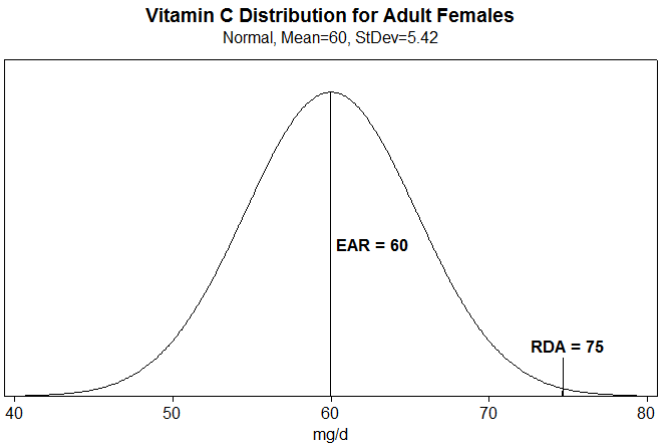
1.151. (a) For car makes (a categorical variable), use either a bar graph or pie chart. For car age (a quantitative variable), use a histogram, stem plot, or boxplot. **(b)** Study time is quantitative, so use a histogram, stem plot, or boxplot. To show change over time, use a time plot (average hours studied against time). **(c)** Use a bar graph or pie chart to show radio station preferences. **(d)** Use a Normal quantile plot to see whether the measurements follow a Normal distribution.

1.152. Answers will vary.

1.153. Answers will vary.

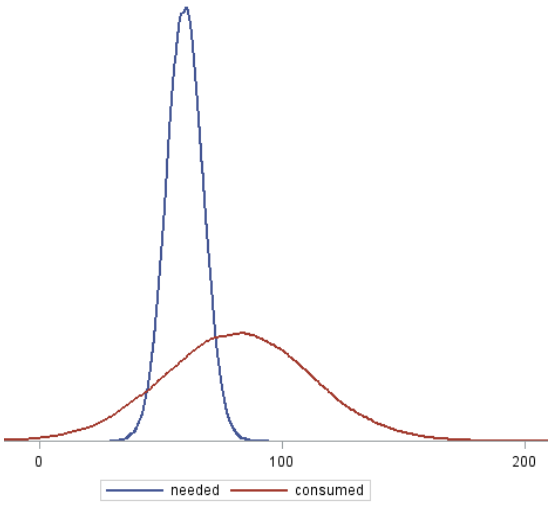
1.154. Answers will vary. Definitions might be as simple as "free time," or "time spent doing something other than studying." For part **(b)**, it might be good to encourage students to discuss practical difficulties; for example, if we ask Sally to keep a log of her activities, the time she spends filling it out presumably reduces her available "leisure time."

1.155. (a) RDA = 75 = 60 +$z\sigma$. From Table A, z = 2.00. We have 15 = 2σ, so σ = 7.5. **(b)** The distribution is shown below. Note that the UL is far off into the upper tail at 2000 mg/d.

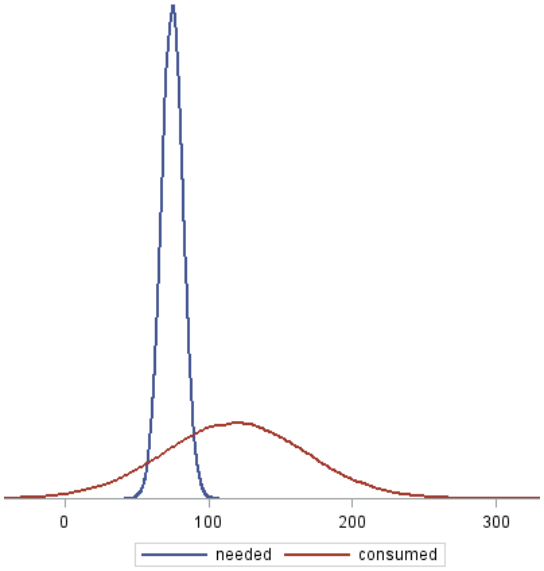


1.156. RDA = 90 = 75 + 2σ. σ = 7.5.

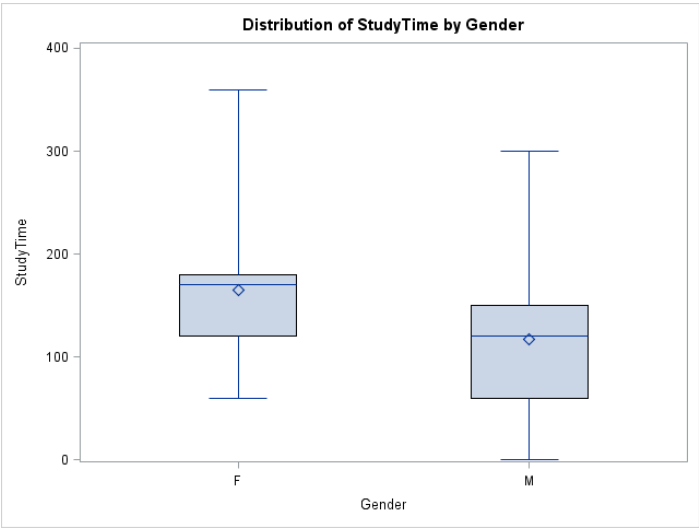
1.157. (a) We were given that the mean is 84.1 and the 50th percentile was 79. In a Normal distribution, these should be equal; one option would be to average these and estimate μ = (79 + 84.1)/2 = 81.55. The 5th percentile is 42 = 81.55 – 1.645σ. This implies σ = 24.04. The 95th percentile is 142 = 81.55 + 1.645σ. This implies σ = 36.75. If we average the two estimates, we would have σ = 30.4. **(b)** The graph is shown below. **(c)** From the two distributions, over half of women consume more vitamin C than they need, but some consume far less.



1.158. (a) Similarly to the previous exercise, we estimate the men's mean as 118.1. The 5th percentile is $55 = 118.1 - 1.645\sigma$, implying $\sigma = 38.36$. The 95th percentile is $217 = 118.1 + 1.645\sigma$, implying $\sigma = 60.12$. If we average the two estimates, we would have $\sigma = 49.24$. **(b)** The graph is shown below. **(c)** The results for men are similar to those for women; over half consume more than is needed, but some consume far less.



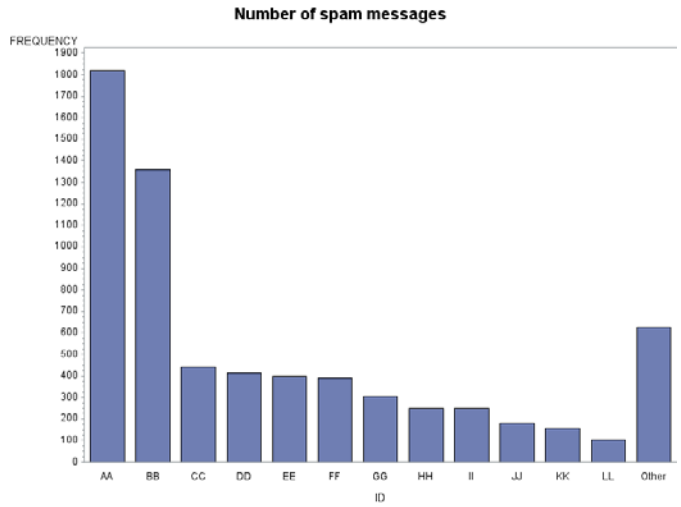
1.159. (a) Not only are most responses multiples of 10, but many are multiples of 30 and 60. Most people will "round" their answers when asked to give an estimate like this; in fact, the most striking answers are ones such as 115, 170, or 230. The students who claimed 360 min (6 hours) and 300 min (5 hours) may have been exaggerating. (Some students might also "consider suspicious" the student who claimed to study 0 min per night. *As a teacher, I can easily believe that such students exist, and I suspect that some of your students might easily accept that claim as well.*) **(b)** Women seem to generally study more (or claim to), as none claim less than 60 min per night. The center (median) for women is 170; for men the median is 120 min. **(c)** The boxplots are given. Opinions will vary.



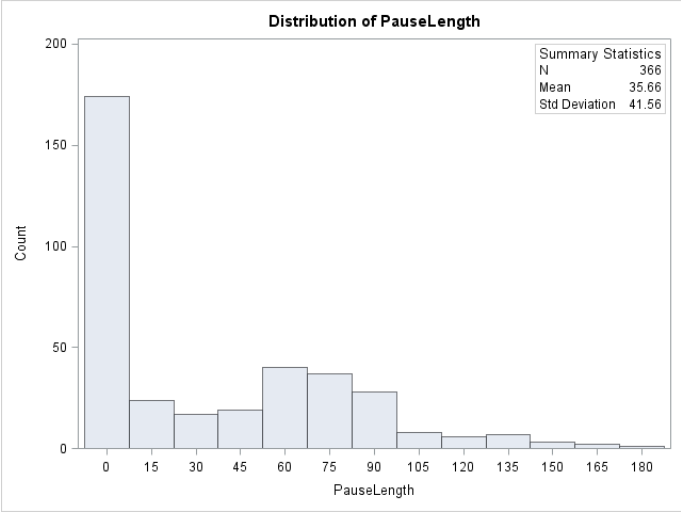
1.160. Gender and automobile preference are categorical. Age and household income are quantitative.

1.161. No and no. It is easy to imagine examples of many different data sets with mean 0 and standard deviation 1, for example, {−1,0,1} and {−2,0,0,0,0,0,0,0,2}. Likewise, for any given five numbers $a \le b \le c \le d \le e$ (not all the same), we can create many data sets with that five-number summary simply by taking those five numbers and adding some additional numbers in between them, for example (in increasing order): 10, , 20, , , 30, , , 40, , 50. As long as the number in the first blank is between 10 and 20, and so on, the five-number summary will be 10, 20, 30, 40, 50.

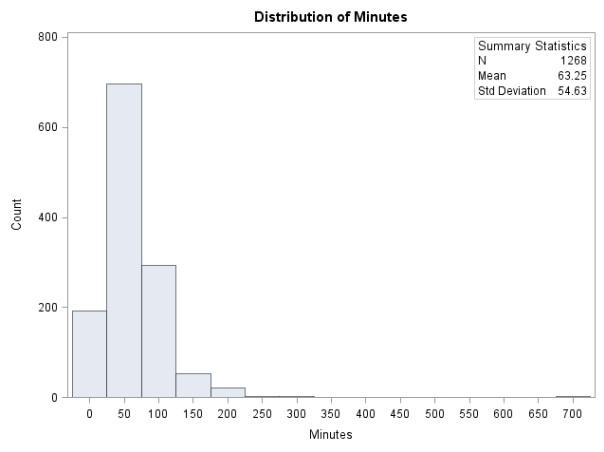
1.162. The given counts total 6067 spam messages, so the other members of the department received 6693 − 6067 = 626 spam messages. We can add that into the data file and create a bar graph as shown; a pie chart might be appropriate, but there would be 13 slices. Users AA and BB receive by far the most spam messages. Perhaps they visit unsafe websites.

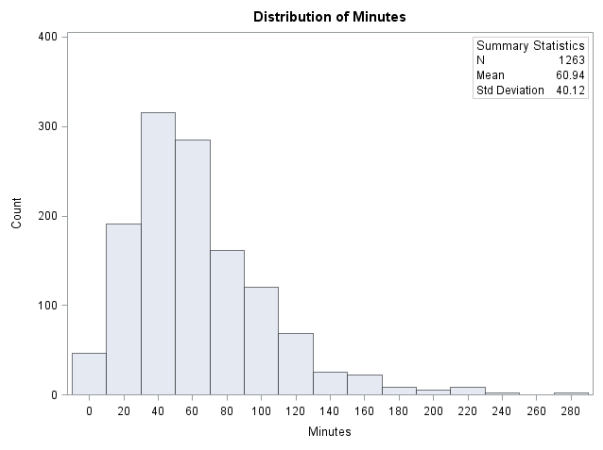


1.163. $\bar{x}$ = 35.66, s = 41.56. Min = 0, Q_1 = 1, M = 11.5, Q_3 = 68, Max = 181. On average, the band pauses for 35.66 seconds; however, for a large portion of the time, they don't pause at all. The distribution, as shown in the histogram below, is strongly right-skewed and shows that sometimes the band pauses for as much as 181 seconds or 3 min before playing the final note.



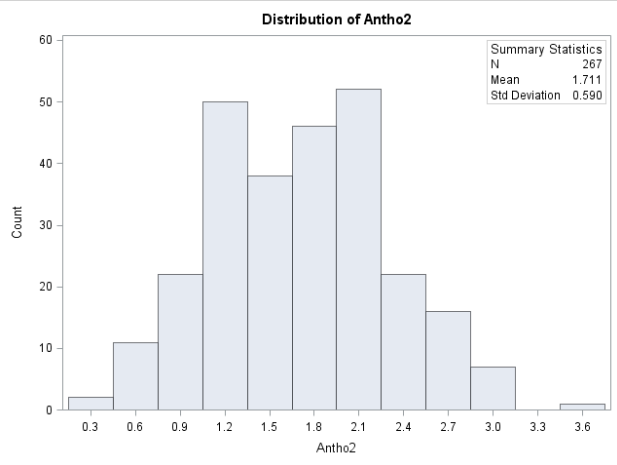
1.164. $\overline{x}$ = 63.25, s = 54.63. Min = 0, Q_1 = 31, M = 55, Q_3 = 80, Max = 713. The histogram below shows a right-skew, as such the five number summary is likely a better description for this dataset. The median length of time each student spends in the help room is 55 min, or about 1 hour. But there are extremes with two students reporting 0 min as well as two more students reporting 485 and 713 min, or roughly 6 and 12 hours. Another histogram without these two high outliers is also shown below.

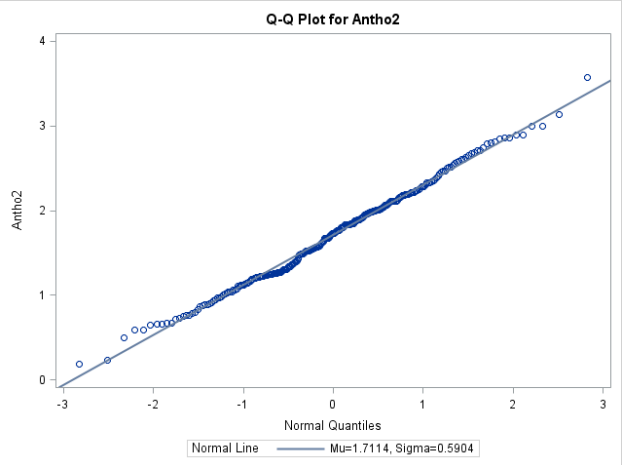




1.165. Antho1 is normally distributed from the histogram and normal quantile plot. $\overline{x}$ = 1.630, s = 0.521.

1.166. Antho2 is normally distributed from the histogram and normal quantile plot. $\overline{x}$ = 1.711, s = 0.590.





1.167. Antho3 is right-skewed; it also has a potential high outlier. Min = 0.0546, Q_1 = 0.4506, M = 0.6781, Q_3 = 1.1282, Max = 6.3109.

1.168. Antho4 is strongly right-skewed; it also has two potential high outliers. Min = 0.01539, Q_1 = 0.0784, M = 0.1423, Q_3 = 0.6975, Max = 4.2295.