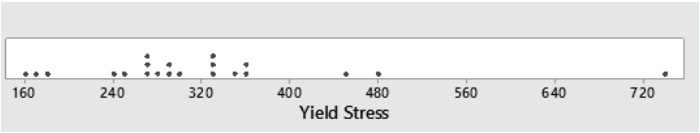
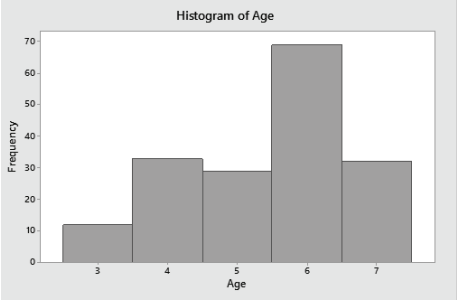
Chapter 2 Solutions

2.1. **(a)** The distribution is unimodal and essentially symmetric except for a high outlier. The center is approximately 291 MPa. The spread is from 164.0 to 740.2 MPa. **(b)** $\overline{x}$ = 319.25 MPa; Median = 290.7 MPa. The large outlier, as seen in the dotplot above, is causing the mean to be greater than the median.



2.2. Because the distribution is strongly right-skewed, we expect the mean to be larger than the median. Therefore, 0.015 is the mean and 0.009 the median.

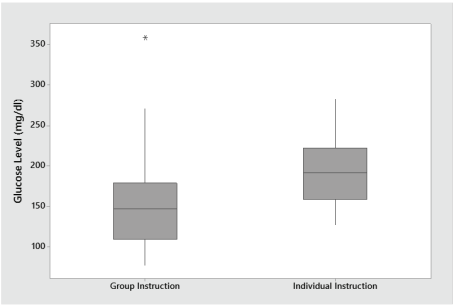
2.3. **(a)** The median age is 6 years. **(b)** The mean age is 5.4 years. **(c)** Since the mean (5.4 years) is less than the median (6 years), we expect the histogram to be left skewed. The histogram to the right validates this fact.



2.4. **(a)** Five-number summary: Min = 0.17, Q_1 = 1.35, M = 1.86, Q_3 = 2.52, Max = 12.8 **(b)** Range = Max - Min = 12.8 - 0.17 = 12.63. IQR = 2.52 - 1.35 = 1.17. **(c)** We do not have enough information to compute the standard deviation of this sample of blood mercury levels. We would need the exact values of the data points to compute the standard deviation.

2.5. **(a)** Five-number summary: Min = 164.0, Q_1 = 260.90, M = 290.7, Q_3 = 354.95, Max = 740.2 (MPa). **(b)** $\overline{x}$ = 319.25 MPa and s = 124.9 MPa **(c)** The five number summary gives more information about the distribution of silk yield stresses. The distance from Q_3 to the Max is much greater than the distance from Q_1 to the Min which is reflected by the high outlier observed in the dotplot in Exercise 2.1.

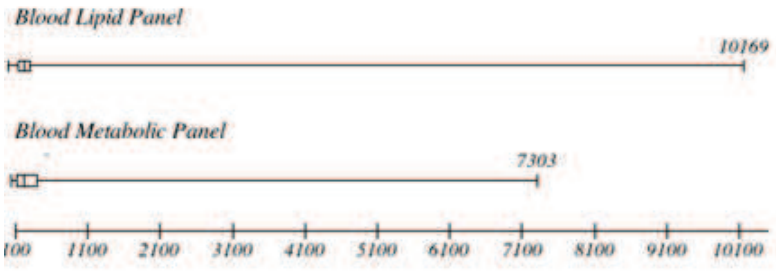
2.6. **(a)** Five-number summary for the group instruction data set: Min = 78, Q_1 = 112, M = 147.5, Q_3 = 172, Max = 359. Five-number summary for the individual instruction data set: Min = 128, Q_1 = 159.5, M = 191.5, Q_3 = 222, Max = 283. **(b)** Side-by-side boxplot shown on the right. Results from group instruction are much more variable than those from individual instruction. However, fasting plasma glucose levels tend to be lower after group instruction than after individual instruction.



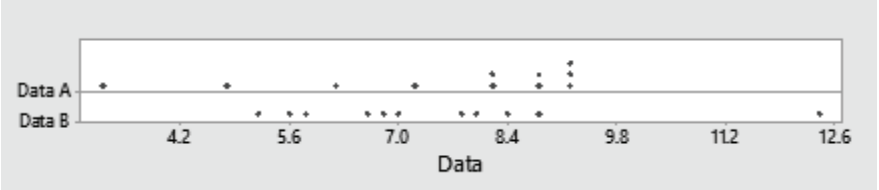
(c) The mean of the group instruction data set is 163.3 and the standard deviation is 70.7. The mean of the individual instruction data set is 190.1 and the standard deviation is 41.9. The mean and standard deviation do not provide information about

the shape of the two distributions. These statistics provide information about the center of the data (using the mean) and how tightly the data is clustered around the mean. **(d)** By adding a symbol representing the mean and errors bars representing one standard deviation above and below the mean to the dotplot, we can see how tightly the data points are clustered around the mean. This is similar to the boxplots produced in part (b) where we can see how tightly the data is clustered around the median by measuring the distance from the median to the first and third quartiles.

2.7. **(a)** The distribution of price for blood lipid panels is roughly the same as blood metabolic panels. Both distributions are heavily skewed to the right. **(b)** Since the two distributions are heavily skewed to the right, the mean and standard deviation would be poor choices to use as summary statistics to cite in a news report, as they are heavily influenced by skewness and outliers. The five-number summary would be a better choice to cite in the news report.



2.8. Data set A: Mean = 7.501, Standard deviation = 2.032; Data set B: Mean = 7.501, Standard deviation = 2.031. Thus, data set A and B have the same mean and almost the same standard deviation. However, data set A appears to be skewed to the left, whereas data set B appears to be fairly symmetric with one large outlier as shown in the dotplots.



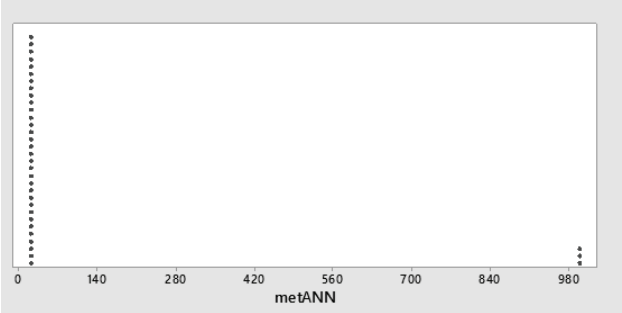
2.9. $IQR = Q_3 - Q_1 = 354.95 - 260.90 = 94.05$, $1.5 \times IQR = 141.075$. Values below 119.825 (that's $Q_1 - 141.075$) or above 496.025 (that's $Q_3 + 141.075$) would be suspected outliers. Only one value qualifies: 740.2 MPa, the largest yield stress value and clear outlier on the dotplot in the solution to Exercise 2.1.

2.10. For group instruction: $IQR = Q_3 - Q_1 = 172 - 112 = 60$, $1.5 \times IQR = 90$. Values below 22 ($Q_1 - 90$) or above 262 ($Q_3 + 90$) would be suspected outliers. There is one suspected outlier: 359 mg/dl, the highest fasting plasma glucose level in the group instruction data set. For individual instruction: $IQR = Q_3 - Q_1 = 222 - 159.5 = 62.5$, $1.5 \times IQR = 93.75$. Values below 65.75 ($Q_1 - 93.75$) or above 315.75 ($Q_3 + 93.75$) would be suspected outliers. There are no suspected outliers in the individual instruction data set.

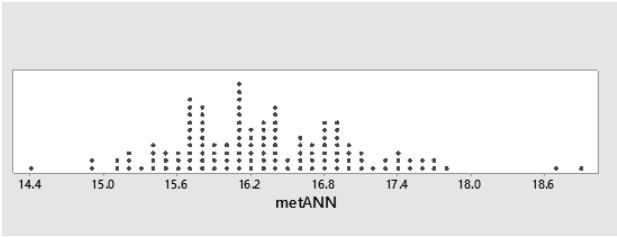
2.11. **(a)** Blood pressure cuff too small, blood pressure cuff used over clothing, arms, back or feet unsupported, not resting for a few minutes, talking, bladder full or empty, and room temperature would lead to data points that we would want to discard in a study of blood pressure in healthy adults. We would discard these data points because they are the result of human error in experimentation or data collection. All of these

factors can be held constant by the doctor's office in the data collection process. **(b)** The remaining factors: emotional state, smoking recently, and consuming alcohol recently are factors that may lead to unusual blood pressure values that should not be ignored in a study of blood pressure in healthy adults. These factors are a function of the healthy adult, and not factors related to the data collection process. **(c)** If a manufacturer of a blood pressure cuff wanted to estimate the variability of readings obtained at the doctor's office, suspicious data points resulting from emotional state, smoking recently, and consuming alcohol recently should be discarded. This is different than the answer in part (a) because the purpose of the study has changed from studying blood pressure in healthy adults to studying variability in readings obtained at the doctor's office. Thus, we are now interested in the variability of the measurements due to the data collection process.

2.12. (a) For the metANN variable, the mean is 88.1 and the median is 16.3. The mean is much larger than the median. **(b)** 999.9 degrees Celsius is not a meteorologically plausible value for the mean annual temperature in Los Angeles. These values illustrate human error in recording information. **(c)** It is not likely NASA repeatedly made the same data entry error in recording temperatures in Los Angeles. The most likely explanation is 999.9 was used as a special code for years that did not have an annual mean temperature recorded.

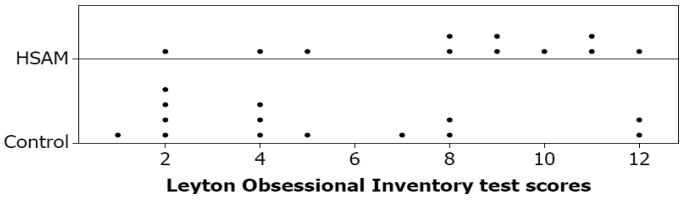


(d) For the cleaned-up metANN variable, the mean is 16.310 and the median is 16.230. The mean and median are much closer together now that the outliers are removed. The updated dotplot for the cleaned-up metANN variable reveals a much more symmetric distribution than in part (b).



2.13. State: How do individuals with and without HSAM compare on a test of obsessional symptoms?

Plan: We need to compare the distributions, including appropriate measures of center and spread.



Solve: Dotplots are shown on the right. Based on these, $\overline{x}$ and s are reasonable choices. For the HSAM group, $\overline{x} = 8.09$ $s = 3.18$; for the control group, $\overline{x} = 5.21$ $s = 3.66$.

Conclude: The means and the dotplots appear to suggest that individuals with and without HSAM have a similar range of scores on the Leyton Obsessional Inventory test but that individuals with HSAM have higher scores, on average.

2.14. (c) The mean is $\bar{x} = 6.28$.

2.15. (b) The median is 6.00.

2.16. (c) The interquartile range is $IQR = 7.3 - 5.0 = 2.3$.

2.17. (c) The mean is pulled in the direction of the skew.

2.18. (b) Half the observations lie between the quartiles.

2.19. (c) A boxplot is a picture of the five-number summary.

2.20. (a) Standard deviations can be any nonnegative number.

2.21. (b) The standard deviation is $s \doteq 1.52$.

2.22. (b) s is measured in the same units as the data.

2.23. (a) The median is resistant to outliers.

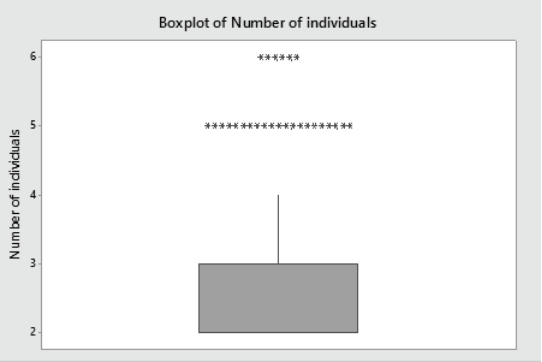
2.24. With a strong right-skew and a high outlier influencing the mean but not the median, we expect the mean to be larger than the median. The mean fatty acid ratio is $\bar{x} = 0.7607$, whereas the median is $M = 0.075$.

2.25. The distribution of ages ranges from 0 to 100 years and is clearly bimodal, with one peak around 5 to 10 years and another peak around 45 to 50 years. A unique set of numerical summary values is not appropriate, because it would mask the fact that there are two distinct age groups most affected by Lyme disease.

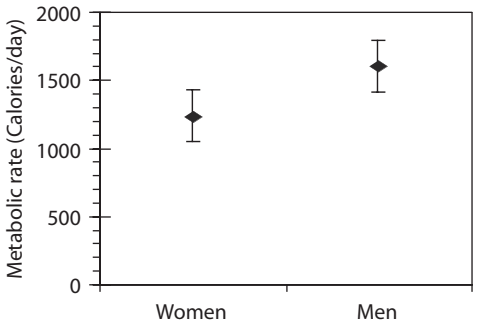
2.26. **(a)** The mean (green arrow) moves along with the moving point (in fact, it moves in the same direction as the moving point, at one third the speed). At the same time, as long as the moving point remains to the right of the other two, the median (red arrow) points to the middle point (the right-most nonmoving point). **(b)** The mean follows the moving point as before. When the moving point passes the right-most fixed point, the median slides along with it until the moving point passes the left-most fixed point, then the median stays there.

2.27. $M = 14$, $Q_1 = 13$, and $Q_3 = 15$ years old: We can use the frequencies shown in the histogram to reconstruct the (sorted) data list of 691 age values; it begins with roughly 40 elevens, then roughly 75 twelves, then roughly 160 thirteens, etc. With 691 data points, the median is the 346th value on this list. Because the 346th number on the list is in the 14-year-old class, 14 is the median. The first quartile is the 173rd number on the list, and Q_3 is the 519th number (or 173rd number from the end of the list).

2.28. **(a)** Five number summary: Min = 2, $Q_1 = 2$, M = 2, $Q_3 = 3$, Max = 6. The boxplot is unusual since the values of the minimum, Q_1, and the median are the same (2). Also, by the 1.5 x IQR rule, any value above 4.5 is an outlier, so all of the laughing groups of size 5 and 6 are outliers as shown in the modified boxplot. **(b)** Since the boxplot is skewed to the right, we expect the mean laughing group size to be larger than the median laughing group size. **(c)** The mean laughing group size is 2.72. This fits with our expectation in part (b) since the mean (2.72) is larger than the median (2).

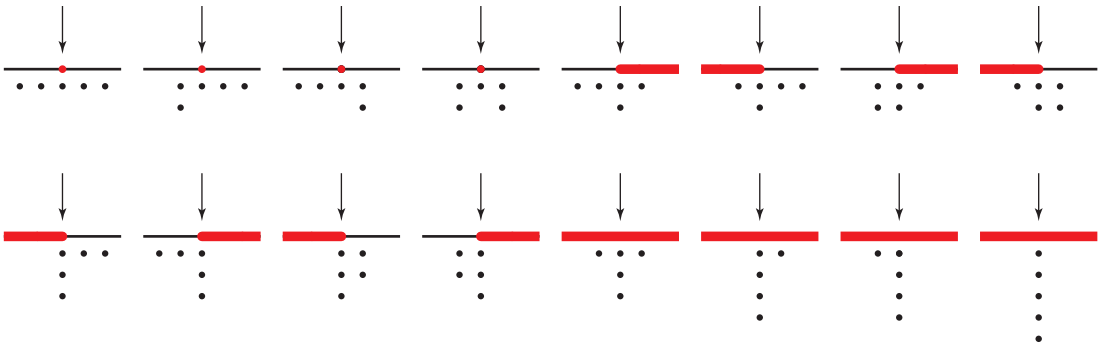


2.29. **(a)** Symmetric distributions with no outliers are best described by their mean and standard deviation. **(b)** Graph is on the right. Because there is no obvious skew or outliers, a simple plot of means plus or minus 1 standard deviation is sufficient for comparing the two groups. The plot shows that the women's metabolic rates are lower on average than the men's.



2.30. **(a)** There are several different answers, depending on the configuration of the first 5 points. *Most students* will likely assume that the first 5 points should be distinct (no repeats), in which case the sixth point *must* be placed at the median. This is because the median of 5 (sorted) points is the third, while the median of 6 points is the average of the third and fourth. If these are to be the same, the third and fourth points of the set of 6 must both equal the third point of the set of 5.

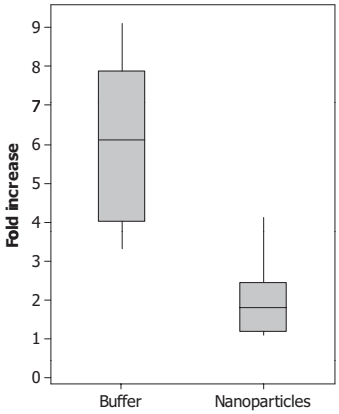
The diagram that follows illustrates all of the possibilities; in each case, the arrow shows the location of the median of the initial 5 points, and the shaded region (or dot) on the line indicates where the sixth point can be placed without changing the median. Notice that there are 4 cases where the median does not change, regardless of the location of the sixth point. (The points need not be equally spaced; these diagrams were drawn that way for convenience.)



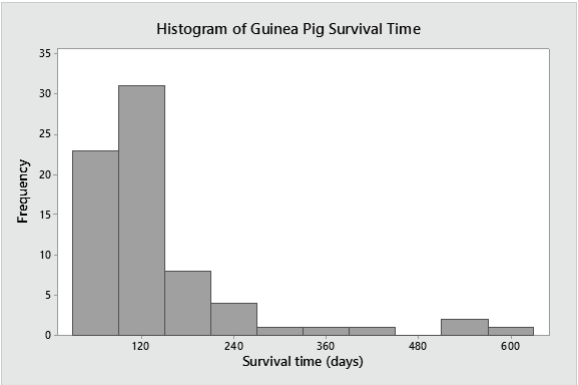
(b) Regardless of the configuration of the first 5 points, if the sixth point is added so as to leave the median unchanged, then in that (sorted) set of 6, the third and fourth points must be equal. One of these 2 points will be the middle (fourth) point of the (sorted) set of 7, no matter where the seventh point is placed.

Note: *If you have a student who illustrates all possible cases above, then it is likely that the student (1) obtained a copy of this solutions manual, (2) should consider a career in writing solutions manuals, (3) has too much time on his or her hands, or (4) both 2 and 3 (and perhaps 1) are true.*

2.31. **(a)** Boxplot shown on the right. For buffer: $\bar{x} = 6.08$ and $M = 6.10$. For nanoparticles: $\bar{x} = 2.02$ and $M = 1.80$. **(b)** The results with the buffer solution are higher overall and much more variable than when using the nanoparticle treatment. The tumor increases with the nanoparticle treatment are right-skewed, with more than half of the tumor not even doubling in volume.

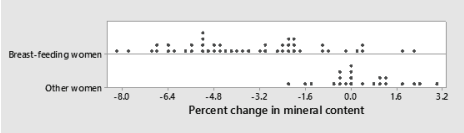


2.32. **(a)** The histogram shows a clear and pronounced right skew. **(b)** Because of the skew, we should choose to summarize the data using the five-number summary rather than the mean and standard deviation (the mean and standard deviation are not resistant measures of center and spread). **(c)** Five-number summary is (43, 82.5, 102.5, 151.5, 598), all measured in days. The difference between Q_3 and the maximum is relatively much larger than the other differences between successive numbers. This indicates a large spread among the high observations—that is, it shows that the data are skewed to the right.



2.33. (a) One possible answer is 1, 1, 1, 1. **(b)** 0, 0, 10, 10. **(c)** For part (a), any set of four identical numbers will have $s = 0$. For part (b), the answer is unique; here is a rough description of why. We want to maximize the "spread-out-ness" of the numbers (which is what standard deviation measures), so 0 and 10 seem to be reasonable choices based on that idea. We also want to make each individual squared deviation—$(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$, $(x_3 - \bar{x})^2$, and $(x_4 - \bar{x})^2$—as large as possible. If we choose 0, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation 7.5^2, but the other three are only 2.5^2. Our best choice is two at each extreme, which makes all four squared deviations equal to 5^2.

2.34. State: Is bone mineral loss greater among the breastfeeding women?
Plan: We need to compare the distributions, including appropriate measures of center and spread.



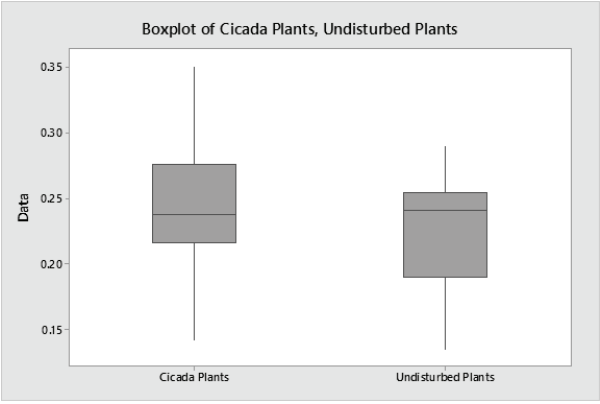
Solve: Shown are two dotplots; it would also be appropriate to produce two histograms or two boxplots (five-number summaries are given below).

Here are numerical summaries; students may give all or just some of these in response to this question.

	$\bar{x}$	Min	Q_1	M	Q_3	Max
BF women	−3.59%	−8.3%	−5.3%	−3.8%	−2.1%	2.2%
Other women	0.31%	−2.2%	−0.4%	−0.05%	1.1%	2.9%

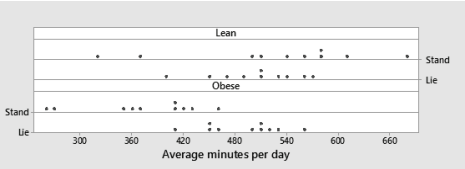
Conclude: Both the graphs and the numerical summaries suggest that there is greater bone mineral loss among the breastfeeding women.

2.35. State: Can dead cicadas serve as fertilzer to increase plant growth?
Plan: We need to compare the distributions, including appropriate measures of center and spread.



Solve: Shown are two boxplots and numerical summaries for cicada plants and undisturbed plants. For cicada plants, five-number summary: Min = 0.109, $Q_1 = 0.217$, $M = 0.238$, $Q_3 = 0.276$, Max = 0.351; $\bar{x} = 0.243$, s = 0.048. For undisturbed plants, five-number summary: Min = 0.135, $Q_1 = 0.190$, $M = 0.241$, $Q_3 = 0.255$, Max = 0.290; $\bar{x} = 0.222$, s = 0.043.
Conclude: Both the boxplots and the values of the standard deviations suggest cicada plants used as fertilizer have slighty more variability in seed mass than undisturbed plants. The mean, first quartile, and third quartile of seed mass when using cicada plants as fertilizer are higher than for undisturbed plants. However, the median seed mass when using cicada plants as fertilizer is slightly lower than for undistributed plants. Based on the boxplots and summary statistics, we may have mild evidence that dead cicadas increase plant growth as measured by seed mass.

2.36. State: How do lean and obese people differ in time spent in activity and in time spent lying down?



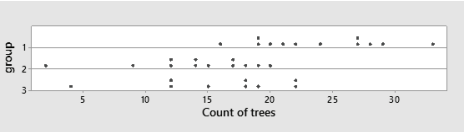
Plan: We will compare each pair of distributions using graphs and numerical summaries.

Solve: On the right are two dotplots; histograms or boxplots could also be used. None of the dotplots show any particular skewness, so either means and standard deviations or five-number summaries would be suitable. All values in the table are in units of minutes.

	$\bar{x}$	s	Min	Q_1	M	Q_3	Max
Lean/Active	525.751	107.121	319.212	504.700	549.522	584.644	677.188
Obese/Active	373.269	67.498	260.244	347.375	388.885	416.531	464.756
Lean/Lying down	501.646	52.045	396.962	467.700	510.291	537.362	567.006
Obese/Lying down	491.743	46.593	412.919	448.856	507.456	521.044	563.300

Conclude: In both the dotplots and the numerical summaries, we observe that lean subjects spent more active time than the obese subjects. There was little difference in time spent lying down.

2.37. State: How does logging affect tree count?



Plan: We need to compare the distributions, including appropriate measures of center and spread.

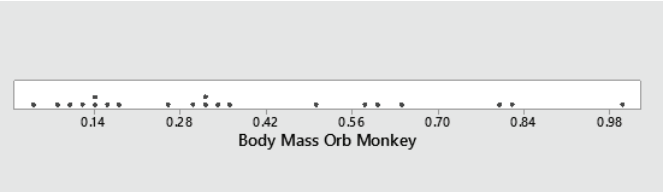
Solve: Dotplots are shown. Based on these, $\bar{x}$ and s are reasonable choices; the means and standard deviations (in units of trees) are given in the table (below).

Conclude: The means and the dotplots appear to suggest that logging reduces the number of trees per plot and that recovery is slow (the 1-year-after and 8-years-after means and dotplots are similar).

Group	$\bar{x}$	s
1	23.7500	5.06548
2	14.0833	4.98102
3	15.7778	5.76146

2.38. (a) From Exercise 2.4, we know IQR = 1.17. $1.5 \times IQR = 1.755$. Values below -0.405 (that's $Q_1 - 1.755$) or above 4.275 (that's $Q_3 + 1.755$) would be suspected outliers. Therefore, the largest value (12.8) would qualify as a suspected outlier. **(b)** No. None of the other quoted values besides the maximum qualify as suspected outliers based on the 1.5 x IQR rule. **(c)** Since none of the other quoted values (including the 95th percentile) qualify as suspected outliers, the distribution of blood mercury levels in this study does include at least one actual outlier. 12.8 is far away from the value of the 95th percentile (4.02).

2.39. Aggressions received are both higher and more variable for females of lower ranks. The 3rd dominance group has a high outlier, and the 4th dominance group has an extreme outlier.



2.40. **(a)** Five-number summary: Min $= 0.04$, $Q_1 = 0.135$, $M = 0.31$, $Q_3 = 0.585$, Max $= 0.99$; $\overline{x} = 0.365$, s $= 0.2736$. The mean is larger than the median. **(b)** Because $IQR = 0.45$, $1.5 \times IQR = 0.675$, spiders with a mass above 1.260 ($Q_1 - 0.675$ is negative and therefore impossible, $Q_3 + 0.675 = 1.260$) would be flagged as potential outliers. There were none. **(c)** Since the mean is larger than the median but there are no suspected outliers, we expect to see a right-skewed distribution. This is confirmed in the dotplot shown above.

2.41. **(a)** Five-number summary: Min $= 17.0$, $Q_1 = 22.5$, $M = 23.5$, $Q_3 = 25.0$, Max $= 31.5$. **(b)** Because $IQR = 2.5$, $1.5 \times IQR = 3.75$, states with a percent of residents under the age of 18 that is below 18.75 ($Q_1 - 3.75$) or greater than 28.75 ($Q_3 + 3.75$) are flagged as potential outliers. These are the two outliers, the District of Columbia (16.8) and Utah (31.5). **(c)** The mean percent residents under the age of 18 in the United States is 23.755, whereas the median is 23.5. The mean and the median are very similar because the distribution is slightly roughly symmetric (and the outliers are located on both sides).

LARGE DATA SET EXERCISES:
ANSWERS GROUPED IN "LARGE DATA SET" CHAPTER

Exercise 2.42 **Elderly health.**