Methods for Describing Sets of Data

Chapter **2**

2.2 In a bar graph, a bar or rectangle is drawn above each class of the qualitative variable corresponding to the class frequency or class relative frequency. In a pie chart, each slice of the pie corresponds to the relative frequency of a class of the qualitative variable.

2.4 First, we find the frequency of the grade A. The sum of the frequencies for all 5 grades must be 200. Therefore, subtract the sum of the frequencies of the other 4 grades from 200. The frequency for grade A is:

$$200 - (36 + 90 + 30 + 28) = 200 - 184 = 16$$

To find the relative frequency for each grade, divide the frequency by the total sample size, 200. The relative frequency for the grade B is 36/200 = .18. The rest of the relative frequencies are found in a similar manner and appear in the table:

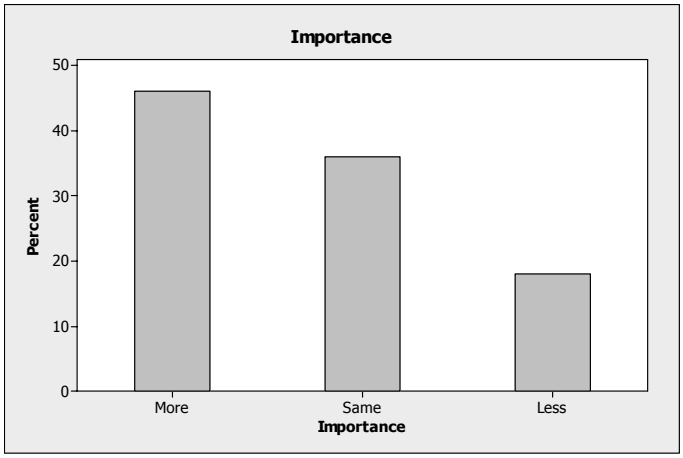
Grade on Statistics Exam	Frequency	Relative Frequency
A: 90–100	16	.08
B: 80– 89	36	.18
C: 65– 79	90	.45
D: 50– 64	30	.15
F: Below 50	28	.14
Total	200	1.00

2.6 a. The graph shown is a pie chart.

b. The qualitative variable described in the graph is opinion on library importance.

c. The most common opinion is more important, with 46.0% of the responders indicating that they think libraries have become more important.

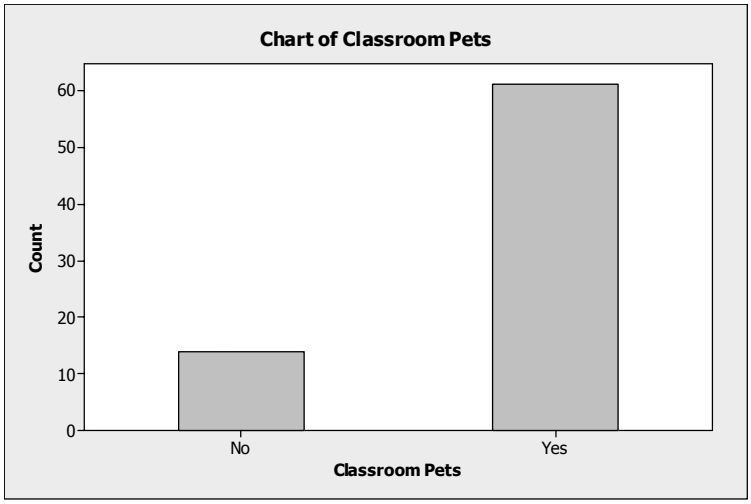
d. Using MINITAB, the Pareto diagram is:

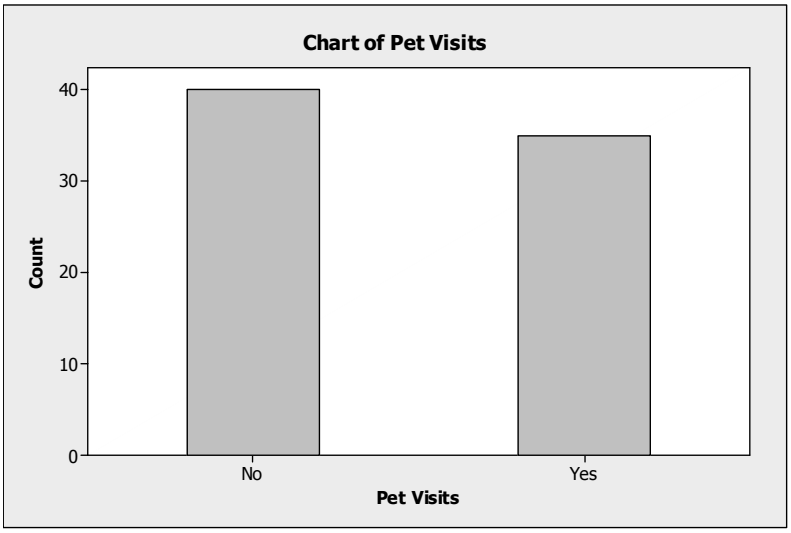


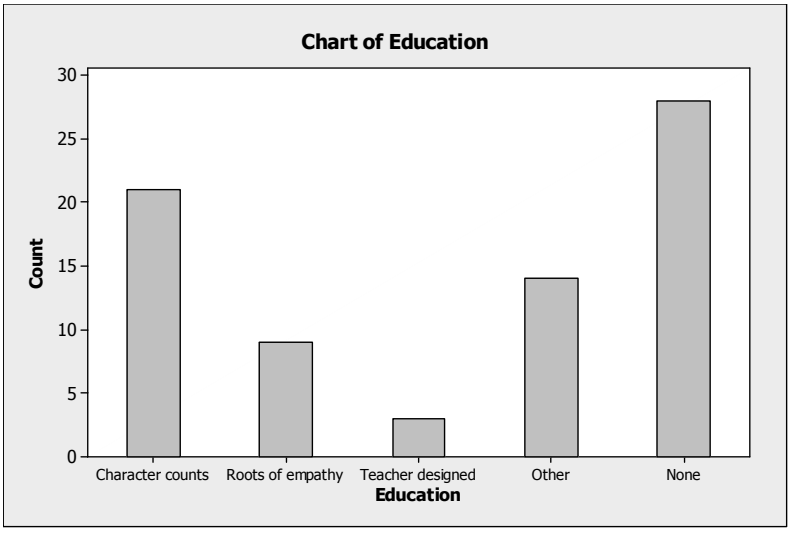
Of those who responded to the question, almost half (46%) believe that libraries have become more important to their community. Only 18% believe that libraries have become less important.

2.8 a. Data were collected on 3 questions. For questions 1 and 2, the responses were either 'yes' or 'no'. Since these are not numbers, the data are qualitative. For question 3, the responses include 'character counts', 'roots of empathy', 'teacher designed', other', and 'none'. Since these responses are not numbers, the data are qualitative.

b. Using MINITAB, bar charts for the 3 questions are:



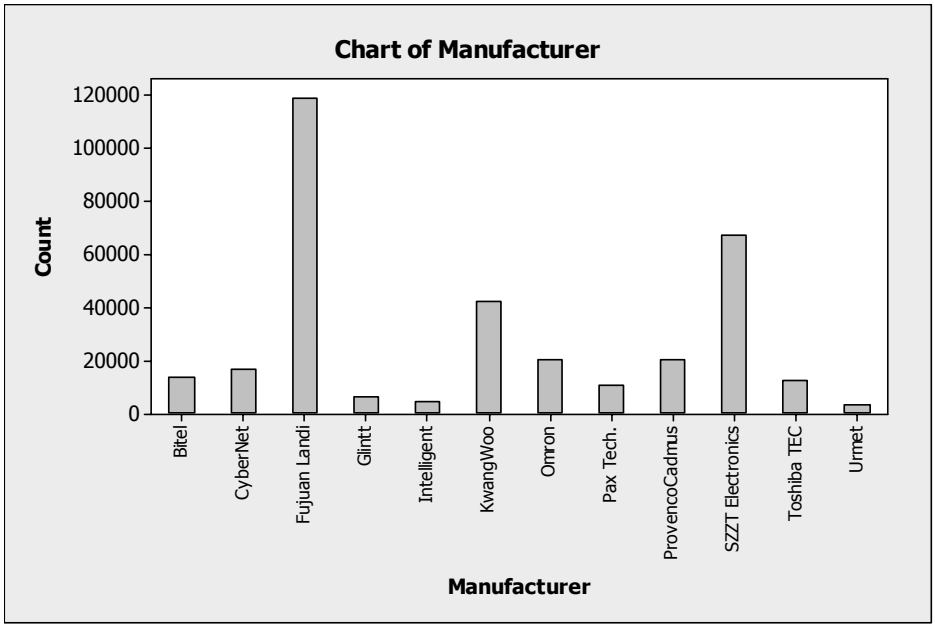




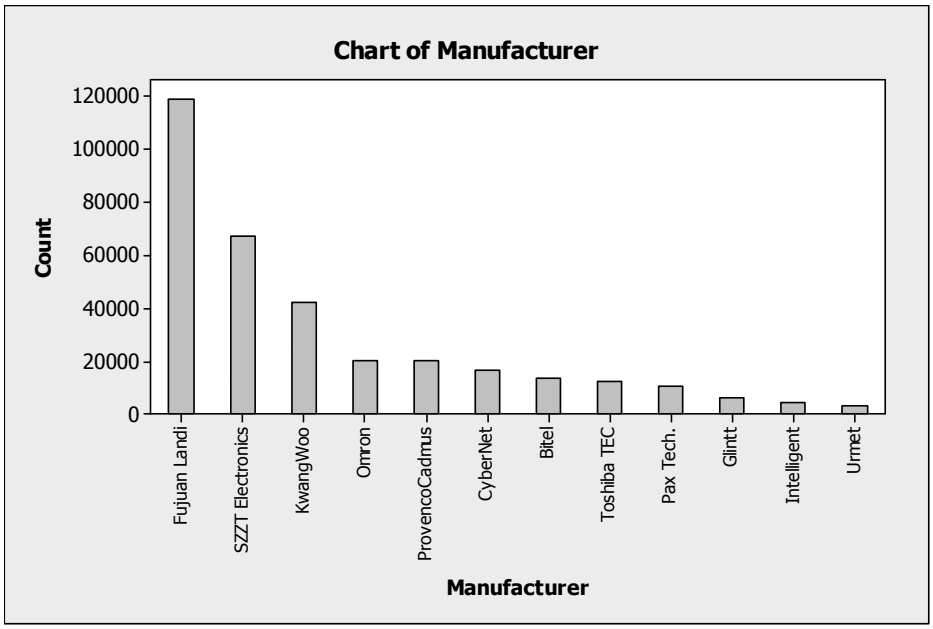
c. Many different things can be written. Possible answers might be: Most of the classroom teachers surveyed (61/75 = .813) keep classroom pets. A little less than half of the surveyed classroom teachers (35/75 = .467) allow visits by pets.

2.10 a. A PIN pad is selected and the manufacturer is determined. Since manufacturer is not a number, the data collected are qualitative.

b. Using MINITAB, the frequency bar chart is:



c. The Pareto chart for the data is:



Most of the PIN pads were shipped by Fujian Landi. They shipped almost twice as many PIN pads as the second highest manufacturer, which was SZZT Electronics. The three manufacturers with the smallest number of Pin pads shipped were Glintt, Intelligent, and Urmet.

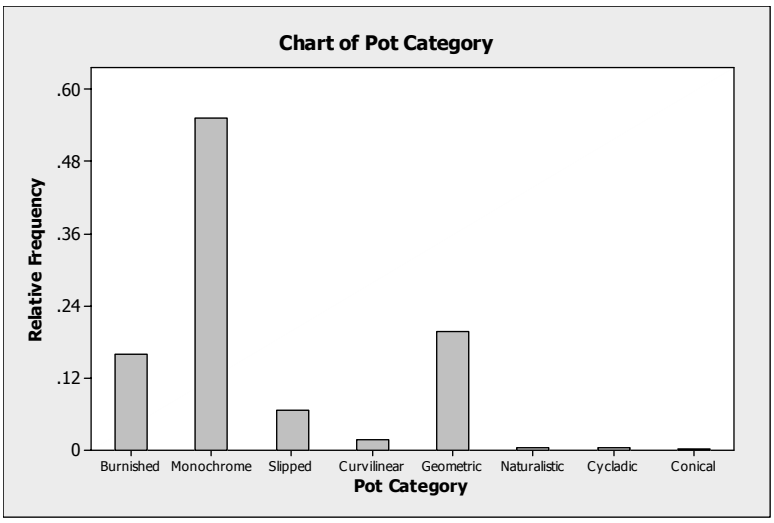
2.12 a. The two qualitative variables graphed in the bar charts are the occupational titles of clan individuals in the continued line and the occupational titles of clan individuals in the dropout line.

b. In the Continued Line, about 63% were in either the high or the middle grade. Only about 20% were in the nonofficial category. In the Dropout Line, only about 22% were in either the high or middle grade while about 64% were in the nonofficial category. The percents in the low grade and provincial official categories were about the same for the two lines.

2.14 Suppose we construct a relative frequency bar chart for this data. This will allow the archaeologists to compare the different categories easier. First, we must compute the relative frequencies for the categories. These are found by dividing the frequencies in each category by the total 837. For the burnished category, the relative frequency is 133 / 837 = .159. The rest of the relative frequencies are found in a similar fashion and are listed in the table.

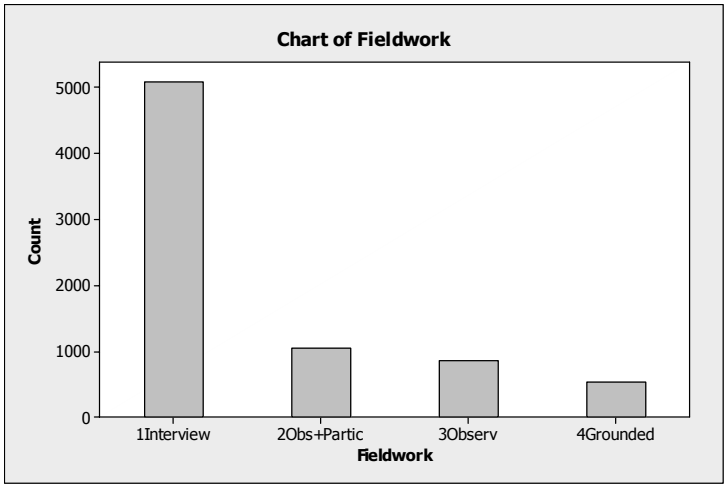
Pot Category	Number Found	Computation	Relative Frequency
Burnished	133	133 / 837	.159
Monochrome	460	460 / 837	.550
Slipped	55	55 / 837	.066
Curvilinear Decoration	14	14 / 837	.017
Geometric Decoration	165	165 / 837	.197
Naturalistic Decoration	4	4 / 837	.005
Cycladic White clay	4	4 / 837	.005
Cononical cup clay	2	2 / 837	.002
Total	837		1.001

A relative frequency bar chart is:



The most frequently found type of pot was the Monochrome. Of all the pots found, 55% were Monochrome. The next most frequently found type of pot was the Painted in Geometric Decoration. Of all the pots found, 19.7% were of this type. Very few pots of the types Painted in naturalistic decoration, Cycladic white clay, and Conical cup clay were found.

2.16 Using MINITAB, a bar graph is:

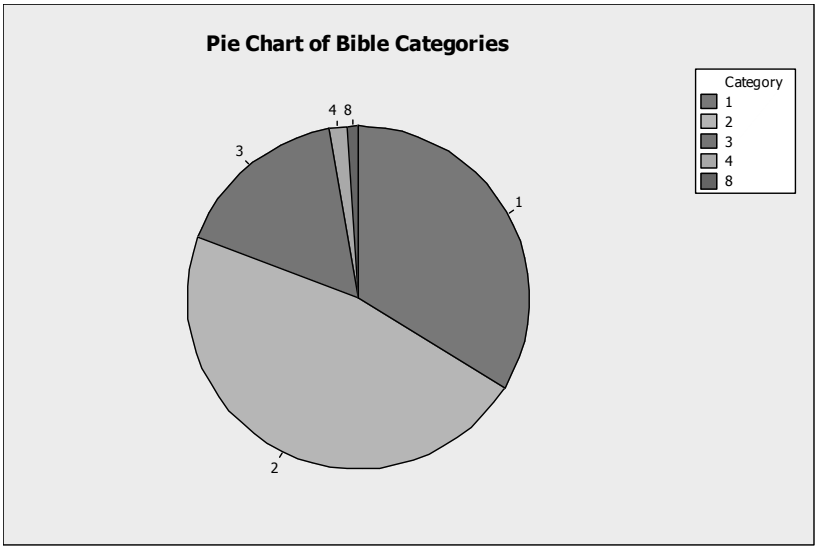


Most of the types of papers found were interviews. There were about twice as many interviews as all other types combined.

2.18 a. There were 1,470 responses that were missing. In addition, 14 responses were 8 = Don't know and 7 responses were 9 = Missing. The missing values were not included, but those responding with an 8 were kept. Therefore, there were only 1333 useable responses. The frequency table is:

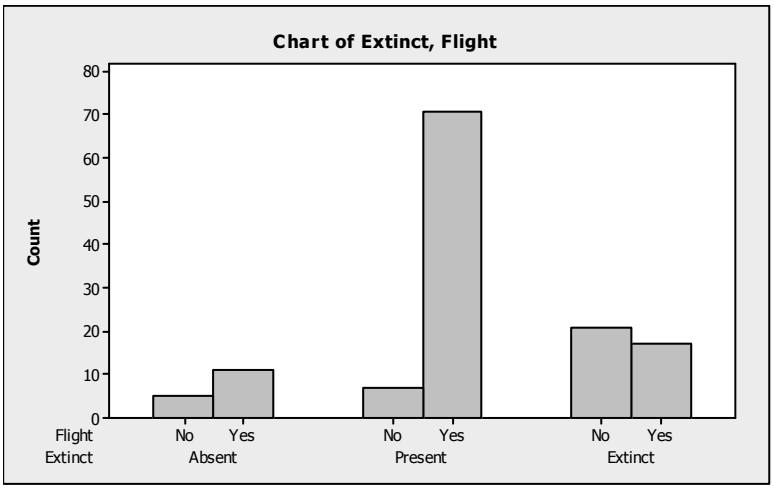
Response	Frequency	Relative Frequency
1	450	450/1333 = .338
2	627	627/1333 = .470
3	219	219/1333 = .164
4	23	23/1333 = .017
8	14	14/1333 = .011
Totals	**1333**	**1.000**

b. Using MINITAB, the pie chart for the data is:



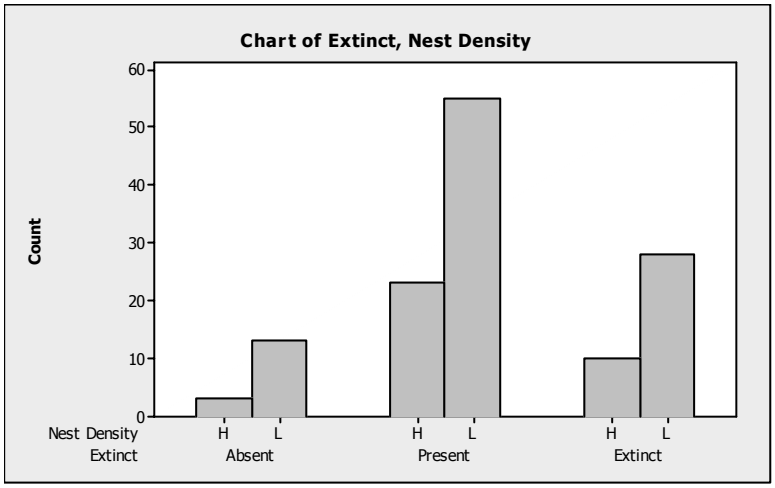
c. The response with the highest frequency is 2, 'the Bible is the inspired word of God but not everything is to be taken literally'. Almost 47% of the respondents selected this answer. About one-third of the respondents answered 1, 'the Bible is the actual word of God and is to be taken literally'. Very few (1.7%) of the respondents chose response 4, 'the Bible has some other origin' and response 8 (1.1%), 'Don't know'.

2.20 Using MINITAB a bar chart for the Extinct status versus flight capability is:



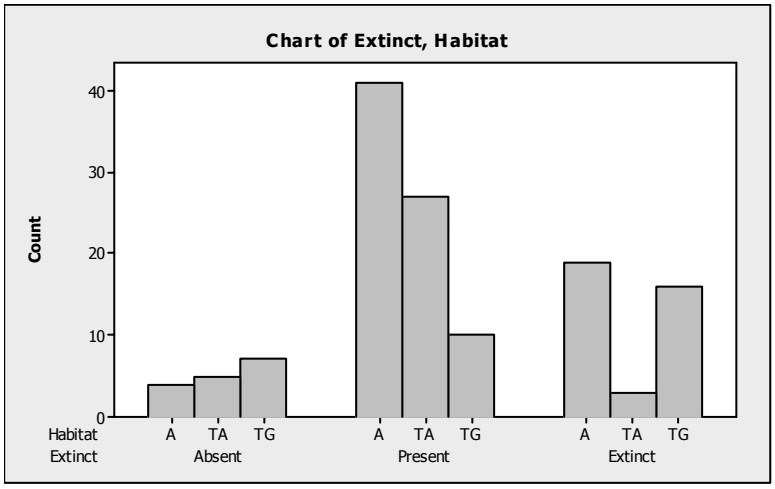
It appears that extinct status is related to flight capability. For birds that do have flight capability, most of them are present. For those birds that do not have flight capability, most are extinct.

The bar chart for Extinct status versus Nest Density is:



It appears that extinct status is not related to nest density. The proportion of birds present, absent, and extinct appears to be very similar for nest density high and nest density low.

The bar chart for Extinct status versus Habitat is:



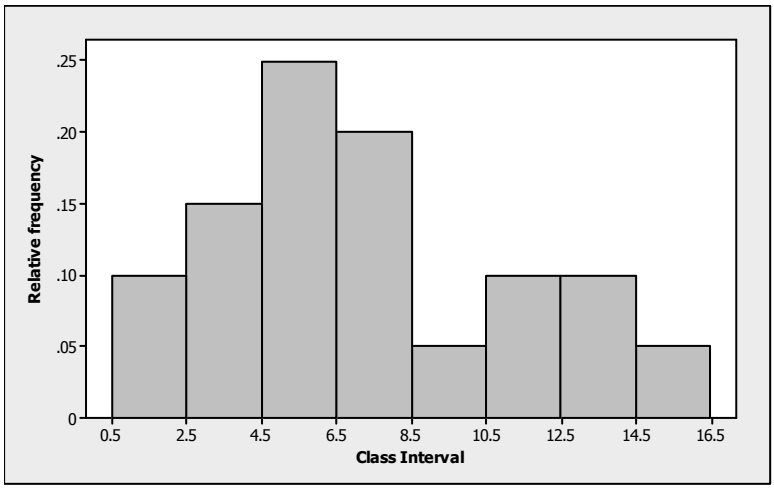
It appears that the extinct status is related to habitat. For those in aerial terrestrial (TA), most species are present. For those in ground terrestrial (TG), most species are extinct. For those in aquatic, most species are present.

2.22 The difference between a bar chart and a histogram is that a bar chart is used for qualitative data and a histogram is used for quantitative data. For a bar chart, the categories of the qualitative variable usually appear on the horizontal axis. The frequency or relative frequency for each category usually appears on the vertical axis. For a histogram, values of the quantitative variable usually appear on the horizontal axis and either frequency or relative frequency usually appears on the vertical axis. The quantitative data are grouped into intervals which appear on the horizontal axis. The number of observations appearing in each interval is then graphed. Bar charts usually leave spaces between the bars while histograms do not.

2.24 In a stem-and-leaf display, the stem is the left-most digits of a measurement, while the leaf is the right-most digit of a measurement.

2.26 As a general rule for data sets containing between 25 and 50 observations, we would use between 7 and 14 classes. Thus, for 50 observations, we would use around 14 classes.

2.28 Using MINITAB, the relative frequency histogram is:

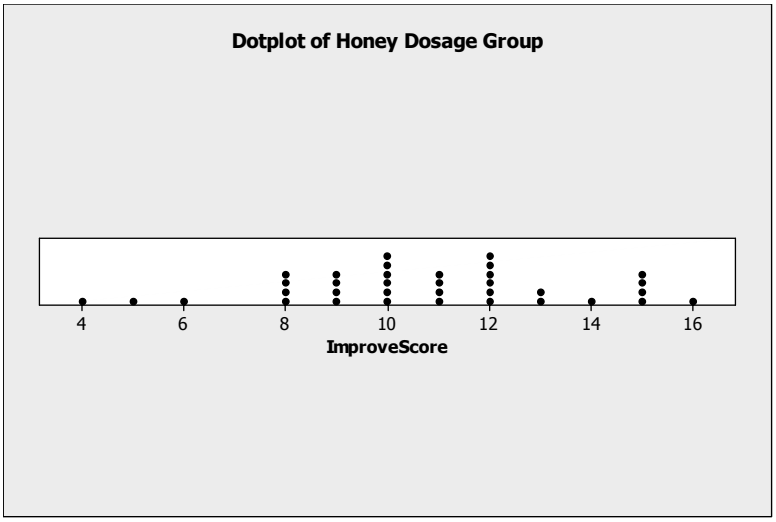


2.30 a. This is a frequency histogram because the number of observations are displayed rather than the relative frequencies.

b. There are 14 class intervals used in this histogram.

c. The total number of measurements in the data set is 49.

2.32 a. Using MINITAB, the dot plot of the honey dosage data is:



b. Both 10 and 12 occurred 6 times in the honey dosage group.

c. From the graph in part c, 8 of the top 11 scores (72.7%) are from the honey dosage group. Of the top 30 scores, 18 (60%) are from the honey dosage group. This supports the conclusions of the researchers that honey may be a preferable treatment for the cough and sleep difficulty associated with childhood upper respiratory tract infection.

2.34 Using MINITAB, the stem-and-leaf display is:

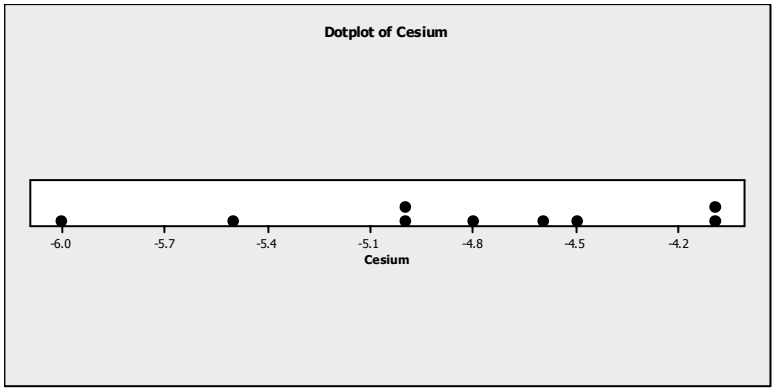
Stem-and-Leaf Display: Depth

```
Stem-and-leaf of Depth  N  = 18
Leaf Unit = 0.10


 2   13  29
 4   14  00
 8   15  7789
(3)  16  125
 7   17  08
 5   18  11
 3   19  347
```

The data in the stem-and-leaf display are displayed to 1 decimal place while the actual data is displayed to 2 decimal places. To 1 decimal place, there are 3 numbers that appear twice – 14.0, 15.7, and 18.1. However, to 2 decimal places, none of these numbers are the same. Thus, no molar depth occurs more frequently in the data.

2.36 a. Using MINITAB, the dot plot for the 9 measurements is:



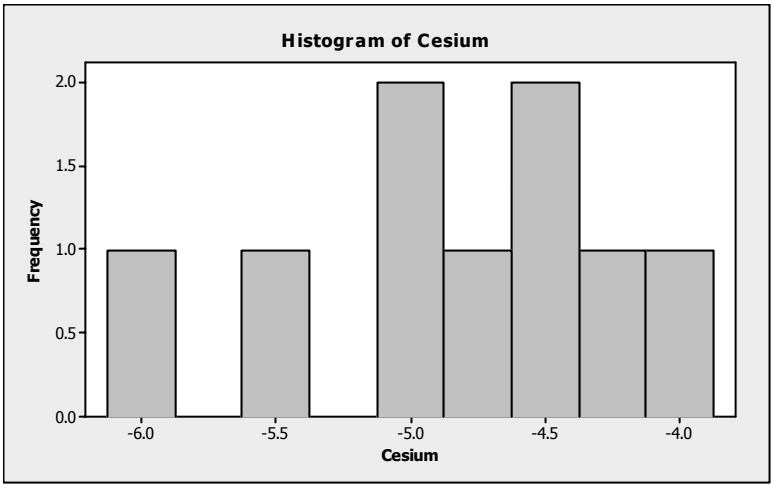
b. Using MINITAB, the stem-and-leaf display is:

Character Stem-and-Leaf Display

```
Stem-and-leaf of Cesium    N  = 9
Leaf Unit = 0.10

    1   -6 0
    2   -5 5
    4   -5 00
   (3)  -4 865
    2   -4 11
```

c. Using MINITAB, the histogram is:



d. The stem-and-leaf display appears to be more informative than the other graphs. Since there are only 9 observations, the histogram and dot plot have very few observations per category.

e. There are 4 observations with radioactivity level of -5.00 or lower. The proportion of measurements with a radioactivity level of -5.0 or lower is 4 / 9 = .444.

2.38 a. Using MINITAB, the stem-and-leaf display is:

Stem-and-Leaf Display: Spider

```
Stem-and-leaf of Spider  N  = 10
Leaf Unit = 10


 1   0  0
 3   0  33
(3)  0  455
 4   0  6 7
 2   0  9
 1   1  1
```

b. The spiders with a contrast value of 70 or higher are in bold type in the stem-and-leaf display in part **a**. There are 3 spiders in this group.

c. The sample proportion of spiders that a bird could detect is 3 / 10 = .3. Thus, we could infer that a bird could detect a crab-spider sitting on the yellow central part of a daisy about 30% of the time.

2.40 a. A stem-and-leaf display of the data using MINITAB is:

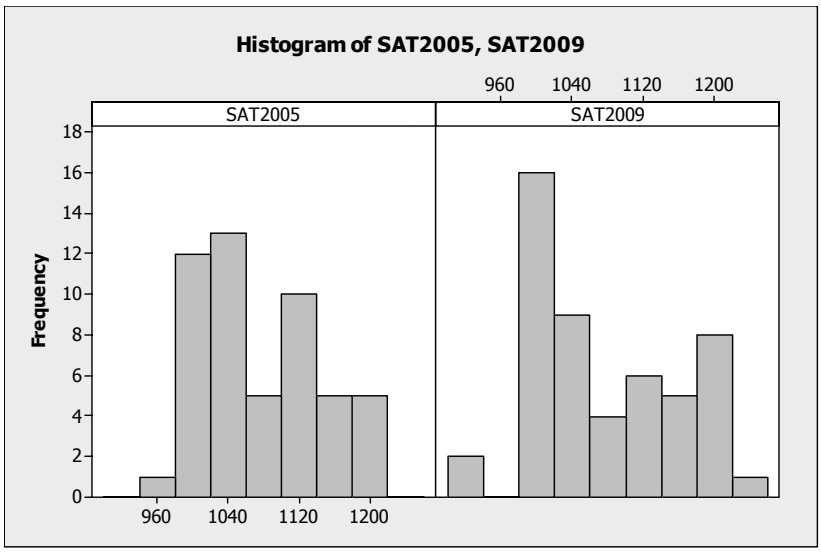
```
Stem-and-leaf of FNE       N  = 25
Leaf Unit = 1.0

 2     0 67
 3     0 8
 6     1 001
10     1 3333
12     1 45
(2)    1 66
11     1 8999
 7     2 0011
 3     2 3
 2     2 45
```

b. The numbers in bold in the stem-and-leaf display represent the bulimic students. Those numbers tend to be the larger numbers. The larger numbers indicate a greater fear of negative evaluation. Thus, the bulimic students tend to have a greater fear of negative evaluation.

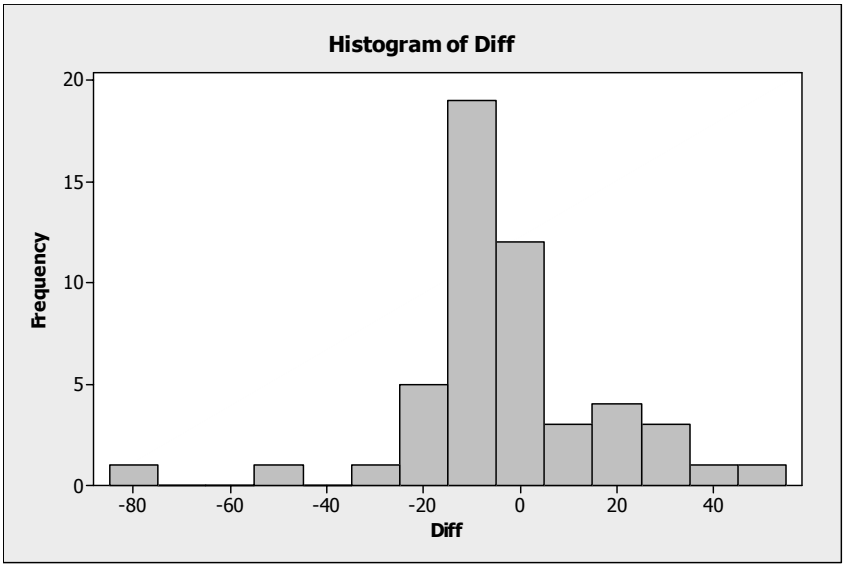
c. A measure of reliability indicates how certain one is that the conclusion drawn is correct. Without a measure of reliability, anyone could just guess at a conclusion.

2.42 a. Using MINITAB, histograms of the two sets of SAT scores are:



It appears that the distributions of both sets of scores are somewhat skewed to the right. However, there appears to be more lower SAT scores for 2009 and more higher SAT scores for 2009 than 2005.

b. Using MINITAB, a histogram of the differences of the 2009 and 2005 SAT scores is:



c. It appears that there are more differences less than 0 than above 0. Thus, it appears that in general, the 2009 SAT scores are lower than the 2005 SAT scores.

d. Wyoming had the largest improvement in SAT scores from 2005 to 2009, with an increase of 48 points.

2.44 a. $\sum x = 5+1+3+2+1 = 12$

b. $\sum x^2 = 5^2+1^2+3^2+2^2+1^2 = 40$

c. $\sum (x-1) = (5-1)+(1-1)+(3-1)+(2-1)+(1-1) = 7$

d. $\sum (x-1)^2 = (5-1)^2+(1-1)^2+(3-1)^2+(2-1)^2+(1-1)^2 = 21$

e. $\left(\sum x\right)^2 = (5+1+3+2+1)^2 = 12^2 = 144 \ = (5 + 1 + 3 + 2 + 1)^2 = 12^2 = 144$

2.46 Using the results from Exercise 2.44,

a. $\sum x^2 - \dfrac{\left(\sum x\right)^2}{5} = 40 - \dfrac{144}{5} = 40 - 28.8 = 11.2$

b. $\sum (x-2)^2 = (5-2)^2+(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2 = 12$

c. $\sum x^2 - 10 = 40 - 10 = 30$

2.48 A measure of central tendency measures the "center" of the distribution while measures of variability measure how spread out the data are.

2.50 The sample mean is represented by $\bar{x}$. The population mean is represented by μ.

2.52 A skewed distribution is a distribution that is not symmetric and not centered around the mean. One tail of the distribution is longer than the other. If the mean is greater than the median, then the distribution is skewed to the right. If the mean is less than the median, the distribution is skewed to the left.

2.54 Assume the data are a sample. The sample mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{3.2 + 2.5 + 2.1 + 3.7 + 2.8 + 2.0}{6} = \frac{16.3}{6} = 2.717$$

The median is the average of the middle two numbers when the data are arranged in order (since $n = 6$ is even). The data arranged in order are: 2.0, 2.1, 2.5, 2.8, 3.2, 3.7. The middle two numbers are 2.5 and 2.8. The median is:

$$\frac{2.5 + 2.8}{2} = \frac{5.3}{2} = 2.65$$

2.56 The median is the middle number once the data have been arranged in order. If n is even, there is not a single middle number. Thus, to compute the median, we take the average of the middle two numbers. If n is odd, there is a single middle number. The median is this middle number.

A data set with 5 measurements arranged in order is 1, 3, 5, 6, 8. The median is the middle number, which is 5.

A data set with 6 measurements arranged in order is 1, 3, 5, 5, 6, 8. The median is the average of the middle two numbers which is $\frac{5+5}{2} = \frac{10}{2} = 5$.

2.58 a. $\bar{x} = \frac{\sum x}{n} = \frac{7 + \cdots + 4}{6} = \frac{15}{6} = 2.5$

Median = $\frac{3+3}{2} = 3$ (mean of 3rd and 4th numbers, after ordering)

Mode = 3

b. $\bar{x} = \frac{\sum x}{n} = \frac{2 + \cdots + 4}{13} = \frac{40}{13} = 3.08$

Median = 3 (7th number, after ordering)

Mode = 3

c. $\bar{x} = \frac{\sum x}{n} = \frac{51 + \cdots + 37}{10} = \frac{496}{10} = 49.6$

Median = $\frac{48+50}{2} = 49$ (mean of 5th and 6th numbers, after ordering)

Mode = 50

2.60 a. From the printout, the sample mean is 50.02, the sample median is 51, and the sample mode is 54. The average age of the 50 most powerful women in business in the U.S. is 50.02 years. The median age is 51. Half of the 50 most powerful women in business in the U.S. are younger than 51 and half are older. The most common age is 54.

b. Since the mean is slightly smaller than the median, the data are skewed slightly to the left.

c. The modal class is the interval with the largest frequency. From the histogram the modal class is 50 to 54.

2.62 a. There are 35 observations in the honey dosage group. Thus, the median is the middle number, once the data have been arranged in order from the smallest to the largest. The middle number is the 18^{th} observation which is 11.

b. There are 33 observations in the DM dosage group. Thus, the median is the middle number, once the data have been arranged in order from the smallest to the largest. The middle number is the 17^{th} observation which is 9.

c. There are 37 observations in the control group. Thus, the median is the middle number, once the data have been arranged in order from the smallest to the largest. The middle number is the 19^{th} observation which is 7.

d. Since the median of the honey dosage group is the highest, the median of the DM groups is the next highest, and the median of the control group is the smallest, we can conclude that the honey dosage is the most effective, the DM dosage is the next most effective, and nothing (control) is the least effective.

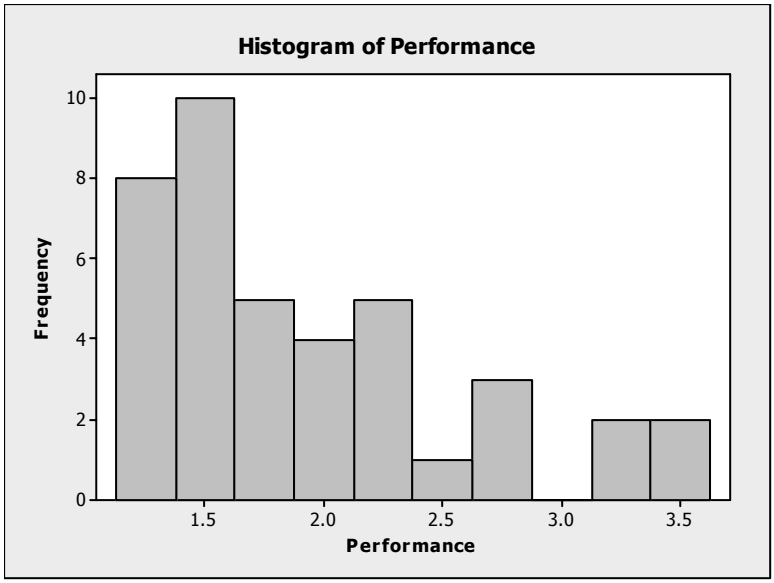
2.64 a. The mean of the driving performance index values is: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{77.07}{40} = 1.927$

The median is the average of the middle two numbers once the data have been arranged in order. After arranging the numbers in order, the 20^{th} and 21^{st} numbers are 1.75 and 1.76. The median is: $\dfrac{1.75+1.76}{2} = 1.755$

The mode is the number that occurs the most frequently and is 1.4.

b. The average driving performance index is 1.927. The median is 1.755. Half of the players have driving performance index values less than 1.755 and half have values greater than 1.755. Three of the players have the same index value of 1.4.

c. Since the mean is greater than the median, the data are skewed to the right. Using MINITAB, a histogram of the data is:



2.66 a. The salaries of all persons employed by a large university are probably skewed to the right. There will be a few individuals with very large salaries (i.e. president, football coach, Dean of the Medical school). However, the majority of the employees will have salaries in a rather small range.

b. The grades on an easy test will probably be skewed to the left. Most students will get very high grades on the test. Since there is an upper limit to the grades (i.e. 100%), there will likely be many grades in this upper range. However, even on an easy test, a few individuals will still not do well.

c. The grades on a difficult test will probably be skewed to the right. Most students will get fairly low grades on the test. However, even on a difficult test, a few individuals will still do quite well.

d. The amounts of time students in your class studied last week will probably be close to symmetric. Some individuals will not study very much, while others will study quite a bit. However, most students will study an average amount of time.

e. The ages of cars on a used car lot will probably be skewed to the left. Most of the cars will be fairly new. However, there will probably be a few fairly old cars.

f. The amounts of time spent by students on a difficult examination will probably be skewed to the left. If there is a maximum time limit, then most students will take that amount of time or close to it. There will probably be a few students who take less time than the maximum allowed.

2.68 a. The mean number of ant species discovered is:

$$\bar{x} = \frac{\sum x}{n} = \frac{3+3+...+4}{11} = \frac{141}{11} = 12.82$$

The median is the middle number once the data have been arranged in order:
3, 3, 4, 4, 4, 5, 5, 5, 7, 49, 52.

The median is 5.

The mode is the value with the highest frequency. Since both 4 and 5 occur 3 times, both 4 and 5 are modes.

b. For this case, we would recommend that the median is a better measure of central tendency than the mean. There are 2 very large numbers compared to the rest. The mean is greatly affected by these 2 numbers, while the median is not.

c. The mean total plant cover percentage for the Dry Steppe region is:

$$\bar{x} = \frac{\sum x}{n} = \frac{40+52+...+27}{5} = \frac{202}{5} = 40.4$$

The median is the middle number once the data have been arranged in order:
27, 40, 40, 43, 52.

The median is 40.

The mode is the value with the highest frequency. Since 40 occurs 2 times, 40 is the mode.

d. The mean total plant cover percentage for the Gobi Desert region is:

$$\bar{x} = \frac{\sum x}{n} = \frac{30+16+...+14}{6} = \frac{168}{6} = 28$$

The median is the mean of the middle 2 numbers once the data have been arranged in order: 14, 16, 22, 30, 30, 56.

The median is $\frac{22+30}{2} = \frac{52}{2} = 26$.

The mode is the value with the highest frequency. Since 30 occurs 2 times, 30 is the mode.

e. Yes, the total plant cover percentage distributions appear to be different for the 2 regions. The percentage of plant coverage in the Dry Steppe region is much greater than that in the Gobi Desert region.

2.70 a. The mean number of power plants is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+2+4+...+3}{20} = \frac{78}{20} = 3.9$$

The median is the mean of the middle 2 numbers once the data have been arranged in order: 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 7, 9, 11

The median is $\frac{3+4}{2} = \frac{7}{2} = 3.5$.

The number 1 occurs 5 times. The mode is 1.

b. Deleting the largest number, 11, the new mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+2+4+...+3}{19} = \frac{67}{19} = 3.526$$

The median is the middle number once the data have been arranged in order:
1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 7, 9

The median is 3.

The number 1 occurs 5 times. The mode is 1.

By dropping the largest measurement from the data set, the mean drops from 3.9 to 3.526. The median drops from 3.5 to 3 and the mode stays the same.

c. Deleting the lowest 2 and highest 2 measurements leaves the following:

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 7

The new mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+1+1+...+7}{16} = \frac{56}{16} = 3.5$$

The trimmed mean has the advantage that some possible outliers have been eliminated.

2.72 The primary disadvantage of using the range to compare variability of data sets is that the two data sets can have the same range and be vastly different with respect to data variation. Also, the range is greatly affected by extreme measures.

2.74 The variance of a data set can never be negative. The variance of a sample is the sum of the *squared* deviations from the mean divided by $n - 1$. The square of any number, positive or negative, is always positive. Thus, the variance will be positive.

The variance is usually greater than the standard deviation. However, it is possible for the variance to be smaller than the standard deviation. If the data are between 0 and 1, the variance will be smaller than the standard deviation. For example, suppose the data set is .8, .7, .9, .5, and .3. The sample mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{.8+.7+.9+.5+.3}{.5} = \frac{3.2}{5} = .64$$

The sample variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{2.28 - \frac{3.2^2}{5}}{5-1} = \frac{2.28 - 2.048}{4} = .058$$

The standard deviation is $s = \sqrt{.058} = .241$

2.76 a. $s^2 = \dfrac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{84 - \frac{20^2}{10}}{10-1} = 4.8889$ $\qquad s = \sqrt{4.8889} = 2.211$

b. $s^2 = \dfrac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{380 - \frac{100^2}{40}}{40-1} = 3.3333$ $\qquad s = \sqrt{3.3333} = 1.826$

c. $s^2 = \dfrac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{18 - \frac{17^2}{20}}{20-1} = .1868$ $\qquad s = \sqrt{.1868} = .432$

2.78 a. Range = 4 – 0 = 4

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{22 - \frac{8^2}{5}}{4-1} = 2.3 \qquad s = \sqrt{2.3} = 1.52$$

b. Range = 6 – 0 = 6

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{63 - \frac{17^2}{7}}{7-1} = 3.619 \qquad s = \sqrt{3.619} = 1.90$$

c. Range = 8 – (–2) = 10

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{145 - \frac{27^2}{9}}{9-1} = 8 \qquad s = \sqrt{8} = 2.828$$

d. Range = 2 – (–3) = 5

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{29 - \frac{(-5)^2}{18}}{18-1} = 1.624 \qquad s = \sqrt{1.624} = 1.274$$

2.80 This is one possibility for the two data sets.

Data Set 1: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Data Set 2: 0, 0, 1, 1, 2, 2, 3, 3, 9, 9

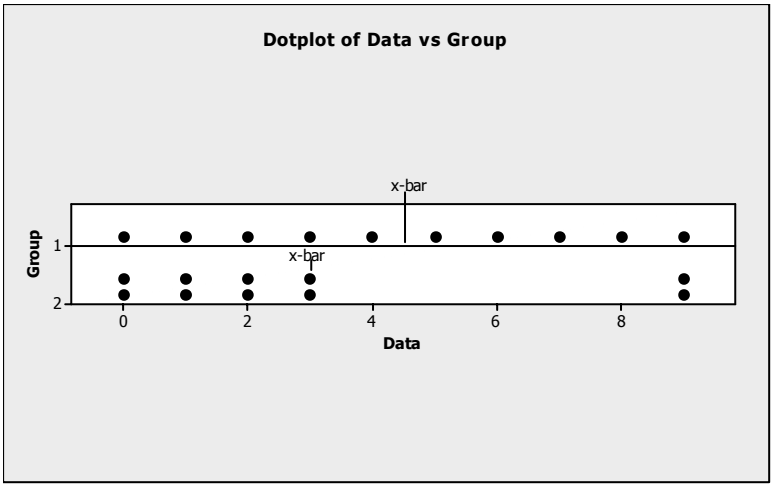
The two sets of data above have the same range = largest measurement – smallest measurement = 9 – 0 = 9.

The means for the two data sets are:

$$\bar{x}_1 = \frac{\sum x}{n} = \frac{0+1+2+3+4+5+6+7+8+9}{10} = \frac{45}{10} = 4.5$$

$$\bar{x}_2 = \frac{\sum x}{n} = \frac{0+0+1+1+2+2+3+3+9+9}{10} = \frac{30}{10} = 3$$

The dot diagrams for the two data sets are shown below.



2.82 a. $$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{226 - \frac{28^2}{5}}{5-1} = \frac{69.2}{4} = 17.3 \qquad s = \sqrt{17.3} = 4.1593$$

b. $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{1213 - \dfrac{55^2}{4}}{4-1} = \dfrac{456.75}{3} = 152.25$ square feet

$s = \sqrt{152.25} = 12.339$ feet

c. $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{59 - \dfrac{(-15)^2}{6}}{6-1} = \dfrac{21.5}{5} = 4.3$ $\qquad s = \sqrt{4.3} = 2.0736$

d. $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{\dfrac{24}{25} - \dfrac{2^2}{6}}{6-1} = \dfrac{.2933}{5} = .0587$ square ounces

$s = \sqrt{.0587} = .2422$ ounce

2.84 a. For those students who earned A, the range is 53 – 24 = 29.

The variance is $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{11,482 - \dfrac{296^2}{8}}{7} = \dfrac{530}{7} = 75.7143$

The standard deviation is $s = \sqrt{s^2} = \sqrt{75.7143} = 8.701$.

b. For those students who earned a B or C, the range is 40 – 16 = 24.

The variance is $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{3,965 - \dfrac{147^2}{6}}{5} = \dfrac{363.5}{5} = 72.7$

The standard deviation is $s = \sqrt{s^2} = \sqrt{72.7} = 8.526$.

c. The students who received A's have a more variable distribution of the number of books read. The range, variance, and standard deviation for this group are greater than the corresponding values for the B-C group

2.86 a. The range is the difference between the largest and smallest observations and is 17.83 – 4.90 = 12.93 meters.

b. The variance is:

$$s^2 = \frac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \frac{1428.64 - \dfrac{126.32^2}{13}}{13-1} = 16.767 \text{ square meters}$$

c. The standard deviation is $s = \sqrt{16.767} = 4.095$ meters.

2.88 a. The maximum age is 64. The minimum age is 28. The range is 64 – 28 = 36.

b. The variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{127135 - \frac{2501^2}{50}}{50-1} = 41.530$$

c. The standard deviation is:

$$s = \sqrt{s^2} = \sqrt{41.53} = 6.444$$

d. Since the standard deviation of the ages of the 50 most powerful women in Europe is 10 years and is greater than that in the U.S. (6.444 years), the age data for Europe is more variable.

e. If the largest age (64) is omitted, then the standard deviation would decrease. The new variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{123039 - \frac{2437^2}{49}}{49-1} = 38.241$$

The new standard deviation is $s = \sqrt{s^2} = \sqrt{38.241} = 6.184$. This is less than the standard deviation with all the observations (s = 6.444).

2.90 Chebyshev's rule can be applied to any data set. The Empirical Rule applies only to data sets that are mound-shaped—that are approximately symmetric, with a clustering of measurements about the midpoint of the distribution and that tail off as one moves away from the center of the distribution.

2.92 Since no information is given about the data set, we can only use Chebyshev's rule.

a. Nothing can be said about the percentage of measurements which will fall between $\bar{x} - s$ and $\bar{x} + s$.

b. At least 3/4 or 75% of the measurements will fall between $\bar{x} - 2s$ and $\bar{x} + 2s$.

c. At least 8/9 or 89% of the measurements will fall between $\bar{x} - 3s$ and $\bar{x} + 3s$.

2.94 a. $\bar{x} = \frac{\sum x}{n} = \frac{206}{25} = 8.24$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{1778 - \frac{206^2}{25}}{25-1} = 3.357 \qquad s = \sqrt{s^2} = 1.83$$

b.

Interval	Number of Measurements in Interval	Percentage
$\bar{x} \pm s$, or (6.41, 10.07)	18	18/25 = .72 or 72%
$\bar{x} \pm 2s$, or (4.58, 11.90)	24	24/25 = .96 or 96%
$\bar{x} \pm 3s$, or (2.75, 13.73)	25	25/25 = 1 or 100%

c. The percentages in part **b** are in agreement with Chebyshev's rule and agree fairly well with the percentages given by the Empirical Rule.

d. Range = 12 − 5 = 7

$s \approx$ range/4 = 7/4 = 1.75

The range approximation provides a satisfactory estimate of s.

2.96 From Exercise 2.60, the sample mean is $\bar{x} = 50.02$. From Exercise 2.88, the sample standard deviation is $s = 6.444$. From Chebyshev's Rule, at least 75% of the ages will fall within 2 standard deviations of the mean. This interval will be:

$$\bar{x} \pm 2s \Rightarrow 50.02 \pm 2(6.444) \Rightarrow 50.02 \pm 12.888 \Rightarrow (37.132, 62.908)$$

2.98 a. If the data are symmetric and mound shaped, then the Empirical Rule will describe the data. About 95% of the observations will fall within 2 standard deviation of the mean. The interval two standard deviations below and above the mean is $\bar{x} \pm 2s \Rightarrow 39 \pm 2(6) \Rightarrow 39 \pm 12 \Rightarrow (27, 51)$. This range would be 27 to 51.

b. To find the number of standard deviations above the mean a score of 51 would be, we subtract the mean from 51 and divide by the standard deviation. Thus, a score of 51 is $\frac{51-39}{6} = 2$ standard deviations above the mean. From the Empirical Rule, about .025 of the drug dealers will have WR scores above 51.

c. By the Empirical Rule, about 99.7% of the observations will fall within 3 standard deviations of the mean. Thus, nearly all the scores will fall within 3 standard deviations of the mean. The interval three standard deviations below and above the mean is $\bar{x} \pm 3s \Rightarrow 39 \pm 3(6) \Rightarrow 39 \pm 18 \Rightarrow (21, 57)$. This range would be 21 to 57.

2.100 a. $\bar{x} \pm 2s \Rightarrow 13.2 \pm 2(19.5) \Rightarrow 13.2 \pm 39 \Rightarrow (-25.8, 52.2)$. Since time cannot be negative, the interval will be $(0, 52.2)$.

b. The number of minutes a student uses a laptop for taking notes each day must be a positive number. The standard deviation is larger than the mean. Thus, even one standard deviation below the mean is a negative number. This implies that the distribution cannot be symmetric.

c. Since we know the distribution of usage times cannot be symmetric, we can use Chebyshev's Rule. We know that at least ¾ or 75% of the observations will be within 2 standard deviations of the mean. Thus, we know that at least 75% of the students have laptop usages between -25.8 and 52.2 minutes per day. Since we know we cannot have negative usages, the interval will be from 0 to 52.2 minutes.

2.102 a. There are 2 observations with missing values for egg length, so there are only 130 useable observations.

$$\bar{x} = \frac{\sum x}{n} = \frac{7{,}885}{130} = 60.65$$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{727{,}842 - \frac{(7{,}885)^2}{130}}{130-1} = \frac{249{,}586.4231}{129} = 1{,}934.7785$$

$$s = \sqrt{s^2} = \sqrt{1{,}934.7785} = 43.99$$

b. The data are not symmetrical or mound-shaped. Thus, we will use Chebyshev's Rule. We know that there are at least 8/9 or 88.9% of the observations within 3 standard deviations of the mean. Thus, at least 88.9% of the observations will fall in the interval:

$$\bar{x} \pm 3s \Rightarrow 60.65 \pm 3(43.99) \Rightarrow 60.65 \pm 131.97 \Rightarrow (-71.32,\ 192.69)$$

Since it is impossible to have negative egg lengths, at least 88.9% of the egg lengths will be between 0 and 192.69.

2.104 If we assume that the distributions are symmetric and mound-shaped, then the Empirical Rule will describe the data. We will compute the mean plus or minus one, two and three standard deviations for both data sets:

Low income:

$$\bar{x} \pm s \Rightarrow 7.62 \pm 8.91 \Rightarrow (-1.29, 16.53)$$

$$\bar{x} \pm 2s \Rightarrow 7.62 \pm 2(8.91) \Rightarrow 7.62 \pm 17.82 \Rightarrow (-10.20, 25.44)$$

$$\bar{x} \pm 3s \Rightarrow 7.62 \pm 3(8.91) \Rightarrow 7.62 \pm 26.73 \Rightarrow (-19.11, 34.35)$$

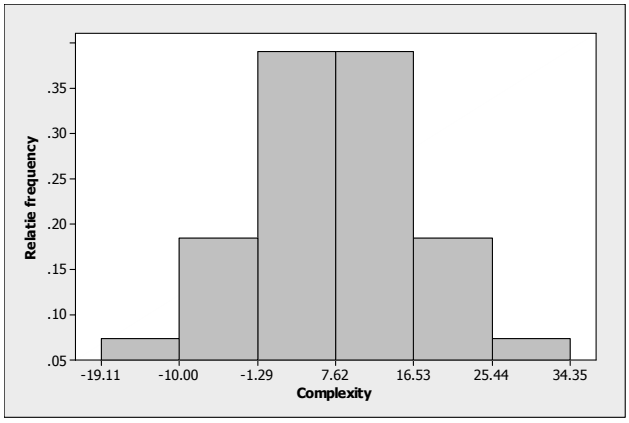
Middle Income:

$$\bar{x} \pm s \Rightarrow 15.55 \pm 12.24 \Rightarrow (3.31, 27.79)$$

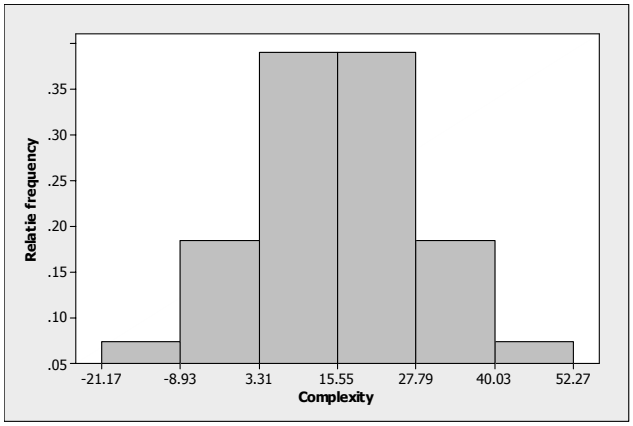
$$\bar{x} \pm 2s \Rightarrow 15.55 \pm 2(12.24) \Rightarrow 15.55 \pm 24.48 \Rightarrow (-8.93, 40.03)$$

$$\bar{x} \pm 3s \Rightarrow 15.55 \pm 3(12.24) \Rightarrow 15.55 \pm 36.72 \Rightarrow (-21.17, 52.27)$$

The histogram for the low income group is as follows:



The histogram for the middle income group is as follows:



The spread of the data for the middle income group is much larger than that of the low income group. The middle of the distribution for the middle income group is 15.55, while the middle of the distribution for the low income group is 7.62. Thus, the middle of the distribution for the middle income group is shifted to the right of that for the low income group.

We might be able to compare the means for the two groups. From the data provided, it looks like the mean score for the middle income group is greater than the mean score for the lower income group.

(Note: From looking at the data, it is rather evident that the distributions are not mound-shaped and symmetric. For the low income group, the standard deviation is larger than the mean. Since the smallest measurement allowed is 0, this indicates that the data set is not symmetric but skewed to the right. A similar argument could be used to indicate that the data set of middle income scores is also skewed to the right.)

2.106 To decide which group the patient is most likely to come from, we will compute the z-score for each group.

Group T: $z = \dfrac{x-\mu}{\sigma} = \dfrac{22.5-10.5}{7.6} = 1.58$

Group V: $z = \dfrac{x-\mu}{\sigma} = \dfrac{22.5-3.9}{7.5} = 2.48$

Group C: $z = \dfrac{x-\mu}{\sigma} = \dfrac{22.5-1.4}{7.5} = 2.81$

The patient is most likely to have come from Group T. The z-score for Group T is $z = 1.58$. This would not be an unusual z-score if the patient was in Group T. The z-scores for the other 2 groups are both greater than 2. We know that z-scores greater than 2 are rather unusual.

2.108 a. The 50th percentile is also called the median.

b. The Q_L is the lower quartile. This is also the 25th percentile or the score which has 25% of the observations less than it.

c. The Q_U is the upper quartile. This is also the 75th percentile or the score which has 75% of the observations less than it.

2.110 For mound-shaped distributions, we can use the Empirical Rule. About 95% of the observations will fall within 2 standard deviations of the mean. Thus, about 95% of the measurements will have z-scores between -2 and 2.

2.112 We first compute z-scores for each x value.

a. $z = \dfrac{x-\mu}{\sigma} = \dfrac{100-50}{25} = 2$

b. $z = \dfrac{x-\mu}{\sigma} = \dfrac{1-4}{1} = -3$

c. $z = \dfrac{x-\mu}{\sigma} = \dfrac{0-200}{100} = -2$

d. $z = \dfrac{x-\mu}{\sigma} = \dfrac{10-5}{3} = 1.67$

The above z-scores indicate that the x value in part **a** lies the greatest distance above the mean and the x value of part **b** lies the greatest distance below the mean.

2.114 The mean score is 283. This is the arithmetic average score of U.S. eighth graders on the mathematics assessment test. The 25th percentile score is 259. This indicates that 25% of the U.S. eighth graders scored 259 or lower on the assessment test. The 75th percentile score is 308. This indicates that 75% of the U.S. eighth graders scored 308 or lower on the assessment test. The 90th percentile score is 329. This indicates that 90% of the U.S. eighth graders scored 329 or lower on the assessment test.

2.116 From Exercise 2.35, $\bar{x} = 95.699$ and $s = 4.963$.

a. The z-score for the Nautilus Explorer is: $z = \dfrac{x - \bar{x}}{s} = \dfrac{74 - 95.699}{4.963} = -4.37$

The score for the Nautilus Explorer is 4.37 standard deviations below the mean for all the cruise ships.

b. The z-score for the Rotterdam is: $z = \dfrac{x - \bar{x}}{s} = \dfrac{92 - 95.699}{4.963} = -0.75$

The score for the Rotterdam is 0.75 standard deviations below the mean for all the cruise ships.

2.118 a. The mean number of books read by students who earned an A grade is:

$$\bar{x} = \frac{\sum x}{n} = \frac{296}{8} = 37$$

From Exercise 2.84, $s = 8.701$.

The z-score for a score of 40 books is $z = \dfrac{x - \bar{x}}{s} = \dfrac{40 - 37}{8.701} = 0.34$. Thus, someone who read 40 books read more than the average number of books, but that number is not very unusual.

b. The mean number of books read by students who earned a B or C grade is:

$$\bar{x} = \frac{\sum x}{n} = \frac{147}{6} = 24.5$$

From Exercise 2.84, $s = 8.526$.

The z-score for a score of 40 books is $z = \dfrac{x - \bar{x}}{s} = \dfrac{40 - 24.5}{8.526} = 1.82$. Thus, someone who read 40 books read many more than the average number of books. Very few students who received a B or a C read more than 40 books.

c. The group of students who earned A's is more likely to have read 40 books. For this group, the z-score corresponding to 40 books is .34. This is not unusual. For the B-C group, the z-score corresponding to 40 books is 1.82. This is close to 2 standard deviations from the mean. This would be fairly unusual.

2.120 Since the 90th percentile of the study sample in the subdivision was .00372 mg/L, which is less than the USEPA level of .015 mg/L, the water customers in the subdivision are not at risk of drinking water with unhealthy lead levels.

2.122 a. If the distribution is mound-shaped and symmetric, then the Empirical Rule can be used. Approximately 68% of the scores will fall within 1 standard deviation of the mean or between 53% ± 15% or between 38% and 68%. Approximately 95% of the scores will fall within 2 standard deviations of the mean or between 53% ± 2(15%) or between 23% and 83%. Approximately all of the scores will fall within 3 standard deviations of the mean or between 53% ± 3(15%) or between 8% and 98%.

b. If the distribution is mound-shaped and symmetric, then the Empirical Rule can be used. Approximately 68% of the scores will fall within 1 standard deviation of the mean or between 39% ± 12% or between 27% and 51%. Approximately 95% of the scores will fall within 2 standard deviations of the mean or between 39% ± 2(12%) or between 15% and 63%. Approximately all of the scores will fall within 3 standard deviations of the mean or between 39% ± 3(12%) or between 3% and 75%.

c. Since the scores on the red exam are shifted to the left of those on the blue exam, a score of 20% is more likely to occur on the red exam than on the blue exam.

2.124 Yes. From the graph in Exercise 2.121 c, we can see that there are 4 observations with z-scores greater than 3. There is then a gap down to 2.18. Those 4 observations are quite different from the rest of the data. After those 4 observations, the data are fairly similar. We know that by ranking the data, we can reduce the influence of outliers. But, by doing this, we loose valuable information.

2.126 The interquartile range is the distance between the upper and lower quartiles.

2.128 For a mound-shaped distribution, the Empirical Rule can be used. Almost all of the observations will fall within 3 standard deviations of the mean. Thus, almost all of the observations will have z-scores between -3 and 3.

2.130 The interquartile range is IQR = $Q_U - Q_L = 85 - 60 = 25$.

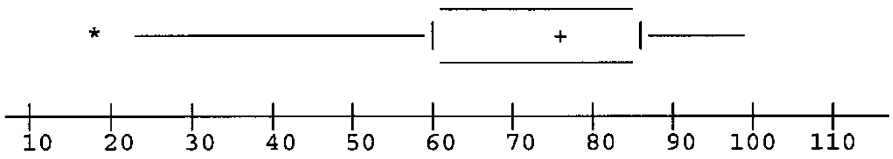
The lower inner fence = $Q_L - 1.5(\text{IQR}) = 60 - 1.5(25) = 22.5$.

The upper inner fence = $Q_U + 1.5(\text{IQR}) = 85 + 1.5(25) = 122.5$.

The lower outer fence = $Q_L - 3(\text{IQR}) = 60 - 3(25) = -15$.

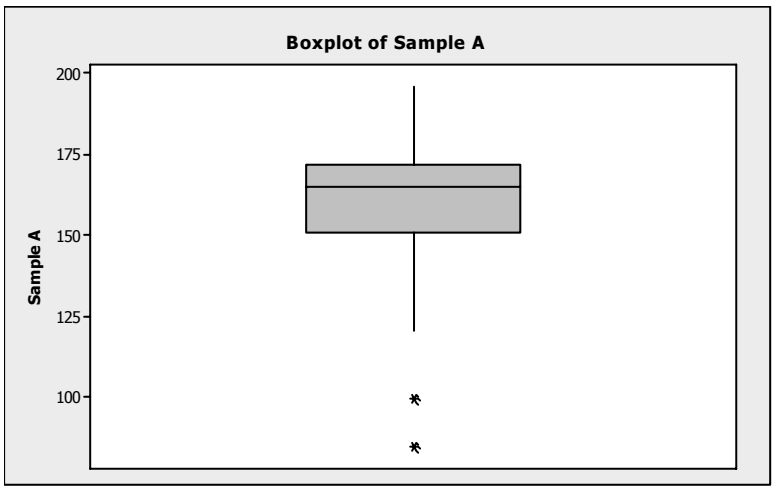
The upper outer fence = $Q_U + 3(\text{IQR}) = 85 + 3(25) = 160$.

With only this information, the box plot would look something like the following:

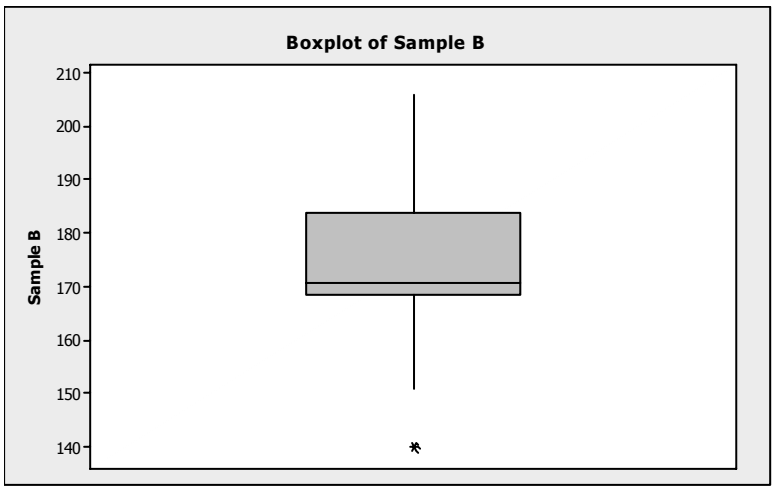


The whiskers extend to the inner fences unless no data points are that small or that large. The upper inner fence is 122.5. However, the largest data point is 100, so the whisker stops at 100. The lower inner fence is 22.5. The smallest data point is 18, so the whisker extends to 22.5. Since 18 is between the inner and outer fences, it is designated with a *. We do not know if there is any more than one data point below 22.5, so we cannot be sure that the box plot is entirely correct.

2.132 a. Using Minitab, the box plot for sample A is given below.



Using Minitab, the box plot for sample B is given below.



b. In sample A, the measurements 84 and 100 are outliers. These measurements fall outside the outer fences.

$$\begin{aligned}\text{Lower outer fence} &= \text{Lower hinge} - 3(\text{IQR}) \\ &\approx 158 - 3(172 - 158) \\ &= 158 - 3(14) \\ &= 116\end{aligned}$$

In addition, 122 and 196 may be outliers. They lie outside the inner fences. In sample B, 140.4 and 206.4 may be outliers. They lie outside the inner fences.

2.134 a. The z-score is $z = \dfrac{x - \bar{x}}{s} = \dfrac{175 - 79}{23} = 4.17$.

b. Yes, we would consider this measurement an outlier. Any observation with a z-score that has an absolute value greater than 3 is considered a highly suspect outlier.

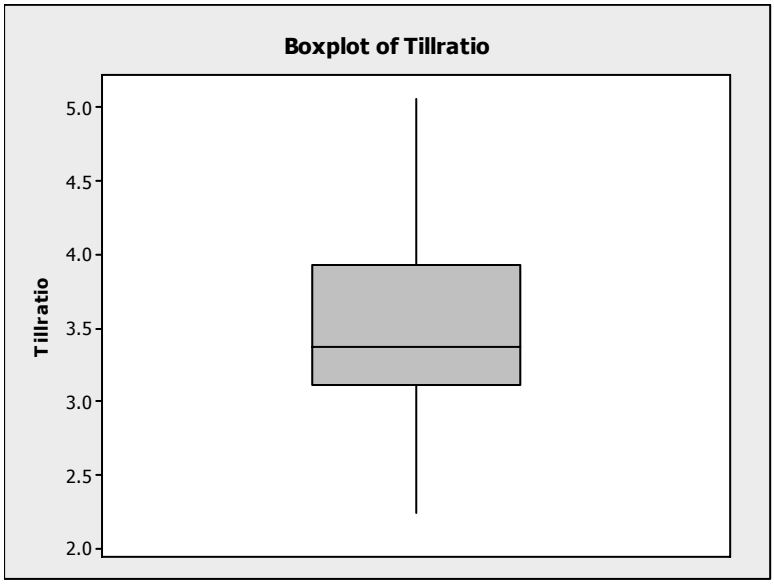
2.136 a. The z-score associated with the largest ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{5.06 - 3.5069}{.63439} = 2.45$

The z-score associated with the smallest ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{2.25 - 3.5069}{.63439} = -1.98$

The z-score associated with the mean ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{3.5069 - 3.5069}{.63439} = 0$

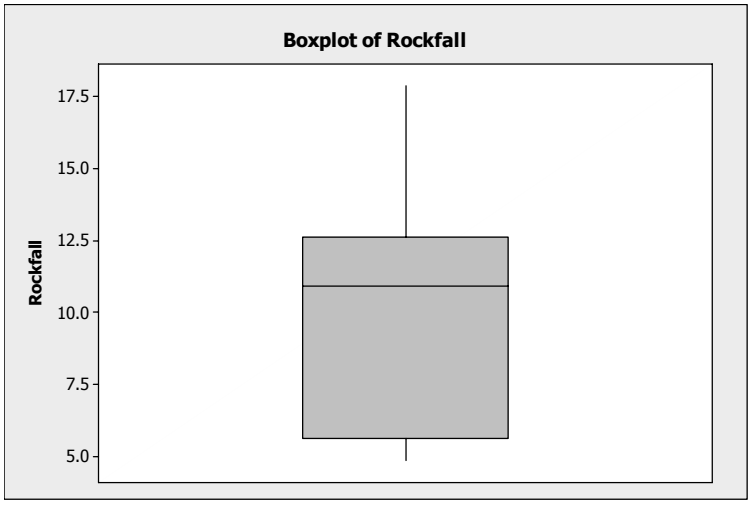
b. Yes, I would consider the z-score associated with the largest ratio to be unusually large. We know if the data are approximately mound-shaped that approximately 95% of the observations will be within 2 standard deviations of the mean. A z-score of 2.45 would indicate that less than 2.5% of all the measurements will be larger than this value.

c. Using MINITAB, the box plot is:



From this box plot, there are no observations marked as outliers.

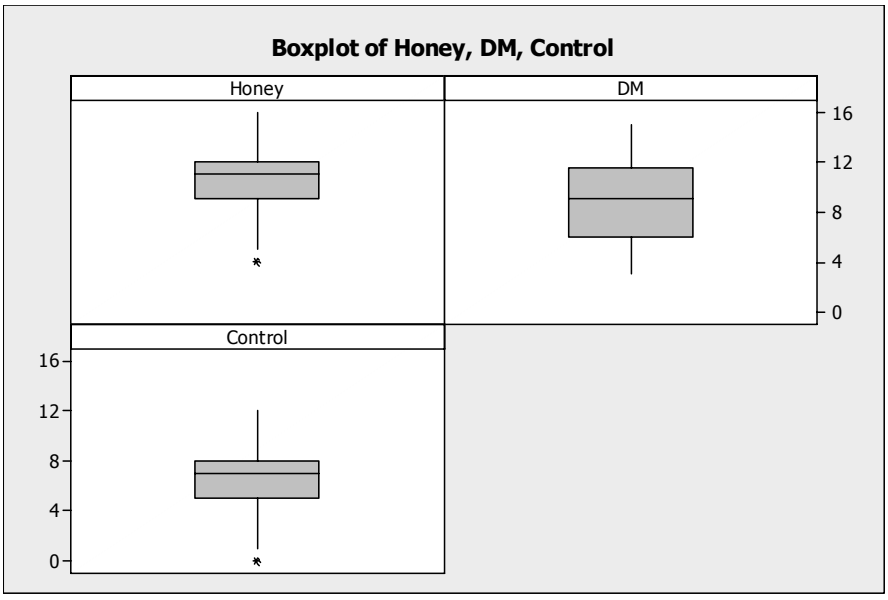
2.138 Using MINITAB, a boxplot of the data is:



From the boxplot, there is no indication that there are any outliers.

We will now use the z-score criterion for determining outliers. From Exercises 2.59 and 2.86, $\bar{x} = 9.72$ and $s = 4.095$. The z-score associated with the minimum value is $z = \dfrac{x - \bar{x}}{s} = \dfrac{4.9 - 9.72}{4.095} = -1.18$ and the z-score associated with the maximum value is $z = \dfrac{x - \bar{x}}{s} = \dfrac{17.83 - 9.72}{4.095} = 1.98$. Neither of these indicates there are any outliers.

2.140 a. Using MINITAB, the boxplots of the three groups are:



b. The median improvement score for the honey dosage group is larger than the median improvement scores for the other two groups. The median improvement score for the DM dosage group is higher than the median improvement score for the control group.

c. Because the interquartile range for the DM dosage group is larger than the interquartile ranges of the other 2 groups, the variability of the DM group is largest. The variability of the honey dosage group and the control group appear to be about the same.

d. There appears to be one outlier in the honey dosage group and one outlier in the control group.

2.142 a. $z = \dfrac{x-\mu}{\sigma} = \dfrac{4-7}{1} = -3$

b. The z-score is low enough to suspect that the librarian's claim is incorrect. Even without any knowledge of the shape of the distribution, Chebyshev's rule states that at least 8/9 of the measurements will fall within 3 standard deviations of the mean (and, consequently, at most 1/9 will be above $z = 3$ or below $z = -3$).

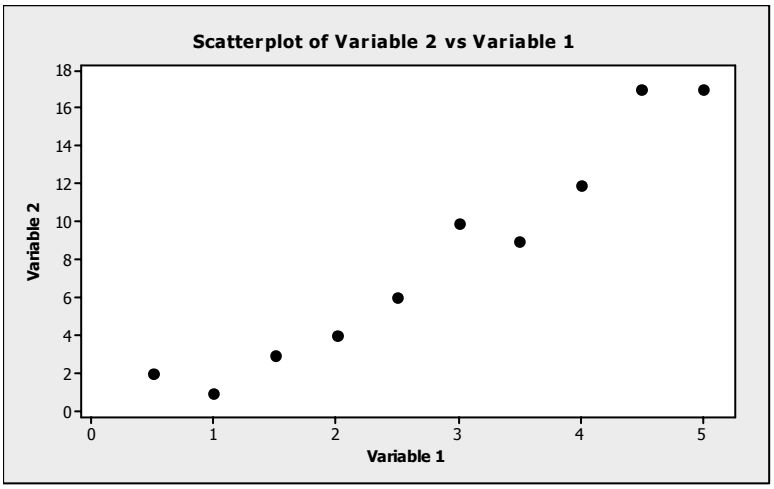
c. The Empirical Rule states that almost none of the measurements should be above $z = 3$ or below $z = -3$. Hence, the librarian's claim is even more unlikely.

d. When $\sigma = 2$, $z = \dfrac{x-\mu}{\sigma} = \dfrac{4-7}{2} = -1.5$

This is not an unlikely occurrence, whether or not the data are mound-shaped. Hence, we would not have reason to doubt the librarian's claim.

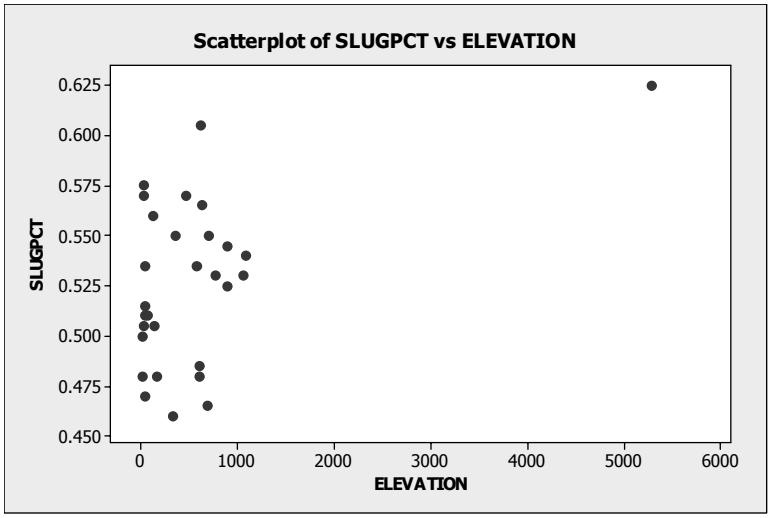
2.144 Scatterplots are useful with quantitative variables.

2.146 Using MINITAB, the scatterplot is as follows:



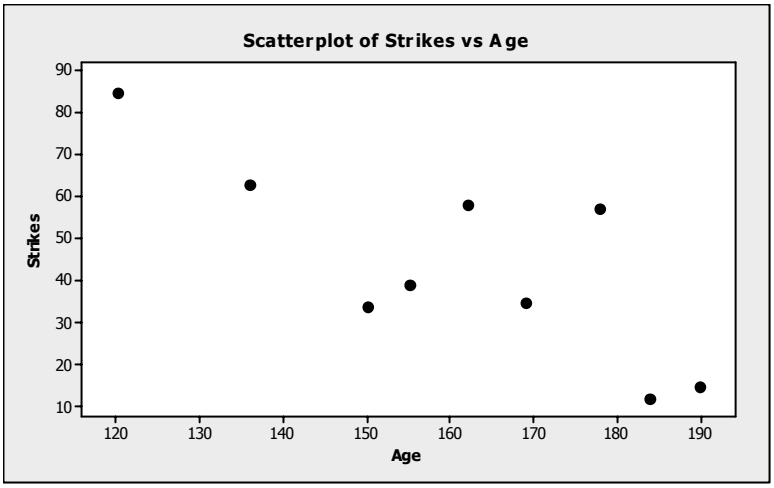
It appears that as variable 1 increases, variable 2 also increases.

2.148 Using MINITAB, a scatter plot of the data is:



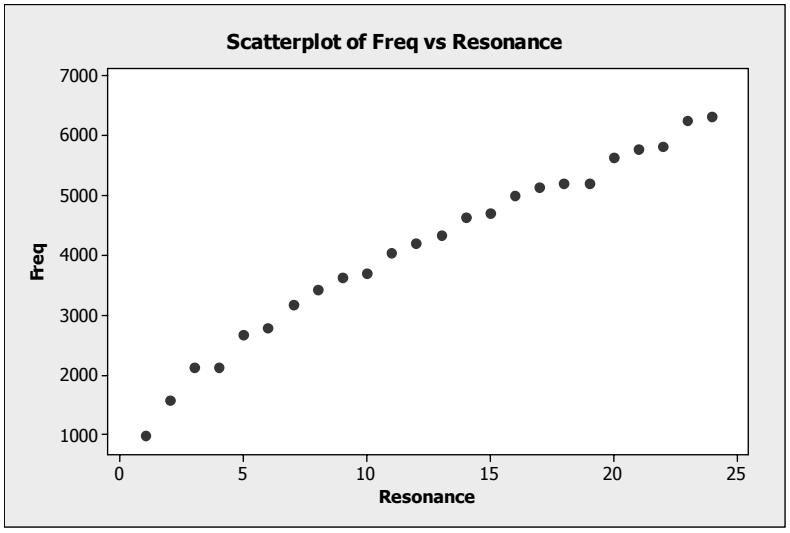
If one uses the one obvious outlier (Denver), then there does appear to be a trend in the data. As the elevation increases, the slugging percentage tends to increase. However, if the outlier is removed, then it does not look like there is a trend to the data.

2.150 a. A scattergram of the data is:



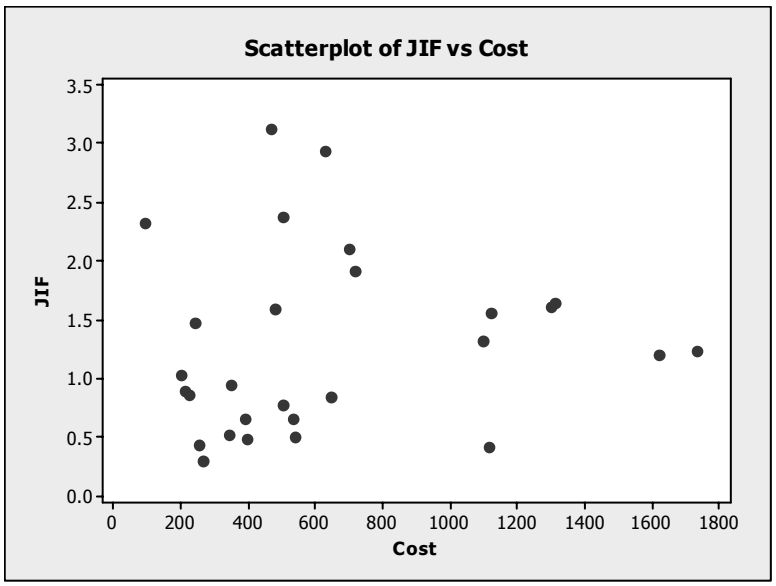
b. There appears to be a trend. As the age increases, the number of strikes tends to decrease.

2.152 Using MINITAB, a scatterplot of the data is:



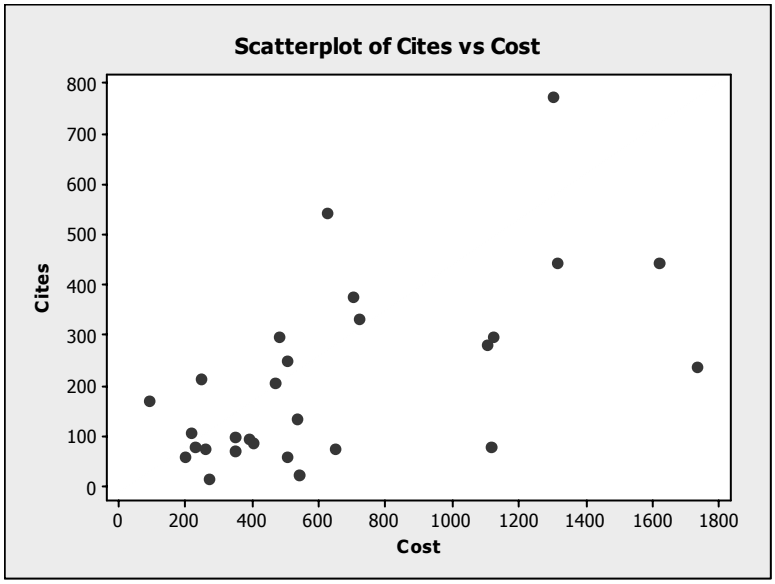
There is an increasing trend and there is very little variation in the plot. This supports the researcher's theory.

2.154 a. Using MINITAB, the scatterplot of JIF and cost is:



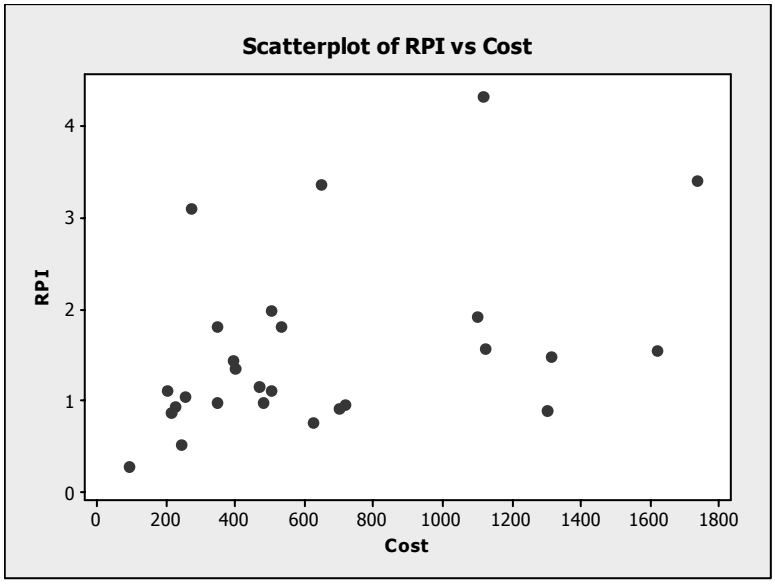
There does not appear to be much of a trend between these two variables.

b. Using MINITAB, the scatterplot of cites and cost is:



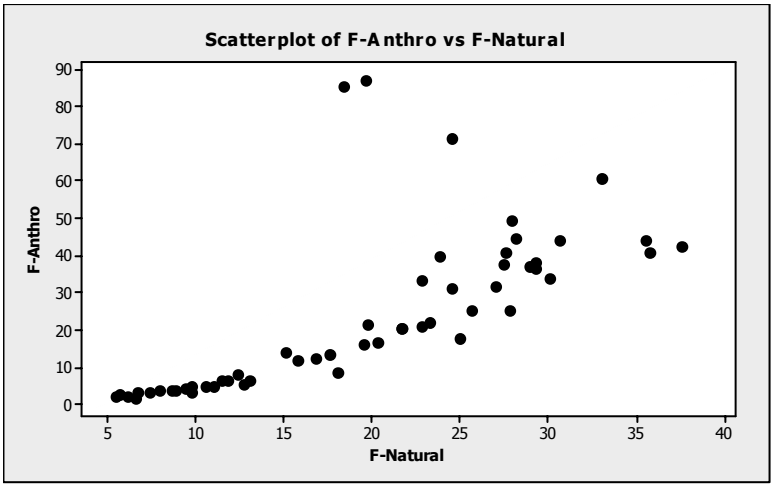
There appears to be a positive linear trend between cites and cost.

c. Using MINITAB, the scatterplot of RPI and cost is:



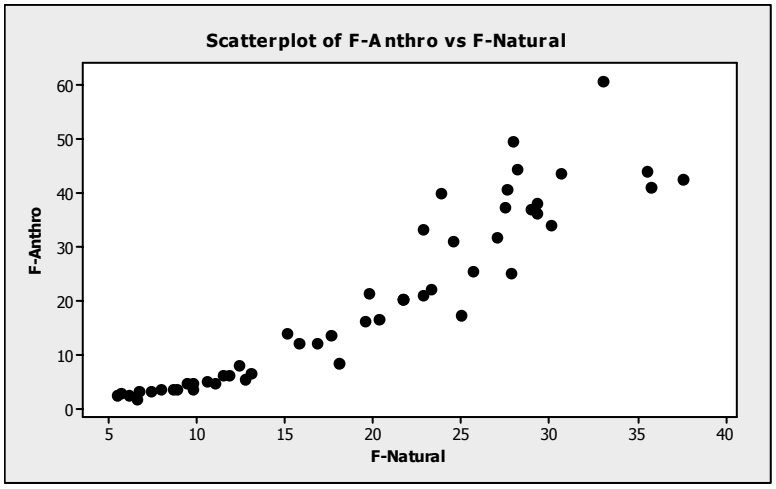
There appears to be a positive linear trend between RPI and cost.

2.156 a. Using MINITAB, a graph of the Anthropogenic Index against the Natural Origin Index is:



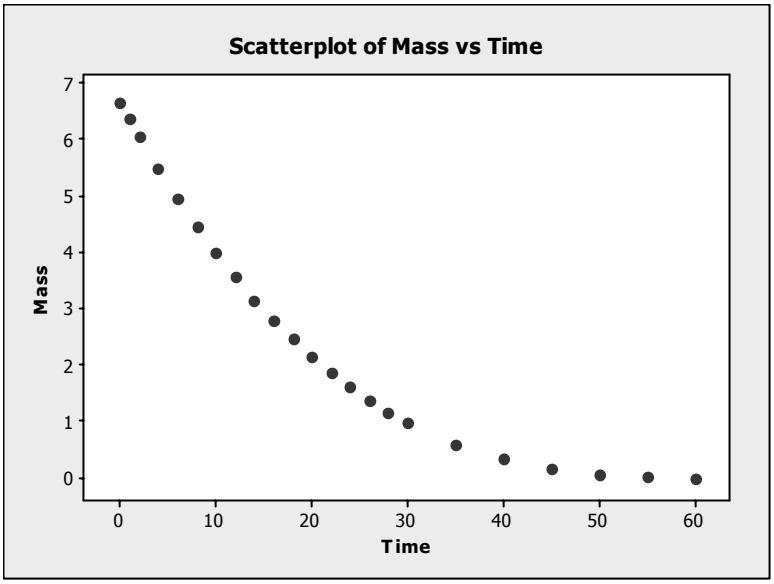
This graph does not support the theory that there is a straight-line relationship between the Anthropogenic Index against the Natural Origin Index. There are several points that do not lie on a straight line.

b. After deleting the three forests with the largest anthropogenic indices, the graph of the data is:



After deleting the 3 data points, the relationship between the Anthropogenic Index against the Natural Origin Index is much closer to a straight line.

2.158 Using MINITAB, a scattergram of the data is:

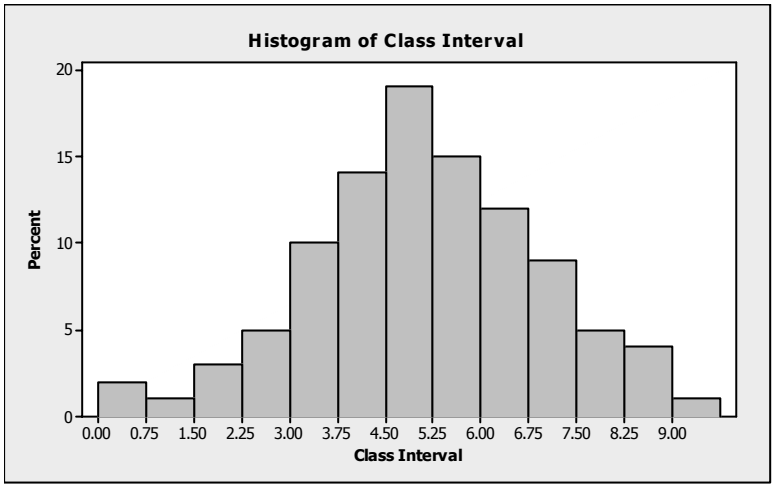


Yes, there appears to be a negative trend in this data. As time increases, the mass tends to decrease. There appears to be a curvilinear relationship. As time increases, mass decreases at a decreasing rate.

2.160 The range can be greatly affected by extreme measures, while the standard deviation is not as affected.

2.162 The z-score approach for detecting outliers is based on the distribution being fairly mound-shaped. If the data are not mound-shaped, then the box plot would be preferred over the z-score method for detecting outliers.

2.164 The relative frequency histogram is:



2.158 From part a of Exercise 2.165, the 3 z-scores are −1, 1 and 2. Since none of these z-scores are greater than 2 in absolute value, none of them are outliers.

From part b of Exercise 2.165, the 3 z-scores are −2, 2 and 4. There is only one z-score greater than 2 in absolute value. The score of 80 (associated with the z-score of 4) would be an outlier. Very few observations are as far away from the mean as 4 standard deviations.

From part c of Exercise 2.165, the 3 z-scores are 1, 3, and 4. Two of these z-scores are greater than 2 in absolute value. The scores associated with the two z-scores 3 and 4 (70 and 80) would be considered outliers.

From part d of Exercise 2.165, the 3 z-scores are .1, .3, and .4. Since none of these z-scores are greater than 2 in absolute value, none of them are outliers.

2.168 $\sigma \approx \text{range}/4 = 20/4 = 5$

2.170 a. $\sum x = 13+1+10+3+3 = 30$

$\sum x^2 = 13^2+1^2+10^2+3^2+3^2 = 288$

$\bar{x} = \sum x = \dfrac{30}{5} = 6$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{288 - \frac{30^2}{5}}{5-1} = \frac{108}{4} = 27 \qquad s = \sqrt{27} = 5.20$$

b. $\sum x = 13+6+6+0 = 25$

$\sum x^2 = 13^2+6^2+6^2+0^2 = 241$

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{25}{4} = 6.25$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{241 - \frac{25^2}{4}}{4-1} = \frac{84.75}{3} = 28.25 \qquad s = \sqrt{28.25} = 5.32$$

c. $\sum x = 1+0+1+10+11+11+15 = 49$

$\sum x^2 = 1^2+0^2+1^2+10^2+11^2+11^2+15^2 = 569$

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{49}{7} = 7$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{569 - \frac{49^2}{7}}{7-1} = \frac{226}{6} = 37.67 \qquad s = \sqrt{37.67} = 6.14$$

d. $\sum x = 3+3+3+3 = 12$

$\sum x^2 = 3^2+3^2+3^2+3^2 = 36$

$$\bar{x} = \frac{\sum x}{n} = \frac{12}{4} = 3$$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{36 - \frac{12^2}{4}}{4-1} = \frac{0}{3} = 0 \qquad s = \sqrt{0} = 0$$

e. a) $\bar{x} \pm 2s \Rightarrow 6 \pm 2(5.2) \Rightarrow 6 \pm 10.4 \Rightarrow (-4.4, 16.4)$

All or 100% of the observations are in this interval.

b) $\bar{x} \pm 2s \Rightarrow 6.25 \pm 2(5.32) \Rightarrow 6.25 \pm 10.64 \Rightarrow (-4.39, 16.89)$

All or 100% of the observations are in this interval.

c) $\bar{x} \pm 2s \Rightarrow 7 \pm 2(6.14) \Rightarrow 7 \pm 12.28 \Rightarrow (-5.28, 19.28)$

All or 100% of the observations are in this interval.

d) $\bar{x} \pm 2s \Rightarrow 3 \pm 2(0) \Rightarrow 3 \pm 0 \Rightarrow (3, 3)$

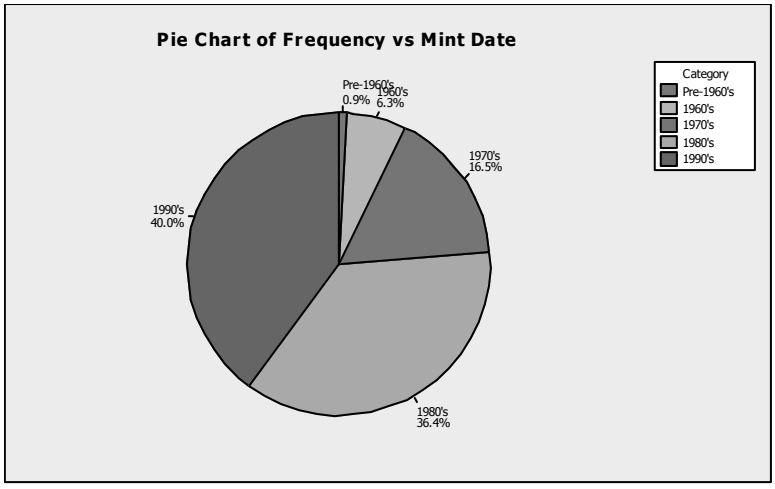
All or 100% of the observations are in this interval.

2.172 a. The experimental unit of interest is a penny.

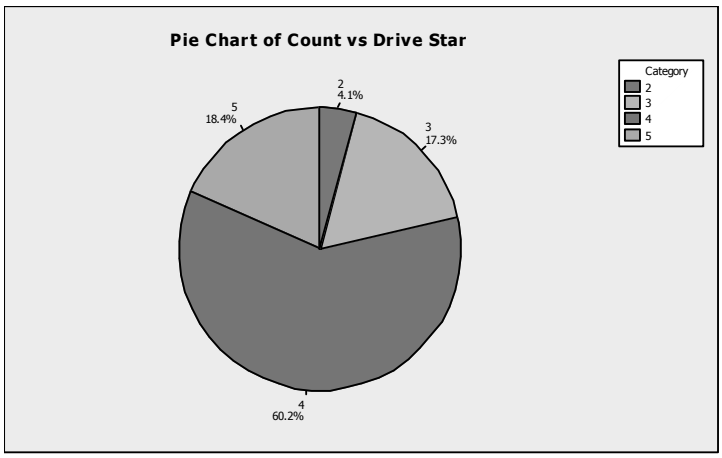
b. The variable measured is the mint date on the penny.

c. The number of pennies that have mint dates in the 1960's is 125. The proportion is found by dividing the number of pennies with mint dates in the 1960's (125) by the total number of pennies (2000). The proportion is 125/2,000 = .0625.

d. Using MINITAB, a pie chart of the data is:



2.174 A pie chart of the data is:



More than half of the cars received 4 star ratings (60.2%). A little less than a quarter of the cars tested received ratings of 3 stars or less.

2.176 a. The mean of the data is $\bar{x} = \dfrac{\sum x}{n} = \dfrac{0+0+0+0+0+1+1+\cdots+5}{14} = \dfrac{20}{14} = 1.429$

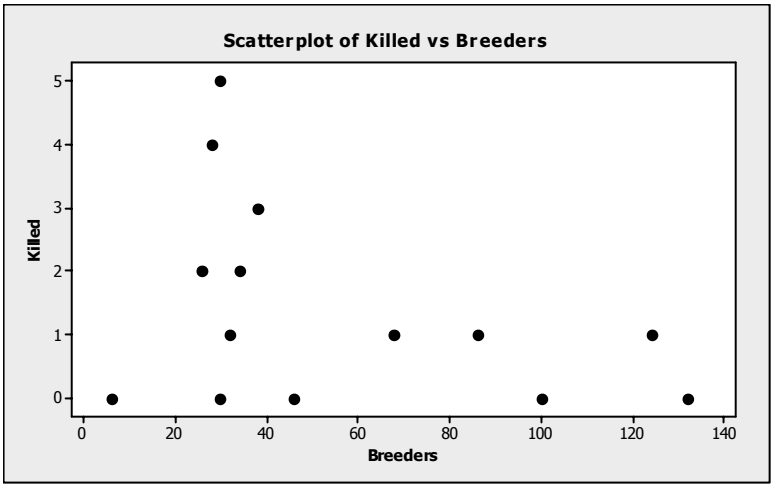
The median is the average of the middle two numbers once the data are arranged in order. The data arranged in order are: 0 0 0 0 0 1 1 1 1 2 2 3 4 5

The middle two numbers are 1 and 1. The median is $\dfrac{1+1}{2} = 1$

The mode is the number occurring the most frequently. In this data set, the mode is 0 because it appears five times, more than any other.

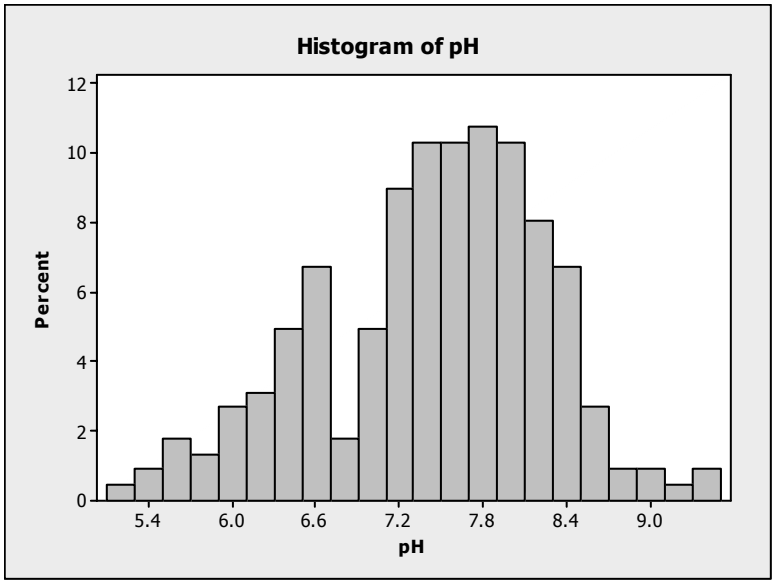
b. The average number of flycatchers killed is 1.429. The median number of flycatchers killed is 1. This means that 50% of the flycatchers killed is less than or equal to 1. The most frequent number of flycatchers killed is 0. Because the mode is the smallest value of the three, the median is the next smallest, and the mean is the largest, the data are skewed to the right. Because the data are skewed, the median is probably a more representative measure for the middle of the data set. Only 5 of the 14 observations are larger than the mean.

c. Using MINITAB, the scatterplot of the data is:



There is a fairly weak negative relationship between the number killed and the number of breeders. As the number of breeders increase, the number of killed tends to decrease.

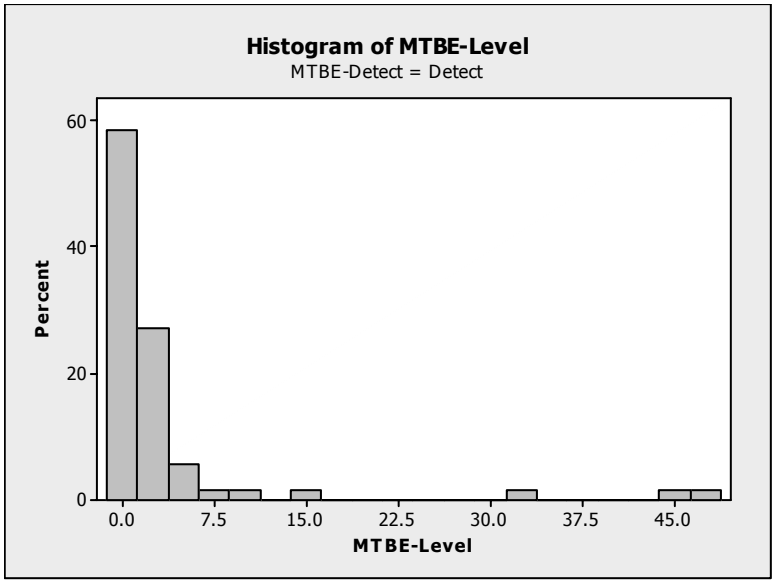
2.178 a. Using MINITAB, a histogram of the data is:



From the graph, it looks like the proportion of wells with ph levels less than 7.0 is:

.005 + .01 + .02 + .015 + .027 + .031 + .05 + .07 + .017 + .05 = .295

b. Using MINITAB, a histogram of the MTBE levels for those wells with detectible levels is:



From the graph, it looks like the proportion of wells with MTBE levels greater than 5 is:

.03 + .01 + .01 + .01 + .01 + .01 + .01 =.09

c. The sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{7.87 + 8.63 + 7.11 + \cdots + 6.33}{223} = \frac{1{,}656.16}{223} = 7.427$$

The variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{12{,}447.9812 - \frac{1{,}656.16^2}{223}}{223\text{-}1} = \frac{148.13391}{222} = .66727$$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{.66727} = .8169$

$\bar{x} \pm 2s \Rightarrow 7.427 \pm 2(.8169) \Rightarrow 7.427 \pm 1.6338 \Rightarrow (5.7932,\ \ 9.0608)$.

From the histogram in part a, the data look approximately mound-shaped. From the Empirical Rule, we would expect about 95% of the wells to fall in this range. In fact, 212 of 223 or 95.1% of the wells have pH levels between 5.7932 and 9.0608.

d. The sample mean of the wells with detectible levels of MTBE is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{.23 + .24 + .24 + \cdots + 48.10}{70} = \frac{240.86}{70} = 3.441$$

The variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{6112.266 - \frac{240.86^2}{70}}{70\text{-}1} = \frac{5283.5011}{69} = 76.5725$$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{76.5725} = 8.7506$

$\bar{x} \pm 2s \Rightarrow 3.441 \pm 2(8.7506) \Rightarrow 3.441 \pm 17.5012 \Rightarrow (-14.0602,\ \ 20.9422)$.

From the histogram in part b, the data do not look mound-shaped. From Chebyshev's Rule, we would expect at least ¾ or 75% of the wells to fall in this range. In fact, 67 of 70 or 95.7% of the wells have MTBE levels between -14.0602 and 20.9422.

2.180 a. If the distribution of scores was symmetric, the mean and median would be equal. The fact that the mean exceeds the median is an indication that the distribution of scores is skewed to the right.

b. It means that 90% of the scores are below 660, and 10% are above 660. (This ignores the possibility of ties, i.e., other people obtaining a score of 660.)

c. If you scored at the 94th percentile, 94% of the scores are below your score, while 6% exceed your score.

2.182 a. For site A, there is no real pattern to the data that would indicate that the data are skewed. For site G, most of the data are concentrated from 250 and up. There are relatively few observations less than 250. This indicates that the data are skewed to the left.

b. For site A, there are 2 modes (two distance intervals with the largest number of observations). Since there is no more than one mode, this would indicate that the data are probably from hearths inside dwellings.

For site G, there is only one mode. This would indicate that the data are probably from open air hearths.

2.184 Using MINITAB, the descriptive statistics are:

Descriptive Statistics: Ammonia

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Ammonia	8	1.4713	0.0640	1.3700	1.4125	1.4900	1.5250	1.5500

The stem-and-leaf display for the data is:

Stem-and-Leaf Display: Ammonia

```
Stem-and-leaf of Ammonia  N  = 8
Leaf Unit = 0.010


 1  13  7
 3  14  12
 4  14  8
 4  15  013
 1  15  5
```

Since the data look fairly mound-shaped, we will use the Empirical Rule. We know that approximately 99.7% of all observations will fall within 3 standard deviation of the mean. For this data, the interval 3 standard deviations below the mean to 3 standard deviations above the mean is:

$$\bar{x} \pm 3s \Rightarrow 1.471 \pm 3(.064) \Rightarrow 1.471 \pm 0.192 \Rightarrow (1.279, 1.663)$$

We would be fairly confident that the ammonia level of a randomly selected day will fall between 1.279 and 1.663 parts per million.

2.186 a. From the histogram, the data do not follow the true mound-shape very well. The intervals in the middle are much higher than they should be. In addition, there are some extremely large velocities and some extremely small velocities. Because the data do not follow a mound-shaped distribution, the Empirical Rule would not be appropriate.

b. Using Chebyshev's rule, at least $1 - 1/4^2$ or $1 - 1/16$ or 15/16 or 93.8% of the velocities will fall within 4 standard deviations of the mean. This interval is:

$$\bar{x} \pm 4s \Rightarrow 27{,}117 \pm 4(1{,}280) \Rightarrow 27{,}117 \pm 5{,}120 \Rightarrow (21{,}997,\ 32{,}237)$$

At least 93.75% of the velocities will fall between 21,997 and 32,237 km per second.

c. Since the data look approximately symmetric, the mean would be a good estimate for the velocity of galaxy cluster A2142. Thus, this estimate would be 27,117 km per second.

2.188 a. The first variable is gender. It has only two values which are not numerical, so it is qualitative. The next variable is group. There are three groups which are not numerical, so group is qualitative. The next variable is DIQ. This variable is measured on a numerical scale, so it is quantitative. The last variable is percent of pronoun errors. This variable is measured on the numerical scale, so it is quantitative.

b. In order to compute numerical descriptive measures, the data must have numbers associated with them. Qualitative variables do not have meaningful numbers associated with them, so one cannot compute numerical measures.

c. The mean of the DIQ scores for the SLI children is:

$$\bar{x} = \frac{\sum x}{n} = \frac{86 + 86 + 94 + ... + 95}{10} = \frac{936}{10} = 93.6$$

The median is the average of the middle two numbers after they have been arranged in order: 84, 86, 86, 87, 89, 94, 95, 98, 107, 110.

The median is $\frac{89 + 94}{2} = \frac{183}{2} = 91.5$

The mode is the value with the highest frequency. Since 86 occurred twice and no other value occurred more than once, the mode is 86.

d. The mean of the DIQ scores for the YND children is:

$$\bar{x} = \frac{\sum x}{n} = \frac{110 + 92 + 92 + ... + 92}{10} = \frac{953}{10} = 95.3$$

The median is the average of the middle two numbers after they have been arranged in order: 86, 90, 90, 92, 92, 92, 96, 100, 105, 110.

The median is $\frac{92 + 92}{2} = \frac{184}{2} = 92$

The mode is the value with the highest frequency. Since 92 occurred three times and no other value occurred more than twice, the mode is 92.

e. The mean of the DIQ scores for the OND children is:

$$\bar{x} = \frac{\sum x}{n} = \frac{110 + 113 + 113 + ... + 98}{10} = \frac{1019}{10} = 101.9$$

The median is the average of the middle two numbers after they have been arranged in order: 87, 92, 94, 95, 98, 108, 109, 110, 113, 113.

The median is $\frac{98 + 108}{2} = \frac{206}{2} = 103$

The mode is the value with the highest frequency. Since 113 occurred twice and no other value occurred more than once, the mode is 113.

f. Of the three groups, the SLI group had the lowest mean DIQ score (93.6), the YND group had a slightly higher mean DIQ score (95.3), while the OND group had the highest mean DIQ score (101.9). Thus, the SLI and the YND groups appear to be fairly similar with regard to DIQ, while the OND group appears to be much higher.

Of the three groups, the SLI group had the lowest median DIQ score (91.5), the YND group had a slightly higher median DIQ score (92), while the OND group had the highest median DIQ score (103). Thus, again, the SLI and the YND groups appear to be fairly similar with regard to DIQ, while the OND group appears to be much higher.

Of the three groups, the SLI group had the lowest mode DIQ score (86), the YND group had a slightly higher mode DIQ score (92), while the OND group had the highest mode IDQ score (113). Thus, again, the SLI and the YND groups appear to be fairly similar with regard to DIQ, while the OND group appears to be much higher.

Since the "centers" for the SLI and YND children are very similar, it appears that one could compute one set of "centers' for these two groups. However, the "centers" for the OND children appear to be much larger than those of the other two groups. One would have to compute a different set of "centers" for this group of children.

g. **YND children**: The mean percentage of pronoun errors is:

$$\bar{x} = \frac{\sum x}{n} = \frac{94.4 + 19.05 + 62.5 + ... + 0}{10} = \frac{468.8}{10} = 46.88$$

The median is the average of the middle 2 numbers once the data have been arranged in order: 0, 0, 18.75, 19.05, 32.43, 55.00, 62.50, 86.67, 94.40, 100.00.

The median is $\frac{32.43 + 55.00}{2} = \frac{87.43}{2} = 43.715$

The mode is the value with the highest frequency. Since 0 occurs 2 times, 0 is the mode.

SLI children: The mean percentage of pronoun errors is:

$$\bar{x} = \frac{\sum x}{n} = \frac{60 + 40 + 31.58 + ... + 0}{10} = \frac{301.71}{10} = 30.171$$

The median is the average of the middle 2 numbers once the data have been arranged in order: 0, 0, 0, 27.27, 31.58, 33.33, 40.00, 42.86, 60.00, 66.67.

The median is $\frac{31.58 + 33.33}{2} = \frac{64.91}{2} = 32.455$

The mode is the value with the highest frequency. Since 0 occurs 3 times, 0 is the mode.

OND children: The mean percentage of pronoun errors is:

$$\bar{x} = \frac{\sum x}{n} = \frac{0 + 0 + 0 + ... + 0}{10} = \frac{0}{10} = 0$$

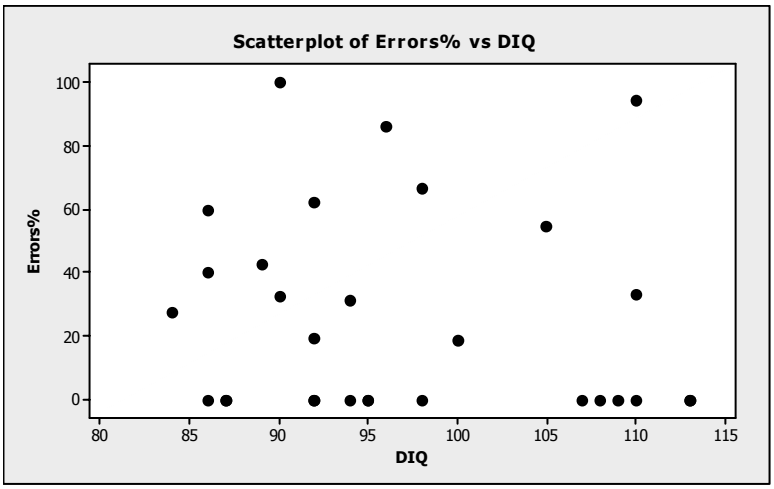
The median is the average of the middle 2 numbers once the data have been arranged in order: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.

The median is $\frac{0+0}{2}=\frac{0}{2}=0$.

The mode is the value with the highest frequency. Since 0 occurs 10 times, 0 is the mode.

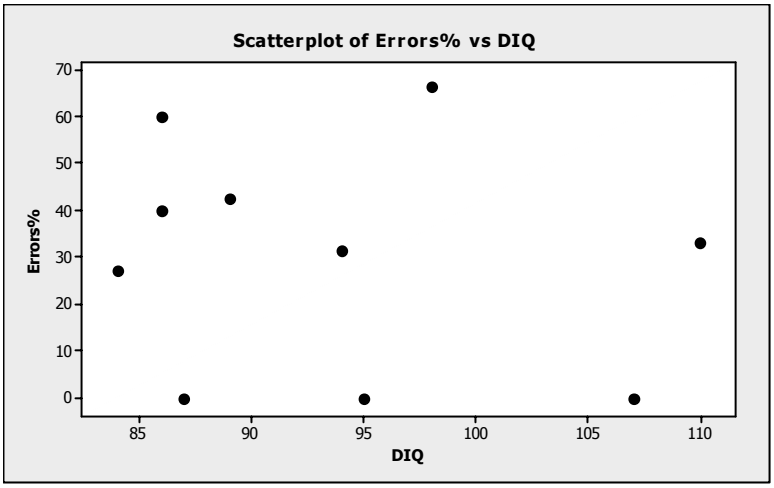
Since none of the means of the 3 groups are close in value and none of the medians are close in value, it appears that three "centers" should be calculated.

h. A scattergram of the data is:



There does not appear to be much of a relationship between deviation intelligence quotient (DIQ) and the percent of pronoun errors. The points are scattered randomly.

i. A plot of the data for the SLI children only is:

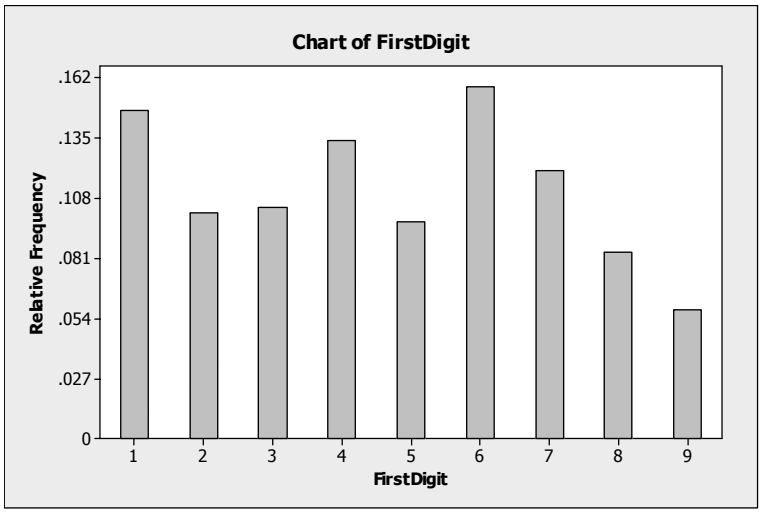


Again, there does not appear to be much of a trend between the DIQ scores and the proper use of pronouns. The data points are randomly scattered.

2.190 The relative frequency for each cell is found by dividing the frequency by the total sample size, $n = 743$. The relative frequency for the digit 1 is 109/743 = .147. The rest of the relative frequencies are found in the same manner and are shown in the table.

First Digit	Frequency	Relative Frequency
1	109	0.147
2	75	0.101
3	77	0.104
4	99	0.133
5	72	0.097
6	117	0.157
7	89	0.120
8	62	0.083
9	43	0.058
Total	743	1.000

Using MINITAB, the relative frequency bar chart is:



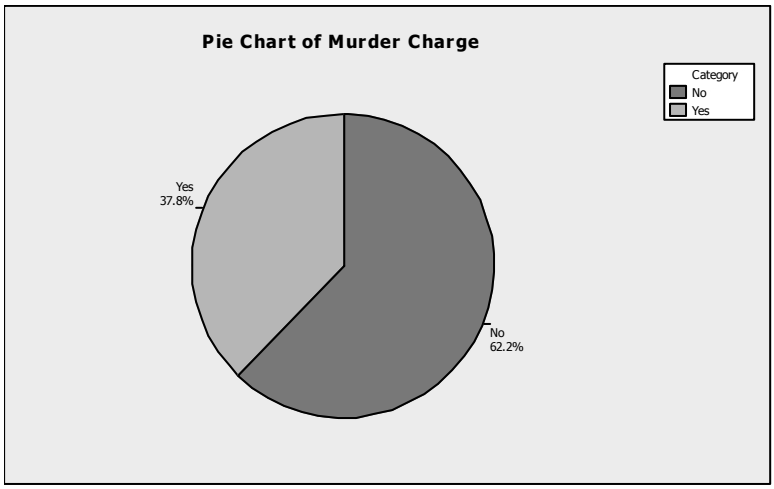
Benford's Law indicates that certain digits are more likely to occur as the first significant digit in a randomly selected number than other digits. The law also predicts that the number "1" is the most likely to occur as the first digit (30% of the time). From the relative frequency bar chart, one might be able to argue that the digits do not occur with the same frequency (the relative frequencies appear to be slightly different). However, the histogram does not support the claim that the digit "1" occurs as the first digit about 30% of the time. In this sample, the number "1" only occurs 14.7% of the time, which is less than half the expected 30% using Benford's Law.

2.192 If the distributions of the standardized tests are approximately mound-shaped, then it would be impossible for 90% of the school districts' students to score above the mean. If the distributions are mound-shaped, then the mean and median are approximately the same. By definition, only 50% of the students would score above the median.

If the distributions are not mound-shaped, but skewed to the left, it would be possible for more than 50% of the students to score above the mean. However, it would be almost impossible for 90% of the students scored above the mean.

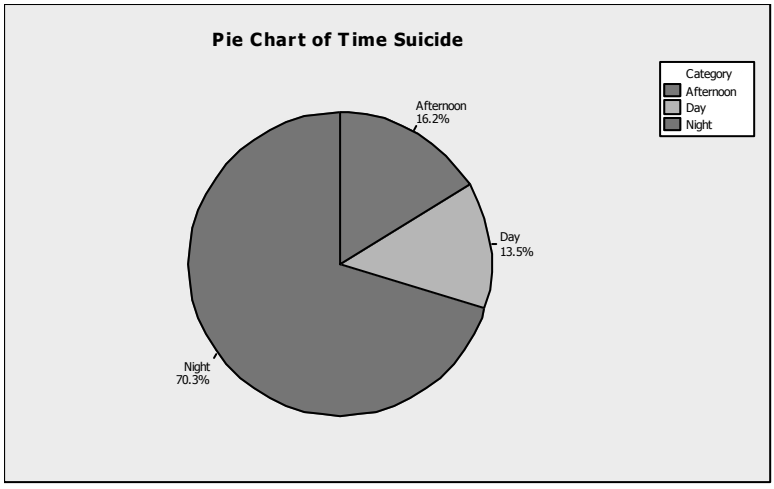
2.194 a. The variable "Days in Jail Before Suicide" is measured on a numerical scale, so it is quantitative. The variables "Marital Status", "Race", "Murder/Manslaughter Charge", and "Time of Suicide" are not measured on a numerical scale, so they are all qualitative. The variable "Year" is measured on a numerical scale, so it is quantitative.

b. Using MINITAB, the pie chart for the data is:



Suicides are more likely to be committed by inmates charged with lesser crimes than by inmates charged with murder/manslaughter. Of the suicides reported, 62.2% are committed by those convicted of a lesser charge.

c. Using MINITAB, the pie chart for the data is:



Suicides are much more likely to be committed at night than any other time. Of the suicides reported, 70.3% were committed at night.

d. Using MINITAB, the descriptive statistics are:

Descriptive Statistics: JAILDAYS

```
Variable   N  Mean  StDev  Minimum    Q1  Median    Q3  Maximum
JAILDAYS  37  41.4   66.7     1.00  4.00    15.0  41.5    309.0
```

The mean length of time an inmate spent in jail before committing suicide is 41.4 days. The median length of time an inmate spent in jail before committing suicide is 15 days. Since the mean is much larger than the median, the data are skewed to the right. Most of those committing suicide, commit it within 15 days of arriving in jail. However, there are a few inmates who spend many more days in jail before committing suicide.

e. First, compute the z-score associated with 200 days:

$$z = \frac{x - \bar{x}}{s} = \frac{200 - 41.4}{66.7} = 2.38$$

Using Chebyshev's rule, we know that at most $1/k^2$ of the observations will fall more than k standard deviations from the mean. For a value of 200, k = 2.38. Thus, at most $1/2.38^2 = .177$ of the observations will fall more than 2.38 standard deviations from the mean. It looks like it would not be that unusual to see someone commit suicide after 200 days since the proportion of times this could happen is at most .177. However, if we look at the data, of the 37 observations, there are only 2 observations of 200 or larger. This proportion is 2/37 = .054. Using this information, it would be rather unusual for an inmate to commit suicide after 200 days.

f. Using MINITAB, the stem-and-leaf plot of the data is:

```
Stem-and-leaf of Year      N = 37
Leaf Unit = 1.0

 1   196  7
 5   196  8889
11   197  000001
13   197  23
16   197  455
18   197  66
(2)  197  99
17   198  000111
11   198  233
 8   198  55555
 3   198  77
 1   198  9
```

From the stem-and-leaf plot, it does not appear that the number of suicides have decreased over time.

2.196 For the first professor, we would assume that most of the grade-points will fall within 3 standard deviations of the mean. This interval would be:

$$\bar{x} \pm 3s \Rightarrow 3.0 \pm 3(.2) \Rightarrow 3.0 \pm .6 \Rightarrow (2.4, 3.6)$$

Thus, if you had the first professor, you would be pretty sure that your grade-point would be between 2.4 and 3.6.

For the second professor, we would again assume that most of the grade-points will fall within 3 standard deviations of the mean. This interval would be:

$$\bar{x} \pm 3s \Rightarrow 3.0 \pm 3(1) \Rightarrow 3.0 \pm 3.0 \Rightarrow (0.0, 6.0)$$

Thus, if you had the second professor, you would be pretty sure that your grade-point would be between 0.0 and 6.0. If we assume that the highest grade-point one could receive is 4.0, then this interval would be (0.0, 4.0). We have gained no information by using this interval, since we know that all grade-points are between 0.0 and 4.0. However, since the standard deviation is so large, compared to the mean, we could infer that the distribution of grade-points in this class is not symmetric, but skewed to the left. There are many high grades, but there are several very low grades.

By taking the first professor, you know you are almost positive that you will get a final grade of at least 2.4, but almost no chance of getting a final grade of 4. By taking the second professor, you know the grades are skewed to the left and that many of the students will get high grades, but also a few will get very low grades.

2.198 The answers to this will vary. Some things that should be included in the discussion are:

From the graph, it is obvious that the amount of money spent on education has increased tremendously over the period from 1966 to 2000 (from about $4.5 billion in 1966 to about $22.5 billion in 2000). However, one should note that the number of students has also increased. It might be better to reflect the amount of money spent as the amount of money spent per student over the years from 1966 to 2000 rather than the total amount spent.

In the description of the exercise, it says that the horizontal line represents the annual average fourth-grade children's reading ability. It also indicates that the fourth-grade reading test scores are designed to have an average of 250 with a standard deviation of 50. Thus, regardless of whether the children's reading abilities increase or decrease, the annual average will always be 250. This line does not give any information about whether the children's reading abilities are improving or not.

In addition, if the reading scores of seventh and twelfth graders and the mathematics scores of fourth graders improved over the same time period, one could conclude that the reading scores of the fourth graders also improved over the same time period.

Thus, this graph does not support the government's position that our children are not making classroom improvements despite federal spending on education. This graph only portrays that the total amount of money spent on education over the time period from 1966 to 2000 increased.