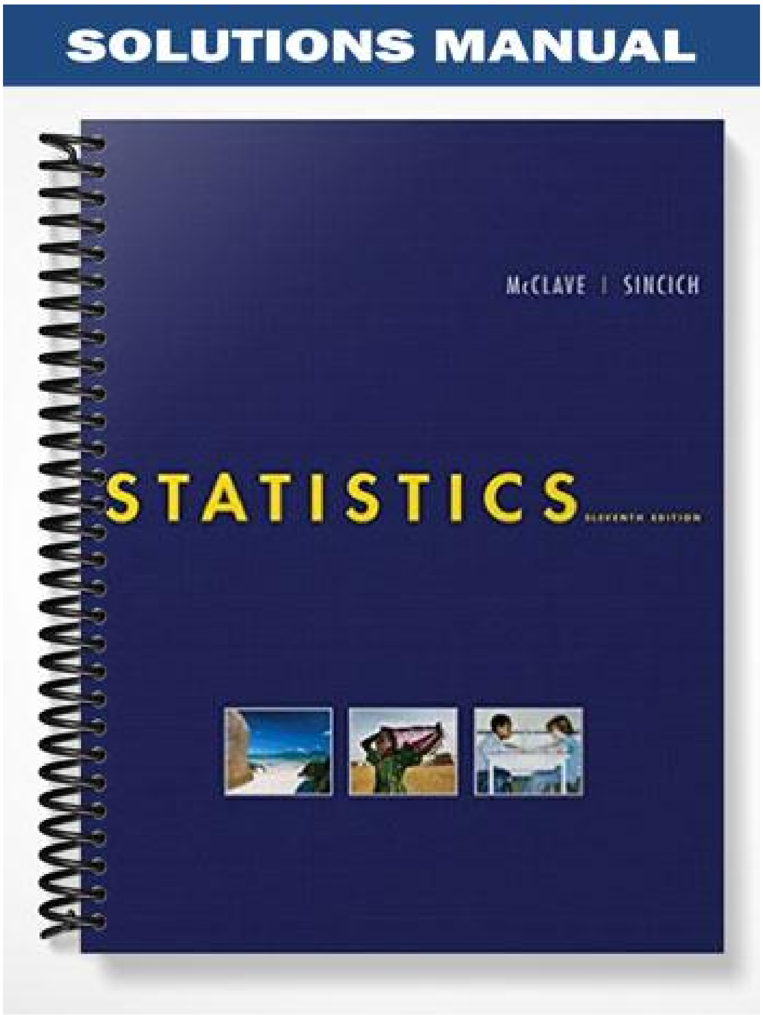


SOLUTIONS MANUAL

McCLAVE | SINCICH

STATISTICS

ELEVENTH EDITION

Chapter

2

Methods for Describing Sets of Data

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:

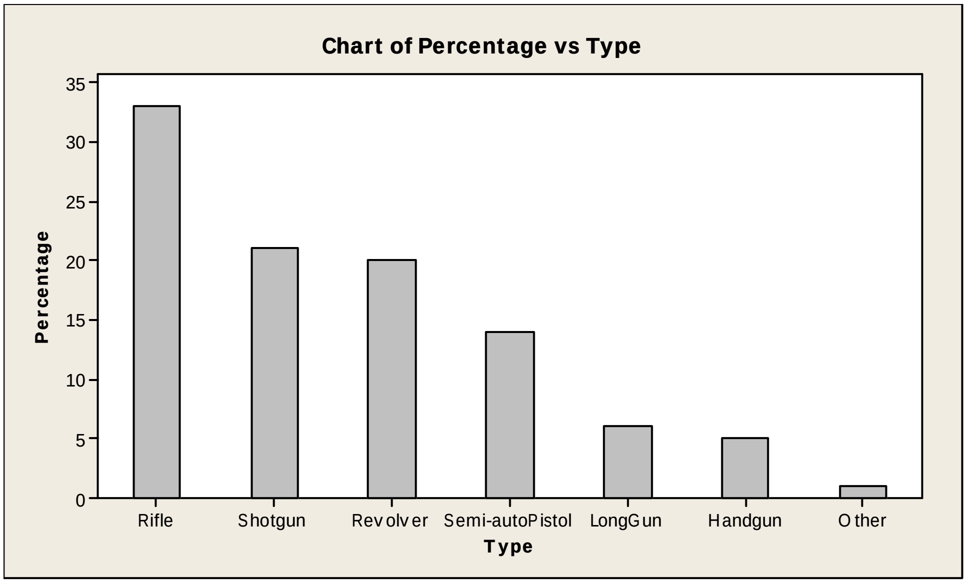
Grade on Statistics Exam	**Frequency**	**Relative Frequency**
A: 90–100	16	.08
B: 80– 89	36	.18
C: 65– 79	90	.45
D: 50– 64	30	.15
F: Below 50	28	.14
Total	200	1.00

2.6 a. The graph shown is a pie chart.

b. The qualitative variable described in the graph is type of firearms owned by adults in the U.S. who own firearms.

c. The most common type of firearm owned is a rifle, with 33.0% of all owners of firearms owning a rifle.

d. The data represented as a Pareto diagram would be:



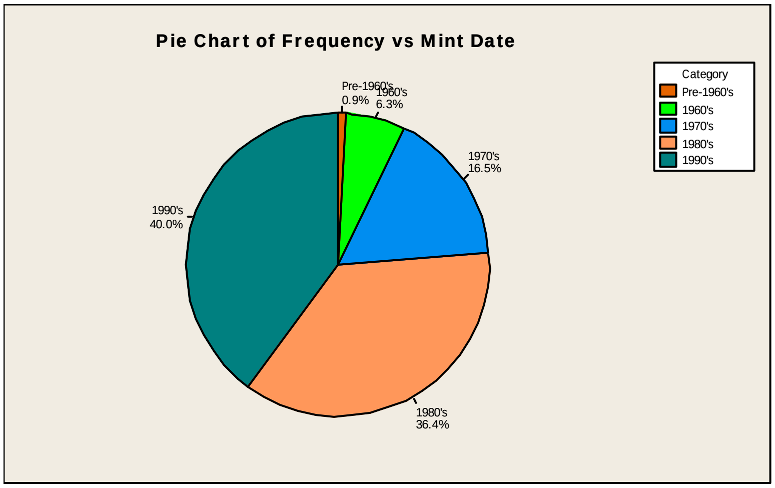
Of those who own firearms, approximately 1/3 own rifles. About 21% and 20% own shotguns and revolvers, respectively. About 14% own semi-automatic pistols while only 6% own long guns and 5% own handguns.

2.8 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.

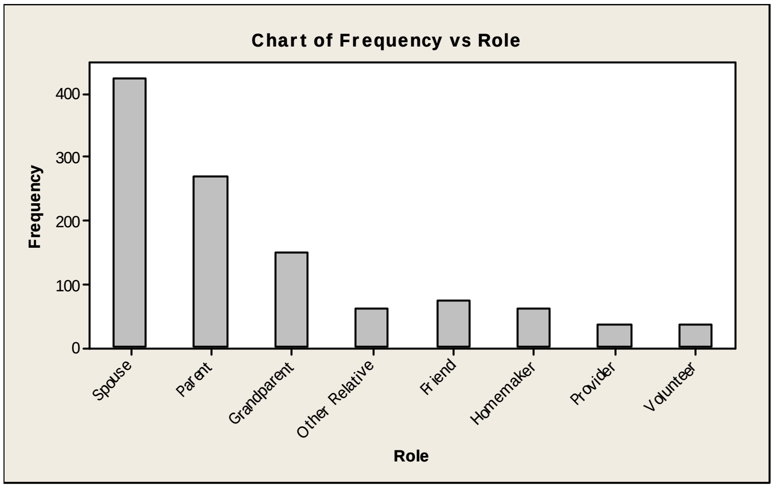
d. Using MINITAB, a pie chart of the data is:



2.10 a. The qualitative variable summarized in the table is the role elderly people feel is the most important later in life. There are 8 categories associated with this variable. They are: Spouse, Parent, Grandparent, Other relative, Friend, Homemaker, Provider, and Volunteer/Club/Church member.

b. The numbers in the table are frequencies because they are whole numbers. Relative frequencies are numbers between 0 and 1.

c. A bar graph of the data is:

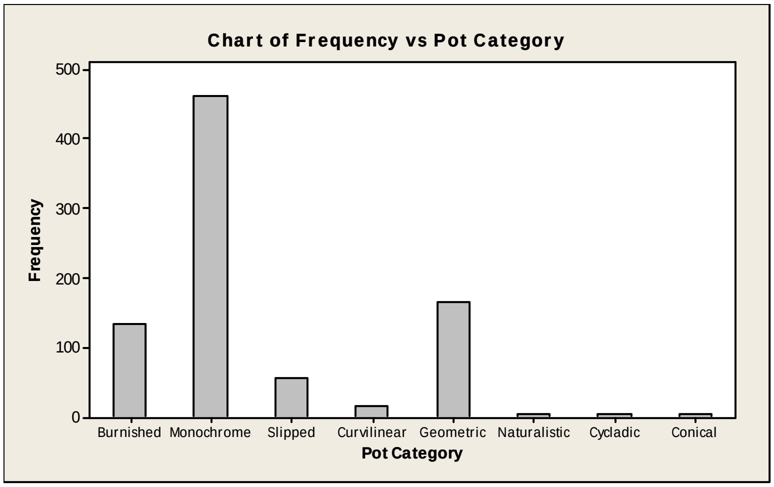


d. The role with the highest percentage of elderly adults is Spouse. The relative frequency is 424 / 1,102 = .385. Multiplying this by 100% gives a percentage of 38.5%. Of all the elderly adults surveyed, 38.5% view their most salient roles as that of spouse.

2.12 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.

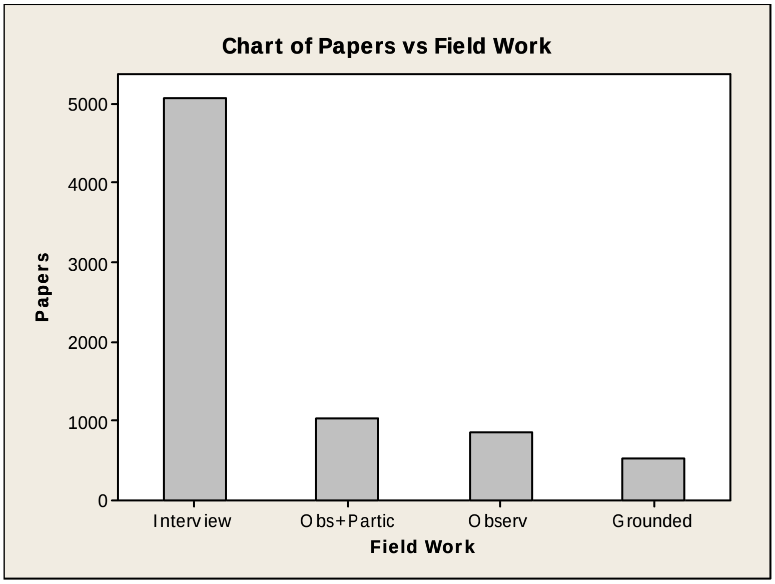
Pot Category	Number Found	Computation	Relative Frequency
Burnished	133	133 / 837	.159
Monochrome	460	460 / 837	.550
Slipped	55	55 / 837	.066
Curvilinear Decoration	14	14 / 837	.017
Geometric Decoration	165	165 / 837	.197
Naturalistic Decoration	4	4 / 837	.005
Cycladic White clay	4	4 / 837	.005
Cononical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

2.14 Using MINITAB, a bar graph is:

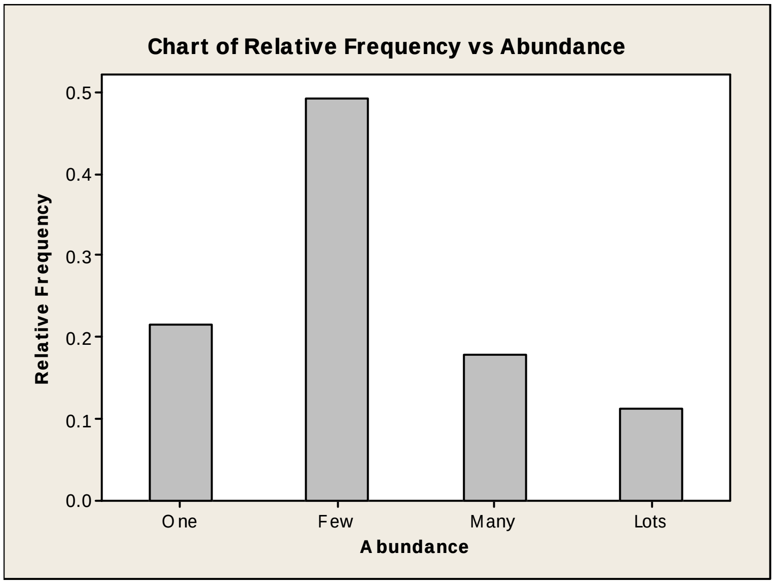


Most of the types of papers found were interviews. There were about twice as many interviews as all other types combined.

2.16 a. To find the relative frequency for the frog abundance category of 'one', one divides the number of recordings for 'one' category (33) by the total number of recordings (152) or 33 / 152 = .217. The rest of the relative frequencies appear in the table:

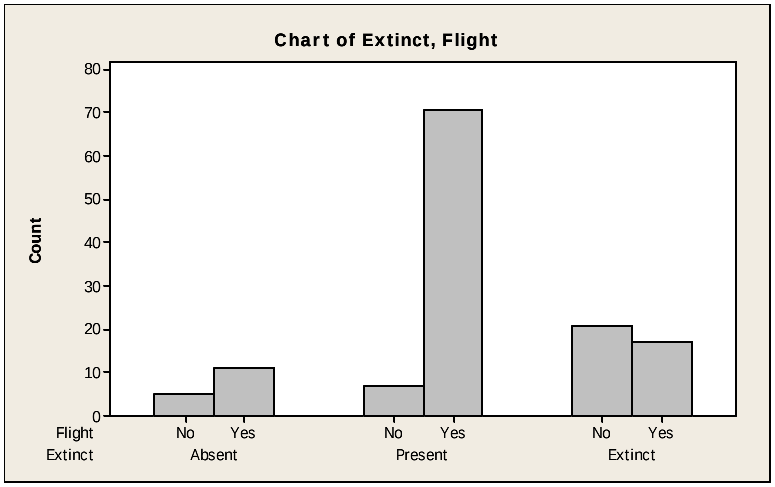
Abundance (number of frogs)	**Frequency Number of Recordings**	**Relative Frequency**
One	33	33 / 152 = .217
Few (2-9)	75	75 / 152 = .493
Many (10-50)	27	27 / 152 = .178
Lots(>50)	17	17 / 152 = .112
Total	**152**	**1.000**

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



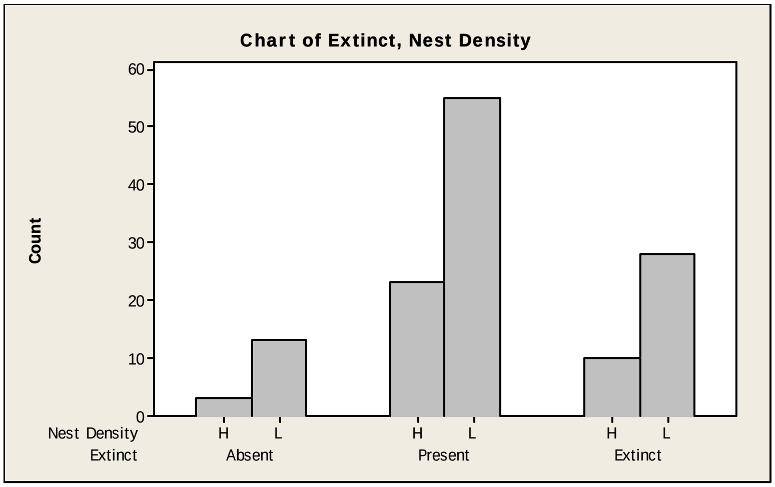
c. The abundance category with the greatest relative frequency is the 'few' category. Its relative frequency is .493.

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



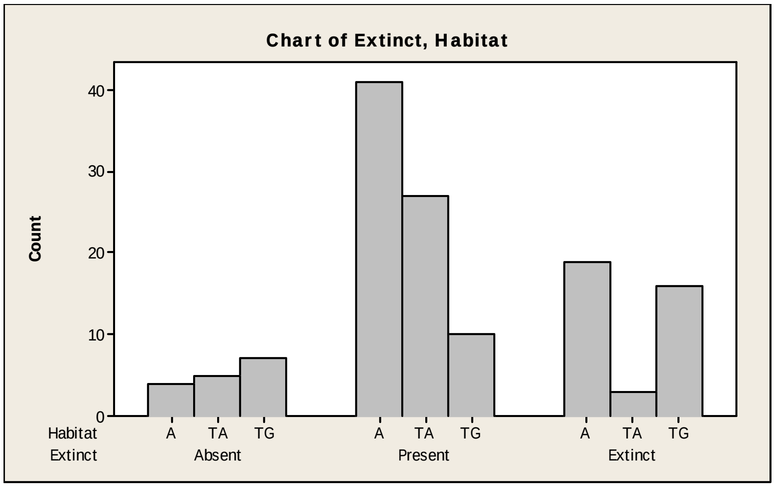
It appears that extinct status is related to flight capability. For birds that do have flight capability, most of them are present. For those birds that do not have flight capability, most are extinct.

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:

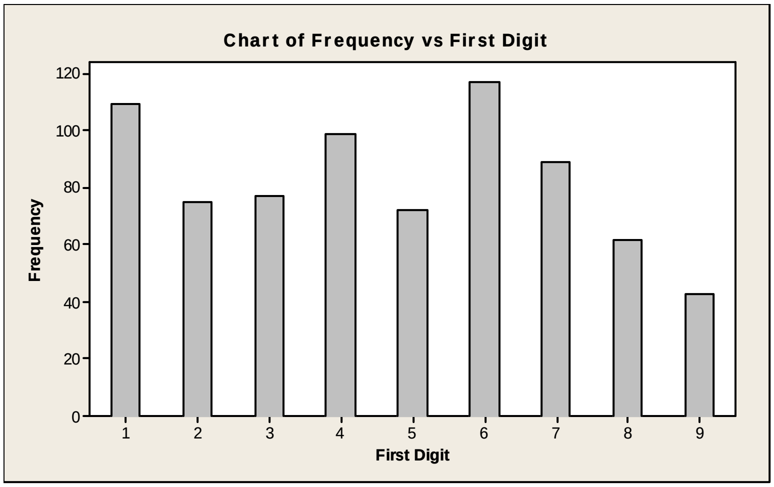


It appears that the extinct status is related to habitat. For those in aerial terrestrial (TA), most species are present. For those in ground terrestrial (TG), most species are extinct. For those in aquatic, most species are present.

2.20 a. 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

The relative frequency bar chart is:

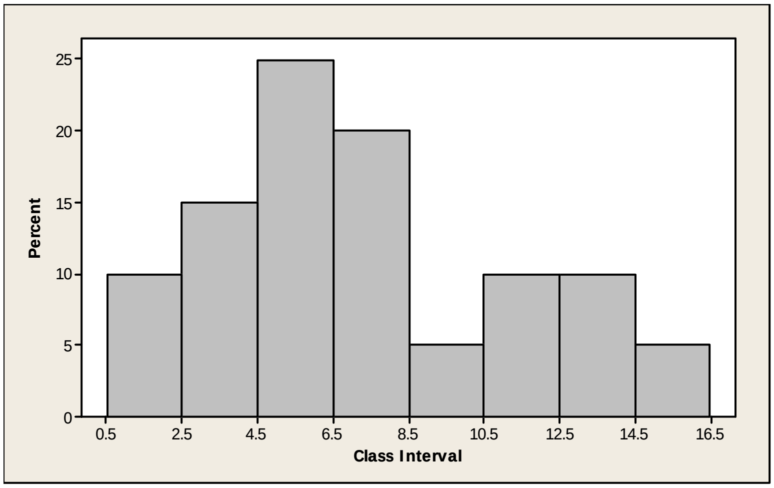


b. Benford's Law indicates that certain digits are more likely to occur as the first significant digit in a randomly selected number than other digits. The law also predicts that the number "1" is the most likely to occur as the first digit (30% of the time). From the relative frequency bar chart, one 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.

2.22 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.

2.24 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.

2.26

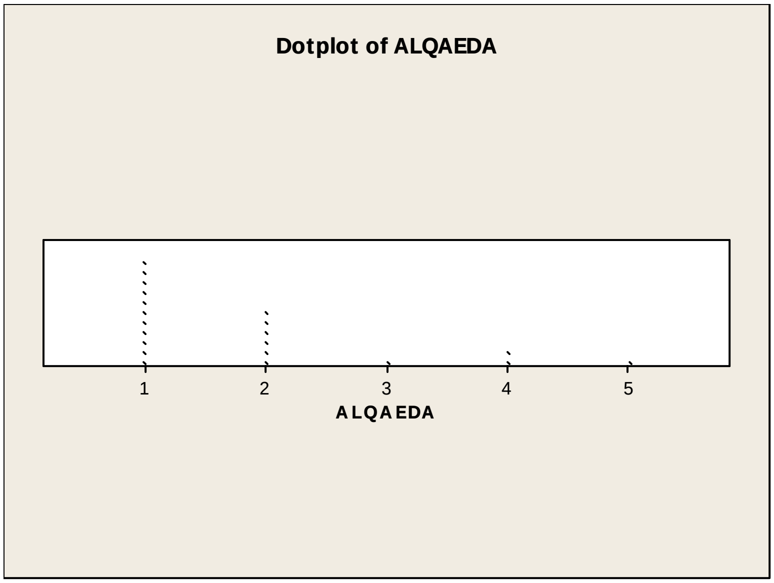


2.28 a. This is a frequency histogram because the number of observations are displayed rather than the relative frequencies.

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

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

2.30 Using MINITAB, the dot plot of the data is:



The most frequent number of attacks per incident is 1. There is only 1 incident with 3 attacks and only 1 incident with 5 attacks.

2.32 a. Using MINITAB, the stem-and-leaf display of the data is:

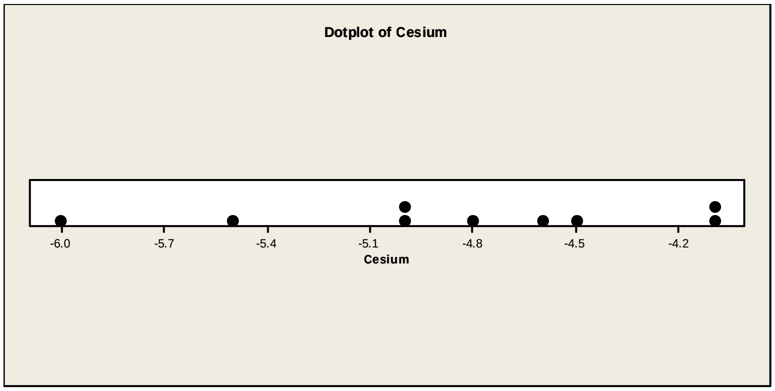
Stem-and-Leaf Display: SCORE

```
Stem-and-leaf of SCORE  N  = 169
Leaf Unit = 1.0


 1     6   2
 1     6
 2     7   2
 3     7   8
 4     8   4
 15    8   66677888899
 56    9   00001111111222222222233333333344444444444
(100)  9   5555555555555555555555666666666666666666777777777777777777888888+
 13   10   0000000000000
```

b. From the stem-and-leaf display, we see that there are only 4 observations with sanitation scores less than the acceptable score of 86. The proportion of ships that have an accepted sanitation standard would be (169 – 4) / 169 = .976.

c. The sanitation score of 84 is in bold in the stem-and-leaf display in part **a**.

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

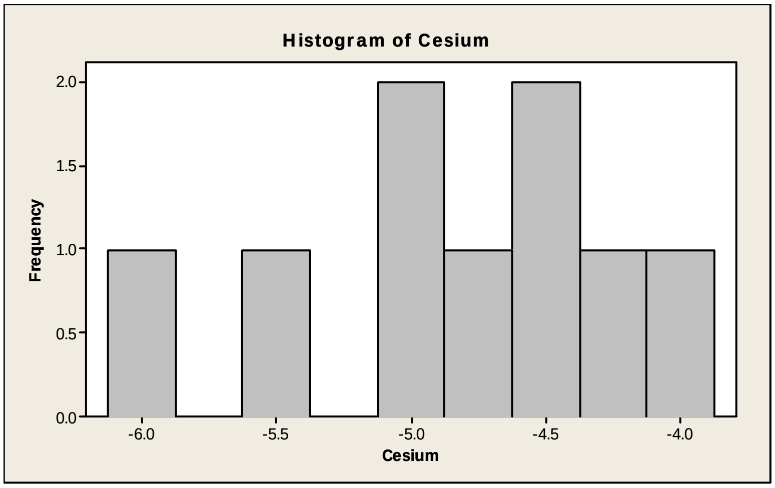


b. Using MINITAB, the stem-and-leaf display is:

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.

2.36 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  67
 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.

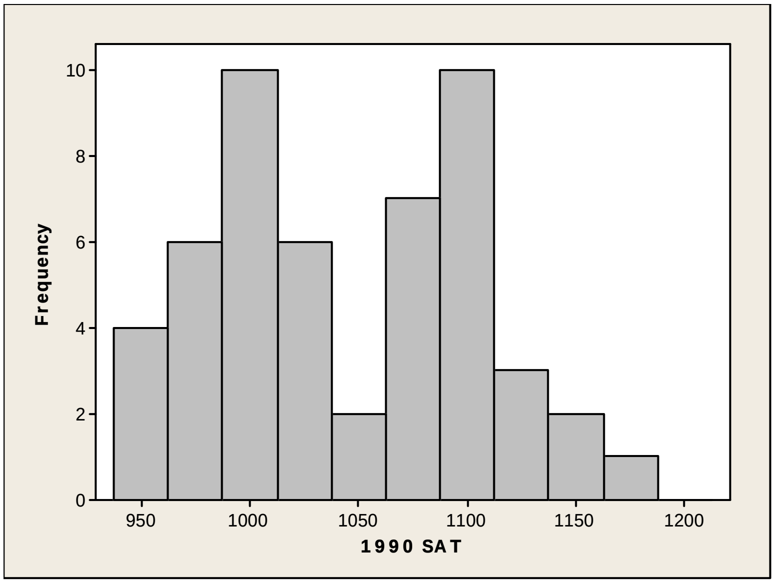
2.38 a. A stem-and-leaf display of the data using MINITAB is:

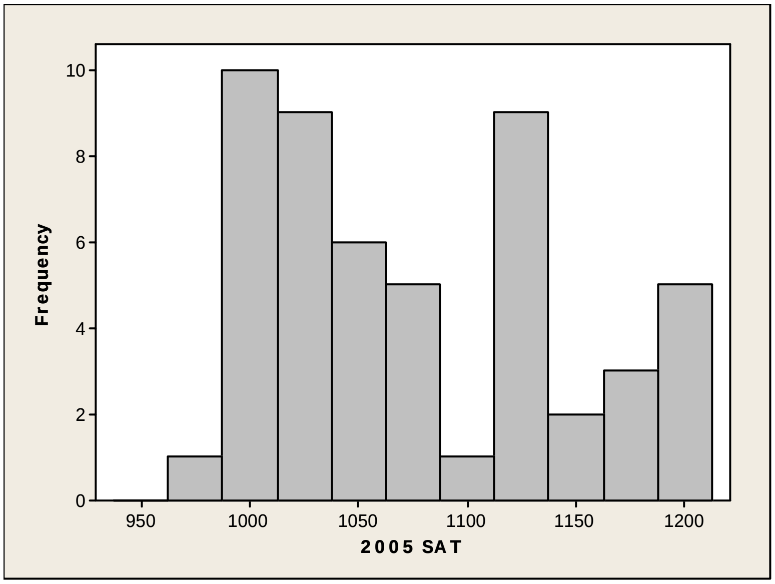
```
Stem-and-leaf of FNE       N  = 25
Leaf Unit = 1.0

  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.

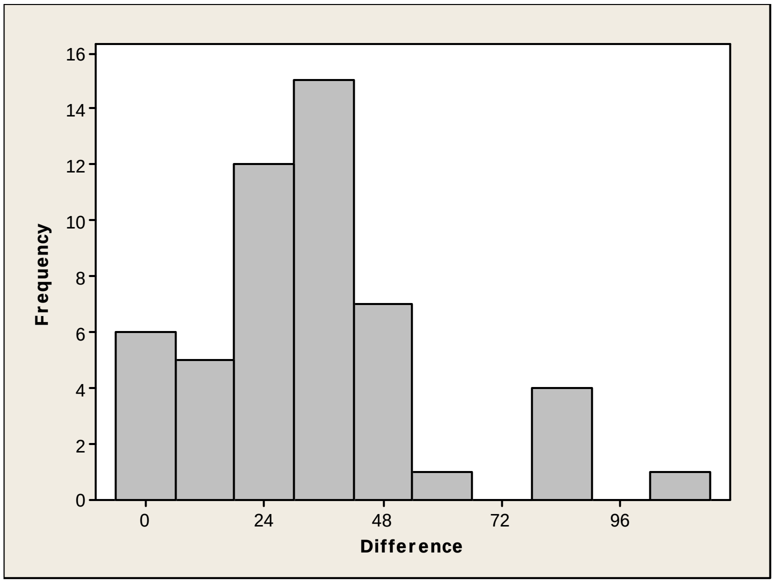
2.40 a. Using MINITAB, the frequency histograms for the 2 years of SAT scores are:





The data from 2005 is shifted to the right of that for 1990.

b. Using MINITAB, the frequency histogram of the differences is:



c. Very few, if any, of the observations of differences are negative. This implies that the SAT scores for 2005 are higher than the SAT scores for 1990 for the most part.

d. Based on the graph, the largest improvement is around 100 points. From the actual data, the real value of the point is 111. The state associated with this large difference is Illinois.

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

2.44 Using the results from Exercise 2.42,

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

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

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

2.46 A measure of central tendency measures the “center” of the distribution while measures of variability measure how spread out the data are.

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

2.50 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.

2.52 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:

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

2.54 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 $\frac{5+5}{2} = \frac{10}{2} = 5$.

2.56 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

2.58 a. The sample mean is:

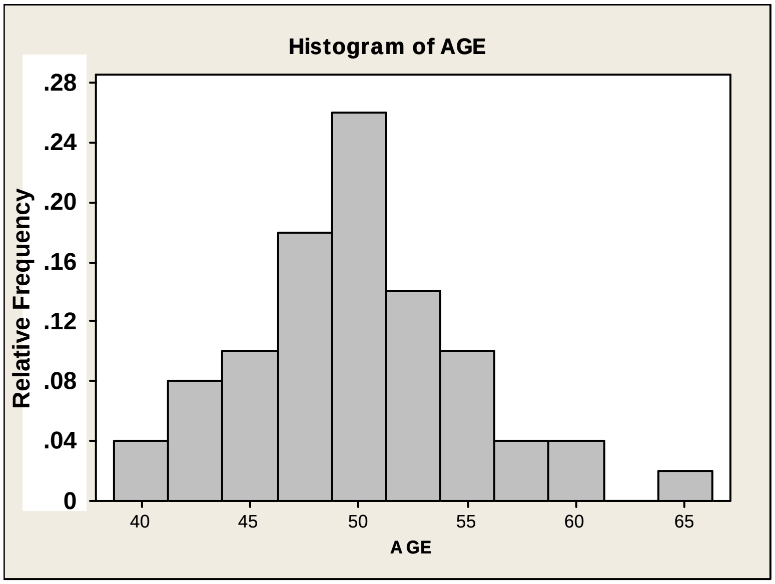
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{49 + 52 + 51 + \cdots + 51}{50} = \frac{2494}{50} = 49.88$$

The sample median is found by finding the average of the 25th and 26th observations once the data are arranged in order. The 25th and 26th observations are 49 and 50. The average of 49 and 50 is $\frac{49 + 50}{2}$ = 49.5. Thus, the sample median is 49.5.

The mode is the observation that occurs the most. It is 51, which occurs 8 times.

b. Since the mean is slightly greater than the median, the data are skewed to the right a little.

c. Using MINITAB, a relative frequency (percent) histogram is:



The modal class is the interval with the largest frequency. From the histogram the modal class is 48.75 to 51.25.

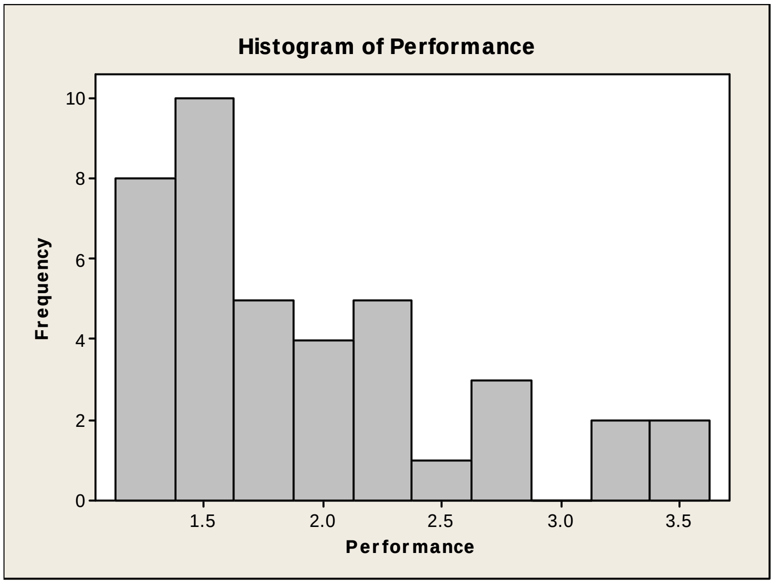
2.60 a. The mean of the driving performance index values is: $\bar{x} = \frac{\sum x}{n} = \frac{77.07}{40} = 1.927$

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 20th and 21st numbers are 1.75 and 1.76. The median is: $\frac{1.75 + 1.76}{2} = 1.755$

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:



2.62 Of all the applications, there could be some with 0 letters, 1 letter, 2 letters, 3 letters, etc. Of all of the possibilities, more applications had 3 letters than had any other number (mode). Since the median was also 3, at least half of the applications had 3 or more letters. The mean was 2.28. Since this is quite a bit smaller than the median, several of the applications must have had 0 or 1 letter.

2.66 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.

2.68 a. The mean number of power plants is:

$$\bar{x} = \frac{\sum x}{n} = \frac{5+3+...+3}{20} = \frac{80}{20} = 4$$

The median is the mean of the middle 2 numbers once the data have been arranged in order: 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 5, 6, 7, 9, 13

The median is $\frac{3+3}{2} = \frac{6}{2} = 3$.

There are 3 numbers that each occur 4 times. They are 1, 2, and 5. Thus, there are 3 modes, 1, 2, and 5.

b. Deleting the largest number, 13, the new mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{5+3+...+3}{19} = \frac{67}{19} = 3.526$$

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

The median is 3.

There are 3 numbers that each occur 4 times. They are 1, 2, and 5. Thus, there are 3 modes, 1, 2, and 5.

By dropping the largest measurement from the data set, the mean drops from 4 to 3.526. The median and the modes stay the same. There is no effect on them.

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

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

The new mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{1+1+\cdots+7}{16} = \frac{56}{16} = 3.5$$

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

2.70 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.

2.72 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 $s = \sqrt{.058} = .241$

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

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

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

2.76 a. Range = 4 – 0 = 4

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

b. Range = 6 – 0 = 6

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

c. Range = 8 – (–2) = 10

$s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{154 - \dfrac{30^2}{18}}{10-1} = 7.111$ $s = \sqrt{7.111} = 2.67$

d. Range = 2 – (–3) = 5

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

2.78 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

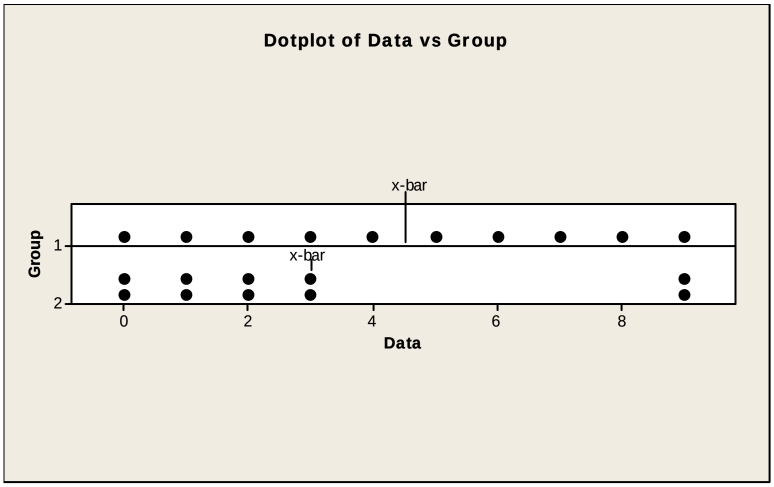
The two sets of data above have the same range = largest measurement – smallest measurement = 9 – 0 = 9.

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.



2.80 a.

$$\sum x = 3 + 1 + 10 + 10 + 4 = 28$$

$$\sum x^2 = 3^2 + 1^2 + 10^2 + 10^2 + 4^2 = 226$$

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{5} = 5.6$$

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

b.

$$\sum x = 8 + 10 + 32 + 5 = 55$$

$$\sum x^2 = 8^2 + 10^2 + 32^2 + 5^2 = 1213$$

$$\bar{x} = \frac{\sum x}{n} = \frac{55}{4} = 13.75 \text{ feet}$$

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

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

c.

$$\sum x = -1 + (-4) + (-3) + 1 + (-4) + (-4) = -15$$

$$\sum x^2 = (-1)^2 + (-4)^2 + (-3)^2 + 1^2 + (-4)^2 + (-4)^2 = 59$$

$$\bar{x} = \frac{\sum x}{n} = \frac{-15}{6} = -2.5$$

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

d. $\sum x = \frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{2}{5}+\frac{1}{5}+\frac{4}{5}=\frac{10}{5} = 2$

$$\sum x^2 = \left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{2}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2 = \frac{24}{25} = .96$$

$$\bar{x} = \frac{\sum x}{n} = \frac{2}{6} = \frac{1}{3} = .33 \text{ ounce}$$

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

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

2.82 a. The range = maximum measurement – minimum measurement. From the printout, the range = 6.7 – 0 = 6.7.

b. From the printout, the variance = .4403.

c. From the printout, the standard deviation is .6636.

d. If the target is these specific 2, 929 aftershocks, then we should use the symbols for the population parameters. The variance would be σ^2 and the standard deviation would be σ.

2.84 a. The maximum age is 64. The minimum age is 39. The range is 64 – 39 = 25.

b. The variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{125{,}764 - \frac{2{,}494^2}{50}}{50-1} = 27.822$$

c. The standard deviation is:

$$s = \sqrt{s^2} = \sqrt{27.822} = 5.275$$

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. (5.275 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:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{121{,}668 - \frac{2{,}430^2}{49}}{49-1} = 24.1633$$

The new standard deviation is $s = \sqrt{s^2} = \sqrt{24.1633} = 4.916$. This is less than the standard deviation with all the observations (s = 5.275).

2.86 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.

2.88 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 $\bar{x} + 3s$.

2.90 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.

	Number of Measurements	
Interval	**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

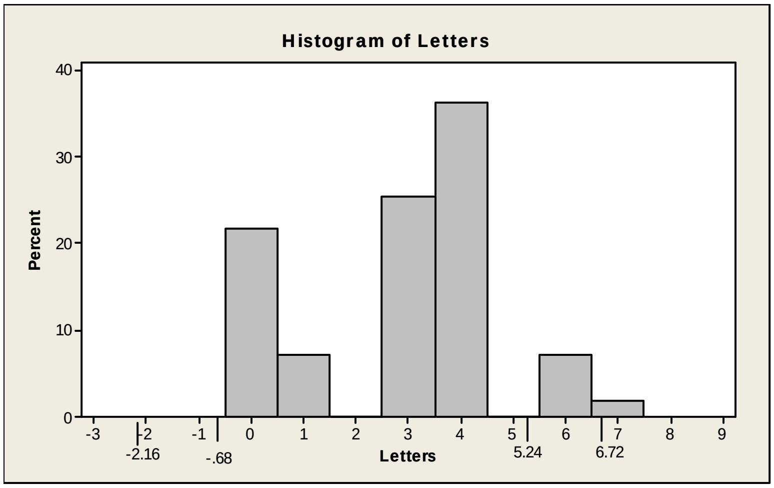
$s \approx$ range/4 = 7/4 = 1.75

The range approximation provides a satisfactory estimate of s.

2.92 From Exercise 2.58, the sample mean is $\bar{x} = 49.88$. From Exercise 2.84, the sample standard deviation is $s = 5.275$. From Chebyshev’s Rule, at least 75% of the ages will fall within 2 standard deviations of the mean. This interval will be:

$$\bar{x} \pm 2s \Rightarrow 49.88 \pm 2(5.275) \Rightarrow 49.88 \pm 10.55 \Rightarrow (39.33,\quad 60.43)$$

2.94 a. Since the mean of the data (2.28) is smaller than the median, the data are probably skewed to the left. Also, there probably aren't too many applications with more than three letters. However, because the standard deviation is more than half of the mean, there must be a couple of very large numbers in the data set. A possible histogram of the data is:



b. Since the histogram drawn is approximately mound-shaped, about 95% of the observations will fall within 2 standard deviations of the mean. This interval is:

$$\bar{x} \pm 2s \Rightarrow 2.28 \pm 2(1.48) \Rightarrow 2.28 \pm 2.96 \Rightarrow (-.68, 5.24)$$

This interval is drawn on the graph in part **a**.

c. Since the histogram drawn is approximately mound-shaped, about all of the observations will fall within 3 standard deviations of the mean. This interval is:

$$\bar{x} \pm 3s \Rightarrow 2.28 \pm 3(1.48) \Rightarrow 2.28 \pm 4.44 \Rightarrow (-2.16, 6.72)$$

The interval is drawn on the graph in part a.

2.96 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.

2.98 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:

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

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

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

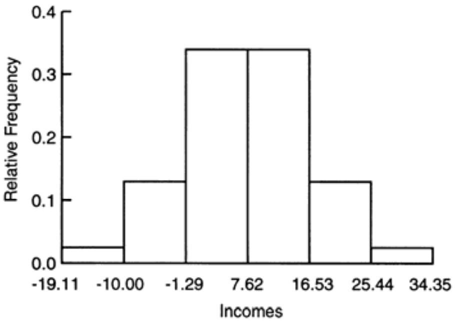
Middle Income:

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

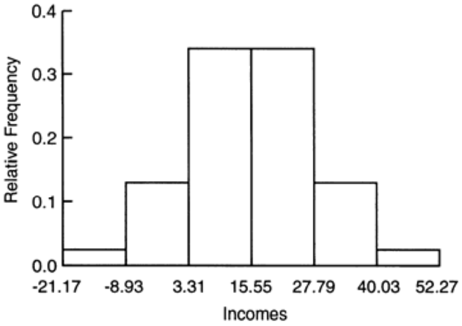
$$\bar{x} \pm 2s \Rightarrow 15.55 \pm 2(12.24) \Rightarrow 15.55 \pm 24.48 \Rightarrow (-8.93, 40.03)$$

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

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.)

2.100 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.

2.102 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.

2.104 The 50th percentile of a data set is the observation that has half of the observations less than it. Another name for the 50th percentile is the median.

2.106 a. $z = \dfrac{x-\bar{x}}{s} = \dfrac{40-30}{5} = 2$ (sample) 2 standard deviations above the mean.

b. $z = \dfrac{x-\mu}{\sigma} = \dfrac{90-89}{2} = .5$ (population) .5 standard deviations above the mean.

c. $z = \dfrac{x-\mu}{\sigma} = \dfrac{50-50}{5} = 0$ (population) 0 standard deviations above the mean.

d. $z = \dfrac{x-\bar{x}}{s} = \dfrac{20-30}{4} = -2.5$ (sample) 2.5 standard deviations below the mean.

2.108 Since the element 40 has a *z*-score of –2 and 90 has a *z*-score of 3,

$$-2 = \frac{40-\mu}{\sigma} \text{ and } 3 = \frac{90-\mu}{\sigma}$$

$$\begin{array}{ll} \Rightarrow -2\sigma = 40 - \mu & \Rightarrow 3\sigma = 90 - \mu \\ \Rightarrow \mu - 2\sigma = 40 & \Rightarrow \mu + 3\sigma = 90 \\ \Rightarrow \mu = 40 + 2\sigma & \end{array}$$

By substitution, $40 + 2\sigma + 3\sigma = 90$

$\Rightarrow 5\sigma = 50$

$\Rightarrow \sigma = 10$

By substitution, $\mu = 40 + 2(10) = 60$

Therefore, the population mean is 60 and the standard deviation is 10.

2.110 The mean score is 279. This is the arithmetic average score of U.S. eighth graders on the mathematics assessment test. The 25th percentile score is 255. This indicates that 25% of the U.S. eighth graders scored 255 or lower on the assessment test. The 75th percentile score is 304. This indicates that 75% of the U.S. eighth graders scored 304 or lower on the assessment test. The 90th percentile score is 324. This indicates that 90% of the U.S. eighth graders scored 324 or lower on the assessment test.

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

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

From Exercise 2.81, *s* = 8.701.

The z-score for a score of 40 books is $z = \frac{x - \overline{x}}{s} = \frac{40 - 37}{8.701} = .345$. 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:

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

From Exercise 2.81, *s* = 8.526.

The *z*-score for a score of 40 books is $z = \frac{x - \overline{x}}{s} = \frac{40 - 24.5}{8.526} = 1.82$. 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 .345. This is not unusual. For the B-C group, the *z*-score corresponding to 40 books id 1.82. This is close to 2 standard deviations from the mean. This would be fairly unusual.

2.114 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.

2.116 a. If the distributions are 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 distributions are 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.

2.118 An observation that is unusually large or small relative to the data values we want to describe is an outlier.

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

2.122 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.

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

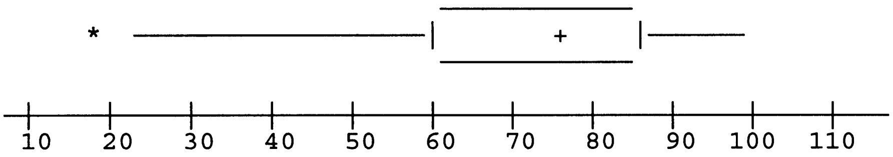
The lower inner fence = $Q_L - 1.5(\text{IQR}) = 60 - 1.5(25) = 22.5$.

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

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

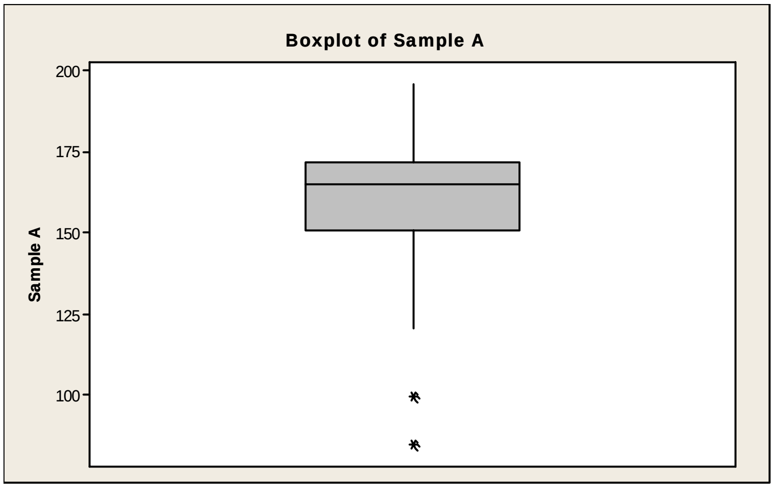
The upper outer fence = $Q_U + 3(\text{IQR}) = 85 + 3(25) = 160$.

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

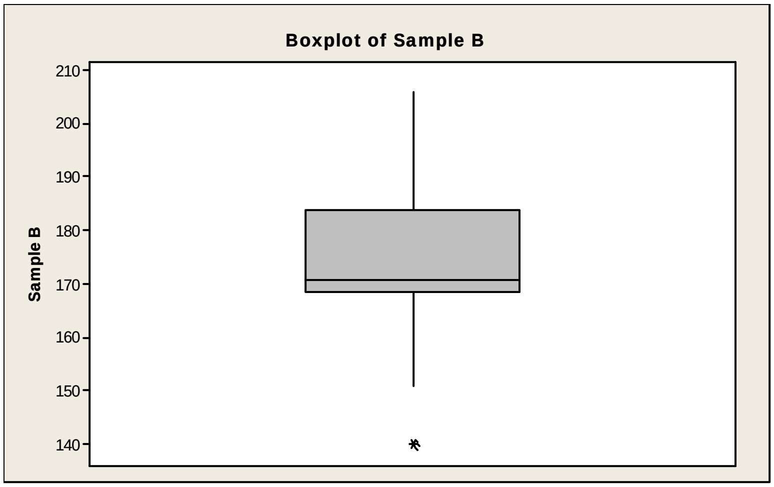


The whiskers extend to the inner fences unless no data points are that small or that large. The upper inner fence is 122.5. However, the largest data point is 100, so the whisker stops at 100. The lower inner fence is 22.5. The smallest data point is 18, so the whisker extends to 22.5. Since 18 is between 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.

2.126 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.

2.128 a. The *z*-score is $z = \dfrac{x - \overline{x}}{s} = \dfrac{175 - 79}{23} = 4.17$.

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.

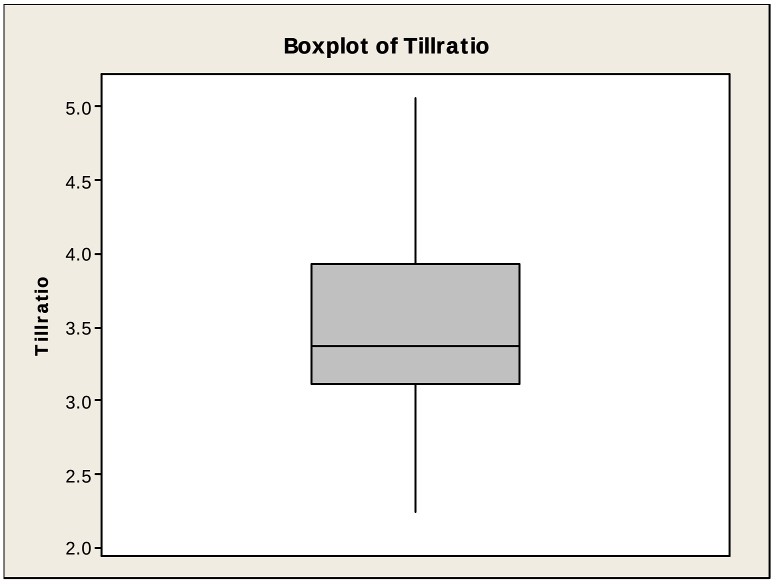
2.130 a. The *z*-score associated with the largest ratio is $z = \dfrac{x - \overline{x}}{s} = \dfrac{5.06 - 3.5069}{.63439} = 2.45$

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

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

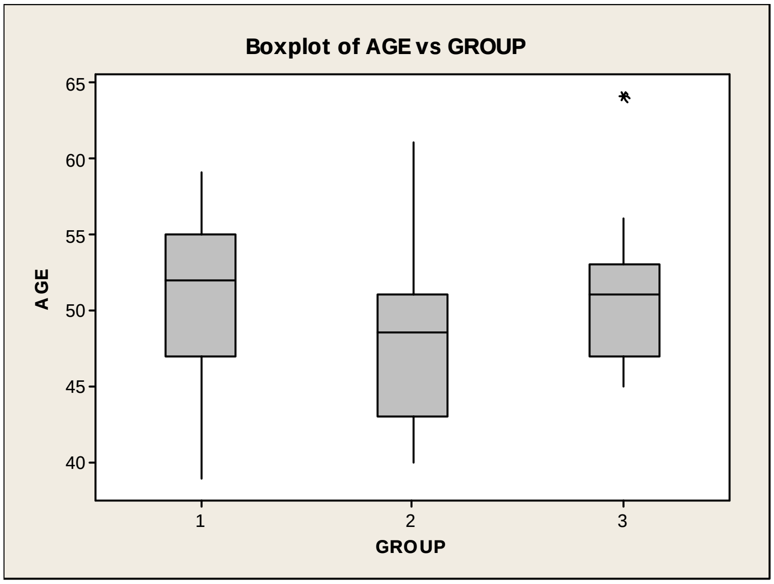
b. Yes, I would consider the *z*-score associated with the largest ratio to be unusually large. We know if the data are approximately mound-shaped that approximately 95% of the observations will be within 2 standard deviations of the mean. A *z*-score of 2.45 would indicate that less than 2.5% of all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

c. Using MINITAB, the box plot is:



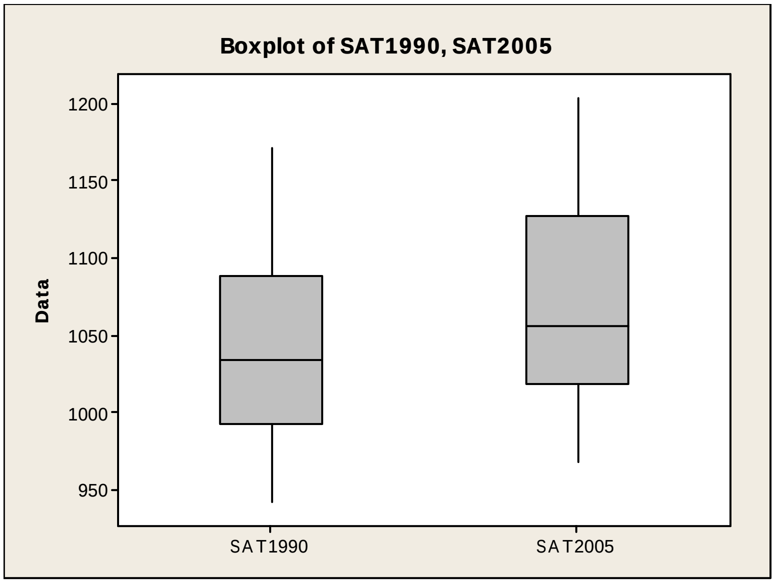
From this box plot, there are no observations marked as outliers.

2.132 Using MINITAB, the box plots are:



From the plots, there appears to be one potential outlier in the third group.

2.134 a. Using MINITAB, the side-by-side box plots are:



b. Using MINITAB, the descriptive statistics are:

Descriptive Statistics: SAT1990, SAT2005

```
Variable   N    Mean  StDev  Minimum      Q1  Median      Q3  Maximum
SAT1990   51  1043.7   59.5    942.0   993.0  1034.0  1089.0   1172.0
SAT2005   51  1077.0   67.9    968.0  1018.0  1056.0  1127.0   1204.0
```

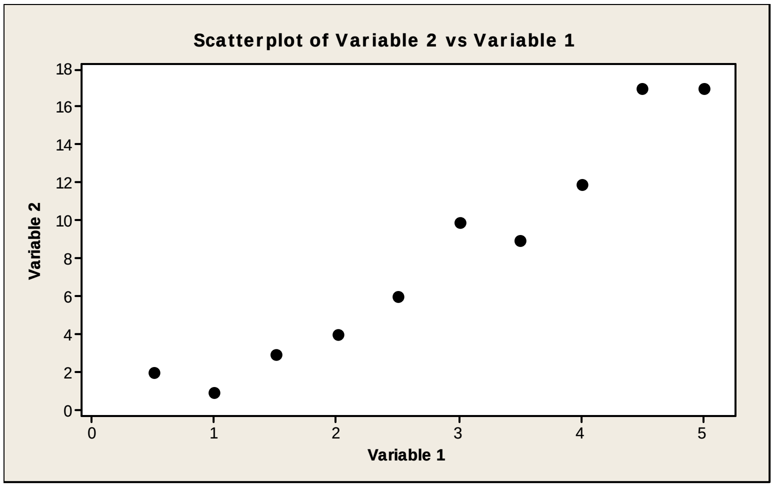
The standard deviation for 1990 is 59.5 and the standard deviation for 2005 is 67.9. Also, the IQR for 1990 is $Q_U – Q_L$ = 1089 – 993 = 96 while the IQR for 2005 is $Q_U – Q_L$ = 1127 – 1018 = 109. Thus, the variability for 2005 is slightly greater than that for 1990.

c. Since there are no observations outside the inner fences for either year, there are no outliers.

2.136 Scatterplots are useful with quantitative variables.

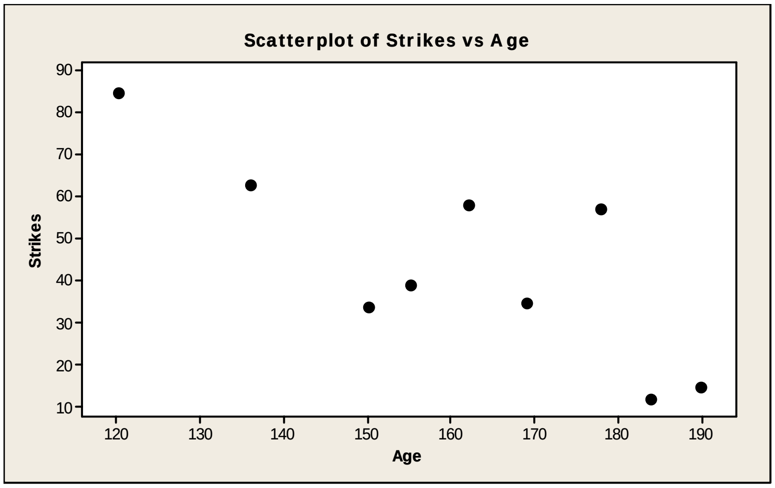
2.138 A positive association between two variables means that as one variable increases, the other variable tends to also increase. A negative association between two variables means that as one variable increases, the other variable tends to decrease.

2.140 Using MINITAB, the scatterplot is as follows:



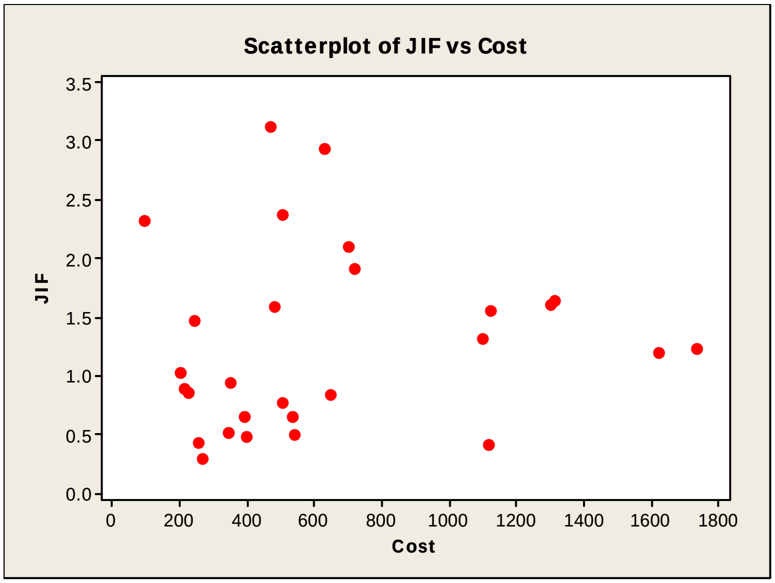
It appears that as variable 1 increases, variable 2 also increases.

2.142 a. A scattergram of the data is:



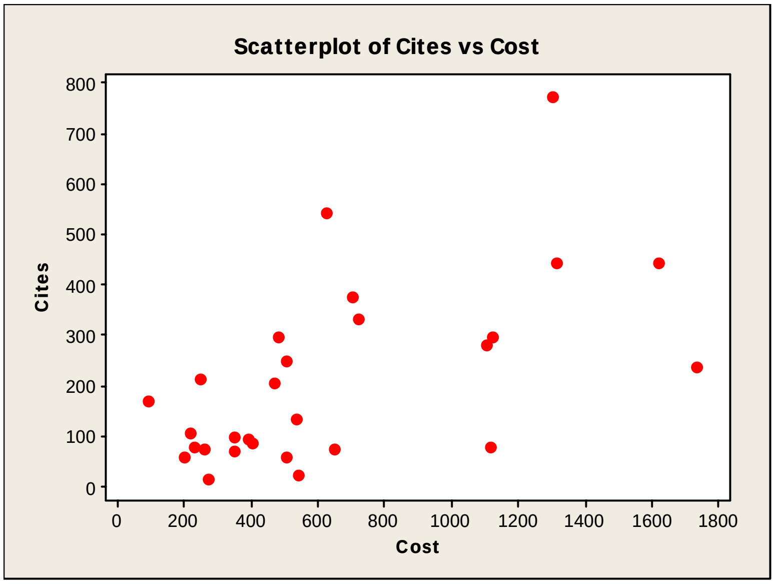
b. There appears to be a trend. As the age increases, the number of strikes tends to decrease.

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



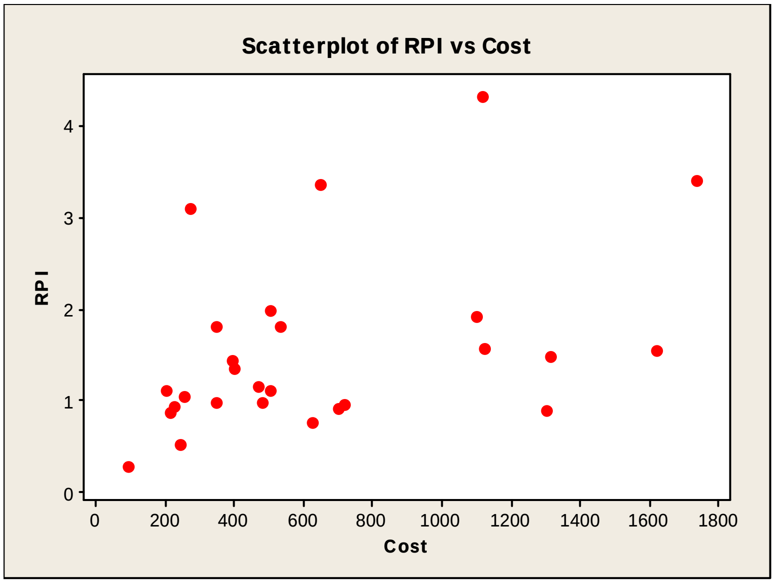
There does not appear to be much of a trend between these two variables.

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



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

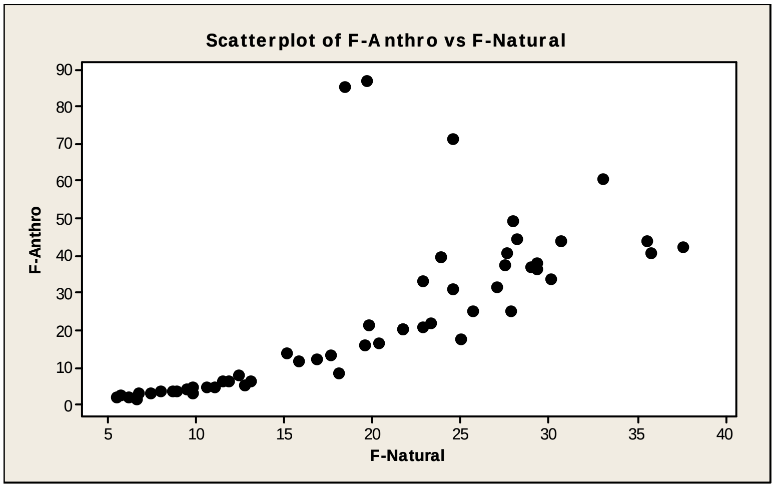
c. Using MINITAB, the scatterplot of RPI and cost is:



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

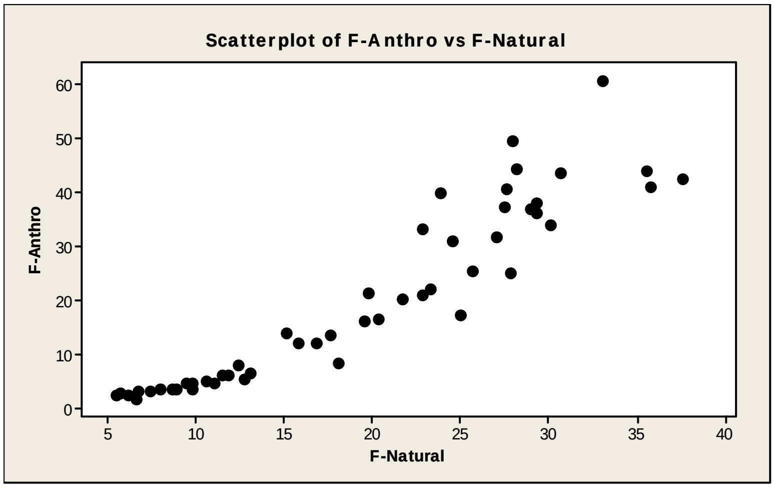
2.146 From the scatterplot, it appears that the data support the theory. Smaller values of hippocampal volume are associated with smaller values of memory scores.

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



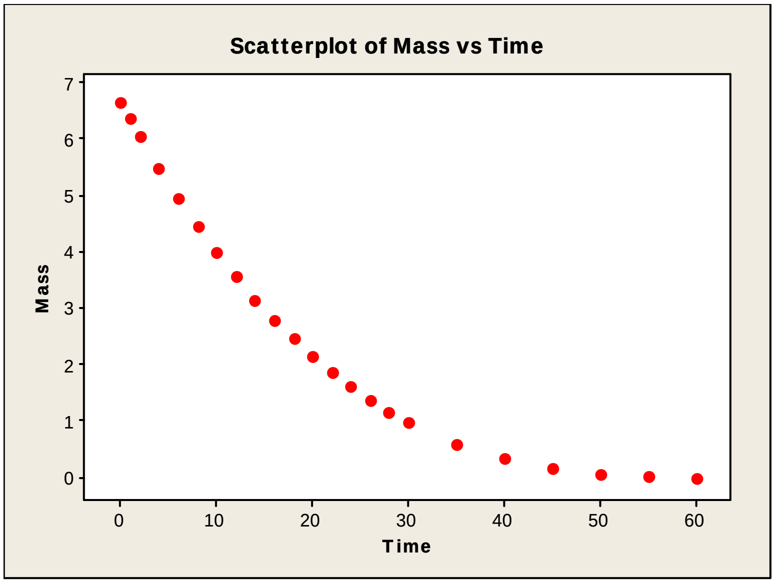
This graph does not support the theory that there is a straight-line relationship between the Anthropogenic Index against the Natural Origin Index. There are several points that do not lie on a straight line.

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.

2.150 Using MINITAB, a scattergram of the data is:

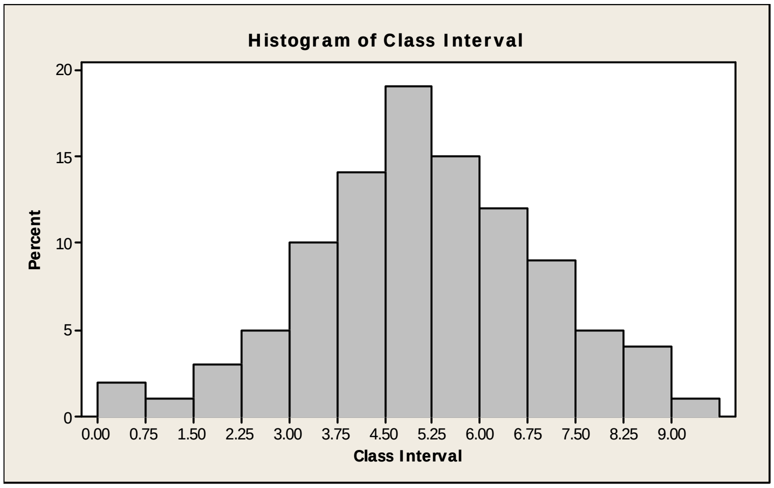


Yes, there appears to be a negative linear trend in this data. As time increases, the mass decreases.

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

2.154 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.

2.156 The relative frequency histogram is:



2.158 From part a of Exercise 2.157, 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 2.157, 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 2.157, 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 2.157, 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.

2.160 $\sigma \approx$ range/4 = 20/4 = 5

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

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

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

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

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

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

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

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

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

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

$$\bar{x} = \frac{\sum x}{n} = \frac{49}{7} = 1$$

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

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.

2.164 a. $\bar{x} = \frac{22}{22} = 1$ $\qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{42 - \frac{22^2}{22}}{22-1} = .9524 \qquad s = .98$

There are two measurements in the interval $\bar{x} \pm s$ or (.02, 1.98), which is 9.1%. This agrees with Chebyshev's rule which says at least $1 - 1/1^2 = 0$ of the observations will fall in this interval.

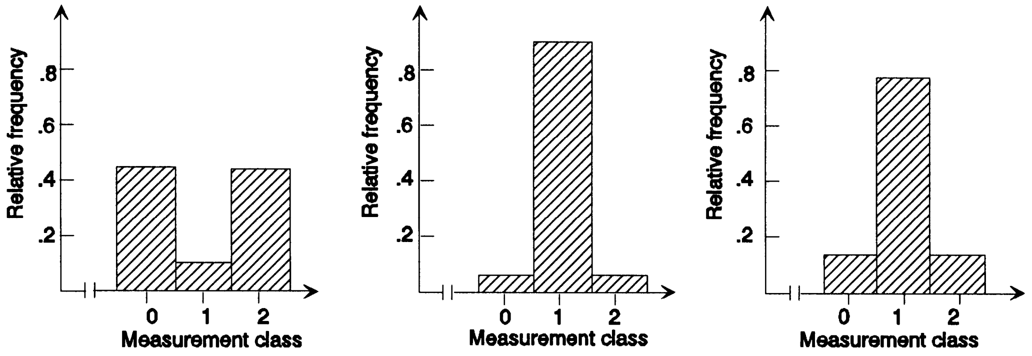
b. $\bar{x} = \frac{42}{42} = 1$ $\qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{52 - \frac{42^2}{42}}{42-1} = .2439 \qquad s = .4939$

There are 32 measurements in the interval $\bar{x} \pm 2s$ or (.01, 1.99), which is 76.2%. This agrees with Chebyshev's rule (at least 75%).

c. $\bar{x} = \frac{56}{56} = 1$ $\qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{62 - \frac{56^2}{56}}{56-1} = .1091$ $\qquad s = \sqrt{.1091} = .33$

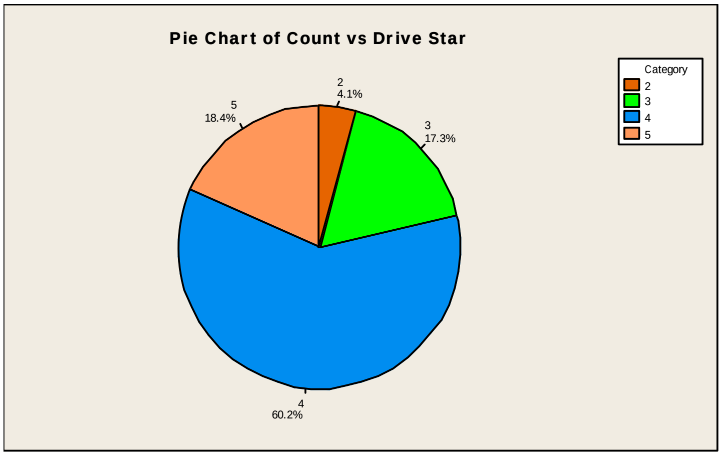
There are 50 measurements in the interval $\bar{x} \pm 3s$ or (.01, 1.99), which is 89.3%. This agrees with Chebyshev's rule (at least 89%).

d. The three histograms are shown below.



Notice that the data need not be mound-shaped for Chebyshev's rule to be appropriate.

2.166 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.

2.168 a. The mean of the data is $= \bar{x} = \frac{\sum x}{n} = \frac{5+4+3+\cdots+0}{14} = \frac{20}{14} = 1.429$

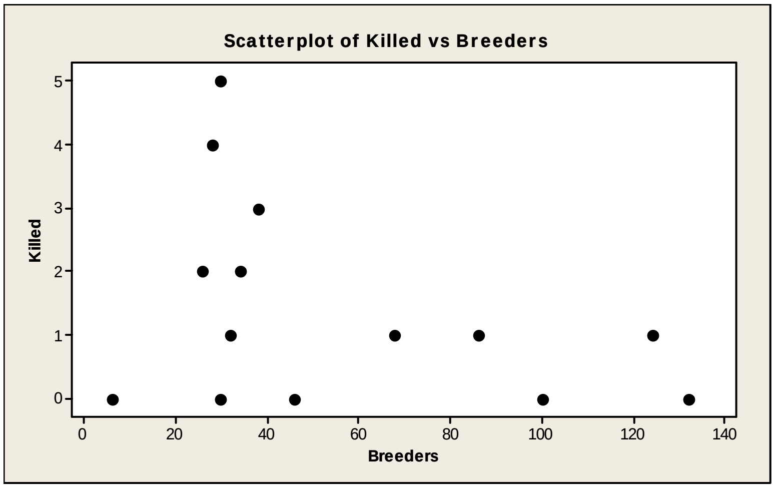
The median is the average of the middle two numbers once the data are arranged in order.

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

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.

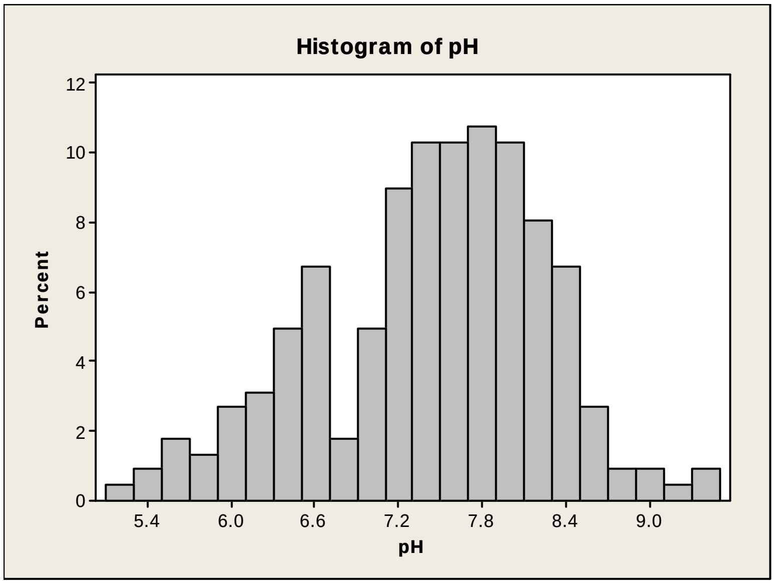
b. Because the mode is the smallest value of the three, the median is the next smallest, and the mean is the largest, the data are skewed to the right. Because the data are skewed, the median is probably a more representative measure for the middle of the data set. Only 5 of the 14 observations are larger 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.

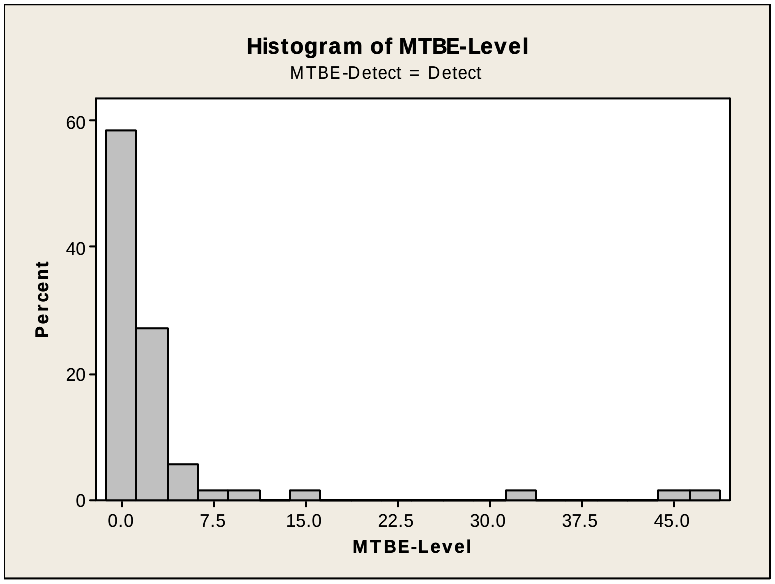
2.170 a. Using MINITAB, a histogram of the data is:



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: $s^2 = \sqrt{s^2} = \sqrt{.66727} = .8169$

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

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:

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

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

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

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.

2.172 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.

2.174 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 nore 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.

2.176 a. For Adults:

		Interval	Percent	Frequency
$\bar{x} \pm s$	6.45 ± 2.89	(3.56, 9.34)	≈68%	46(.68) ≈ 31.3
$\bar{x} \pm 2s$	$6.45 \pm 2(2.89)$	(0.67, 12.23)	≈95%	46(.95) ≈ 43.7

For Adolescents:

		Interval	Percent	Frequency
$\bar{x} \pm s$	10.89 ± 2.48	(8.41, 13.37)	≈68%	19(.68) ≈ 12.9
$\bar{x} \pm 2s$	$10.89 \pm 2(2.48)$	(5.93, 15.85)	≈95%	19(.95) ≈ 18.1

b. See numbers in the above tables.

2.178 a. Using Minitab, the stem-and-leaf display is:

```
Stem-and-leaf of PAF       N=17
Leaf Unit = 1.0

 6   0  000009
 8   1  25
(2)  2  45
 7   3  13
 5   4  0
 4   5
 4   6  2
 3   7  057
```

b. The mean of the data is $= \bar{x} = \frac{\sum x}{n} = \frac{77+33+75+\cdots+31}{17} = \frac{473}{17} = 27.82$

The median is the middle number once the data are arranged in order. The data arranged in order are: 0, 0, 0, 0, 0, 9, 12, 15, 24, 25, 31, 33, 40, 62, 70, 75, 77.

The middle number or the median is 24.

The number occurring most frequently is 0. The mode is 0.

c. The mode corresponds to the smallest number. It does not seem to locate the center of the distribution. Both the mean and the median are in the middle of the stem-and-leaf display. Thus, it appears that both of them locate the center of the data.

d. Range = 77 – 0 = 77.

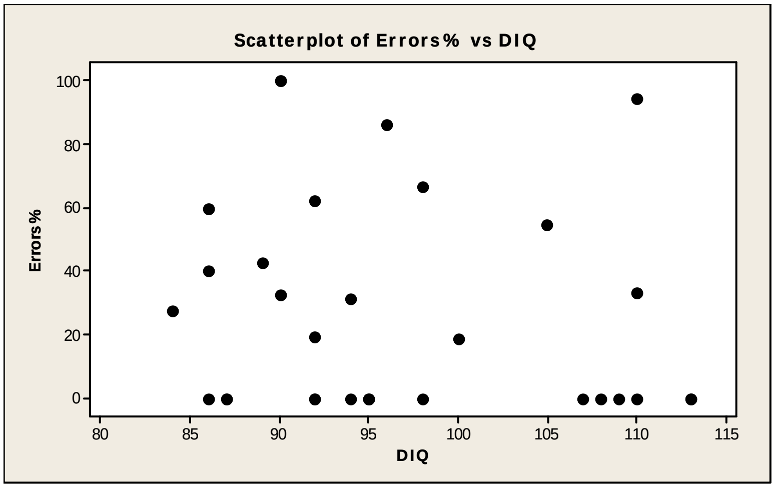
$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{25,599 - \frac{473^2}{17}}{17-1} = \frac{12,438.47059}{16} = 777.4044$$

$$s = \sqrt{s^2} = \sqrt{777.4044} = 27.882$$

e. Since the stem-and-leaf display is not mound-shaped, we must use Chebyshev's Rule to describe the data. We know that at least $1-\frac{1}{k^2}=1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}$ of the observations will fall within 3 standard deviations of the mean.

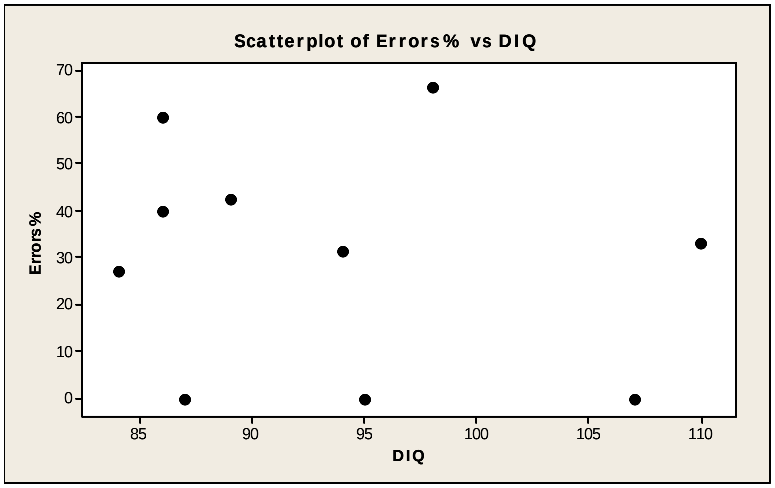
$\bar{x}\pm 3s \Rightarrow 27.82\pm 3(27.882)\Rightarrow 27.82\pm 83.646\Rightarrow(-55.826,\ \ 111.466)$. At least 8/9 of the observations will fall between –55.826 and 111.466.

2.180 a. A scattergram of the data is:



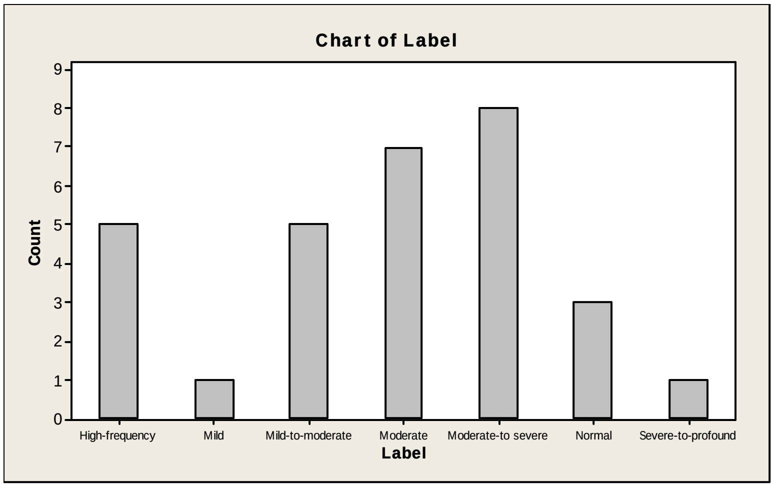
There does not appear to be much of a relationship between deviation intelligence quotient (DIQ) and the percent of pronoun errors. The points are scattered randomly.

b. 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.

2.182 A frequency bar graph is used to depict the data:



From the bar graph, it appears that "Moderate-to-severe loss" is the most prevalent type of hearing loss. "Moderate loss" is the next most prevalent type of hearing loss.

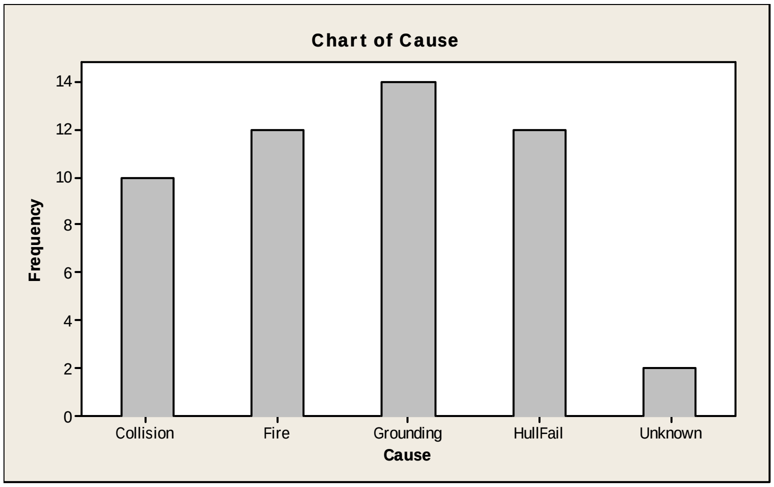
2.184 If the distributions of the standardized tests are approximately mound-shaped, then it would be impossible for 90% of the school districts' students to score above the mean. If the distributions are mound-shaped, then the mean and median are approximately the same. By definition, only 50% of the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

2.186 a. We will use a frequency bar graph to describe the data. First, we must add up the number of spills's under each category. These values are summarized in the following table:

Cause of Spillage	Frequency
Collision	11
Grounding	13
Fire/Explosion	12
Hull Failure	12
Unknown	2
Total	50

The frequency bar graph is:



Because each of the bars are about the same height, it does not appear that one cause is more likely to occur than any other.

b. $\bar{x} = \frac{\sum x}{n} = \frac{2991}{50} = 59.82$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{318,477 - \frac{2,991^2}{50}}{50-1} = \frac{139,555.38}{49} = 2,848.06898$$

$$s = \sqrt{s^2} = \sqrt{2,848.06898} = 53.367$$

Since the data are not mound-shaped, we must use Chebyshev's Rule to describe the data. We know that at least $1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9}$ of the observations will fall within 3 standard deviations of the mean.

$\bar{x} \pm 3s \Rightarrow 59.82 \pm 3(53.367) \Rightarrow 59.82 \pm 160.101 \Rightarrow (-100.281,\ \ 219.921)$. At least 8/9 of the observations will fall between –100.281 and 219.921.

2.188 Since we do not know if the distribution of the heights of the trees is mound-shaped, we need to apply Chebyshev's rule. We know $\mu = 30$ and $\sigma = 3$. Therefore,

$\mu \pm 3\sigma \Rightarrow 30 \pm 3(3) \Rightarrow 30 \pm 9 \Rightarrow (21, 39)$

According to Chebyshev's rule, at least 8/9 or .89 of the tree heights on this piece of land fall within this interval and at most $\frac{1}{9}$ or .11 of the tree heights will fall above the interval.

However, the buyer will only purchase the land if at least $\frac{1000}{5000}$ or .20 of the tree heights are at least 40 feet tall. Therefore, the buyer should not buy the piece of land.

2.190 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.