SOLUTIONS MANUAL
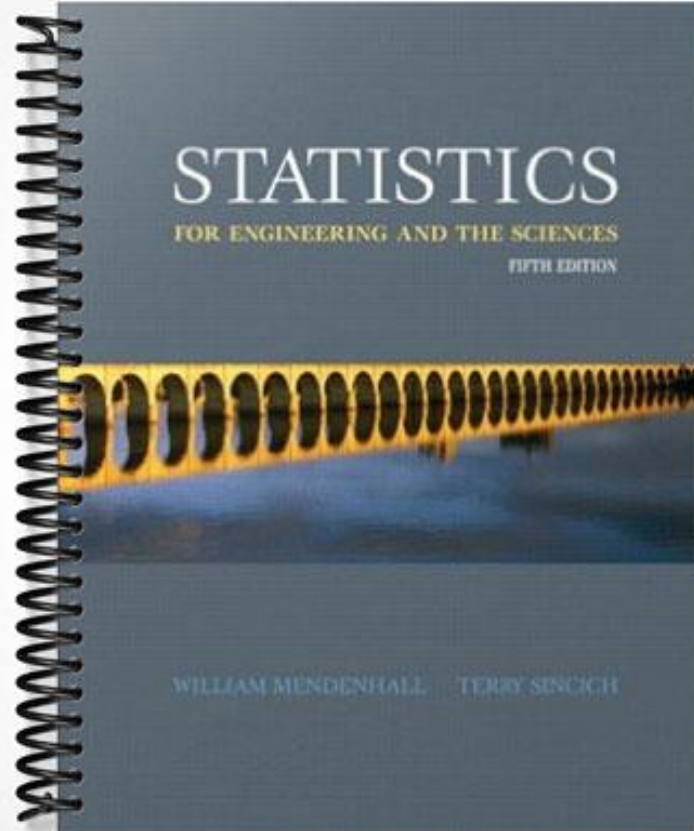
STATISTICS

FOR ENGINEERING AND THE SCIENCES

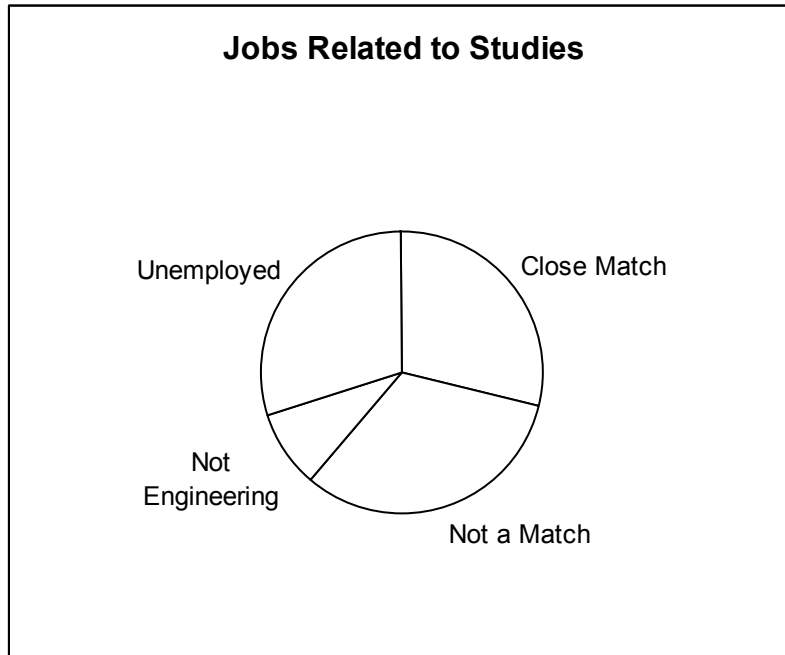### FIFTH EDITION

WILLIAM MENDENHALL   TERRY SINCICH

CHAPTER 2

....

Descriptive Statistics

2.2   The data were entered into Excel and the following pie chart was created:

Jobs Related to Studies

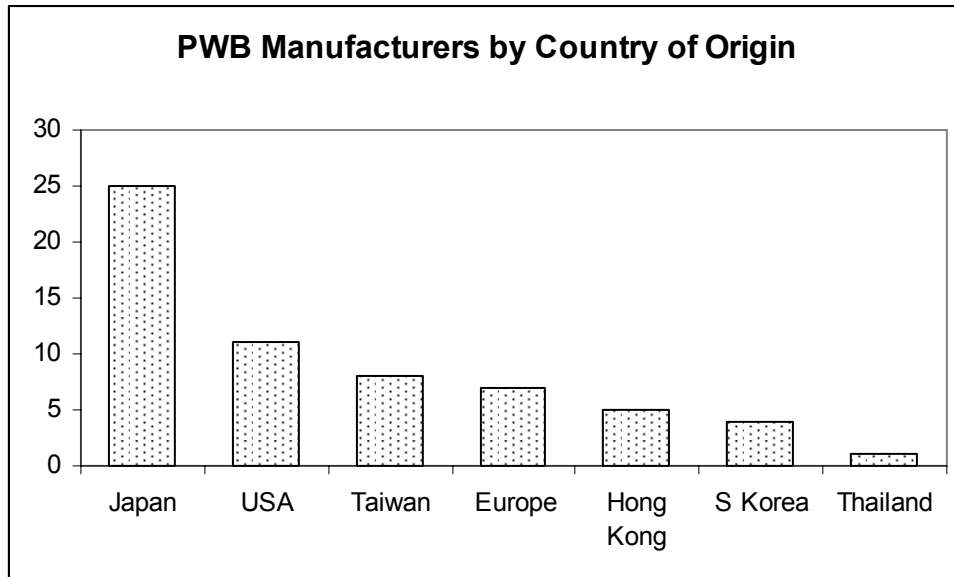Unemployed
Close Match
Not Engineering
Not a Match

The responses were pretty evenly divided between close match, not a match and unemployed.

2.4   The Pareto diagram is shown below:
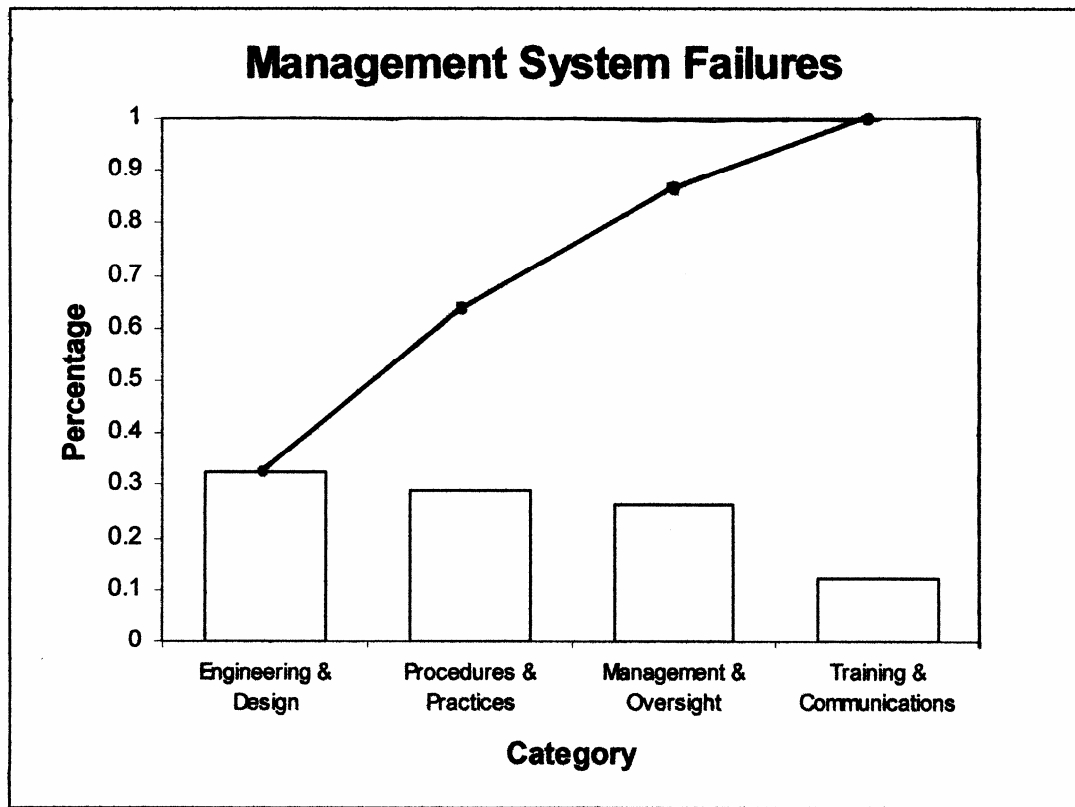
The root causes of the 83 incidences are pretty evenly split between the causes Engineering & Design, Procedures & Practices, and Management & Oversight.

2.6   The countries for all the European PWB manufacturers were combined and the following bar graph was produced:

PWB Manufacturers by Country of Origin

| Country	Value
Japan	25
USA	11
Taiwan	8
Europe	7
Hong Kong	5
S Korea	4
Thailand	1

While not a leading location of PWB manufacturers, we don't see a major cause of concern about the viability of the PWB industry in Europe.

2.8   A pie chart for the data appears below:

Analysis of Software Code

"True" 10%

"False" 90%
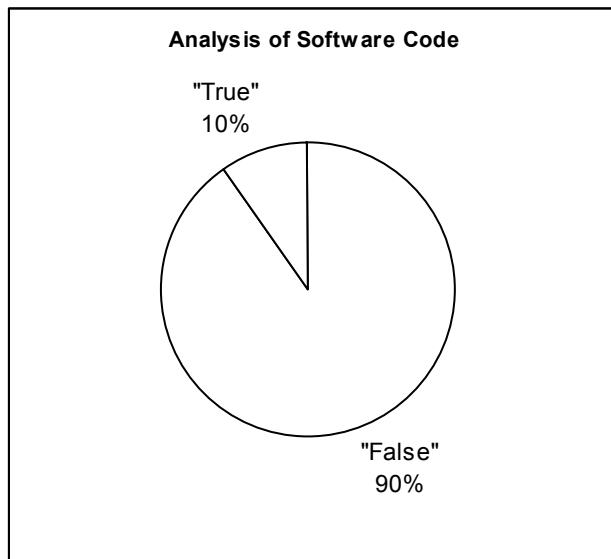
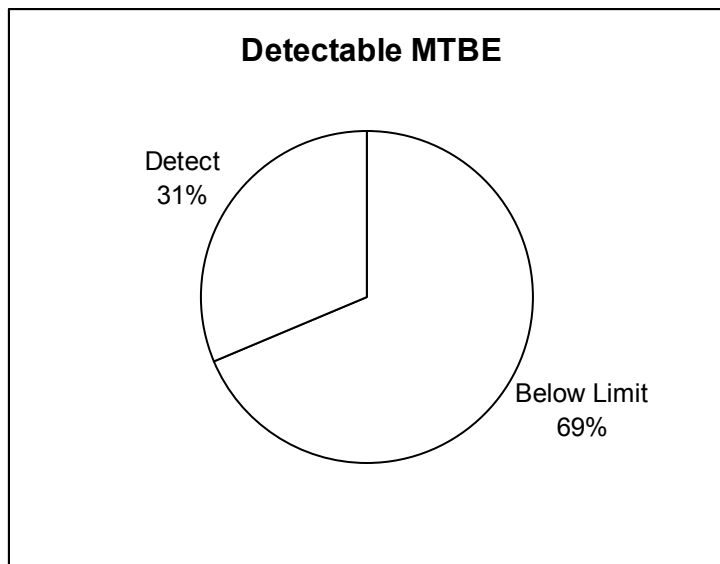It appears that a very high percentage of the software code is defect free.

2.10   A bar graph was used to describe the aquifer variable below:
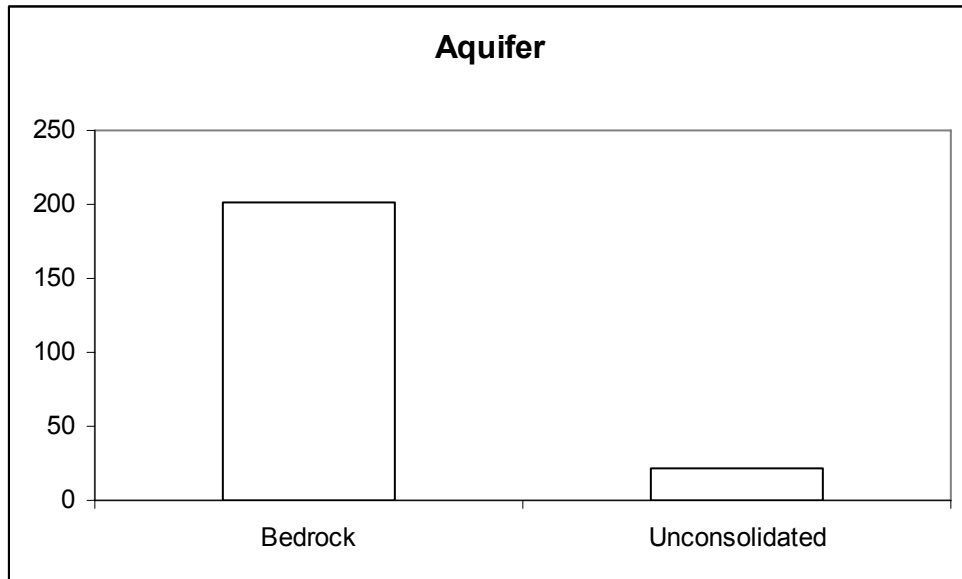
Aquifer

It appears that most of the aquifer types are bedrock

A pie chart was used to describe the detectable levels of MTBE information below:

Detectable MTBE

It appears that twice as many wells are below that detectable level than are above the detectable level.

A bar graph was used to describe the well class variable below:

Well Class
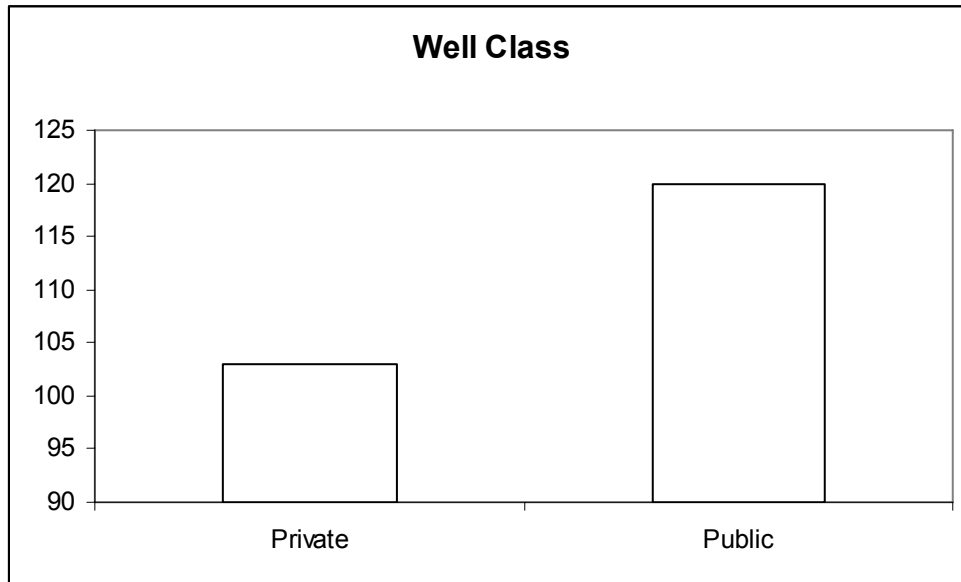
There are more public than private wells in the data set.

2.12   a.   We will allow each number to represent the stems and mark the leaves with an asterisk
              (*). The stem-and-leaf display is:

Stems	Leaves
1	**********
2	*************
3	***********
4	****
5	
6	**

        b.   10 of the 40 or 10/40 = .25 of the asteroid observations resulted in exactly 1 spectral
              image exposure.

2.14   a.   To construct a frequency histogram, first calculate the range by subtracting the smallest
              endpoint of the histogram from the largest.

              To include the largest and smallest values, we will start the histogram at 2.12 and end
   at 10.7625. Range = 10.7625 − 7.95 = 2.8125.

              Next, find the class width for 9 classes by dividing the range by the number of classes:

              Class width = 2.8125/9 = .3125

The first class will begin at 7.95, below the smallest voltage reading. The classes are shown below:

Class	Class Interval	Data Tabulation	Frequency	Relative Frequency
1	7.9500 – 8.2625	|	1	.03
2	8.2625 – 8.5750		0	0
3	8.5750 – 8.8875	‖|	3	.10
4	8.8875 – 9.2000		0	0
5	9.2000 – 9.5125		0	0
6	9.5125 – 9.8250	ⅣⅠ	5	.17
7	9.8250 – 10.1375	ⅣⅠ ⅣⅠ ⅣⅠ	15	.50
8	10.1375 – 10.4500	ⅣⅠ	5	.17
9	10.4500 – 10.7625	|	1	.03
			Totals $n = 30$	1.00

To obtain the class frequency, count the number of observations that fall within each class interval. The class relative frequency is the class frequency divided by the total number of observations (30).
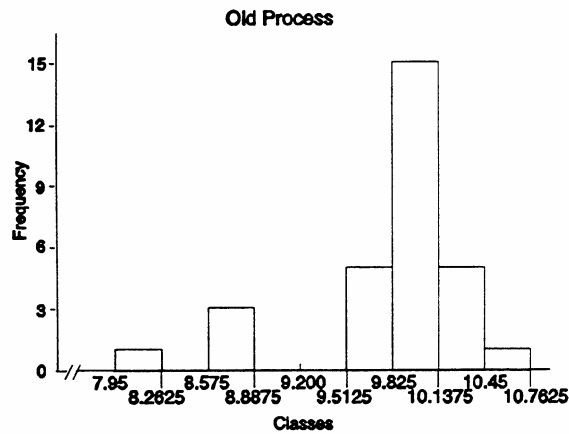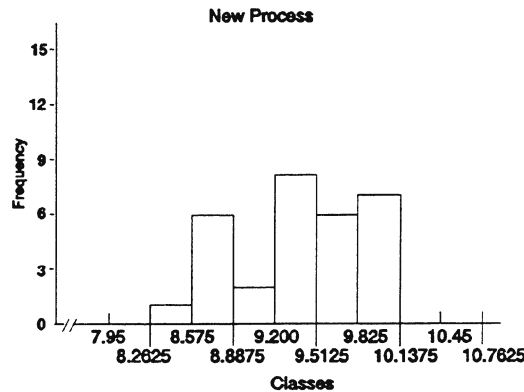

b.

Stems	Leaves
8.	05 72 72 80
9.	98 97 87 80 87 55 95 70 84 80 73 98 84
10.	26 05 29 03 55 26 12 05 15 00 15 02 01

The histogram in part **a** more effectively describes how the data falls.

c.   Using the same classes as in part **a**, our histogram now becomes:

New Process

d.   The new process appears to be worse than the old process. More voltage readings are less than 9.2 volts with the new process than with the old process.

2.16   Statistix was used to generate the following stem-and-leaf plot for the surface roughness data:

```
Stem and Leaf Plot of ROUGH

Leaf Digit Unit = 0.1              Minimum   1.0600
1   0   represents 1.0             Median    2.0400
                                   Maximum   2.6400
          Stem   Leaves
     3     1    001
     5     1    22
     7     1    45
     8     1    7
     9     1    9
    (5)    2    00111
     6     2    23
     4     2    455
     1     2    6
```

The data appear to be pretty evenly spread out across the entire range of the data.

2.18   a.   The stem will be the first decimal value (tenths) and the leaves will be the second and third decimal values (hundredths and thousandths). The stem-and-leaf display is:

Stems	Leaves
.1	(12)
.2	(70)(41)(05)(70) 25 (39)
.3	30 [75]
.4	
.5	(23) [91]
.6	[18]

b.   The three oxon/thion ratios for the clear air days have been marked in boxes above. The ratios for the fog air days have been circled above. The clear air days do seem to produce oxion/thion ratios that are larger than the fog air days. Keep in mind, however, that such a statement cannot be made with any measure of reliability (yet).

2.20   A stem-and-leaf display for the ratios is shown below:

```
          Stem and Leaf Plot of RATIO

    Leaf Digit Unit = 0.1              Minimum   2.2500
    2   2   represents 2.2             Median    3.3750
                                       Maximum   5.0600
              Stem  Leaves
        1      2    2
        5      2    5779
       (9)     3    012223333
       12      3    678889
        6      4    0000
        2      4    5
        1      5    0
```

Only two of the 26 till ratios exceed the value of 4.5, so we would estimate this proportion to be $2/26 = 0.0769$.

2.22   a.   The sample mean radioactivity level is:

$$\bar{y} = \frac{\sum y}{n} = \frac{-43.75}{9} = -4.861$$

The median is the middle observation once they have been ordered. The 5th observation is -4.85. Thus the median is -4.85.

The mode is -5.00.

b.   The average radioactivity level is -4.861. Half of the radioactivity levels are less than -4.85 and half are greater. The modal observation occurred 2 times.

2.24   a.   The mean of the data is $\bar{y} = \frac{\sum y}{n} = \frac{11.77}{8} = 1.4713$.

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

1.37, 1.41, 1.42, 1.48, 1.50, 1.51, 1.53, 1.55

The middle two numbers are 1.48 and 1.50. The median is $\frac{1.48+1.50}{2} = 1.49$

c.   Since the mean is less than the median, the data are somewhat skewed to the left.

2.26   The following measures of central tendency were calculated for the 174 ship sanitation scores:

```
    Descriptive Statistics

    Variable        N       Mean      Median      Mode
    Rating         174     94.420     95.000     97.000
```

The average sanitation score of the 174 ships sampled was 94.420.

Half of the 174 ships sampled had a sanitation score that was below 95 and half had a sanitation score that was above 95.

The most frequently occurring sanitation score among the 174 ships sampled was the score 97.

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

$$\bar{y} = \frac{\sum y}{n} = \frac{3+3+\cdots+4}{11} = \frac{141}{11} = 12.82$$

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:

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

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:

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

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.

2.30 a. The mean number of power plants is:

$$\bar{y} = \frac{y}{n} = \frac{5+3+\cdots+3}{20} = \frac{80}{20} = 4$$

The median is the mean of the middle 2 numbers once the data have been arranged in order:

1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 7, 9, 13

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

The mode is the value with the highest frequency. Since 1 occurs 5 times, 1 is the mode.

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

$$\bar{y} = \frac{\sum y}{n} = \frac{5+3+\cdots+3}{19} = \frac{67}{19} = 3.526$$

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 mode is the value with the highest frequency. Since 1 occurs 5 times, 1 is the mode.

By dropping the largest measurement from the data set, the mean drops from 4 to 3.526. The median drops from 3.5 to 3. There is no effect on the mode.

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:

$$\bar{y} = \frac{\sum y}{n} = \frac{5+3+\cdots+3}{16} = \frac{56}{16} = 3.5$$

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

2.32 a. Range = 13 − 1 = 12

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{498 - \frac{80^2}{20}}{20-1} = \frac{178}{19} = 9.3684$$

$$s = \sqrt{s^2} = \sqrt{9.3584} = 3.061$$

b. Dropping the largest measurement:

Range = 9 − 1 = 8

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{329 - \frac{67^2}{19}}{19-1} = \frac{92.7368421}{18} = 5.1520$$

$$s = \sqrt{s^2} = \sqrt{5.1520} = 2.270$$

By dropping the largest observation from the data set, the range decreased from 12 to 8, the variance decreased from 9.3684 to 5.1520 and the standard deviation decreased from 3.061 to 2.270.

c. Dropping the largest and smallest measurements:

Range = 9 − 1 = 8

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{328 - \frac{66^2}{18}}{18-1} = \frac{86}{17} = 5.0588$$

$$s = \sqrt{s^2} = \sqrt{5.0588} = 2.249$$

By dropping the largest and smallest observations from the data set, the range decreased from 12 to 8, the variance decreased from 9.3684 to 5.0588 and the standard deviation decreased from 3.061 to 2.249.

2.34 Comparing the means of the two distributions, we see that the center of the "true" distribution will be located to the right of the "false distribution when viewed on a number line.

Comparing the standard deviations of the two distributions, we see that the spread of the "true" distribution will be much greater than the spread of the "false distribution. Graphically, this would be represented by a "true" distribution that looked shorter and more spread out than the "false" distribution.

2.36 In mound-shaped symmetric distributions, the Empirical Rule tells us to expect approximately 95% of the data to fall within two standard deviations of the mean.

$$\bar{y} \pm 2s \Rightarrow 19.5 \pm 2(4.7) \Rightarrow (10.1,\ 28.9)$$

We expect approximately 95% of the SNR values in the population to fall between 10.1 and 28.9.

We would not expect to see an SNR value outside of this interval. Therefore, a value of 30 would not be expected.

2.38   a.   Since no information is given about the distribution of the velocities of the Winchester bullets, we can only use Chebyshev's Rule to describe the data. We know that at least 3/4 of the velocities will fall within the interval:

$$\bar{y} \pm 2s \Rightarrow 936 \pm 2(10) \Rightarrow 936 \pm 20 \Rightarrow (916,\ 956)$$

Also, at least 8/9 of the velocities will fall within the interval:

$$\bar{y} \pm 3s \Rightarrow 936 \pm 3(10) \Rightarrow 936 \pm 30 \Rightarrow (906,\ 966)$$

     b.   Since a velocity of 1,000 is much larger than the largest value in the second interval in part **a**, it is very unlikely that the bullet was manufactured by Winchester.

2.40   a.   From the printout, $\bar{y} = 2.425$ and $s = 1.259$

     b. $\bar{y} \pm s \Rightarrow 2.425 \pm 1.259 \Rightarrow (1.166,\ 3.684)$

        $\bar{y} \pm 2s \Rightarrow 2.425 \pm 2(1.259) \Rightarrow 2.425 \pm 2.518 \Rightarrow (-0.093,\ 4.943)$

        $\bar{y} \pm 3s \Rightarrow 2.425 \pm 3(1.259) \Rightarrow 2.425 \pm 3.777 \Rightarrow (-1.352,\ 6.202)$

     c.   24 observations fall in the interval $\bar{y} \pm s$ or 24/40 = .60. The Empirical Rule says there should be approximately .68 of the measurements within 1 standard deviation of the mean. This is fairly close.

        38 observations fall in the interval $\bar{y} \pm 2s$ or 38/40 = .95. The Empirical Rule says there should be approximately .95 of the measurements within 2 standard deviations of the mean. This agrees with the Empirical Rule.

        40 observations fall in the interval $\bar{y} \pm 3s$ or 40/40 = 1.00. The Empirical Rule says approximately all of the measurements should fall within 3 standard deviations of the mean. Again, this agrees with the Empirical Rule.

2.42   a.   The lower quartile is the observation with rank equal to:

$$\ell = 1/4(n + 1) = 1/4(40 + 1) = 10.25 \Rightarrow P = 10$$

$$Q_L = y_{(10)} = 1$$

The upper quartile is the observation with rank equal to:

$$u = 3/4(n + 1) = 3/4(40 + 1) = 30.75 \Rightarrow u = 31$$

$$Q_U = y_{(31)} = 3$$

25% of the observations fall at or below 1 and 75% of the observations fall at or below 3.

b.   The z-score for an observation is $z = \dfrac{y - \bar{y}}{s}$

$$\bar{y} = \frac{\sum y}{n} = \frac{97}{40} = 2.425$$

$$s^2 = \frac{\sum y - \dfrac{\left(\sum y\right)^2}{n}}{n-1} = \frac{297 - \dfrac{(97)^2}{40}}{40 - 1} = 1.584$$

$$s = \sqrt{s^2} = \sqrt{1.584} = 1.259$$

The z-score for 6 is $z = \dfrac{6 - 2.425}{1.259} = 2.84$

This observation falls 2.84 standard deviations above the mean of 2.425.

2.44   a.   Using SAS, the output is:

```
            EGG LENGTH                                    1

                    The UNIVARIATE Procedure
                      Variable:  length

                          Moments

N                        130    Sum Weights               130
Mean             60.6538462    Sum Observations         7885
Std Deviation    43.9861168    Variance           1934.77847
Skewness          2.1091362    Kurtosis           4.70026314
Uncorrected SS       727842    Corrected SS       249586.423
Coeff Variation  72.5199136    Std Error Mean     3.85783765

                   Basic Statistical Measures

        Location                        Variability

    Mean      60.65385    Std Deviation           43.98612
    Median    49.50000    Variance                    1935
    Mode      35.00000    Range                 220.00000
                          Interquartile Range    32.00000
```

```
                    Tests for Location: Mu0=0

          Test              -Statistic-      -----p Value------

          Student's t     t  15.72224      Pr > |t|    <.0001
          Sign            M        65      Pr >= |M|   <.0001
          Signed Rank     S    4257.5      Pr >= |S|   <.0001


                         Quantiles (Definition 5)

                         Quantile       Estimate

                         100% Max          236.0
                         99%               218.0
                         95%               160.0
                         90%               122.0
                         75% Q3             67.0
                         50% Median         49.5
                         25% Q1             35.0
                         10%                23.0
                         5%                 19.0
                         1%                 16.0
                         0% Min             16.0


                           EGG LENGTH                   2

                         The UNIVARIATE Procedure
                           Variable:  length

                         Extreme Observations

                 ----Lowest----           ----Highest---

                 Value       Obs          Value       Obs

                 16.0        102            195        128
                 16.0         90            205        123
                 17.0        101            216        122
                 18.0        103            218        129
                 18.5        100            236        130


                            Missing Values

                                   -----Percent Of-----
                 Missing                           Missing
                   Value       Count     All Obs       Obs

                     .           2        1.52       100.00
```

The 10th percentile egg length is 23.

b. From the printout, $\bar{y} = 60.65$ and $s = 43.99$. The z-score corresponding to the Moas, P. australis bird species' egg length is:

$$z = \frac{y - \bar{y}}{s} = \frac{205 - 60.65}{43.99} = 3.28$$

The *z*-score for Moas, P. australis is 3.28 standard deviations above the mean. This is a fairly large value for a *z*-score. This indicates that the egg length for Moas, P. australis could be very unusual.

2.46   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.

2.48   We will use SAS to compute the percentiles. The output is:

```
                    The UNIVARIATE Procedure
                      Variable:  LEFTEYE

                            Moments

   N                        25    Sum Weights               25
   Mean                -0.1544    Sum Observations       -3.86
   Std Deviation    0.19676721    Variance          0.03871733
   Skewness         -4.5220839    Kurtosis          21.7019614
   Uncorrected SS       1.5252    Corrected SS        0.929216
   Coeff Variation    -127.4399   Std Error Mean    0.03935344

                    Basic Statistical Measures

          Location                      Variability

       Mean     -0.15440    Std Deviation          0.19677
       Median   -0.11000    Variance               0.03872
       Mode     -0.17000    Range                  1.03000
                            Interquartile Range    0.08000

   NOTE: The mode displayed is the smallest of 4 modes with a count of 3.

                    Tests for Location: Mu0=0

          Test             -Statistic-    -----p Value------

          Student's t     t  -3.92342    Pr > |t|    0.0006
          Sign            M     -12.5    Pr >= |M|   <.0001
          Signed Rank     S    -162.5    Pr >= |S|   <.0001

                    Quantiles (Definition 5)

                    Quantile       Estimate

                    100% Max         -0.04
                    99%              -0.04
                    95%              -0.06
                    90%              -0.06
                    75% Q3           -0.08
                    50% Median       -0.11
                    25% Q1           -0.16
                    10%              -0.20
                    5%               -0.21
                    1%               -1.07
                    0% Min           -1.07
```

```
                    Extreme Observations

            ----Lowest----          ----Highest---

            Value        Obs         Value        Obs

            -1.07          7         -0.07         23
            -0.21         18         -0.06          2
            -0.20         11         -0.06         12
            -0.17         22         -0.06         21
            -0.17         17         -0.04         16
```

a. From the printout above, the 10th percentile is -0.20. Thus, 10% of the cylinder power measurements are below -0.20 and 90% are above -0.20.

b. From the printout above, the 95th percentile is -0.06. Thus, 95% of the cylinder power measurements are below -0.06 and 5% are above -0.06.

c. $z = \dfrac{y - \bar{y}}{s} = \dfrac{-1.07 - (-.1544)}{.19677} = -4.65$. A power measurement of -1.07 is 4.65 standard deviations below the mean. Since this z-score is so small, it is an extremely unlikely value to observe. The cylinder value of -1.07 is an extreme value.

2.50 a. From the data, median $= 2.000$, $Q_L = 1.250$, and $Q_U = 3.000$. The interquartile range is IQR $= Q_U - Q_L = 3.000 - 1.250 = 1.750$. The inner and outer fences are located a distance of $1.5(\text{IQR}) = 1.5(1.75) = 2.625$ and $3(\text{IQR}) = 3(1.75) = 5.25$ below Q_L and above Q_U, respectively. Values between the inner and outer fences are suspect outliers and are designated by *. Highly suspect outliers lie outside the outer fences and are designated by o. The closest points to the inner fences which are still inside the inner fences are marked by x and whiskers are drawn between these points and the box. The box plot is shown below:
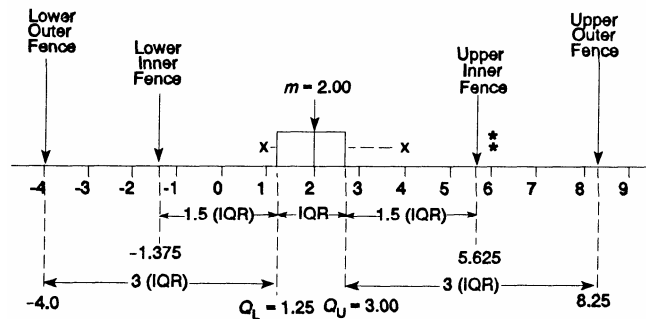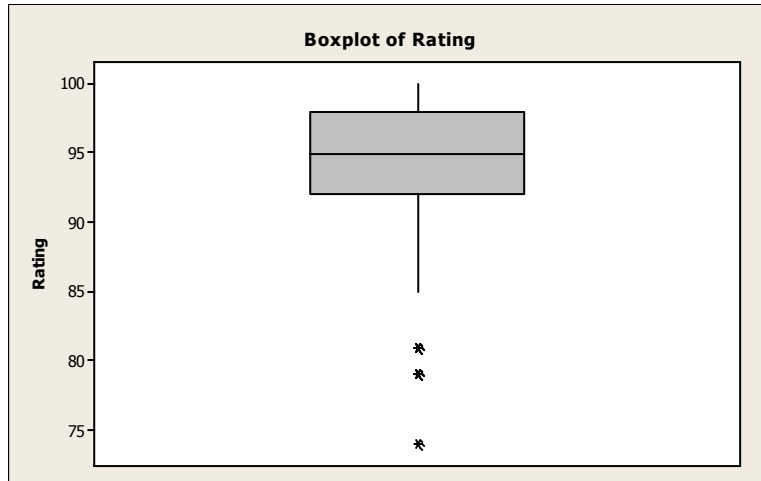

b. From Exercise 2.40, $\bar{y} = 2.425$ and $s = 1.259$. The z-score corresponding to the suspect outliers is:

$$z = \frac{6 - 2.425}{1.259} = 2.84$$

Since this value is fairly close to 3, the two observations with the value 6 are suspect outliers.

2.52   a.   Using MINITAB, the box plot of the data is:

Boxplot of Rating

The descriptive statistics are:

Descriptive Statistics: Rating

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Rating	174	94.420	95.000	94.814	4.380	0.332

Variable	Minimum	Maximum	Q1	Q3
Rating	74.000	100.000	92.000	98.000

The median is 95, the upper quartile is $Q_U = 98$ and the lower quartile is $Q_L = 92$. The interquartile range is $IQR = Q_U - Q_L = 98 - 92 = 6$.

The lower inner fence is $Q_L - 1.5(IQR) = 92 - 1.5(6) = 92 - 9 = 83$.

The upper inner fence is $Q_U + 1.5(IQR) = 98 + 1.5(6) = 98 + 9 = 107$.

The lower outer fence is $Q_L - 3(IQR) = 92 - 3(6) = 92 - 18 = 74$.

The upper outer fence is $Q_U + 3(IQR) = 98 + 3(6) = 98 + 18 = 116$.

There are 5 observations below the lower inner fence. These are suspect outliers. The observations have values 74, 79, 79, 81, and 81. There are no observations below the lower outer fence.

b.   The z-scores corresponding to these points are:

$$74: \quad z = \frac{y - \bar{y}}{s} = \frac{74 - 94.42}{4.38} = -4.66 \qquad 79: \quad z = \frac{y - \bar{y}}{s} = \frac{79 - 94.42}{4.38} = -3.52$$

$$81: \quad z = \frac{y - \bar{y}}{s} = \frac{81 - 94.42}{4.38} = -3.06$$

c.   According to the box plot, there are 5 suspect outliers. According to the z-score method, all of the suspect outliers have z-scores more than 3 standard deviations from the mean. They would all be considered outliers.

2.54   Statistix was used to get the following information:

```
Breakdown for TRANERR

              PROJECTIONS
INTRINSICS                No        Yes       Total

No        N                 1          5          6
          Mean    0.0000034     25.160     20.967
          SD              M      6.8002     11.937

Yes       N                 5          0          5
          Mean       1.6200          M     1.6200
          SD         0.7918          M     0.7918

Total     N                 6          5         11
          Mean       1.3500     25.160     12.173
          SD         0.9690      6.8002     13.175

Cases Included 11     Missing Cases 0
```

a.   For trials with perturbed intrinsics but no perturbed projections, we find: $\bar{y}$ = 1.62 and s = 0.7918.

b.   For trials with perturbed projections but no perturbed intrinsics, we find: $\bar{y}$ = 25.16 and s = 6.8002.

c.   To determine if the value is an outlier, we calculate the z-score for the observation.

For trials with perturbed intrinsics but no perturbed projections, we find:

$$z = \frac{y - \bar{y}}{s} = \frac{4.5 - 1.62}{0.7918} = 3.637$$

For trials with perturbed projections but no perturbed intrinsics, we find:

$$z = \frac{y - \bar{y}}{s} = \frac{4.5 - 25.16}{6.8002} = -3.038$$

This observation is considered an outlier in both these situations. We would not expect it to come from either of these camera perturbations. If we had to choose, we would select the trials with perturbed projections but no perturbed intrinsics since the z-score for that camera perturbation is the smallest (in absolute value).

2.56   a.   The variable "reason" is qualitative because the responses are categories. There are 7 values for the variable – Infant, Child, Medical, Infant & Medical, Child & Medical, Infant & Child, and Infant, Child & Medical.

b.   To compute the relative frequencies, divide the frequencies (or number of requests) by the total sample size of 30,337. The relative frequency for Infant is 1,852 / 30,337 = .061. The rest of the relative frequencies are computed in the table.

Reason	Number of Requests	Computation	Relative Frequency
Infant	1,852	1,852 / 30,337	.061
Child	17,148	17,148 / 30,337	.565
Medical	8,377	8,377 / 30,337	.276
Infant & Medical	44	44 / 30,337	.001
Child & Medical	903	903 / 30,337	.030
Infant & Child	1,878	1,878 / 30,337	.062
Infant & Child & Medical	135	135 / 30,337	.004
Total	30,337		.999

c.   A bar chart for the data is:
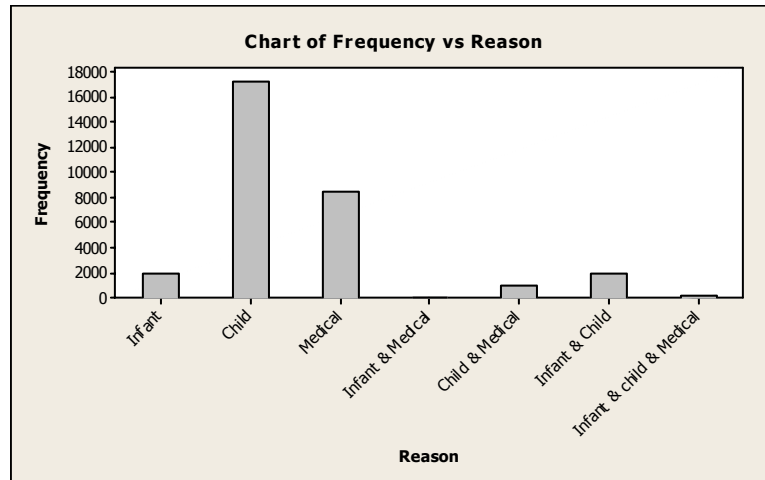
d.   The proportion of car owners who requested on-off bag switches who gave Medical as one of the reasons is $(8,377 + 44 + 903 + 135) / 30,337 = 9,459 / 30,337 = .312$.

2.58   a.   Mean = 2.1197: the average aftershock of these 2929 earthquake aftershocks was 2.1197 measured on the Richter scale.

Median = 2.000: half of the 2929 earthquake aftershocks registered above 2.000 on the Richter scale and half registered below 2.000 on the Richter scale.

Q1 = 1.700: 25% of the 2929 earthquake aftershocks registered below 1.700 on the Richter scale and 75% registered above 1.700 on the Richter scale.

Q3 = 2.400: 75% of the 2929 earthquake aftershocks registered below 2.400 on the Richter scale and 25% registered above 2.400 on the Richter scale.

$s = 0.6636$: Assuming a mound-shaped symmetric distribution, the Empirical Rule tells us that approximately 95% of the aftershocks register between $\bar{y} \pm 2s \Rightarrow 2.1197 \pm 2(0.6636) \Rightarrow (0.7925, 3.4469)$ on the Richter scale.

b. Statistix constructed the following box plot for the aftershock magnitude data:
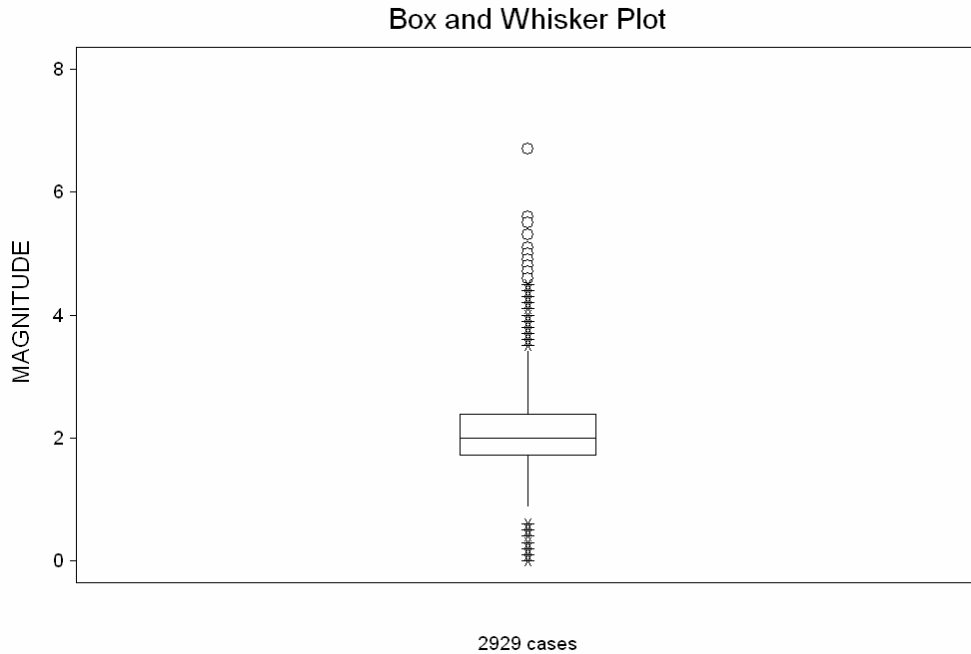
We can see that there are number suspects (shown as the *'s) and highly suspect outliers (shown as o's) that are identified in this box plot.

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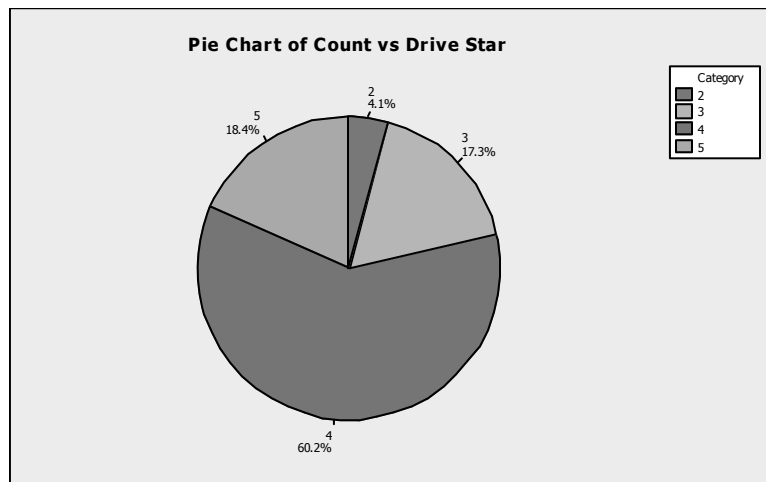
More than half of the cars received 4 star ratings (60.2%). A little less than a quarter of the cars tested received ratings of 3 stars or less.

2.62   Excel was used to generate the following summary statistics:

RATIO	
Mean	3.50731
Standard Error	0.12444
Median	3.375
Mode	4.09
Standard Deviation	0.63454
Sample Variance	0.40264
Kurtosis	0.2953
Skewness	0.27643
Range	2.81
Minimum	2.25
Maximum	5.06
Sum	91.19
Count	26

a.   Mean = 3.50731: the average of the 26 till ratios is 3.50731.
Median = 3.375: half of the 26 till ratios fall above 3.375, half fall below 3.375.
Mode = 4.09: the most frequently occurring till ratio is the value 4.09.

b.   Range = 2.81: the difference between the largest and the smallest till ratio is 2.81.

Variance = 0.40264: the variance has no useful interpretation. It is the calculation that allows us to determine the value of the standard deviation.

Standard Deviation = 0.63454: Assuming a mound-shaped symmetric distribution, the Empirical Rule tells us that approximately 95% of the till ratios fall between $\bar{y} \pm 2s \Rightarrow$ $3.50731 \pm 2(0.63454) \Rightarrow (2.23823, 4.77639)$.

c.   Statistix constructed the following box plot for the till ratio data:

Box and Whisker Plot
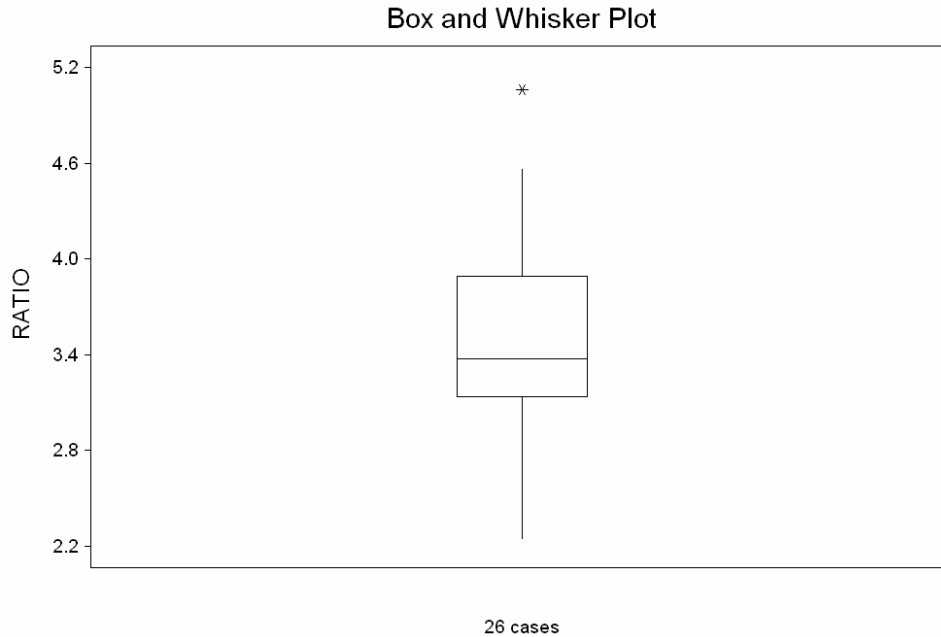
We see one suspect outlier (5.06) identified in the box plot.

2.64   a.   To convert a frequency histogram to a relative frequency histogram, we must first
divide each of the frequencies by the sum of all the frequencies, which is $1 + 1 + 3 + 4$
$+ 10 + 7 + 5 + 4 + 3 + 2 + 3 + 2 + 3 + 2 = 50$. The relative frequency table is:

Class Interval	Frequency	Relative Frequency
5.5−15.5	1	1/50 = .02
15.5−25.5	1	1/50 = .02
25.5−35.5	3	3/50 = .06
35.5−45.5	4	4/50 = .08
45.5−55.5	10	10/50 = .20
55.5−65.5	7	7/50 = .14
65.5−75.5	5	5/50 = .10
75.5−85.5	4	4/50 = .08
85.5−95.5	3	3/50 = .06
95.5−105.5	2	2/50 = .04
145.5−155.5	3	3/50 = .06
165.5−175.5	2	2/50 = .04
175.5−185.5	3	3/50 = .06
195.5−205.5	2	2/50 = .04

The relative frequency histogram is:
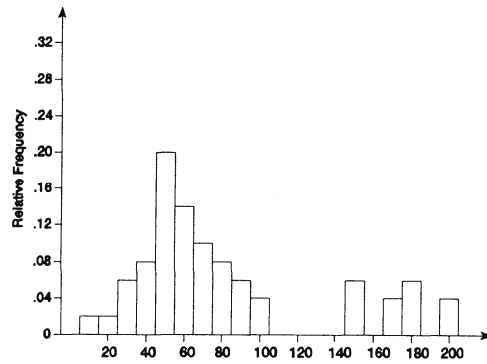
b.   It would be very unusual to observe a drill chip with a length of at least 190 mm. There are only 2 out of 50 drill chips that are 190 mm or longer. The proportion of drill chips with lengths of at least 190 mm is .04.

2.66   a.   To construct a stem and leaf display for the PCB levels of rural soil samples, we will use the digits to the left of the decimal point as the stems and the digits to the right of the decimal point as the leaves.

Stems	Leaves	Frequency	Relative Frequency
1	0 5 6 8	4	.286
2		0	.000
3	5	1	.071
4		0	.000
5	3	1	.071
6			
7			
8	1 2	2	.143
9	0 7 8	3	.214
10			
11			
12	0	1	.071
13			
14			
15	0	1	.071
⋮			
23	0	1	.071

Totals $n = 14$ .998

b. For the urban samples, the stem-and-leaf display is:

Stems	Leaves	Frequency	Relative Frequency
11	0 0	2	.133
12	0	1	.067
13	0	1	.067
14			
15			
16	0	1	.067
17		0	.000
18	0 0	2	.133
19			
20			
21	0	1	.067
22	0	1	.067
23		0	.000
24	0	1	.067
!			
29	0	1	.067
!			
49	0	1	.067
!			
94	0	1	.067
!			
107	0	1	.067
!			
141	0	1	.067
		Totals $n = 15$	1.003

| | | | Relative |
Stems	Leaves	Frequency	Frequency
1	0 5 6 8	4	.133
2		0	.000
3	5	1	.034
4		0	.000
5	3	1	.034
6			
7			
8	1 2	2	.069
9	0 7 8	3	.103
10			
11	⓪⓪	2	.069
12	⓪0	1	.069
13	⓪	1	.034
14			.000
15	0		.034
16	⓪	1	.034
17		0	.000
18	⓪⓪	2	.069
19			
20			
21	⓪	1	.034
22	⓪	1	.034
23	0	0	.034
24	⓪	1	.034
!			
29	⓪	1	
!			
49	⓪	1	.034
!			
94	⓪	1	.034
!			
107	⓪	1	.034
!			
141	⓪	1	.034

Totals $n = 29$ .993

The graph supports the researchers' claim that a significant difference exists between the PCB levels for urban and rural areas. The rural PCB levels are almost all less than the lowest reading for the urban areas.

2.68   a.   mean $= \bar{y} = \sum \dfrac{y}{n} = \dfrac{5891}{50} = 117.82$

median $= \dfrac{117 + 118}{2} = 117.5$,   the average of the middle two observations, once the observations have been arranged in order.

modes = 97, 112, 124, 128, and 131, all of which appear 3 times.

b.   range = $150 - 88 = 62$

$$\text{variance} = s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{705119 - \frac{(5891)^2}{50}}{49} = \frac{11041.38}{49}$$

$$= 225.3343$$

standard deviation $= s = \sqrt{225.3343} = 15.01$

c. $\bar{y} \pm s \Rightarrow 117.82 \pm (15.01) \Rightarrow (102.81,\ 132.83)$

The number of observations that appear in this interval is 31. This is close to the 68% or $.68(50) \approx 34$ given by the Empirical Rule.

$\bar{y} \pm 2s \Rightarrow 117.82 \pm 2(15.01) \Rightarrow 117.82 \pm 30.02 \Rightarrow (87.80,\ 147.84)$

The number of observations that appear in this interval is 49. This is close to the 95% or $.95(50) \approx 48$ given by the Empirical Rule.

$\bar{y} \pm 3s \Rightarrow 117.82 \pm 3(15.01) \Rightarrow 117.82 \pm 45.03 \Rightarrow (72.79,\ 162.85)$

The number of observations that appear in this interval is 50. This agrees with $\approx 100\%$ or $1.00(50) = 50$ given by the Empirical Rule.

d.   To construct a box plot, we first must make some preliminary calculations. First, we compute Q_L, and Q_U. The lower quartile, Q_L, is the data point with the rank of $(n + 1)/4 = (50 + 1)/4 = 12.75 \approx 13$. The 13th ranked data point is 109. The upper quartile, Q_U, is the data point with the rank of $3(n + 1)/4 = 3(50 + 1)/4 = 38.25 \approx 38$. The 38th ranked data point is 131. The interquartile range, IQR, is $Q_U - Q_L = 131 - 109 = 22$. The inner fences are located $1.5(\text{IQR}) = 1.5(22) = 33$ below Q_L and above Q_U. The inner fences are $109 - 33 = 76$ and $131 + 33 = 164$. The outer fences are located $3(\text{IQR}) = 3(22) = 66$ below Q_L and above Q_U. The outer fences are $109 - 66 = 43$ and $131 + 66 = 197$. The box plot is shown below.
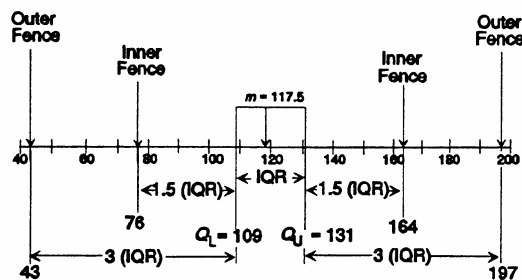
There are no suspect outliers in the data set because no data points lie outside the inner fence.

e. The 70th percentile is the data point that has a rank of $.70(n) = .70(50) = 35$. The 35th data point is 128. Seventy percent of the data points have a value less than or equal to 128.

2.70 a. Using Minitab, a stem-and-leaf display of the data is:

```
Stem-and-leaf of VELOCITY      N = 51
Leaf Unit = 100

 1     18 4
 3     18 79
12     19 001112444
18     19 566788
20     20 12
21     20 7
21     21
23     21 99
(5)    22 11344
23     22 5666777777889
10     23 001222344
 1     23
 1     24
 1     24 9
```

b. From this stem-and-leaf display, it is fairly obvious that there are two different distributions since there are two groups of data.

c. Since there appears to be two distributions, we will compute two sets of numerical descriptive measures. We will call the group with the smaller velocities A1775A and the group with the larger velocities A1775B.

For A1775A:

$$\bar{y} = \frac{\sum y}{n} = \frac{408,707}{21} = 19{,}462.2$$

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{7{,}960{,}019{,}531 - \frac{408{,}707^2}{21}}{21-1} = 283{,}329.3$$

$$s = \sqrt{283{,}329.3} = 532.29$$

For A1775B:

$$\bar{y} = \frac{\sum y}{n} = \frac{685{,}154}{30} = 22{,}838.5$$

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{15{,}656{,}992{,}942 - \frac{685{,}154^2}{30}}{30-1} = 314{,}694.83$$

$$s = \sqrt{314{,}694.83} = 560.98$$

d. To determine which of the two clusters this observation probably belongs to, we will compute *z*-scores for this observation for each of the two clusters.

For A1775A:
$$z = \frac{20{,}000 - 19{,}462.2}{532.29} = 1.01$$

Since this *z*-score is so small, it would not be unlikely that this observation came from this cluster.

For A1775B:

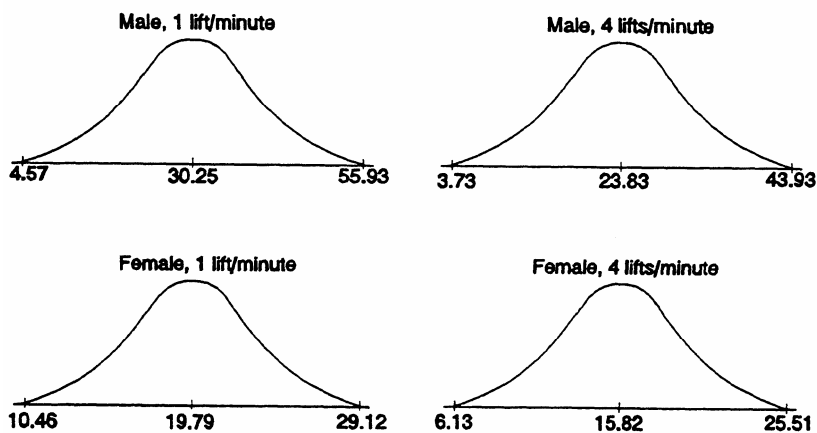$$z = \frac{20{,}000 - 22{,}838.5}{560.98} = -5.06$$

Since this *z*-score is so large (in magnitude), it would be very unlikely that this observation came from this cluster.

Thus, this observation probably came from the cluster A1775A.

2.72 a. To sketch the distributions, we need to calculate the intervals $\bar{y} \pm s$, $\bar{y} \pm 2s$, and $\bar{y} \pm 3s$ for each of the four scenarios.

Sex	Lifts/Minute	$\bar{y} \pm s$	$\bar{y} \pm 2s$	$\bar{y} \pm 3s$
M	1	21.69−38.81	13.13−47.37	4.47−55.93
	4	17.13−30.53	10.43−37.23	3.73−43.93
F	1	16.68−22.90	13.57−26.01	10.46−29.12
	4	12.59−19.05	9.36−22.28	6.13−25.51

The distributions are sketched below:

Male, 1 lift/minute
4.57   30.25   55.93

Male, 4 lifts/minute
3.73   23.83   43.93

Female, 1 lift/minute
10.46   19.79   29.12

Female, 4 lifts/minute
6.13   15.82   25.51

b. The $\bar{y} \pm 2s$ intervals are calculated in part **a**. We expect approximately 95% of the measurements to fall in these calculated intervals.

c. The average male could be expected to safely lift the 25 kilogram box at a rate of 4 lifts per minute based on the $\bar{y} \pm 2s$ interval calculated (10.43−37.23). The average female could not be expected to safely lift the 25 kilogram box at a rate of 4 lifts per minute based on the $\bar{y} \pm 2s$ interval calculated (9.36−22.28).