Chapter 2
Methods for Describing Sets of Data

2.1 First, we find the frequency of the grade A. The sum of the frequencies for all five grades must be 200. Therefore, subtract the sum of the frequencies of the other four 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

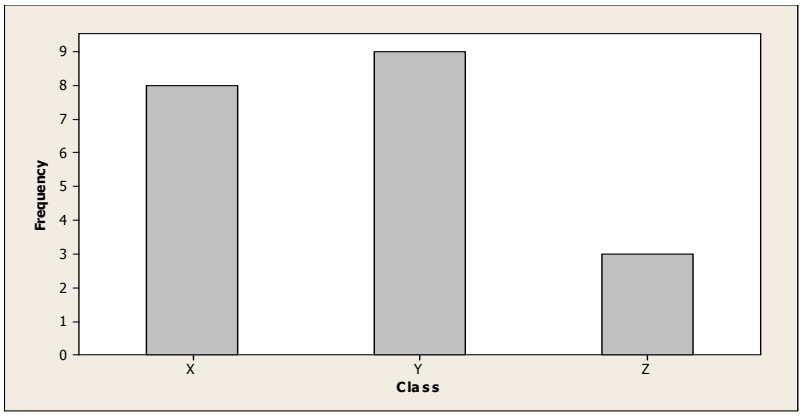
2.2 a. To find the frequency for each class, count the number of times each letter occurs. The frequencies for the three classes are:

Class	Frequency
X	8
Y	9
Z	3
Total	20

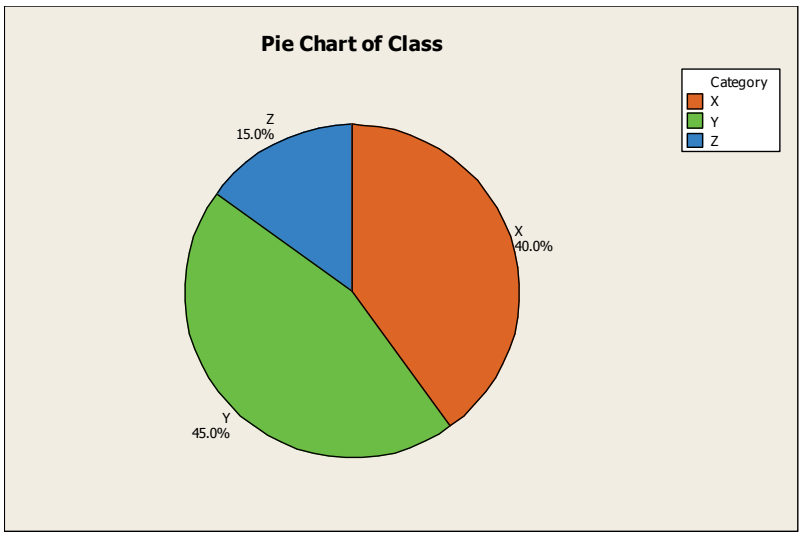
b. The relative frequency for each class is found by dividing the frequency by the total sample size. The relative frequency for the class X is 8/20 = .40. The relative frequency for the class Y is 9/20 = .45. The relative frequency for the class Z is 3/20 = .15.

Class	Frequency	Relative Frequency
X	8	.40
Y	9	.45
Z	3	.15
Total	20	1.00

c. The frequency bar chart is:



d. The pie chart for the frequency distribution is:



2.3 a. The type of graph is a bar graph.

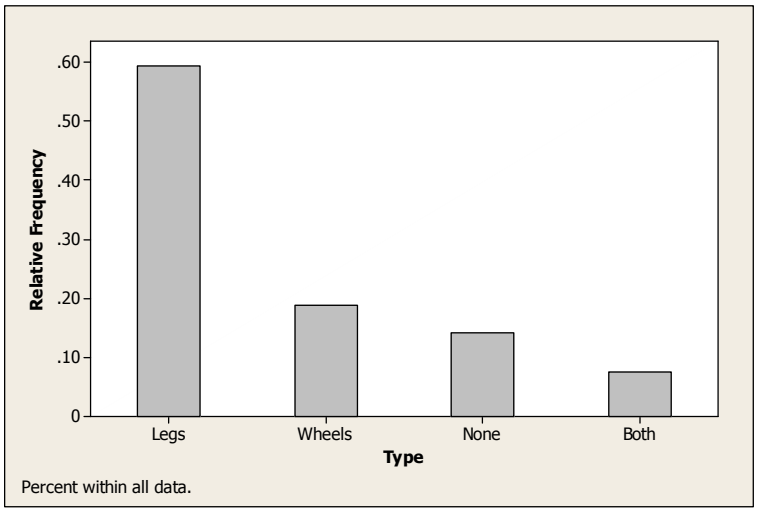
b. The variable measured for each of the robots is type of robotic limbs.

c. From the graph, the design used the most is the "legs only" design.

d. The relative frequencies are computed by dividing the frequencies by the total sample size. The total sample size is $n = 106$. The relative frequencies for each of the categories are:

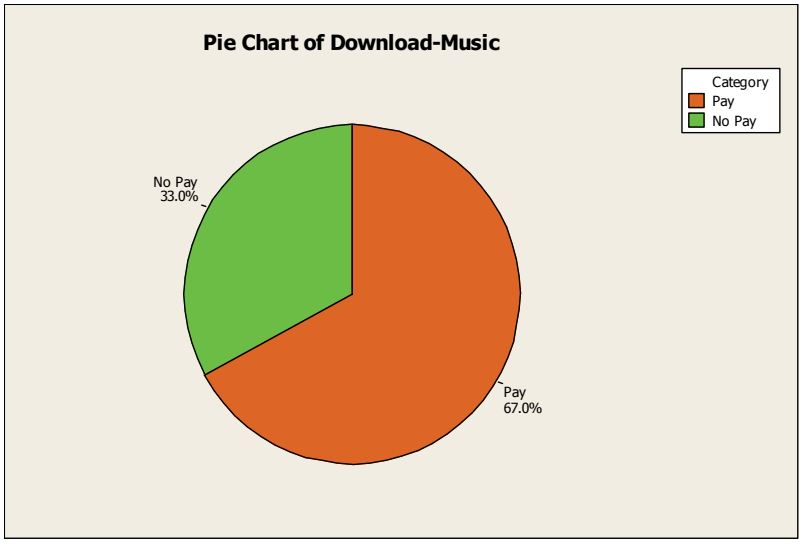
Type of Limbs	Frequency	Relative Frequency
None	15	15/106 = .142
Both	8	8 / 106 = .075
Legs ONLY	63	63/106 = .594
Wheels ONLY	20	20/106 = .189
Total	106	1.000

e. Using MINITAB, the Pareto diagram is:

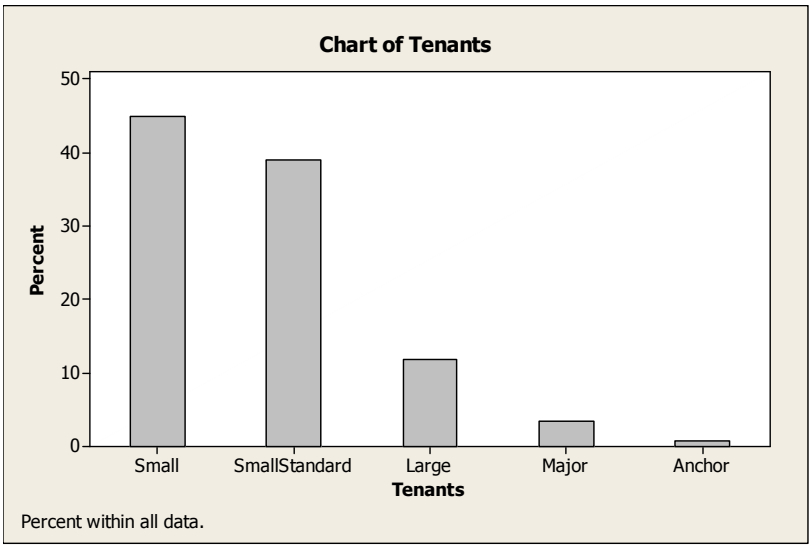


2.4 a. From the pie chart, 50.4% or .504 of the sampled adults living in the U.S. use the internet and pay to download music. From the data, 506 out of 1,003 adults or 506/1,003 = .504 of sampled adults in the U.S. use the internet and pay to download music. These two results agree.

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



2.5 Using MINITAB, the Pareto diagram for the data is:

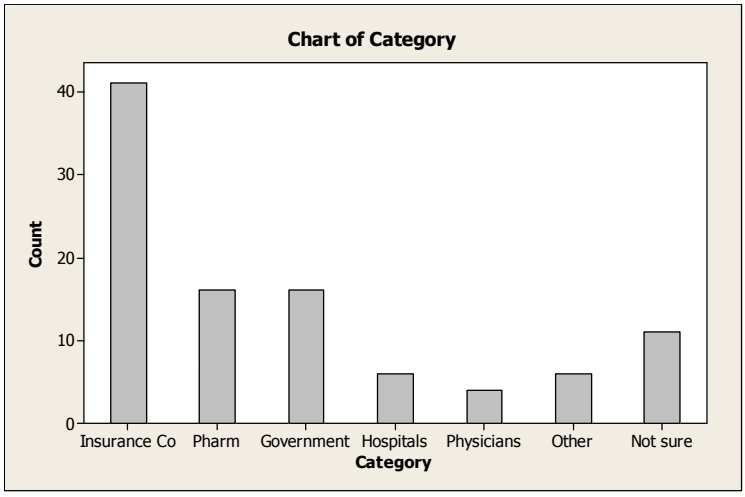


Most of the tenants in UK shopping malls are small or small standard. They account for approximately 84% of all tenants ([711 + 819]/1,821 = .84). Very few (less than 1%) of the tenants are anchors.

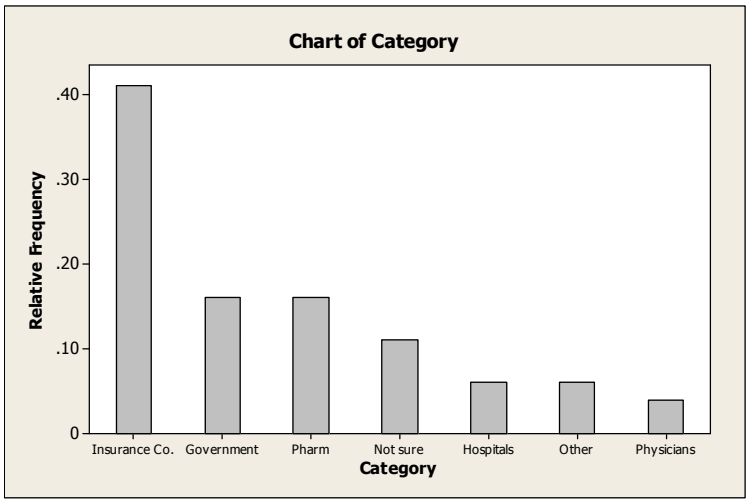
2.6 a. The relative frequency for each response category is found by dividing the frequency by the total sample size. The relative frequency for the category "Insurance Companies" is 869/2119 = .410. The rest of the relative frequencies are found in a similar manner and are reported in the table.

Most responsible for rising health-care costs	***Number responding***	***Relative Frequencies***
Insurance companies	869	869/2119 = .410
Pharmaceutical companies	339	339/2119 = .160
Government	338	338/2119 = .160
Hospitals	127	127/2119 = .060
Physicians	85	85/2119 = .040
Other	128	128/2119 = .060
Not at all sure	233	233/2119 = .110
TOTAL	2,119	1.000

b. Using MINITAB, the relative frequency bar chart is:



c. Using MINITAB, the Pareto diagram is:

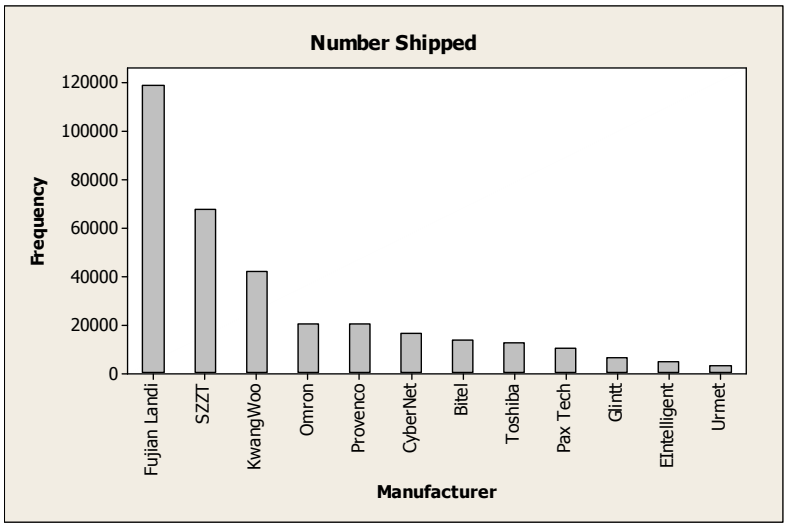


Most American adults in the sample (41%) believe that the Insurance companies are the most responsible for the rising costs of health care. The next highest categories are Government and Pharmaceutical companies with about 16% each. Only 4% of American adults in the sample believe physicians are the most responsible for the rising health care costs.

2.7 a. Since the variable measured is manufacturer, the data type is qualitative.

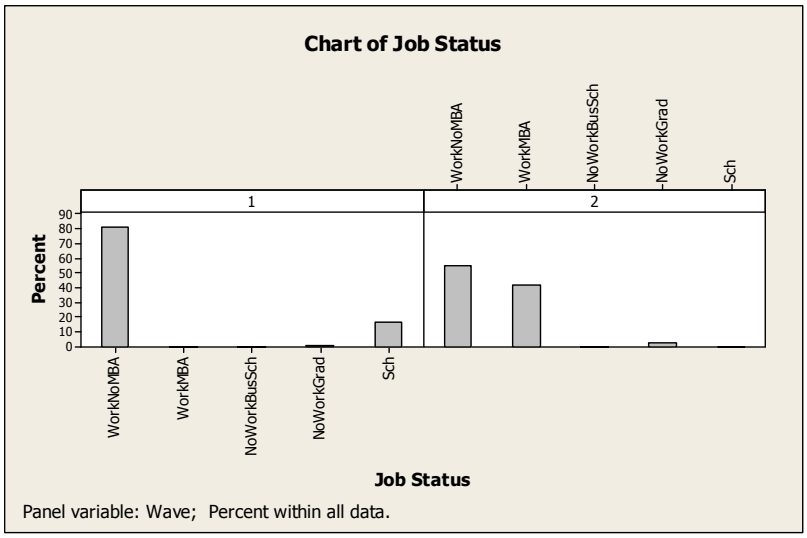
b. Using MINITAB, a frequency bar chart for the data is:

c. Using MINITAB, the Pareto diagram is:



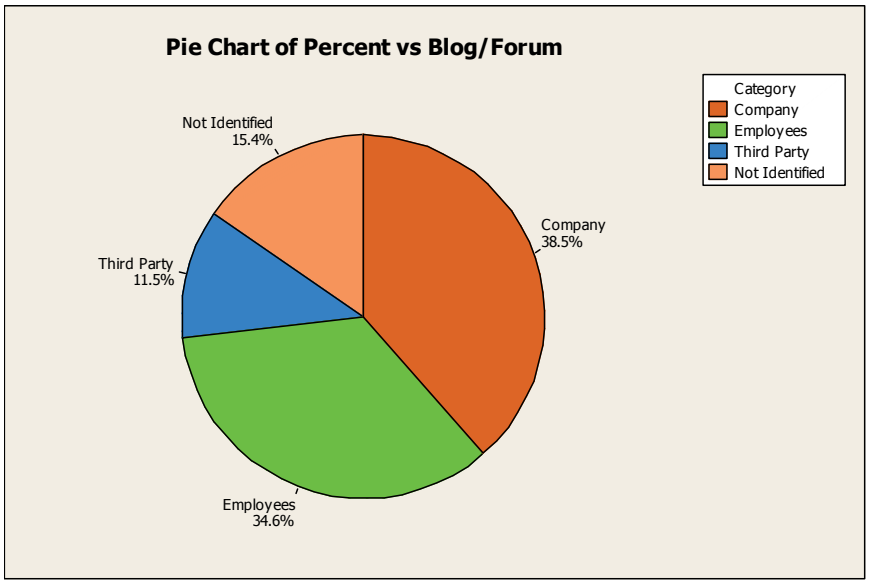
Most PIN pads shipped in 2007 were manufactured by either Fujian Landi or SZZT Electronics. These two categories make up (119,000 + 67,300)/334,039= 186,300/334,039 = .558 of all PIN pads shipped in 2007. Urmet shipped the fewest number of PIN pads among these 12 manufacturers.

2.8 Using MINITAB, the bar graphs of the 2 waves is:



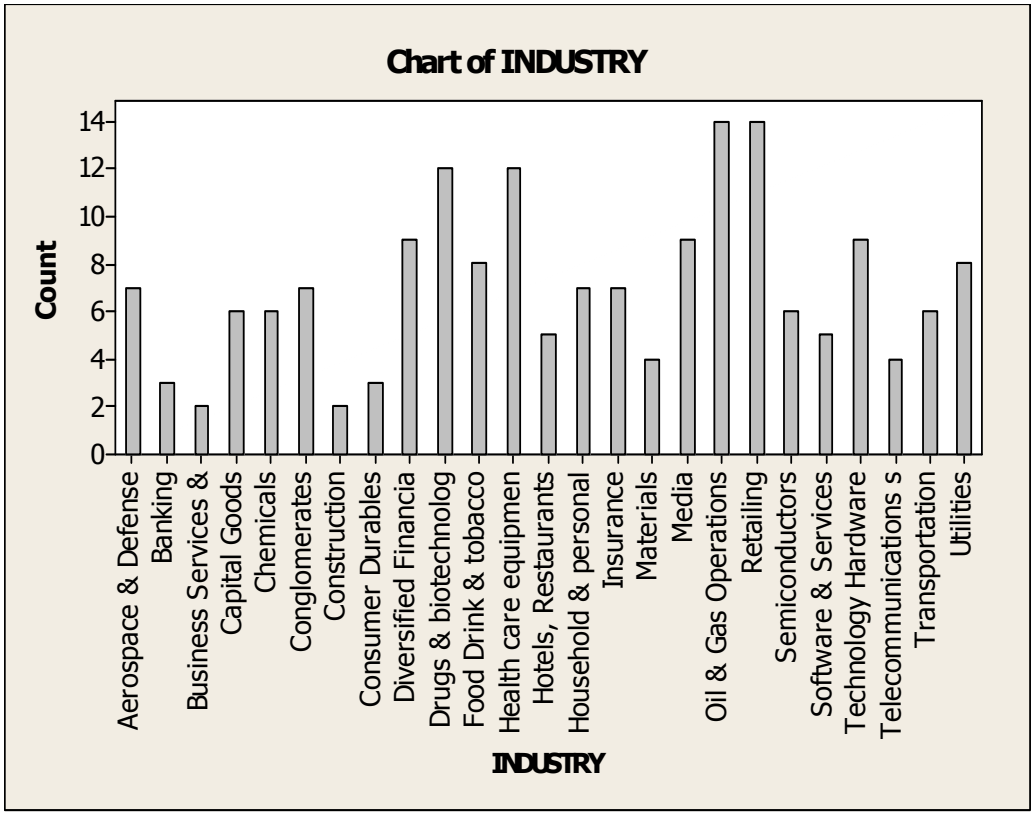
In wave 1, most of those taking the GMAT were working (2657/3244 =.819) and none had MBA's. About 20% were not working but were in either a 4-year institution or other graduate school ([36 + 551]/3244 = .181). In wave 2, almost all were now working ([1787 + 1372]/3244 = .974). Of those working, more than half had MBA's (1787/[1787 + 1372] = .566). Of those not working, most were in another graduate school.

2.9 Using MINITAB, the pie chart is:



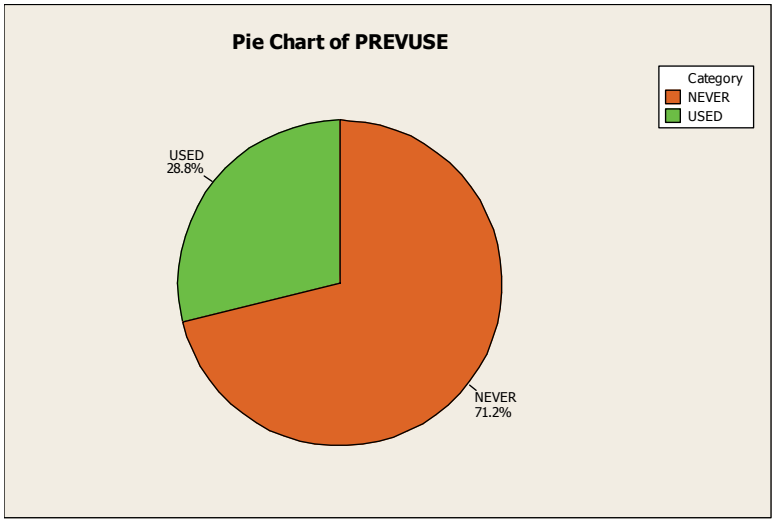
Companies and Employees represent (38.5 + 34.6 = 73.1) slightly more than 73% of the entities creating blogs/forums. Third parties are the least common entity.

2.10 Using MINITAB, a bar chart of the data is:



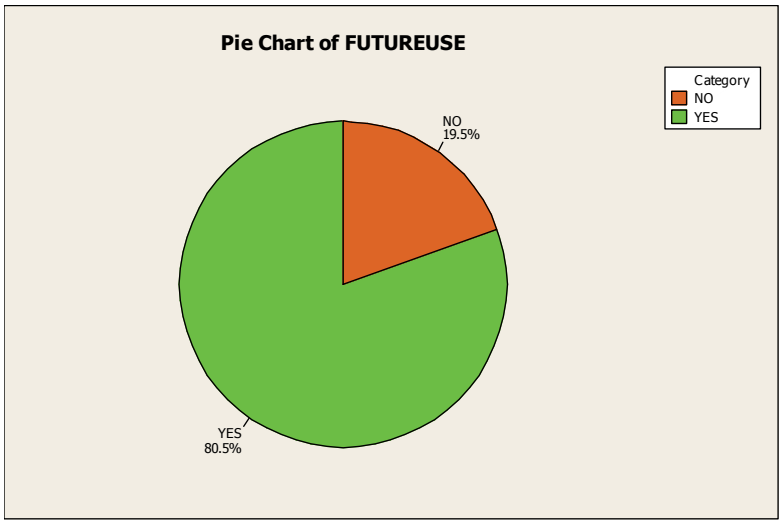
Industries with the highest frequencies include Oil & Gas Operations, Retailing, Drugs & biotechnologies, and Health care equipment. Industries with the smallest frequencies include Business Services, Construction, Banking, and Consumer Durables.

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



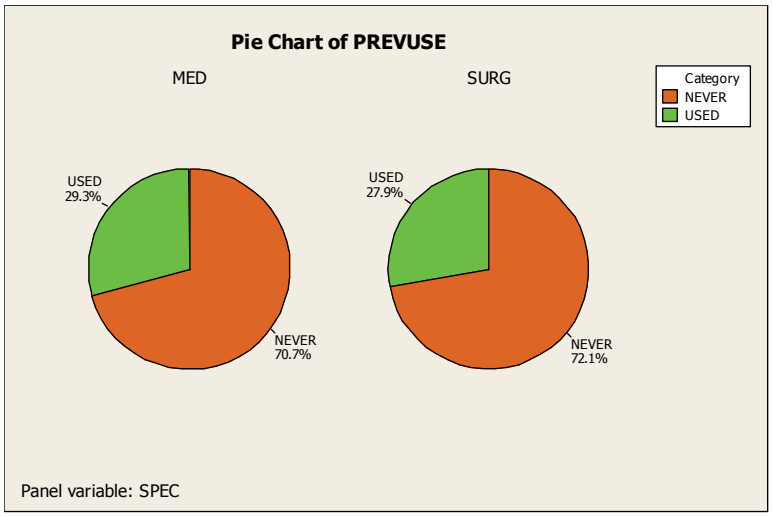
From the chart, 71.2% or .712 of the sampled physicians have never used ethics consultation.

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



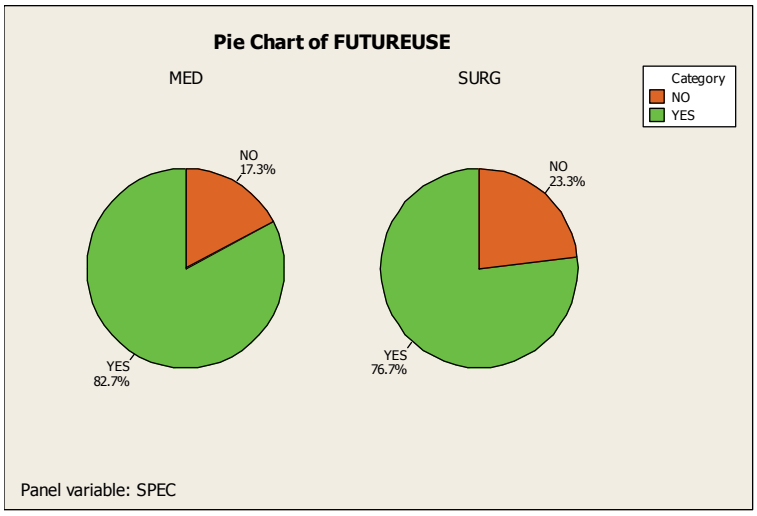
From the chart, 19.5% or .195 of the sampled physicians state that they will not use the services in the future.

c. Using MINITAB, the side-by-side pie charts are:



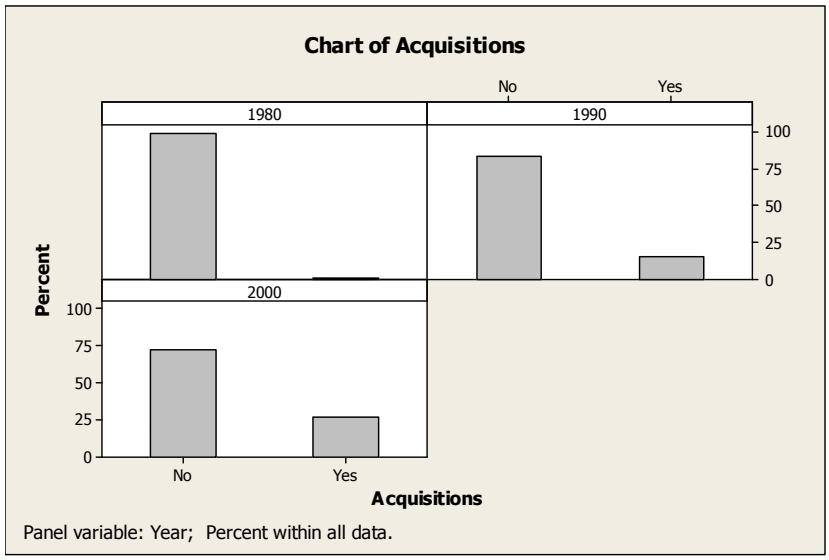
The proportion of medical practitioners who have never used ethics consultation is .707. The proportion of surgical practitioners who have never used ethics consultation is .721. These two proportions are almost the same.

d. Using MINITAB, the side-by-side pie charts are:



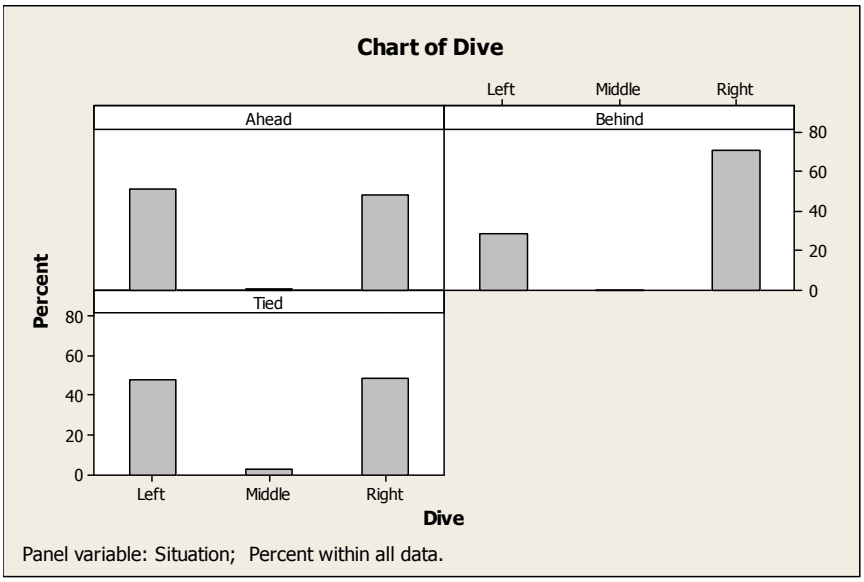
The proportion of medical practitioners who will not use ethics consultation in the future is .173. The proportion of surgical practitioners who will not use ethics consultation in the future is .233. The proportion of surgical practitioners who will not use ethics consultation in the future is greater than that of the medical practitioners.

2.12 Using MINITAB, the side-by-side bar graphs are:



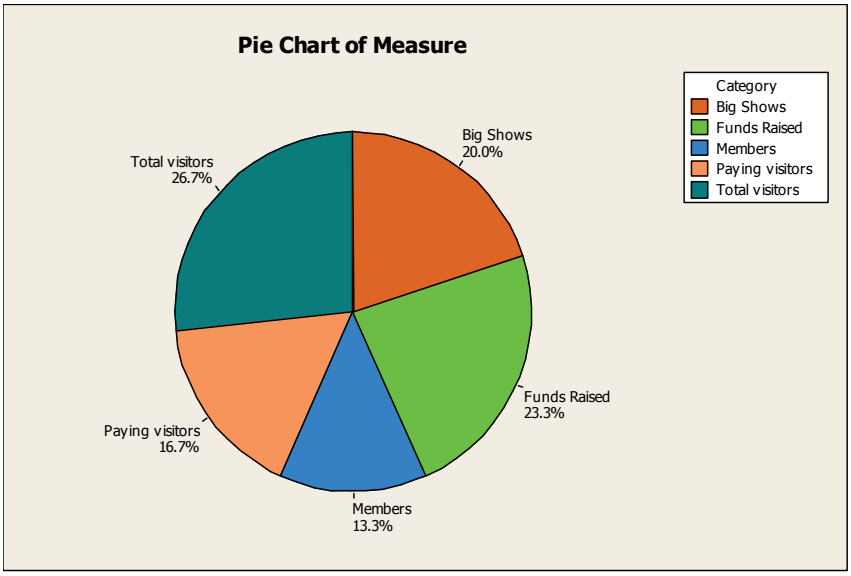
In 1980, very few firms had acquisitions $(18/1{,}963 = .009)$. By 1990, the proportion of firms having acquisitions increased to $350/2{,}197 = .159$. By 2000, the proportion of firms having acquisitions increased to $748/2{,}778 = .269$.

2.13 Using MINITAB, the side-by-side bar graphs are:



From the graphs, it appears that if the team is either tied or ahead, the goal-keepers tend to dive either right or left with equal probability, with very few diving in the middle. However, if the team is behind, then the majority of goal-keepers tend to dive right (71%).

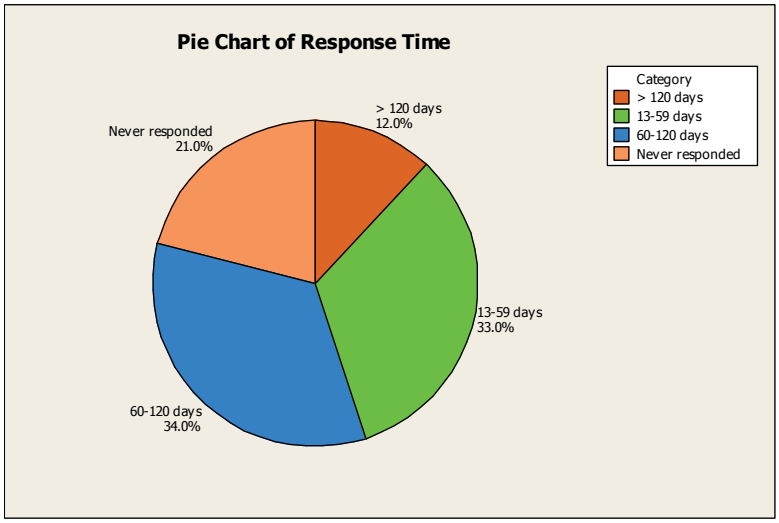
2.14 Using MINITAB, a pie chart of the data is:



Since the sizes of the slices are close to each other, it appears that the researcher is correct. There is a large amount of variation within the museum community with regard to performance measurement and evaluation.

2.15 a. The variable measured by Performark is the length of time it took for each advertiser to respond back.

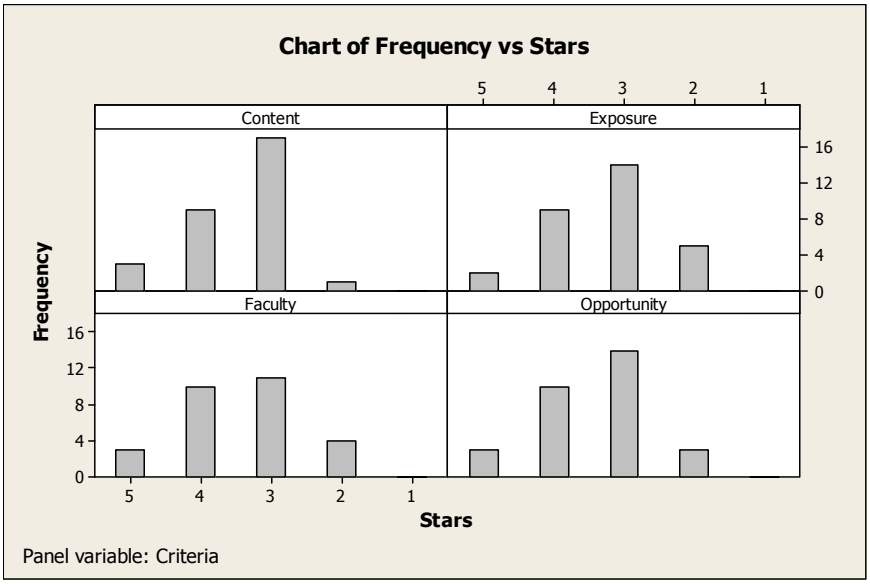
b. The pie chart is:



c. Twenty-one percent or $.21 \times 17{,}000 = 3{,}570$ of the advertisers never respond to the sales lead.

d. The information from the pie chart does not indicate how effective the "bingo cards" are. It just indicates how long it takes advertisers to respond, if at all.

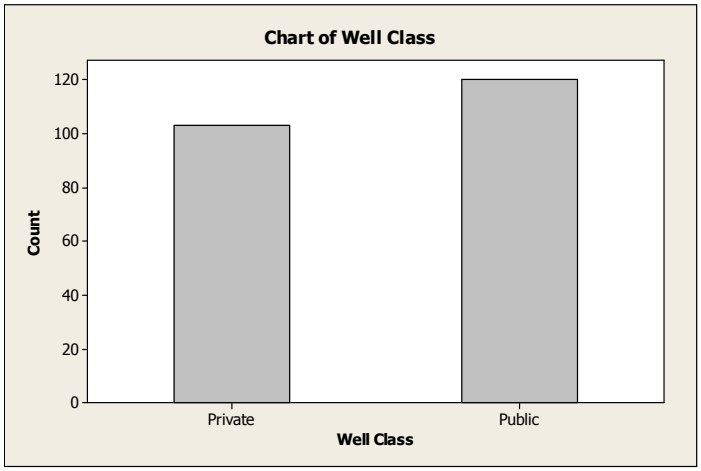
2.16 a. Using MINITAB, the side-by-side graphs are:

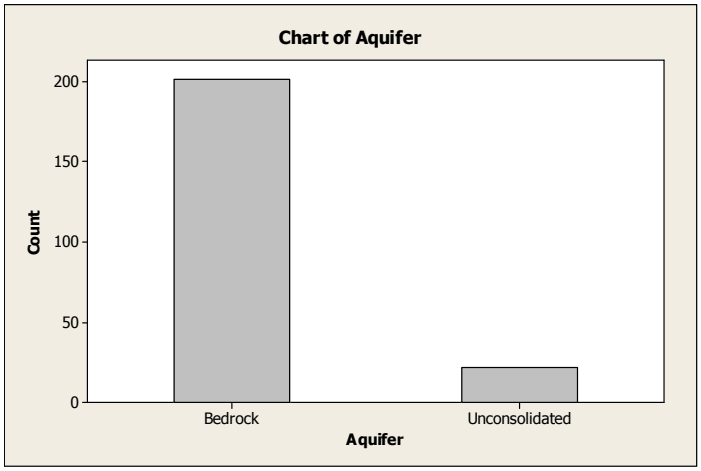


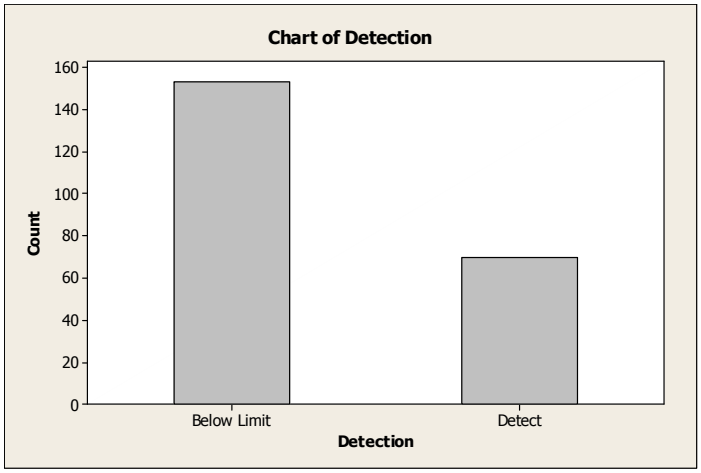
From these graphs, one can see that very few of the top 30 MBA programs got 5-stars in any criteria. In addition, about the same number of programs got 4 stars in each of the 4 criteria. The biggest difference in ratings among the 4 criteria was in the number of programs receiving 3-stars. More programs received 3-stars in Course Content than in any of the other criteria. Consequently, fewer programs received 2-stars in Course Content than in any of the other criteria.

b. Since this chart lists the rankings of only the top 30 MBA programs in the world, it is reasonable that none of these best programs would be rated as 1-star on any criteria.

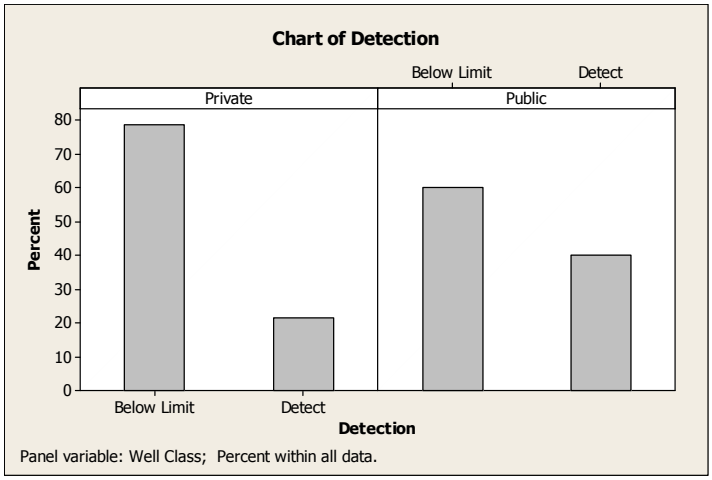
2.17 a. Using MINITAB, bar charts for the 3 variables are:



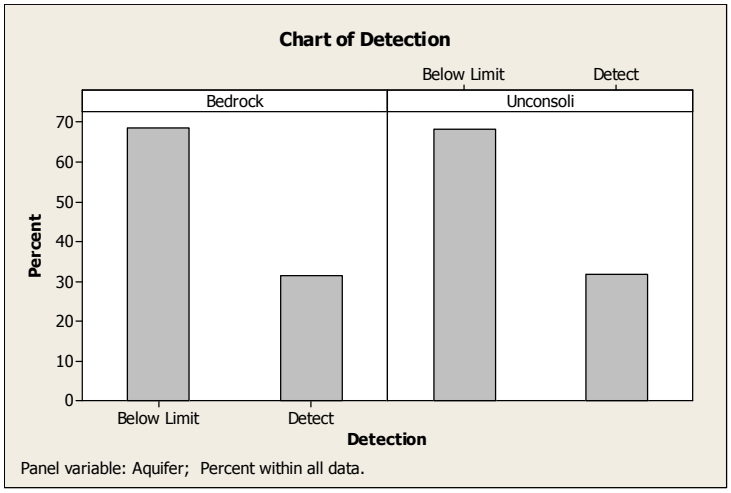




b. Using MINITAB, the side-by-side bar chart is:

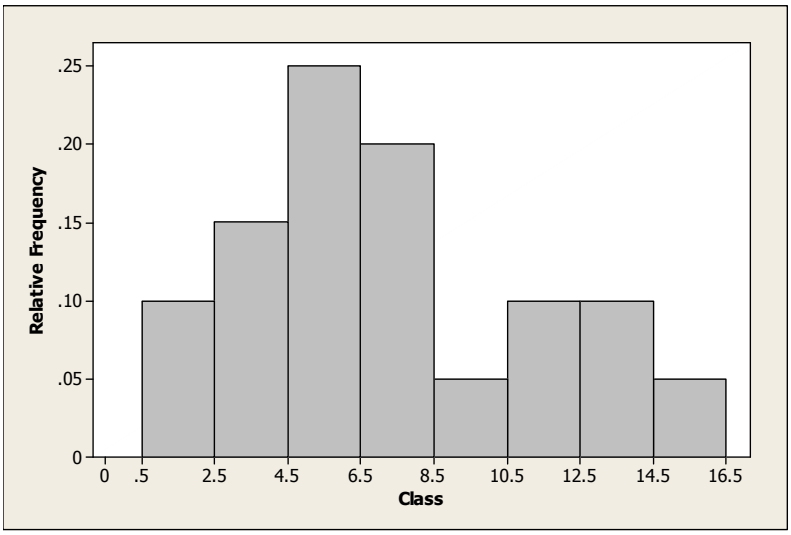


c. Using MINITAB, the side-by-side bar chart is:



d. From the bar charts in parts a-c, one can infer that most aquifers are bedrock and most levels of MTBE were below the limit ($\approx 2/3$). Also the percentages of public wells verses private wells are relatively close. Approximately 80% of private wells are not contaminated, while only about 60% of public wells are not contaminated. The percentage of contaminated wells is about the same for both types of aquifers ($\approx 30\%$).

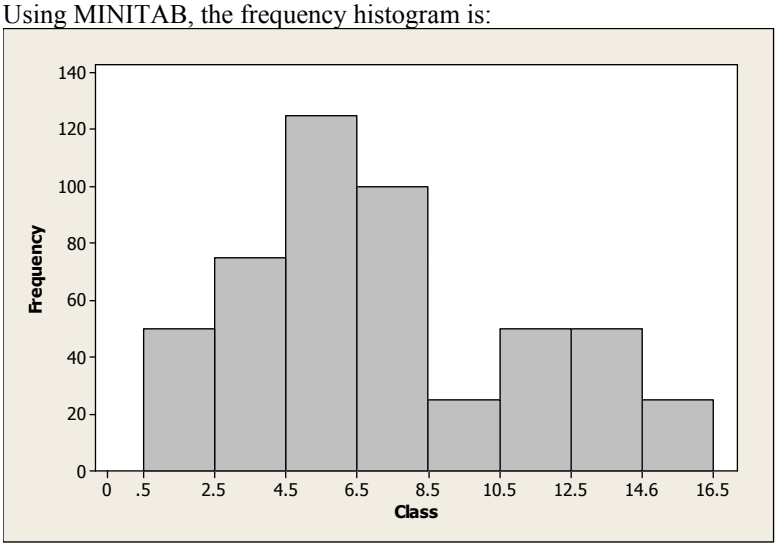
2.18 Using MINITAB, the relative frequency histogram is:



2.19 To find the number of measurements for each measurement class, multiply the relative frequency by the total number of observations, $n = 500$. The frequency table is:

Measurement Class	Relative Frequency	Frequency
.5 – 2.5	.10	500(.10) = 50
2.5 – 4.5	.15	500(.15) = 75
4.5 – 6.5	.25	500(.25) = 125
6.5 – 8.5	.20	500(.20) = 100
8.5 – 10.5	.05	500(.05) = 25
10.5 – 12.5	.10	500(.10) = 50
12.5 – 14.5	.10	500(.10) = 50
14.5 – 16.5	.05	500(.05) = 25
		500

Using MINITAB, the frequency histogram is:

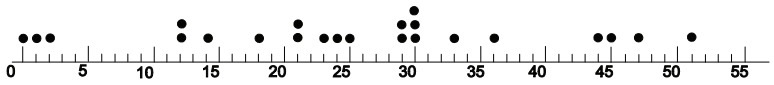


2.20 a. The original data set has 1 + 3 + 5 + 7 + 4 + 3 = 23 observations.

b. For the bottom row of the stem-and-leaf display:

The stem is 0.
The leaves are 0, 1, 2.
Assuming that the data are up to two digits, rounded off to the nearest whole number, the numbers in the original data set are 0, 1, and 2.

c. Again, assuming that the data are up to two digits, rounded off to the nearest whole number, the dot plot corresponding to all the data points is:



2.21 a. This is a frequency histogram because the number of observations is graphed for each interval rather than the relative frequency.

b. There are 14 measurement classes.

c. There are 49 measurements in the data set.

2.22 a. The measurement class 10 – 20 has the highest proportion of respondents.

b. The approximate proportion of the 144 organizations that reported a percentage monetary loss from malicious insider actions less than 20% is .30 + .38 = .68.

c. The approximate proportion of the 144 organizations that reported a percentage monetary loss from malicious insider actions greater than 60% is .07 + .03 + .04 + .05 = .19.

d. The approximate proportion of the 144 organizations that reported a percentage monetary loss from malicious insider actions between 20% and 30% is .11. Therefore about .11(144) = 15.84 or 16 of the 144 organizations reported a percentage monetary loss from malicious insider actions between 20% and 30%.

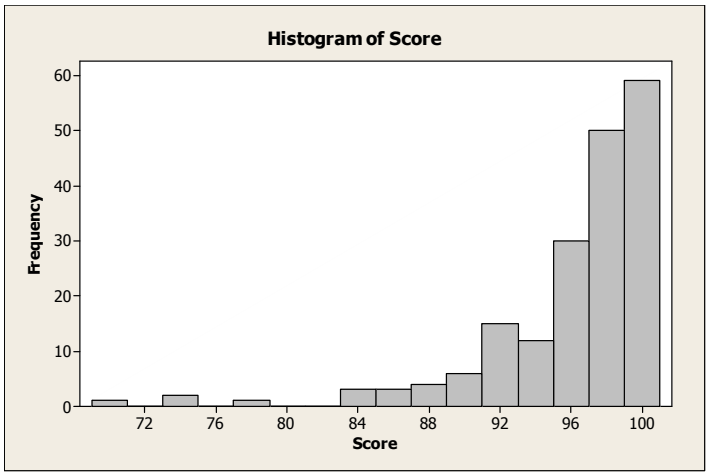
2.23 a. Since the label on the vertical axis is Percent, this is a relative frequency histogram. We can divide the percents by 100% to get the relative frequencies.

b. Summing the percents represented by all of the bars above 100, we get approximately 12%.

2.24 a. Using MINITAB, the stem-and-leaf display and histogram are:

Stem-and-Leaf Display: Score

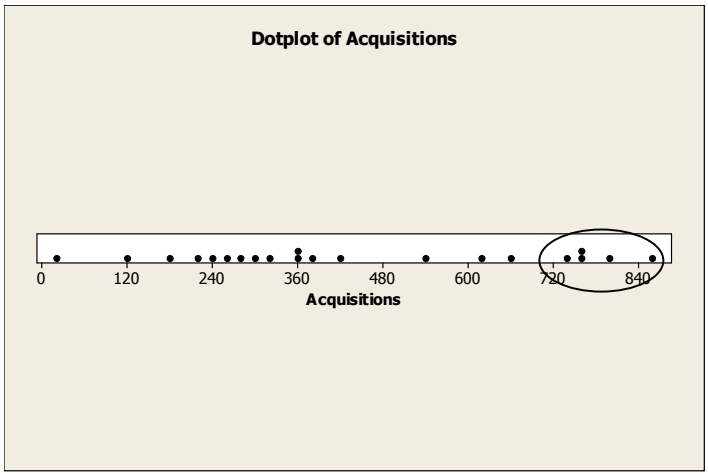
```
Stem-and-leaf of Score  N  = 186
Leaf Unit = 1.0

 1    6   9
 1    7
 2    7   3
 3    7   4
 3    7
 4    7   8
 4    8
 5    8   3
 7    8   44
 11   8   6667
 17   8   888999
 24   9   0001111
 41   9   22222222222333333
 62   9   444444555555555555555
(37)  9   66666666666666677777777777777777777777
 87   9   88888888888888888888888888889999999999999999999999999999999
 27   10  000000000000000000000000000
```



b. From the stem-and-leaf display, there are only 7 observations with sanitation scores less than 86. The proportion of ships with accepted sanitation standards is $(186 - 7)/186 = 179/186 = .962$.

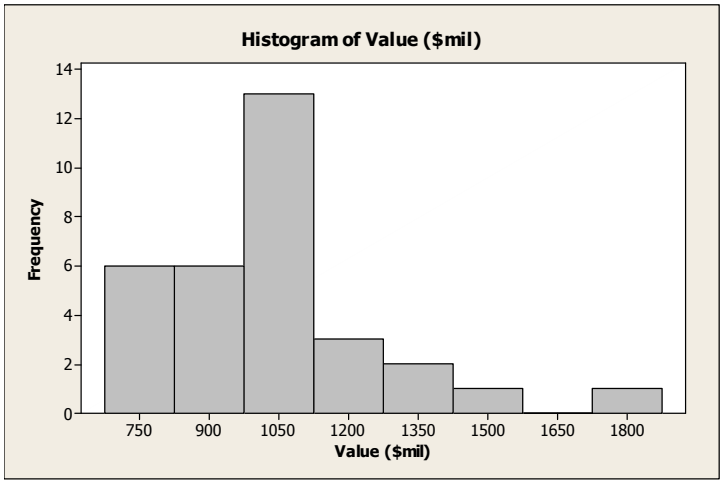
c. The score of 69 is highlighted in the stem-and-leaf display.

2.25 a. Using MINITAB, a dot plot of the data is:

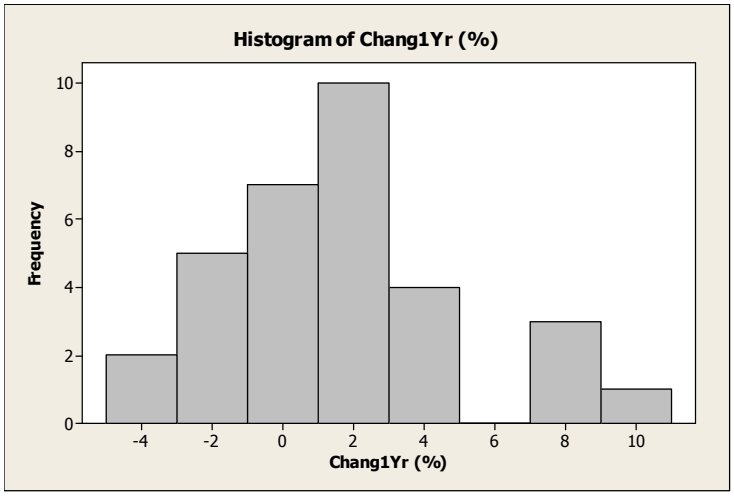


b. By looking at the dot plot, one can conclude that the years 1996-2000 had the highest number of firms with at least one acquisition. The lowest number of acquisitions in that time frame (748) is almost 100 higher than the highest value from the remaining years.

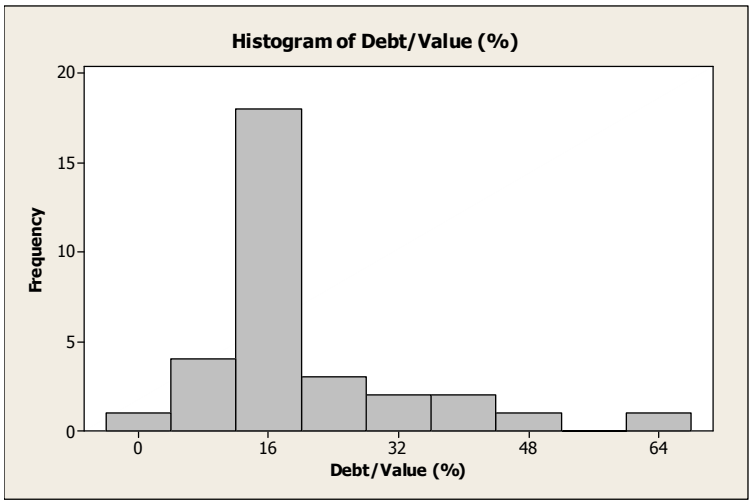
2.26 a. Using MINITAB, a histogram of the current values of the 32 NFL teams is:



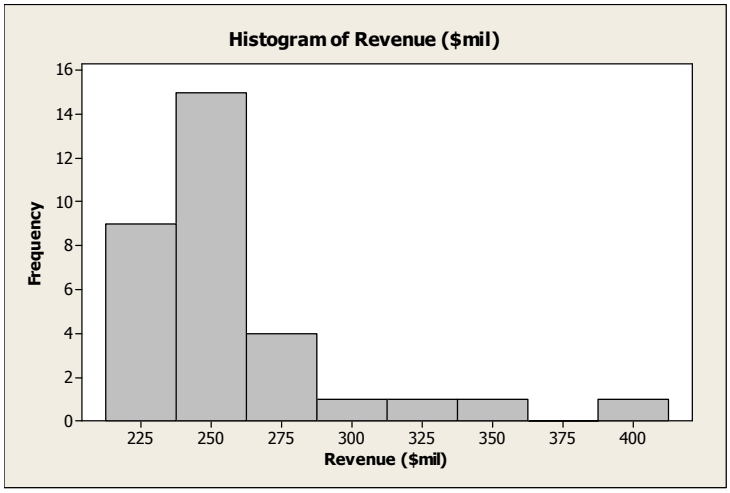
b. Using MINITAB, a histogram of the 1-year change in current value for the 32 NFL teams is:



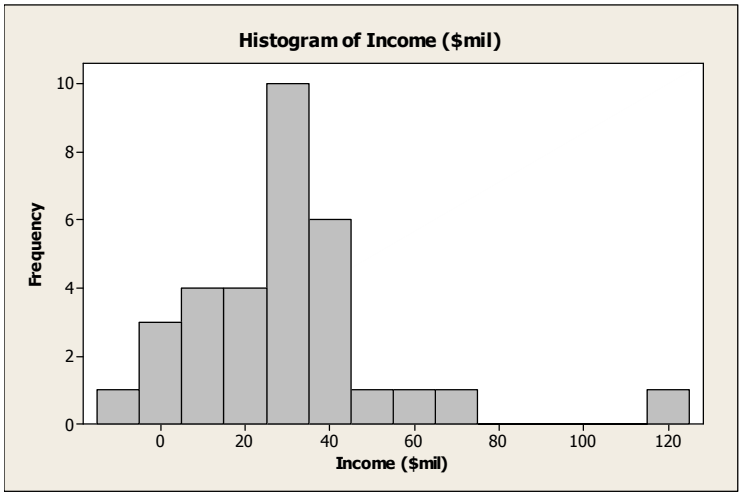
c. Using MINITAB, a histogram of the debt-to-value ratios for the 32 NFL teams is:



d. Using MINITAB, a histogram of the annual revenues for the 32 NFL teams is:

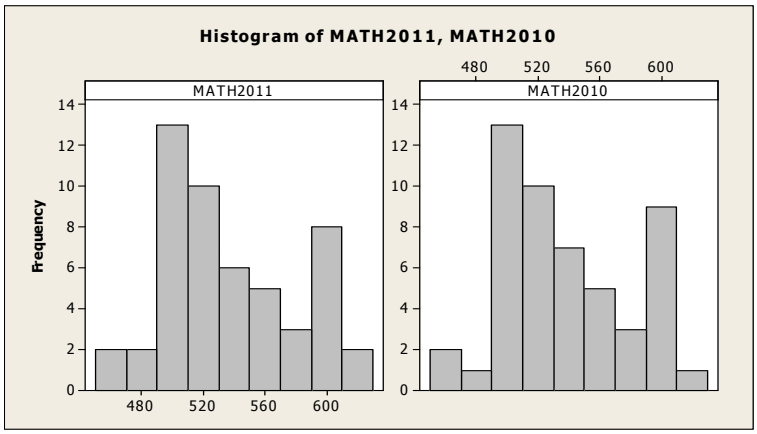


e. Using MINITAB, a histogram of the operating incomes for the 32 NFL teams is:



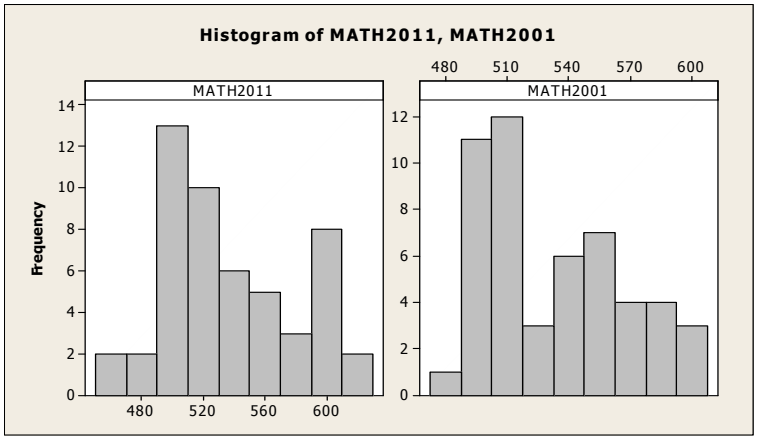
f. For all of the histograms, there is 1 team that has a very high score. The Dallas Cowboys have the largest values for current value, annual revenues, and operating income. However, the New York Giants have the highest 1-year change, while the New York Jets have the highest debt-to-value ratio. All of the graphs except the one showing the 1-Yr Value Changes are skewed to the right.

2.27 a. Using MINITAB, the frequency histograms for 2011 and 2010 SAT mathematics scores are:



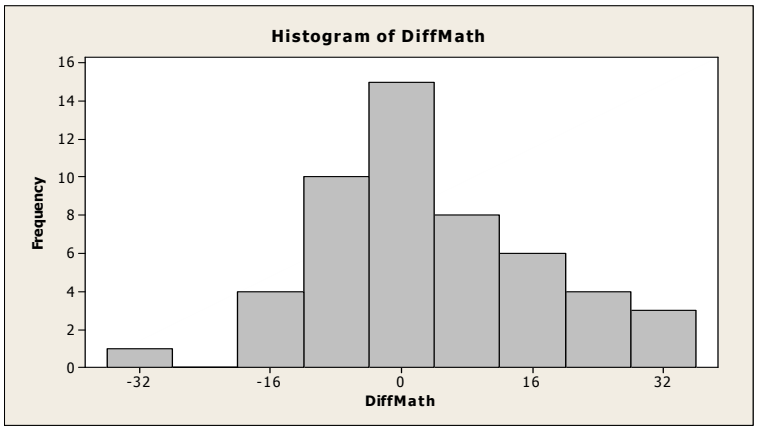
It appears that the scores have not changed very much at all. The graphs are very similar.

b. Using MINITAB, the frequency histograms for 2011 and 2001 SAT mathematics scores are:



It appears that the scores have shifted to the right. The scores in 2011 appear to be somewhat better than the scores in 2011.

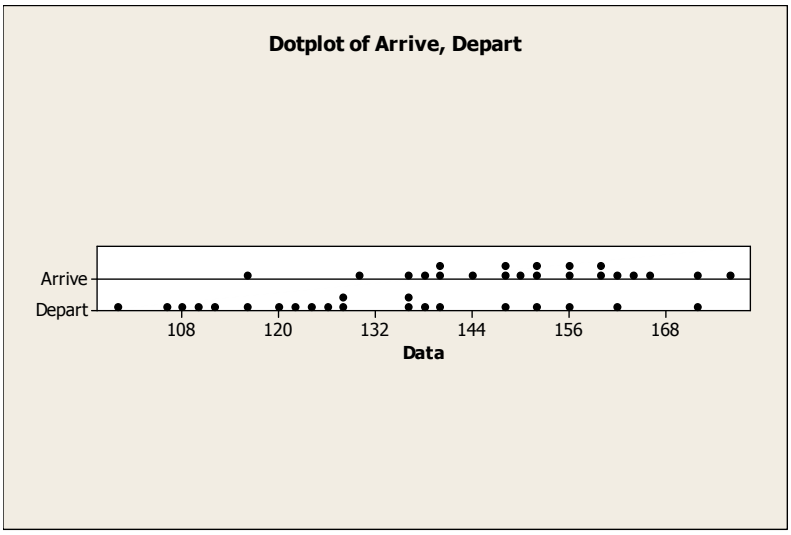
c. Using MINITAB, the frequency histogram of the differences is:



From this graph of the differences, we can see that there are more observations to the right of 0 than to the left of 0. This indicates that, in general, the scores have improved since 2001.

d. From the graph, the largest improvement score is in the neighborhood of 32. The actual largest score is 32 and it is associated with Michigan.

2.28 Using MINITAB, the two dot plots are:



Yes. Most of the numbers of items arriving at the work center per hour are in the 135 to 165 area. Most of the numbers of items departing the work center per hour are in the 110 to 140 area. Because the number of items arriving is larger than the number of items departing, there will probably be some sort of bottleneck.

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

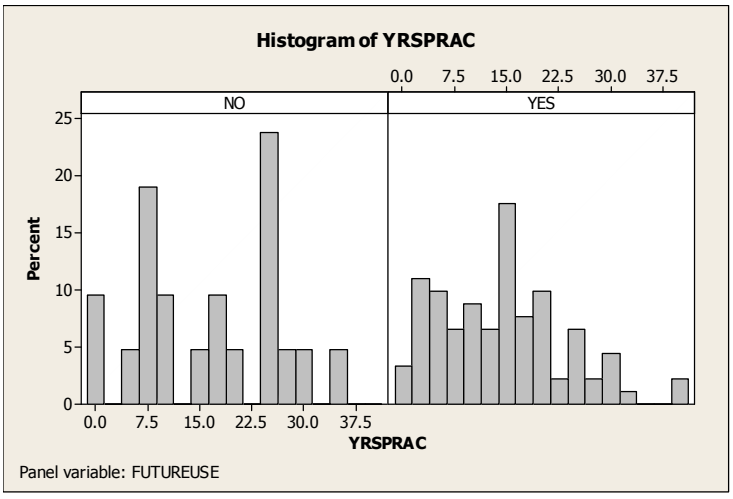
Stem-and-Leaf Display: Dioxide

```
Stem-and-leaf of Dioxide  N  = 16
Leaf Unit = 0.10


 5    0  12234
 7    0  55
(2)   1  34
 7    1
 7    2  44
 5    2
 5    3  3
 4    3
 4    4  0000
```

The highlighted values are values that correspond to water specimens that contain oil. There is a tendency for crude oil to be present in water with lower levels of dioxide as 6 of the lowest 8 specimens with the lowest levels of dioxide contain oil.

2.30 Yes, we would agree with the statement that honey may be the preferable treatment for the cough and sleep difficulty associated with childhood upper respiratory tract infection. For those receiving the honey dosage, 14 of the 35 children (or 40%) had improvement scores of 12 or higher. For those receiving the DM dosage, only 9 of the 33 (or 24%) children had improvement scores of 12 or higher. For those receiving no dosage, only 2 of the 37 children (or 5%) had improvement scores of 12 or higher. In addition, the median improvement score for those receiving the honey dosage was 11, the median for those receiving the DM dosage was 9 and the median for those receiving no dosage was 7.

2.31 Using MINITAB, the relative frequency histograms of the years in practice for the two groups of doctors are:



The researchers hypothesized that older, more experienced physicians will be less likely to use ethics consultation in the future. From the histograms, approximately 38% of the doctors that said "no" have more than 20 years of experience. Only about 19% of the doctors that said "yes" had more than 20 years of experience. This supports the researchers' assertion.

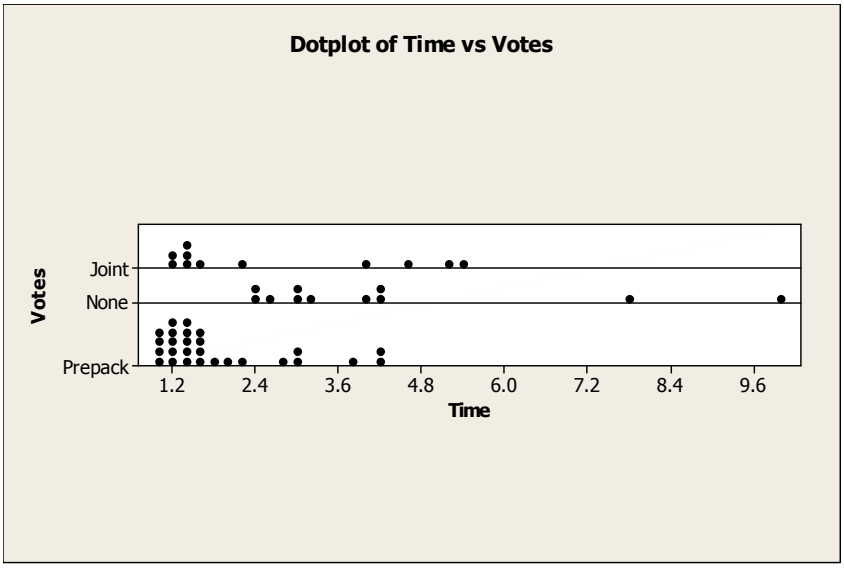
2.32 a. Using MINITAB, the stem-and-leaf display is as follows, where the stems are the units place and the leaves are the decimal places:

Stem-and-Leaf Display: Time

```
Stem-and-leaf of Time  N  = 49
Leaf Unit = 0.10


(26)  1   00001122222344444445555679
 23   2   11446799
 15   3   002899
 9    4   11125
 4    5   24
 2    6
 2    7   8
 1    8
 1    9
 1    10  1
```

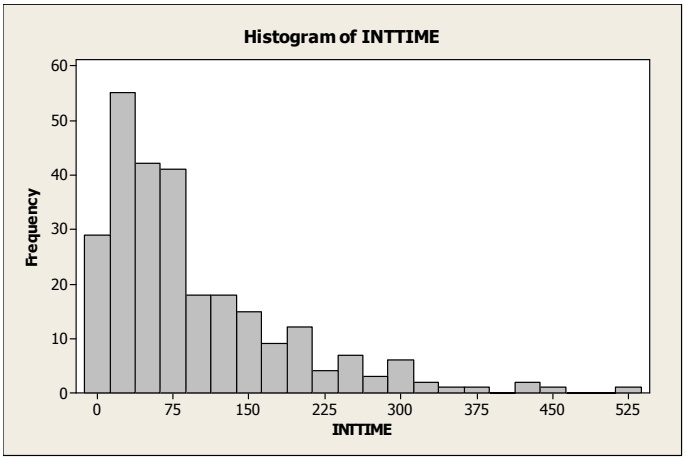
b. A little more than half (26/49 = .53) of all companies spent less than 2 months in bankruptcy. Only two of the 49 companies spent more than 6 months in bankruptcy. It appears that, in general, the length of time in bankruptcy for firms using "prepacks" is less than that of firms not using prepacks."

c. A dot diagram will be used to compare the time in bankruptcy for the three types of "prepack" firms:



d. The highlighted times in part **a** correspond to companies that were reorganized through a leverage buyout. There does not appear to be any pattern to these points. They appear to be scattered about evenly throughout the distribution of all times.

2.33 Using MINITAB, the histogram of the data is:



This histogram looks very similar to the one shown in the problem. Thus, there appears that there was minimal or no collaboration or collusion from within the company. We could conclude that the phishing attack against the organization was not an inside job.

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

Stem-and-Leaf Display: Time

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

    3     3 239
    7     4 3499
   (7)    5 0011469
   11     6 34458
    6     7 13
    4     8 26
    2     9 5
    1    10 2
```

The numbers in bold represent delivery times associated with customers who subsequently did not place additional orders with the firm. Since there were only 2 customers with delivery times of 68 days or longer that placed additional orders, I would say the maximum tolerable delivery time is about 65 to 67 days. Everyone with delivery times less than 67 days placed additional orders.

2.35 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:

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

2.36 a. $\bar{x} = \dfrac{\sum x}{n} = \dfrac{85}{10} = 8.5$

b. $\bar{x} = \dfrac{400}{16} = 25$

c. $\bar{x} = \dfrac{35}{45} = .778$

d. $\bar{x} = \dfrac{242}{18} = 13.44$

2.37 The mean and median of a symmetric data set are equal to each other. The mean is larger than the median when the data set is skewed to the right. The mean is less than the median when the data set is skewed to the left. Thus, by comparing the mean and median, one can determine whether the data set is symmetric, skewed right, or skewed left.

2.38 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 five measurements arranged in order is 1, 3, 5, 6, 8. The median is the middle number, which is 5.

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

2.39 Assume the data are a sample. The mode is the observation that occurs most frequently. For this sample, the mode is 15, which occurs three times.

The sample mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{18+10+15+13+17+15+12+15+18+16+11}{11} = \frac{160}{11} = 14.545$$

The median is the middle number when the data are arranged in order. The data arranged in order are: 10, 11, 12, 13, 15, 15, 15, 16, 17, 18, 18. The middle number is the 6th number, which is 15.

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

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

Mode = 3

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

Median = 3 (7th number, after ordering)

Mode = 3

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

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

Mode = 50

2.41 a. For a distribution that is skewed to the left, the mean is less than the median.

b. For a distribution that is skewed to the right, the mean is greater than the median.

c. For a symmetric distribution, the mean and median are equal.

2.42 a. The mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{9 + (-.1) + (-1.6) + 14.6 + 16.0 + 7.7 + 19.9 + 9.8 + 3.2 + 24.8 + 17.6 + 10.7 + 9.1}{13} = \frac{140.7}{13} = 10.82$$

The average annualized percentage return on investment for 13 randomly selected stock screeners is 10.82.

b. Since the number of observations is odd, the median is the middle number once the data have been arranged in order. The data arranged in order are:

-1.6 -.1 3.2 7.7 9.0 9.1 9.8 10.7 14.6 16.0 17.6 19.9 24.8

The middle number is 9.8 which is the median. Half of the annualized percentage returns on investment are below 9.8 and half are above 9.8.

2.43 a. The mean amount exported on the printout is 653. This means that the average amount of money per market from exporting sparkling wine was \$653,000.

b. The median amount exported on the printout is 231. Since the median is the middle value, this means that half of the 30 sparkling wine export values were above \$231,000 and half of the sparkling wine export values were below \$231,000.

c. The mean 3-year percentage change on the printout is 481. This means that in the last three years, the average change is 481%, which indicates a large increase.

d. The median 3-year percentage change on the printout is 156. Since the median is the middle value, this means that half, or 15 of the 30 countries' 3-year percentage change values were above 156% and half, or 15 of the 30 countries' 3-year percentage change values were below 156%.

2.44 a. The sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.72 + 2.50 + 2.16 + \cdots + 1.95}{20} = \frac{37.62}{20} = 1.881$$

The sample average surface roughness of the 20 observations is 1.881.

b. The median is found as the average of the 10th and 11th observations, once the data have been ordered. The ordered data are:

1.06 1.09 1.19 1.26 1.27 1.40 1.51 1.72 1.95 2.03 2.05 2.13 2.13 2.16 2.24 2.31 2.41 2.50 2.57 2.64

The 10th and 11th observations are 2.03 and 2.05. The median is:

$$\frac{2.03+2.05}{2}=\frac{4.08}{2}=2.04$$

The middle surface roughness measurement is 2.04. Half of the sample measurements were less than 2.04 and half were greater than 2.04.

c. The data are somewhat skewed to the left. Thus, the median might be a better measure of central tendency than the mean. The few small values in the data tend to make the mean smaller than the median.

2.45 a. The mean is $\bar{x}=\frac{\sum x}{n}=\frac{1,680,927+885,182+881,777+\cdots+563,967}{20}=\frac{15,192,021}{20}=759,601.05$.
The average research expenditures for the top 20 ranked universities is 759,601.05 thousand dollars.

b. Since the number of observations is even, the median is the average of the middle 2 numbers once the data have been arranged in order. Since the data are already arranged in order, the median is $\frac{702,592+688,225}{2}=695,408.5$.

Half of the institutions have a research expenditure less than 695,408.5 thousand dollars and half have research expenditures greater than 695,408.5 thousand dollars.

c. No, the mean from part a would not be a good measure for the center of the distribution for all American universities. The data in part a come from only the top 20 universities. These universities would not be representative of all American universities.

2.46 a. The mean is 67.755. The statement is accurate.

b. The median is 68.000. The statement is accurate.

c. The mode is 64. The statement is not accurate. A better statement would be: "The most common reported level of support for corporate sustainability for the 992 senior managers was 64.

d. Since the mean and median are almost the same, the distribution of the 992 support levels should be fairly symmetric. The histogram in Exercise 2.23 is almost symmetric.

2.47 a. The median is the middle number (18th) once the data have been arranged in order because $n = 35$ is odd. The honey dosage data arranged in order are:

4,5,6,8,8,8,8,9,9,9,9,10,10,10,10,10,10,11,11,11,11,12,12,12,12,12,12,13,13,14,15,15,15,15,16

The 18th number is the median = 11.

b. The median is the middle number (17^{th}) once the data have been arranged in order because $n = 33$ is odd. The DM dosage data arranged in order are:

3,4,4,4,4,4,4,6,6,6,7,7,7,7,7,8,9,9,9,9,9,10,10,10,11,12,12,12,12,12,13,13,15

The 17^{th} number is the median = 9.

c. The median is the middle number (19^{th}) once the data have been arranged in order because $n = 37$ is odd. The No dosage data arranged in order are:

0,1,1,1,3,3,4,4,5,5,5,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,9,9,9,9,10,11,12,12

The 19^{th} number is the median = 7.

d. Since the median for the Honey dosage is larger than the other two, it appears that the honey dosage leads to more improvement than the other two treatments.

2.48 a. The mean dioxide level is $\bar{x} = \frac{3.3 + 0.5 + 1.3 + \cdots + 4.0}{16} = \frac{29}{16} = 1.81$. The average dioxide amount is 1.81.

b. Since the number of observations is even, the median is the average of the middle 2 numbers once the data are arranged in order. The data arranged in order are:

0.1 0.2 0.2 0.3 0.4 0.5 0.5 1.3 1.4 2.4 2.4 3.3 4.0 4.0 4.0 4.0

The median is $\frac{1.3 + 1.4}{2} = \frac{2.7}{2} = 1.35$. Half of the dioxide levels are below 1.35 and half are above 1.35.

c. The mode is the number that occurs the most. For this data set the mode is 4.0. The most frequent level of dioxide is 4.0.

d. Since the number of observations is even, the median is the average of the middle 2 numbers once the data are arranged in order. The data arranged in order are:

0.1 0.3 1.4 2.4 2.4 3.3 4.0 4.0 4.0 4.0

The median is $\frac{2.4 + 3.3}{2} = \frac{5.7}{2} = 2.85$.

e. Since the number of observations is even, the median is the average of the middle 2 numbers once the data are arranged in order. The data arranged in order are:

0.2 0.2 0.4 0.5 0.5 1.3

The median is $\frac{0.4 + 0.5}{2} = \frac{0.9}{2} = 0.45$.

f. The median level of dioxide when crude oil is present is 0.45. The median level of dioxide when crude oil is not present is 2.85. It is apparent that the level of dioxide is much higher when crude oil is not present.

2.49 a. Skewed to the right. There will be a few people with very high salaries such as the president and football coach.

b. Skewed to the left. On an easy test, most students will have high scores with only a few low scores.

c. Skewed to the right. On a difficult test, most students will have low scores with only a few high scores.

d. Skewed to the right. Most students will have a moderate amount of time studying while a few students might study a long time.

e. Skewed to the left. Most cars will be relatively new with a few much older.

f. Skewed to the left. Most students will take the entire time to take the exam while a few might leave early.

2.50 a. The sample means is:

$$\bar{x} = \frac{\sum x}{n} = \frac{3.58 + 3.48 + 3.27 + \cdots + 1.17}{40} = \frac{77.07}{40} = 1.927$$

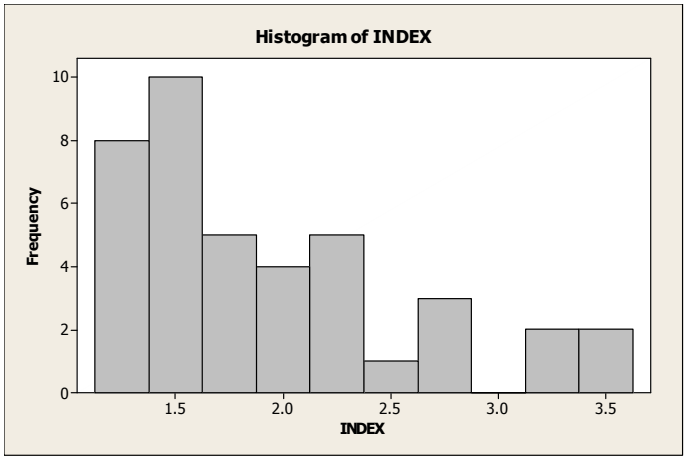
The median is found as the 20th and 21st observations, once the data have been ordered. The 20th and 21st observations are 1.75 and 1.76. The median is:

$$\frac{1.75 + 1.76}{2} = \frac{3.51}{2} = 1.755$$

The mode is the number that occurs the most and is 1.4, which occurs 3 times.

b. The sample average driving performance index is 1.927. The median driving performance index is 1.755. Half of all driving performance indexes are less than 1.755 and half are higher. The most common driving performance index value is 1.4.

c. Since the mean is larger than the median, the data are skewed to the right. Using MINITAB, a histogram of the driving performance index values is:



2.51 The mean is 141.31 hours. This means that the average number of semester hours per candidate for the CPA exam is 141.31 hours. The median is 140 hours. This means that 50% of the candidates had more than 140 semester hours of credit and 50% had less than 140 semester hours of credit. Since the mean and median are so close in value, the data are probably not skewed, but close to symmetric.

2.52 a. Using MINITAB, the output is:

Descriptive Statistics: YRSPRAC

Variable	N	N*	Mean	Minimum	Median	Maximum	Mode	N for Mode
YRSPRAC	112	6	14.598	1.000	14.000	40.000	14, 20, 25	9

The mean is 14.598. The average length of time in practice for this sample is 14.598 years. The median is 14. Half of the physicians have been in practice less than 14 years and half have been in practice longer than 14 years. There are 3 modes: 14, 20, and 25. The most frequent years in practice are 14, 20, and 25 years.

b. Using MINITAB, the results are:

Descriptive Statistics: YRSPRAC

Variable	FUTUREUSE	N	N*	Mean	Minimum	Median	Maximum	Mode	N for Mode
YRSPRAC	NO	21	2	16.43	1.00	18.00	35.00	25	5
	YES	91	4	14.176	1.000	14.000	40.000	14, 20	8

The mean for the physicians who would refuse to use ethics consultation in the future is 16.43. The average time in practice for these physicians is 16.43 years. The median is 18. Half of the physicians who would refuse ethics consultation in the future have been in practice less than 18 years and half have been in practice more than 18 years. The mode is 25. The most frequent years in practice for these physicians is 25 years.

c. From the results in part b, the mean for the physicians who would use ethics consultation in the future is 14.176. The average time in practice for these physicians is 14.176 years. The median is 14. Half of the physicians who would use ethics consultation in the future have been in practice less than 14 years and half have been in practice more than 14 years. There are 2 modes: 14 and 20. The most frequent years in practice for these physicians are 14 and 20 years.

d. The results in parts b and c confirm the researchers' theory. The mean, median and mode of years in practice are larger for the physicians who would refuse to use ethics consultation in the future than those who would use ethics consultation in the future.

2.53 For the "Joint exchange offer with prepack" firms, the mean time is 2.6545 months, and the median is 1.5 months. Thus, the average time spent in bankruptcy for "Joint" firms is 2.6545 months, while half of the firms spend 1.5 months or less in bankruptcy.

For the "No prefiling vote held" firms, the mean time is 4.2364 months, and the median is 3.2 months. Thus, the average time spent in bankruptcy for "No prefiling vote held" firms is 4.2364 months, while half of the firms spend 3.2 months or less in bankruptcy.

For the "Prepack solicitation only" firms, the mean time is 1.8185 months, and the median is 1.4 months. Thus, the average time spent in bankruptcy for "Prepack solicitation only" firms is 1.8185 months, while half of the firms spend 1.4 months or less in bankruptcy.

Since the means and medians for the three groups of firms differ quite a bit, it would be unreasonable to use a single number to locate the center of the time in bankruptcy. Three different "centers" should be used.

2.54 a. The sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+2+4+\ldots+3}{20} = \frac{78}{20} = 3.90$$

The sample median is found by finding the average of the 10^{th} and 11^{th} observations once the data are arranged in order. The data arranged in order are:

1 1 1 1 1 2 2 3 3 3 4 4 4 5 5 5 6 7 9 11

The 10^{th} and 11^{th} observations are 3 and 4. The average of these two numbers (median) is:

$$\text{median} = \frac{3+4}{2} = \frac{7}{2} = 3.5$$

The mode is the observation appearing the most. For this data set, the mode is 1, which appears 5 times.

b. Eliminating the largest number which is 11 results in the following:

The sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+2+4+\ldots+3}{19} = \frac{67}{19} = 3.53$$

The sample median is found by finding the middle observation once the data are arranged in order. The data arranged in order are:

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

The 10^{th} observation is 3. The median is 3

The mode is the observations appearing the most. For this data set, the mode is 1, which appears 5 times.

By dropping the largest number, the mean is reduced from 4.05 to 3.68. The median is reduced from 3.5 to 3. There is no effect on the mode.

c. The data arranged in order are:

1 1 1 1 1 2 2 3 3 3 4 4 4 5 5 5 6 7 9 11

If we drop the lowest 2 and largest 2 observations we are left with:

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

The sample 10% trimmed mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+1+2+\ldots+7}{16} = \frac{56}{16} = 3.5$$

The advantage of the trimmed mean over the regular mean is that very large and very small numbers that could greatly affect the mean have been eliminated.

2.55 a. Due to the "elite" superstars, the salary distribution is skewed to the right. Since this implies that the median is less than the mean, the players' association would want to use the median.

b. The owners, by the logic of part **a**, would want to use the mean.

2.56 a. 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.

b. The sample variance is the sum of the squared deviations of the observations from the sample mean divided by the sample size minus 1. The population variance is the sum of the squared deviations of the values from the population mean divided by the population size.

c. 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: $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{2.28 - \dfrac{3.2^2}{13}}{5-1} = \dfrac{.232}{4} = .058$

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

2.57 a. Range = 4 – 0 = 4

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

b. Range = 6 – 0 = 6

$$s^2 = \frac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \frac{63 - \dfrac{17^2}{7}}{7-1} = 3.619 \qquad s = \sqrt{3.619} = 1.9$$

c. Range = 8 – (–2) = 10

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

d. Range = 1 – (–3) = 4

$$s^2 = \frac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \frac{25.04 - \dfrac{(-6.8)^2}{17}}{17-1} = 1.395 \qquad s = \sqrt{1.395} = 1.18$$

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

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

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

2.59 a. $\sum x = 3+1+10+10+4 = 28$ $\quad \sum x^2 = 3^2+1^2+10^2+10^2+4^2 = 226$

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

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

b. $\sum x = 8+10+32+5 = 55$ $\quad \sum x^2 = 8^2+10^2+32^2+5^2 = 1213$

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

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

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

c. $\sum x = -1+(-4)+(-3)+1+(-4)+(-4) = -15$ $\quad \sum x^2 = (-1)^2+(-4)^2+(-3)^2+1^2+(-4)^2+(-4)^2 = 59$

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

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

d. $\sum x = \frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{2}{5}+\frac{1}{5}+\frac{4}{5} = \frac{10}{5} = 2$ $\quad \sum x^2 = \left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{2}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2 = \frac{24}{25} = .96$

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

$$s^2 = \frac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \frac{\dfrac{24}{25} - \dfrac{2^2}{6}}{6-1} = \frac{.2933}{5} = .0587 \text{ square ounce} \qquad s = \sqrt{.0587} = .2422 \text{ ounce}$$

2.60 a. Range = 42 – 37 = 5

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{7935 - \frac{199^2}{5}}{5-1} = 3.7 \qquad s = \sqrt{3.7} = 1.92$$

b. Range = 100 – 1 = 99

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{25,795 - \frac{303^2}{9}}{9-1} = 1,949.25 \qquad s = \sqrt{1,949.25} = 44.15$$

c. Range = 100 – 2 = 98

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{20,033 - \frac{295^2}{8}}{8-1} = 1,307.84 \qquad s = \sqrt{1,307.84} = 36.16$$

2.61 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

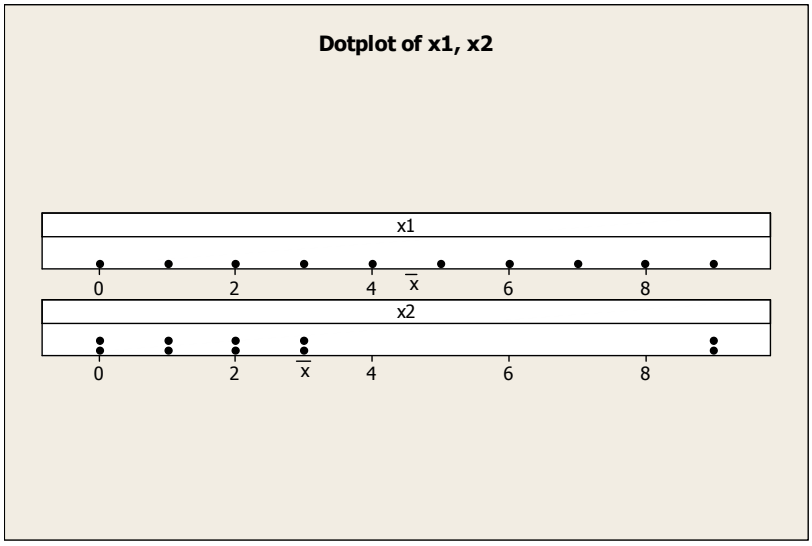
The two sets of data above have the same range = largest measurement – smallest measurement = 9 – 0 = 9.

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.



2.62 This is one possibility for the two data sets.

Data Set 1: 1, 1, 2, 2, 3, 3, 4, 4, 5, 5
Data Set 2: 1, 1, 1, 1, 1, 5, 5, 5, 5, 5

$$\bar{x}_1 = \frac{\sum x}{n} = \frac{1+1+2+2+3+3+4+4+5+5}{10} = \frac{30}{10} = 3$$

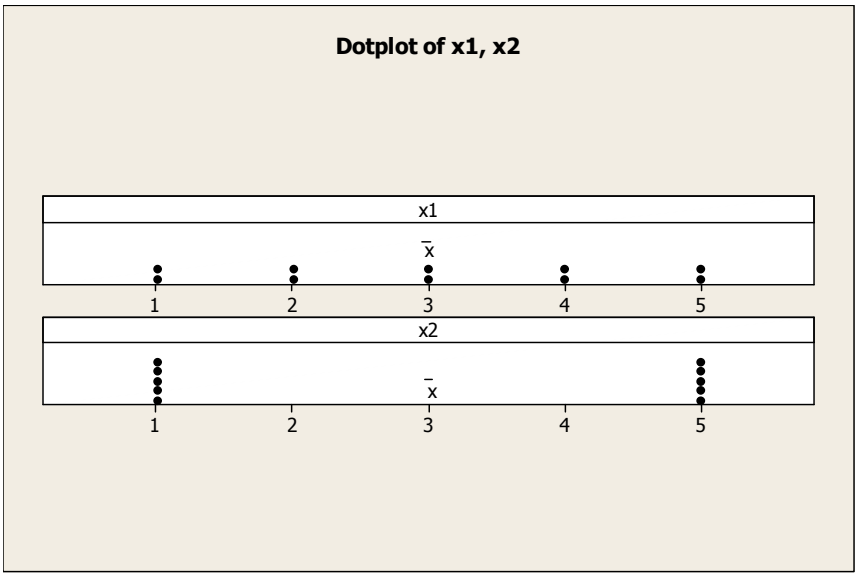
$$\bar{x}_2 = \frac{\sum x}{n} = \frac{1+1+1+1+1+5+5+5+5+5}{10} = \frac{30}{10} = 3$$

Therefore, the two data sets have the same mean. The variances for the two data sets are:

$$s_1^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{110 - \frac{30^2}{10}}{9} = \frac{20}{9} = 2.2222$$

$$s_2^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{130 - \frac{30^2}{10}}{9} = \frac{40}{9} = 4.4444$$

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



2.63 a. Range = 3 – 0 = 3

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{15 - \frac{7^2}{5}}{5-1} = 1.3 \qquad s = \sqrt{1.3} = 1.14$$

b. After adding 3 to each of the data points,

Range = 6 – 3 = 3

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{102 - \frac{22^2}{5}}{5-1} = 1.3 \qquad s = \sqrt{1.3} = 1.14$$

c. After subtracting 4 from each of the data points,

Range = –1 – (–4) = 3

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{39 - \frac{(-13)^2}{5}}{5-1} = 1.3 \qquad s = \sqrt{1.3} = 1.14$$

d. The range, variance, and standard deviation remain the same when any number is added to or subtracted from each measurement in the data set.

2.64 a. The range is the difference between the maximum and minimum values. The range $= 24.8 - (-1.6) = 26.4$. The units of measurement are percents.

b. The variance is

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{2236.41 - \frac{140.7^2}{13}}{13-1} = \frac{2236.41 - 1522.8069}{12} = \frac{713.6031}{12} = 59.4669$$

The units are square percents.

c. The standard deviation is $s = \sqrt{59.4669} = 7.7115$. The units are percents.

2.65 a. The range is the difference between the largest observation and the smallest observation. From the printout, the largest observation is \$4,852 thousand and the smallest observation is \$70 thousand. The range is:

$R = \$4,852 - \$70 = \$4,882$ thousand

b. From the printout, the standard deviation is s = \$1,113 thousand.

c. The variance is the standard deviation squared. The variance is:

$s^2 = 1,113^2 = 1,238,769$ million dollars squared

2.66 a. The sample variance of the honey dosage group is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{4295 - \frac{375^2}{35}}{35-1} = \frac{277.142857}{34} = 8.1512605$$

The standard deviation is: $s = \sqrt{8.1512605} = 2.855$

b. The sample variance of the DM dosage group is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{2631 - \frac{275^2}{33}}{33-1} = \frac{339.33333}{32} = 10.604167$$

The standard deviation is: $s = \sqrt{10.604167} = 3.256$

c. The sample variance of the control group is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{1881 - \frac{241^2}{37}}{37-1} = \frac{311.243243}{36} = 8.6456456$$

The standard deviation is: $s = \sqrt{8.6456456} = 2.940$

d. The group with the most variability is the group with the largest standard deviation, which is the DM group. The group with the least variability is the group with the smallest standard deviation, which is the honey group.

2.67 a. The range is 155. The statement is accurate.

b. The variance is 722.036. The statement is not accurate. A more accurate statement would be: "The variance of the levels of supports for corporate sustainability for the 992 senior managers is 722.036."

c. The standard deviation is 26.871. If the units of measure for the two distributions are the same, then the distribution of support levels for the 992 senior managers has less variation than a distribution with a standard deviation of 50. If the units of measure for the second distribution is not known, then we cannot compare the variation in the two distributions by looking at the standard deviations alone.

d. The standard deviation best describes the variation in the distribution. The range can be greatly affected by extreme measures. The variance is measured in square units, which is hard to interpret. Thus, the standard deviation is the best measure to describe the variation.

2.68 a. Using MINITAB, the results are:

Descriptive Statistics: YRSPRAC

Variable	N	N*	Mean	StDev	Variance	Range
YRSPRAC	112	6	14.598	9.161	83.918	39.000

The range is 39. The difference between the largest years in practice and the smallest years in practice is 39 years. The variance is 83.918 square years. The standard deviation is 9.161 years.

b. Using MINITAB, the results are:

Descriptive Statistics: YRSPRAC

Variable	FUTUREUSE	N	N*	Mean	StDev	Variance	Range
YRSPRAC	NO	21	2	16.43	10.05	100.96	34.00
	YES	91	4	14.176	8.950	80.102	39.000

For the physicians who would refuse to use ethics consultation in the future, the standard deviation is 10.05 years.

c. For the physicians who would use ethics consultation in the future, the standard deviation is 8.95 years.

d. The variation in the length of time in practice for the physicians who would refuse to use ethics consultation in the future is greater than that for the physicians who would use ethics consultation in the future.

2.69 a. The range is the largest observation minus the smallest observation or 11 – 1 = 10.

The variance is: $s^2 = \dfrac{\sum_i x_i^2 - \dfrac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \dfrac{450 - \dfrac{78^2}{20}}{20-1} = 7.6737$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{7.6737} = 2.77$

b. The largest observation is 11. It is deleted from the data set. The new range is: 9 – 1 = 8.

The variance is: $s^2 = \dfrac{\sum_i x_i^2 - \dfrac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \dfrac{329 - \dfrac{67^2}{19}}{19-1} = 5.1520$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{5.1520} = 2.27$

When the largest observation is deleted, the range, variance and standard deviation decrease.

c. The largest observation is 11 and the smallest is 1. When these two observations are deleted from the data set, the new range is: 9 – 1 = 8.

The variance is: $s^2 = \dfrac{\sum_i x_i^2 - \dfrac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \dfrac{328 - \dfrac{66^2}{18}}{18-1} = 5.0588$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{5.0588} = 2.25$

When the largest and smallest observations are deleted, the range, variance and standard deviation decrease.

2.70 a. A worker's overall time to complete the operation under study is determined by adding the subtask-time averages.

Worker A

The average for subtask 1 is: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{211}{7} = 30.14$

The average for subtask 2 is: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{21}{7} = 3$

Worker A's overall time is 30.14 + 3 = 33.14.

Worker B

The average for subtask 1 is: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{213}{7} = 30.43$

The average for subtask 2 is: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{29}{7} = 4.14$

Worker B's overall time is 30.43 + 4.14 = 34.57.

b. **Worker A**

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{6455 - \frac{211^2}{7}}{7-1}} = \sqrt{15.8095} = 3.98$$

Worker B

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{6487 - \frac{213^2}{7}}{7-1}} = \sqrt{.9524} = .98$$

c. The standard deviations represent the amount of variability in the time it takes the worker to complete subtask 1.

d. **Worker A**

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{67 - \frac{21^2}{7}}{7-1}} = \sqrt{.6667} = .82$$

Worker B

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{147 - \frac{29^2}{7}}{7-1}} = \sqrt{4.4762} = 2.12$$

e. I would choose workers similar to worker B to perform subtask 1. Worker B has a slightly higher average time on subtask 1 (A: $\bar{x} = 30.14$, B: $\bar{x} = 30.43$). However, Worker B has a smaller variability in the time it takes to complete subtask 1 (part **b**). He or she is more consistent in the time needed to complete the task.

I would choose workers similar to Worker A to perform subtask 2. Worker A has a smaller average time on subtask 2 (A: $\bar{x} = 3$, B: $\bar{x} = 4.14$). Worker A also has a smaller variability in the time needed to complete subtask 2 (part **d**).

2.71 a. The unit of measurement of the variable of interest is dollars (the same as the mean and standard deviation). Based on this, the data are quantitative.

b. Since no information is given about the shape of the data set, we can only use Chebyshev's Rule.

\$900 is 2 standard deviations below the mean, and \$2100 is 2 standard deviations above the mean. Using Chebyshev's Rule, at least 3/4 of the measurements (or $3/4 \times 200 = 150$ measurements) will fall between \$900 and \$2100.

$600 is 3 standard deviations below the mean and $2400 is 3 standard deviations above the mean. Using Chebyshev's Rule, at least 8/9 of the measurements (or $8/9 \times 200 \approx 178$ measurements) will fall between $600 and $2400.

$1200 is 1 standard deviation below the mean and $1800 is 1 standard deviation above the mean. Using Chebyshev's Rule, nothing can be said about the number of measurements that will fall between $1200 and $1800.

$1500 is equal to the mean and $2100 is 2 standard deviations above the mean. Using Chebyshev's Rule, at least 3/4 of the measurements (or $3/4 \times 200 = 150$ measurements) will fall between $900 and $2100. It is possible that all of the 150 measurements will be between $900 and $1500. Thus, nothing can be said about the number of measurements between $1500 and $2100.

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

2.73 According to the Empirical Rule:

a. Approximately 68% of the measurements will be contained in the interval $\bar{x} - s$ to $\bar{x} + s$.

b. Approximately 95% of the measurements will be contained in the interval $\bar{x} - 2s$ to $\bar{x} + 2s$.

c. Essentially all the measurements will be contained in the interval $\bar{x} - 3s$ to $\bar{x} + 3s$.

2.74 a. $\bar{x} = \dfrac{\sum x}{n} = \dfrac{206}{25} = 8.24$

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

b.

Interval	Number of Measurements in Interval	Percentage
$\bar{x} \pm s$, or (6.41, 10.07)	18	$18/25 = .72$ or 72%
$\bar{x} \pm 2s$, or (4.58, 11.90)	24	$24/25 = .96$ or 96%
$\bar{x} \pm 3s$, or (2.75, 13.73)	25	$25/25 = 1.00$ 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. $\text{Range} = 12 - 5 = 7$ and $s \approx \dfrac{\text{Range}}{4} = \dfrac{7}{4} = 1.75$

The range approximation provides a satisfactory estimate of $s = 1.83$ from part **a**.

2.75 Using Chebyshev's Rule, at least 8/9 of the measurements will fall within 3 standard deviations of the mean. Thus, the range of the data would be around 6 standard deviations. Using the Empirical Rule, approximately 95% of the observations are within 2 standard deviations of the mean. Thus, the range of the data would be around 4 standard deviations. We would expect the standard deviation to be somewhere between Range/6 and Range/4.

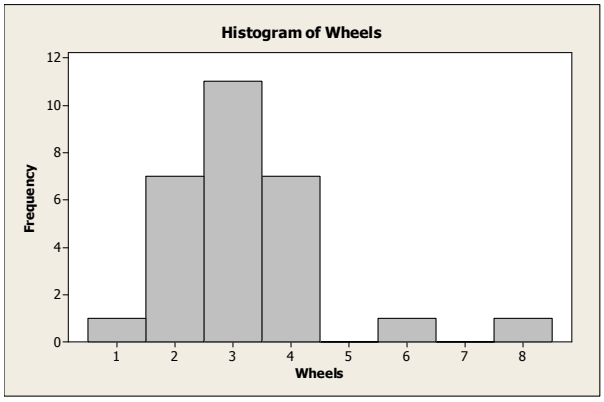
For our data, the $\text{range} = 760 - 135 = 625$.

The $\dfrac{\text{Range}}{6} = \dfrac{625}{6} = 104.17$ and $\dfrac{\text{Range}}{4} = \dfrac{625}{4} = 156.25$.

Therefore, I would estimate that the standard deviation of the data set is between 104.17 and 156.25.

It would not be feasible to have a standard deviation of 25. If the standard deviation were 25, the data would span 625/25 = 25 standard deviations. This would be extremely unlikely.

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



Since the distribution is skewed to the right, it is not mound-shaped and it is not symmetric.

b. Using MINITAB, the results are:

Descriptive Statistics: Wheels

```
Variable   N   Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Wheels    28  3.214  1.371    1.000  2.000   3.000  4.000    8.000
```

The mean is 3.214 and the standard deviation is 1.371.

c. The interval is: $\bar{x} \pm 2s \Rightarrow 3.214 \pm 2(1.371) \Rightarrow 3.214 \pm 2.742 \Rightarrow (0.472, 5.956)$.

d. According to Chebyshev's rule, at least 75% of the observations will fall within 2 standard deviations of the mean.

e. According to the Empirical Rule, approximately 95% of the observations will fall within 2 standard deviations of the mean.

f. Actually, 26 of the 28 or 26/28 = .929 of the observations fall within the interval. This value is close to the 95% that we would expect with the Empirical Rule.

2.77 a. The interval $\bar{x} \pm 2s$ will contain at least 75% of the observations. This interval is $\bar{x} \pm 2s \Rightarrow 3.11 \pm 2(.66) \Rightarrow 3.11 \pm 1.32 \Rightarrow (1.79,\ 4.43)$.

b. No. The value 1.25 does not fall in the interval $\bar{x} \pm 2s$. We know that at least 75% of all observations will fall within 2 standard deviations of the mean. Since 1.25 falls more than 2 standard deviations from the mean, it would not be a likely value to observe.

2.78 a. Using Chebyshev's Rule, at least 75% of the observations will fall within 2 standard deviations of the mean.

$\bar{x} \pm 2s \Rightarrow 4.25 \pm 2(12.02) \Rightarrow 4.25 \pm 24.04 \Rightarrow (-19.79,\ 28.29)$ or (0, 28.29) since we cannot have a negative number blogs.

b. We would expect the distribution to be skewed to the right. We know that we cannot have a negative number of blogs/forums. Even 1 standard deviation below the mean is a negative number. We would assume that there are a few very large observations because the standard deviation is so big compared to the mean.

2.79 a. The 2 standard deviation interval around the mean is:

$\bar{x} \pm 2s \Rightarrow 141.31 \pm 2(17.77) \Rightarrow 141.31 \pm 35.54 \Rightarrow (105.77,\ \ 176.85)$

b. Using Chebyshev's Theorem, at least ¾ of the observations will fall within 2 standard deviations of the mean. Thus, at least ¾ of first-time candidates for the CPA exam have total credit hours between 105.77 and 176.85.

c. In order for the above statement to be true, nothing needs to be known about the shape of the distribution of total semester hours.

2.80 a. Since the data are mound-shaped and symmetric, we know from the Empirical Rule that approximately 95% of the observations will fall within 2 standard deviations of the mean. This interval will be: $\bar{x} \pm 2s \Rightarrow 39 \pm 2(6) \Rightarrow 39 \pm 12 \Rightarrow (27,\ 51)$.

b. We know that approximately .05 of the observations will fall outside the range 27 to 51. Since the distribution of scores is symmetric, we know that half of the .05 or .025 will fall above 51.

c. We know from the Empirical Rule that approximately 99.7% (essentially all) of the observations will fall within 3 standard deviations of the mean. This interval is: $\bar{x} \pm 3s \Rightarrow 39 \pm 3(6) \Rightarrow 39 \pm 18 \Rightarrow (21,\ 57)$.

2.81 a. The sample mean is: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{17{,}800}{186} = 95.699$

The sample variance is: $s^2 = \dfrac{\sum_{i=1}^{n} x^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \dfrac{1{,}707{,}998 - \dfrac{17{,}800^2}{186}}{186-1} = 24.6332$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{24.6332} = 4.9632$

b. $\bar{x} \pm s \Rightarrow 95.699 \pm 4.963 \Rightarrow (90.736,\ 100.662)$

$\bar{x} \pm 2s \Rightarrow 95.699 \pm 2(4.963) \Rightarrow 95.699 \pm 9.926 \Rightarrow (85.773,\ \ 105.625)$

$\bar{x} \pm 3s \Rightarrow 95.699 \pm 3(4.963) \Rightarrow 95.699 \pm 14.889 \Rightarrow (80.810,\ \ 110.558)$

c. There are 166 out of 186 observations in the first interval. This is $(166/186)\times 100\% = 89.2\%$. There are 179 out of 186 observations in the second interval. This is $(179/186)\times 100\% = 96.2\%$. There are 182 out of 186 observations in the second interval. This is $(182/186)\times 100\% = 97.8\%$.

The percentages for the first 2 intervals are much larger than we would expect using the Empirical Rule. The Empirical Rule indicates that approximately 68% of the observations will fall within 1 standard deviation of the mean. It also indicates that approximately 95% of the observations will fall within 2 standard deviations of the mean. Chebyshev's Theorem says that at least ¾ or 75% of the observations will fall within 2 standard deviations of the mean and at least 8/9 or 88.9% of the observations will fall within 3 standard deviations of the mean. It appears that our observed percentages agree with Chebyshev's Theorem better than the Empirical Rule.

2.82 a. The interval is: $\bar{x} \pm 2s \Rightarrow 13.2 \pm 2(19.5) \Rightarrow 13.2 \pm 39 \Rightarrow (-25.8,\ 52.2)$ or (0, 52.2) since we cannot have negative number of minutes.

b. Since this interval contains negative numbers, we know that the distribution cannot be symmetric. One cannot have negative values for time spent on a laptop computer.

c. Since we know the data are not symmetric, we must use Chebyshev's Rule. At least ¾ or 75% of the observations will fall between -25.8 and 52.2 or between 0 and 52.2 minutes.

2.83 The sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{240.9 + 248.8 + 215.7 + \cdots + 238.0}{10} = \frac{2347.4}{10} = 234.74$$

The sample variance deviation is:

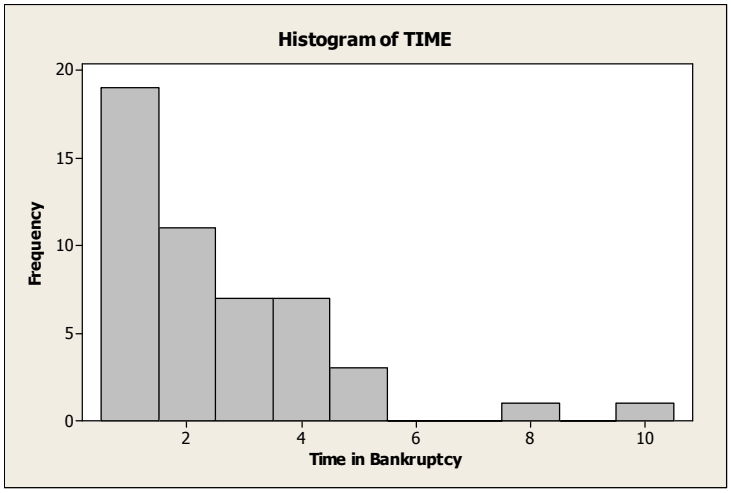
$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{551{,}912.1 - \dfrac{2347.4^2}{10}}{9} = \frac{883.424}{9} = 98.1582$$

The sample standard deviation is: $\sqrt{s^2} = \sqrt{98.1582} = 9.91$

The data are fairly symmetric, so we can use the Empirical Rule. We know from the Empirical Rule that almost all of the observations will fall within 3 standard deviations of the mean. This interval would be:

$\bar{x} \pm 3s \Rightarrow 234.74 \pm 3(9.91) \Rightarrow 234.74 \pm 29.73 \Rightarrow (205.01,\ \ 264.47)$

2.84 a. Using MINITAB, the frequency histogram for the time in bankruptcy is:



The Empirical Rule is not applicable because the data are not mound shaped.

b. Using MINITAB, the descriptive measures are:

Descriptive Statistics: TIME

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TIME	49	2.549	1.828	1.000	1.350	1.700	3.500	10.100

From Chebyshev's Theorem, we know that at least 75% of the observations will fall within 2 standard deviations of the mean. This interval is:

$\bar{x} \pm 2s \Rightarrow 2.549 \pm 2(1.828) \Rightarrow 2.549 \pm 3.656 \Rightarrow (-1.107,\ \ 6.205)$ or (0, 6.205) since we cannot have negative months.

c. There are 47 of the 49 observations within this interval. The percentage would be $(47/49) \times 100\% = 95.9\%$. This agrees with Chebyshev's Theorem (at least 75%). It also agrees with the Empirical Rule (approximately 95%).

d. From the above interval we know that about 95% of all firms filing for prepackaged bankruptcy will be in bankruptcy between 0 and 6.2 months. Thus, we would estimate that a firm considering filing for bankruptcy will be in bankruptcy up to 6.2 months.

2.85 a. The interval $\bar{x} \pm 2s$ for the flexed arm group is $\bar{x} \pm 2s \Rightarrow 59 \pm 3(4) \Rightarrow 59 \pm 12 \Rightarrow (47, 71)$. The interval for the extended are group is $\bar{x} \pm 2s \Rightarrow 43 \pm 3(2) \Rightarrow 43 \pm 6 \Rightarrow (37, 49)$. We know that at least 8/9 or 88.9% of the observations will fall within 3 standard deviations of the mean using Chebyshev's Rule. Since these 2 intervals barely overlap, the information supports the researchers' theory. The shoppers from the flexed arm group are more likely to select vice options than the extended arm group.

b. The interval $\bar{x} \pm 2s$ for the flexed arm group is $\bar{x} \pm 2s \Rightarrow 59 \pm 2(10) \Rightarrow 59 \pm 20 \Rightarrow (39, 79)$. The interval for the extended are group is $\bar{x} \pm 2s \Rightarrow 43 \pm 2(15) \Rightarrow 43 \pm 30 \Rightarrow (13, 73)$. Since these two intervals overlap almost completely, the information does not support the researcher's theory. There does not appear to be any difference between the two groups.

2.86 a. Yes. The distribution of the buy-side analysts is fairly flat and skewed to the right. The distribution of the sell-side analysts is more mound shaped and is not spread out as far as the buy-side distribution. Since the buy-side distribution is more spread out, the variance of the buy-side distribution will be larger than the variance of the sell-side distribution. Because the buy-side distribution is skewed to the right, the mean will be pulled to the right. Thus, the mean of the buy-side distribution will be greater than the mean of the sell-side distribution.

b. Since the sell-side distribution is fairly mound-shaped, we can use the Empirical Rule. The Empirical Rule says that approximately 95% of the observations will fall within 2 standard deviations of the mean. The interval for the sell-side distribution would be:

$$\bar{x} \pm 2s \Rightarrow -.05 \pm 2(.85) \Rightarrow -.05 \pm 1.7 \Rightarrow (-1.75, \quad 1.65)$$

Since the buy-side distribution is skewed to the right, we cannot use the Empirical Rule. Thus, we will use Chebyshev's Rule. We know that at least $(1 - 1/k^2)$ will fall within k standard deviations of the mean. If we choose $k = 4$, then $(1-1/4^2) = .9375$ or 93.75%. This is very close to 95% requested in the problem. The interval for the buy-side distribution to contain at least 93.75% of the observations would be: $\bar{x} \pm 4s \Rightarrow .85 \pm 4(1.93) \Rightarrow .85 \pm 7.72 \Rightarrow (-6.87, \quad 8.57)$

Note: This interval will contain *at least* 93.75% of the observations. It may contain more than 93.75% of the observations.

2.87 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 $8/9 = .89$ of the tree heights on this piece of land fall within this interval and at most $1/9 = .11$ of the tree heights will fall above the interval. However, the buyer will only purchase the land if at least $\frac{1000}{5000} = .20$ of the tree heights are at least 40 feet tall. Therefore, the buyer should not buy the piece of land.

2.88 a. Since we do not have any idea of the shape of the distribution of SAT-Math score changes, we must use Chebyshev's Theorem. We know that at least 8/9 of the observations will fall within 3 standard deviations of the mean. This interval would be:

$$\bar{x} \pm 3s \Rightarrow 19 \pm 3(65) \Rightarrow 19 \pm 195 \Rightarrow (-176, \ 214)$$

Thus, for a randomly selected student, we could be pretty sure that this student's score would be anywhere from 176 points below his/her previous SAT-Math score to 214 points above his/her previous SAT-Math score.

b. Since we do not have any idea of the shape of the distribution of SAT-Verbal score changes, we must use Chebyshev's Theorem. We know that at least 8/9 of the observations will fall within 3 standard deviations of the mean. This interval would be:

$$\bar{x} \pm 3s \Rightarrow 7 \pm 3(49) \Rightarrow 7 \pm 147 \Rightarrow (-140, \ 154)$$

Thus, for a randomly selected student, we could be pretty sure that this student's score would be anywhere from 140 points below his/her previous SAT-Verbal score to 154 points above his/her previous SAT-Verbal score.

c. A change of 140 points on the SAT-Math would be a little less than 2 standard deviations from the mean. A change of 140 points on the SAT-Verbal would be a little less than 3 standard deviations from the mean. Since the 140 point change for the SAT-Math is not as big a change as the 140 point on the SAT-Verbal, it would be most likely that the score was a SAT-Math score.

2.89 We know $\mu = 25$ and $\sigma = 1$. Therefore, $\mu \pm 2\sigma \Rightarrow 25 \pm 2(.1) \Rightarrow 25 \pm .2 \Rightarrow (24.8,\ 25.2)$

The machine is shut down for adjustment if the contents of two consecutive bags fall more than 2 standard deviations from the mean (i.e., outside the interval (24.8, 25.2)). Therefore, the machine was shut down yesterday at 11:30 (25.23 and 25.25 are outside the interval) and again at 4:00 (24.71 and 25.31 are outside the interval).

2.90 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.

2.91 Using the definition of a percentile:

	Percentile	Percentage Above	Percentage Below
a.	75th	25%	75%
b.	50th	50%	50%
c.	20th	80%	20%
d.	84th	16%	84%

2.92 Q_L corresponds to the 25th percentile. Q_M corresponds to the 50th percentile. Q_U corresponds to the 75th percentile.

2.93 We first compute *z*-scores for each *x* value.

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

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

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

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

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

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

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

$$\Rightarrow -2\sigma = 40-\mu \qquad \Rightarrow 3\sigma = 90-\mu$$
$$\Rightarrow \mu - 2\sigma = 40 \qquad \Rightarrow \mu + 3\sigma = 90$$
$$\Rightarrow \mu = 40 + 2\sigma$$

By substitution, $40 + 2\sigma + 3\sigma = 90 \Rightarrow 5\sigma = 50 \Rightarrow \sigma = 10$ and $\mu = 40 + 2(10) = 60$.

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

2.95 The mean score of U.S. eighth-graders on a mathematics assessment test is 283. This is the average score. The 25th percentile is 259. This means that 25% of the U.S. eighth-graders score below 259 on the test and 75% score higher. The 75th percentile is 308. This means that 75% of the U.S. eighth-graders score below 308 on the test and 25% score higher. The 90th percentile is 329. This means that 90% of the U.S. eighth-graders score below 329 on the test and 10% score higher.

2.96 a. The *z*-score is $z = \frac{x-\bar{x}}{s} = \frac{30-39}{6} = -1.5$. A score of 30 is 1.5 standard deviations below the mean.

b. Since the data are mound-shaped and symmetric and 39 is the mean, .5 of the sampled drug dealers will have WR scores below 39.

c. If 5% of the drug dealers have WR scores above 49, then 95% will have WR scores below 49. Thus, 49 will be the 95th percentile.

2.97 A median starting salary of $41,100 indicates that half of the University of South Florida graduates had starting salaries less than $41,100 and half had starting salaries greater than $41,100. At mid-career, half of the University of South Florida graduates had a salary less than $71,100 and half had salaries greater than $71,100. At mid-career, 90% of the University of South Florida graduates had salaries under $131,000 and 10% had salaries greater than $131,000.

2.98 a. From Exercise 2.81, $\bar{x} = 95.699$ and $s = 4.963$. The *z*-score for an observation of 74 is:

$$z = \frac{x-\bar{x}}{s} = \frac{74-95.699}{4.963} = -4.37$$

This *z*-score indicates that an observation of 74 is 4.37 standard deviations below the mean. Very few observations will be lower than this one.

b. The z-score for an observation of 98 is:

$$z = \frac{x-\bar{x}}{s} = \frac{92-95.699}{4.963} = -0.75$$

This *z*-score indicates that an observation of 92 is .75 standard deviations below the mean. This score

is not an unusual observation in the data set.

2.99 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.

2.100 The *z*-score associated with a score of 155 is $z = \frac{x - \bar{x}}{s} = \frac{155 - 67.755}{26.871} = 3.25$. This score would not be considered a typical level of support. It is 3.25 standard deviations above the mean. Very few observations would be above this value.

2.101 a. The 10th percentile is the score that has at least 10% of the observations less than it. If we arrange the data in order from the smallest to the largest, the 10th percentile score will be the .10(75) = 7.5 or 8th observation. When the data are arranged in order, the 8th observation is 0. Thus, the 10th percentile is 0.

b. The 95th percentile is the score that has at least 95% of the observations less than it. If we arrange the data in order from the smallest to the largest, the 95th percentile score will be the .95(75) = 71.25 or 72nd observation. When the data are arranged in order, the 72nd observation is 21. Thus, the 95th percentile is 21.

c. The sample mean is: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{393}{75} = 5.24$

The sample variance is: $s^2 = \frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \frac{5943 - \frac{393^2}{75}}{75-1} = 52.482$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{52.482} = 7.244$

The z-score for a county with 48 Superfund sites is: $z = \frac{x - \bar{x}}{s} = \frac{48 - 5.24}{7.244} = 5.90$

d. Yes. A score of 48 is almost 6 standard deviations from the mean. We know that for any data set almost all (at least 8/9 using Chebyshev's Theorem) of the observations are within 3 standard deviations of the mean. To be almost 6 standard deviations from the mean is very unusual.

2.102 a. Since the data are approximately mound-shaped, we can use the Empirical Rule.
On the blue exam, the mean is 53% and the standard deviation is 15%. We know that approximately 68% of all students will score within 1 standard deviation of the mean. This interval is:

$\bar{x} \pm s \Rightarrow 53 \pm 15 \Rightarrow (38,\ 68)$

About 95% of all students will score within 2 standard deviations of the mean. This interval is:
$\bar{x} \pm 2s \Rightarrow 53 \pm 2(15) \Rightarrow 53 \pm 30 \Rightarrow (23,\ 83)$

About 99.7% of all students will score within 3 standard deviations of the mean. This interval is:

$\bar{x} \pm 3s \Rightarrow 53 \pm 3(15) \Rightarrow 53 \pm 45 \Rightarrow (8,\ 98)$

b. Since the data are approximately mound-shaped, we can use the Empirical Rule.

On the red exam, the mean is 39% and the standard deviation is 12%. We know that approximately 68% of all students will score within 1 standard deviation of the mean. This interval is:

$\bar{x} \pm s \Rightarrow 39 \pm 12 \Rightarrow (27,\ 51)$

About 95% of all students will score within 2 standard deviations of the mean. This interval is:

$\bar{x} \pm 2s \Rightarrow 39 \pm 2(12) \Rightarrow 39 \pm 24 \Rightarrow (15,\ 63)$

About 99.7% of all students will score within 3 standard deviations of the mean. This interval is:

$\bar{x} \pm 3s \Rightarrow 39 \pm 3(12) \Rightarrow 39 \pm 36 \Rightarrow (3,\ 75)$

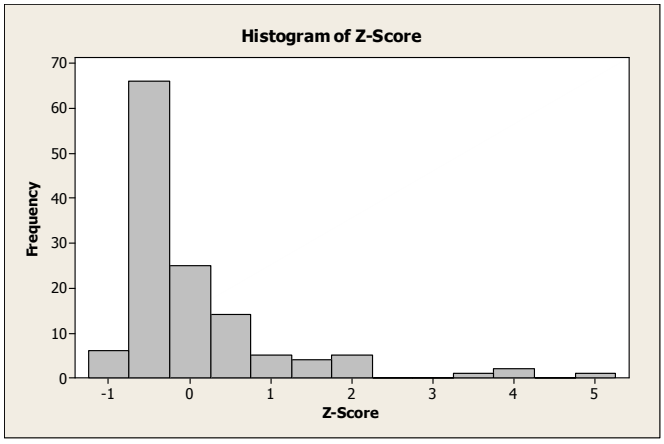
c. The student would have been more likely to have taken the red exam. For the blue exam, we know that approximately 95% of all scores will be from 23% to 83%. The observed 20% score does not fall in this range. For the red exam, we know that approximately 95% of all scores will be from 15% to 63%. The observed 20% score does fall in this range. Thus, it is more likely that the student would have taken the red exam.

2.103 a. The z-score for Harvard is $z = 5.08$. This means that Harvard's productivity score was 5.08 standard deviations above the mean. This is extremely high and extremely unusual.

b. The z-score for Howard University is $z = -.85$. This means that Howard University's productivity score was .85 standard deviations below the mean. This is not an unusual z-score.

c. Yes. Other indicators that the distribution is skewed to the right are the values of the highest and lowest z-scores. The lowest z-score is less than 1 standard deviation below the mean while the highest z-score is 5.08 standard deviations above the mean.

Using MINITAB, the histogram of the z-scores is:



This histogram does imply that the data are skewed to the right.

2.104 a. From the problem, $\mu = 2.7$ and $\sigma = .5$

$$z = \frac{x-\mu}{\sigma} \Rightarrow z\sigma = x-\mu \Rightarrow x = \mu + z\sigma$$

For $z = 2.0$, $x = 2.7 + 2.0(.5) = 3.7$

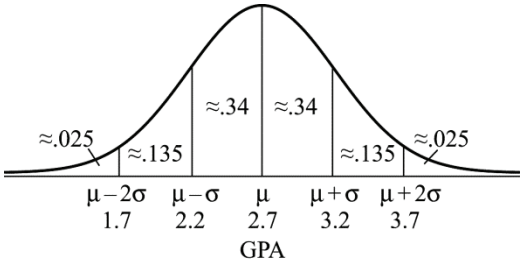
For $z = -1.0$, $x = 2.7 - 1.0(.5) = 2.2$

For $z = .5$, $x = 2.7 + .5(.5) = 2.95$

For $z = -2.5$, $x = 2.7 - 2.5(.5) = 1.45$

b. For $z = -1.6$, $x = 2.7 - 1.6(.5) = 1.9$

c. If we assume the distribution of GPAs is approximately mound-shaped, we can use the Empirical Rule.



From the Empirical Rule, we know that ≈.025 or ≈2.5% of the students will have GPAs above 3.7 (with $z = 2$). Thus, the GPA corresponding to summa cum laude (top 2.5%) will be greater than 3.7 ($z > 2$).

We know that ≈.16 or 16% of the students will have GPAs above 3.2 ($z = 1$). Thus, the limit on GPAs for cum laude (top 16%) will be greater than 3.2 ($z > 1$).

We must assume the distribution is mound-shaped.

2.105 Not necessarily. Because the distribution is highly skewed to the right, the standard deviation is very large. Remember that the *z*-score represents the number of standard deviations a score is from the mean. If the standard deviation is very large, then the *z*-scores for observations somewhat near the mean will appear to be fairly small. If we deleted the schools with the very high productivity scores and recomputed the mean and standard deviation, the standard deviation would be much smaller. Thus, most of the *z*-scores would be larger because we would be dividing by a much smaller standard deviation. This would imply a bigger spread among the rest of the schools than the original distribution with the few outliers.

2.106 To determine if the measurements are outliers, compute the *z*-score.

a. $z = \frac{x-\bar{x}}{s} = \frac{65-57}{11} = .727$ Since the *z*-score is less than 3, this would not be an outlier.

b. $z = \frac{x-\bar{x}}{s} = \frac{21-57}{11} = -3.273$ Since the *z*-score is greater than 3 in absolute value, this would be an outlier.

c. $z = \dfrac{x - \bar{x}}{s} = \dfrac{72 - 57}{11} = 1.364$ Since the z-score is less than 3, this would not be an outlier.

d. $z = \dfrac{x - \bar{x}}{s} = \dfrac{98 - 57}{11} = 3.727$ Since the z-score is greater than 3 in absolute value, this would be an outlier.

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

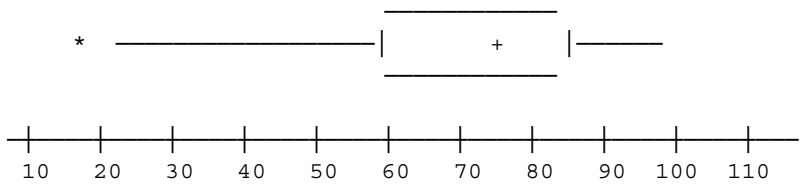
The lower inner fence $= Q_L - 1.5(IQR) = 60 - 1.5(25) = 22.5$.

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

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

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

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.

2.108 a. Median is approximately 4.

b. Q_L is approximately 3 (Lower Quartile)

Q_U is approximately 6 (Upper Quartile)

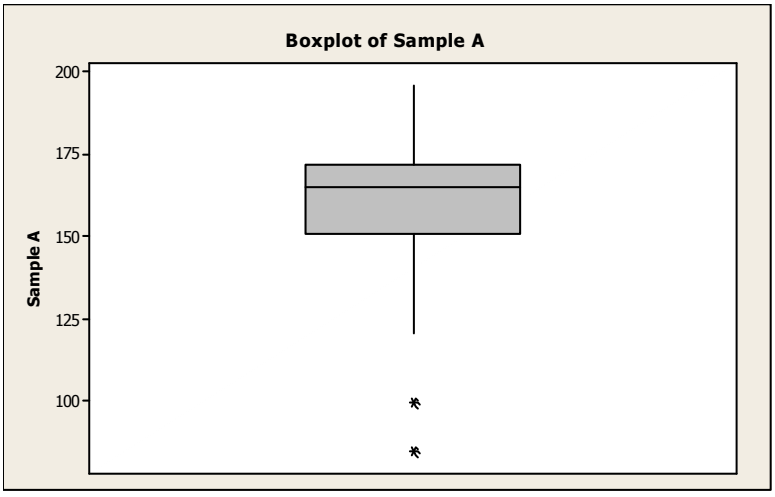
c. $IQR = Q_U - Q_L \approx 6 - 3 = 3$

d. The data set is skewed to the right since the right whisker is longer than the left, there is one outlier, and there are two potential outliers.

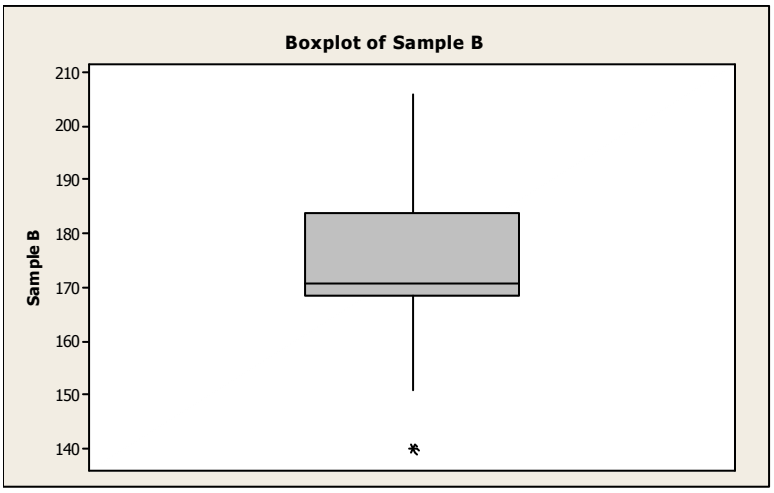
e. 50% of the measurements are to the right of the median and 75% are to the left of the upper quartile.

f. The upper inner fence is $Q_U + 1.5(IQR) = 6 + 1.5(3) = 10.5$. The upper outer fence is $Q_U + 3(IQR) = 6 + 3(3) = 15$. Thus, there are two suspect outliers, 12 and 13. There is one highly suspect outlier, 16.

2.109 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 measurement 84 is an outlier. This measurement falls outside the lower outer fence.

Lower outer fence = Lower hinge $-3(IQR) \approx 150 - 3(172 - 150) = 150 - 3(22) = 84$
Lower inner fence = Lower hinge $-1.5(IQR) \approx 150 - 1.5(22) = 117$
Upper inner fence = Upper hinge $+1.5(IQR) \approx 172 + 1.5(22) = 205$

In addition, 100 may be an outlier. It lies outside the inner fence.

In sample B, 140 and 206 may be outliers. The point 140 lies outside the inner fence while the point 206 lies right at the inner fence.

Lower outer fence = Lower hinge $-3(IQR) \approx 168 - 3(184 - 169) = 168 - 3(15) = 123$
Lower inner fence = Lower hinge $-1.5(IQR) \approx 168 - 1.5(15) = 145.5$
Upper inner fence = Upper hinge $+1.5(IQR) \approx 184 + 1.5(15) = 206.5$

2.110 a. The approximate 25th percentile PASI score before treatment is 10. The approximate median before treatment is 15. The approximate 75th percentile PASI score before treatment is 28.

b. The approximate 25th percentile PASI score after treatment is 3. The approximate median after treatment is 5. The approximate 75th percentile PASI score after treatment is 7.5.

c. Since the 75th percentile after treatment is lower than the 25th percentile before treatment, it appears that the ichthyotherapy is effective in treating psoriasis.

2.111 a. The average expenditure per full-time employee is \$6,563. The median expenditure per employee is \$6,232. Half of all expenditures per employee were less than \$6,232 and half were greater than \$6,232. The lower quartile is \$5,309. Twenty-five percent of all expenditures per employee were below \$5,309. The upper quartile is \$7,216. Seventy-five percent of all expenditures per employee were below \$7,216.

b. $IQR = Q_U - Q_L = \$7,216 - \$5,309 = \$1,907$.

c. The interquartile range goes from the 25th percentile to the 75th percentile. Thus, $.5 = .75 - .25$ of the 1,751 army hospitals have expenses between \$5,309 and \$7,216.

2.112 a. From the printout, $\bar{x} = 52.334$ and $s = 9.224$.

The highest salary is 75 (thousand).

The z-score is $z = \dfrac{x - \bar{x}}{s} = \dfrac{75 - 52.334}{9.224} = 2.46$

Therefore, the highest salary is 2.46 standard deviations above the mean.

The lowest salary is 35.0 (thousand).

The z-score is $z = \dfrac{x - \bar{x}}{s} = \dfrac{35.0 - 52.334}{9.224} = -1.88$

Therefore, the lowest salary is 1.88 standard deviations below the mean.

The mean salary offer is 52.33 (thousand).

The z-score is $z = \dfrac{x - \bar{x}}{s} = \dfrac{52.33 - 52.334}{9.224} = 0$

The z-score for the mean salary offer is 0 standard deviations from the mean.

No, the highest salary offer is not unusually high. For any distribution, at least 8/9 of the salaries should have z-scores between −3 and 3. A z-score of 2.46 would not be that unusual.

b. Since no salaries are outside the inner fences, none of them are suspect or highly suspect outliers.

2.113 a. The z-score is: $z = \dfrac{x - \bar{x}}{s} = \dfrac{160 - 141.31}{17.77} = 1.05$

Since the z-score is not large, it is not considered an outlier.

b. *Z*-scores with values greater than 3 in absolute value are considered outliers. An observation with a *z*-score of 3 would have the value:

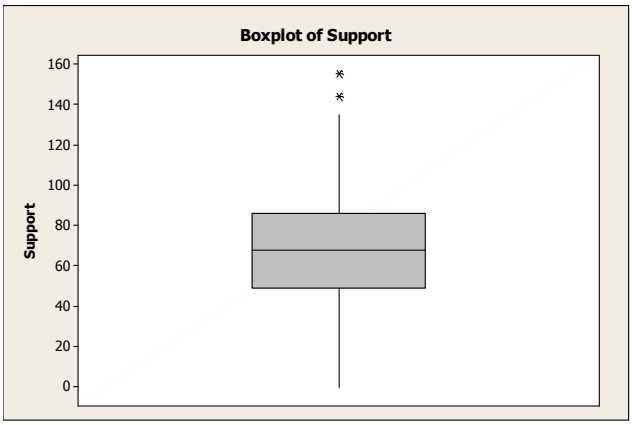
$$z = \frac{x - \bar{x}}{s} \Rightarrow 3 = \frac{x - 141.31}{17.77} \Rightarrow 3(17.77) = x - 141.31 \Rightarrow 53.31 = x - 141.31 \Rightarrow x = 194.62$$

An observation with a *z*-score of −3 would have the value:

$$z = \frac{x - \bar{x}}{s} \Rightarrow -3 = \frac{x - 141.31}{17.77} \Rightarrow -3(17.77) = x - 141.31 \Rightarrow -53.31 = x - 141.31 \Rightarrow x = 88.00$$

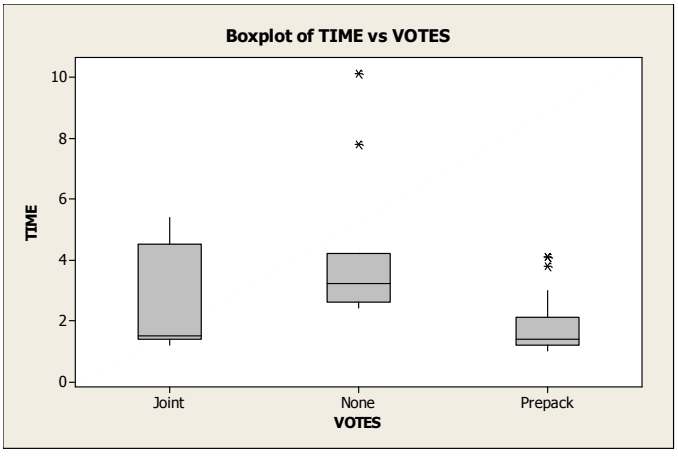
Thus any observation of semester hours that is greater than or equal to 194.62 or less than or equal to 88 would be considered an outlier.

2.114 From Exercise 2.100, $\bar{x} = 67.755$ and $s = 26.87$. Using MINITAB, a boxplot of the data is:



From the boxplot, the support level of 155 would be an outlier. From Exercise 2.100, we found the *z*-score associated with a score of 155 as $z = \frac{x - \bar{x}}{s} = \frac{155 - 67.755}{26.871} = 3.25$. Since this *z*-score is greater than 3, the observation 155 is considered an outlier.

2.115 a. Using MINITAB, the boxplots for each type of firm are:



b. The median bankruptcy time for Joint firms is about 1.5. The median bankruptcy time for None firms is about 3.2. The median bankruptcy time for Prepack firms is about 1.4.

c. The range of the "Prepack" firms is less than the other two, while the range of the "None" firms is the largest. The interquartile range of the "Prepack" firms is less than the other two, while the interquartile range of the "Joint" firms is larger than the other two.

d. No. The interquartile range for the "Prepack" firms is the smallest which corresponds to the smallest standard deviation. However, the second smallest interquartile range corresponds to the "None" firms. The second smallest standard deviation corresponds to the "Joint" firms.

e. Yes. There is evidence of two outliers in the "Prepack" firms. These are indicated by the two *'s. There is also evidence of two outliers in the "None" firms. These are indicated by the two *'s.

2.116 a. From Exercise 2.101, $\bar{x} = 5.24$, $s^2 = 52.482$, and $s = 7.244$.

We will use 3 standard deviations from the mean as the cutoff for outliers. *Z*-scores with values greater than 3 in absolute value are considered outliers. An observation with a *z*-score of 3 would have the value:

$$z = \frac{x - \bar{x}}{s} \Rightarrow 3 = \frac{x - 5.24}{7.244} \Rightarrow 3(7.244) = x - 5.24 \Rightarrow 21.732 = x - 5.24 \Rightarrow x = 26.972$$

An observation with a *z*-score of -3 would have the value:

$$z = \frac{x - \bar{x}}{s} \Rightarrow -3 = \frac{x - 5.24}{7.244} \Rightarrow -3(7.244) = x - 5.24 \Rightarrow -21.732 = x - 5.24 \Rightarrow x = -16.492$$

Thus, any observation that is greater than 26.972 or less than -16.492 would be considered an outlier. In this data set there would be 1 outlier: 48.

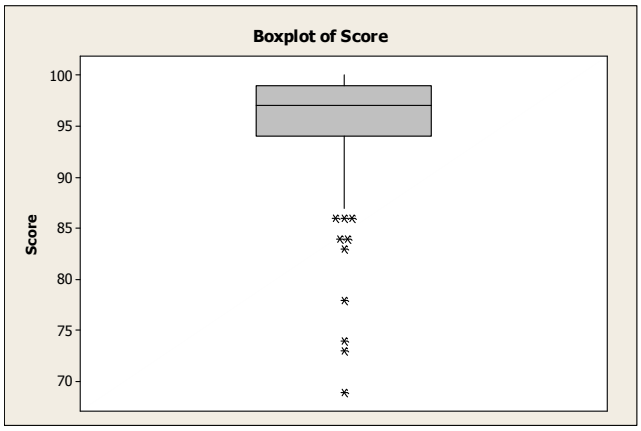
b. Deleting the observation 48, the sample mean is: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{345}{74} = 4.66$

The sample variance is: $s^2 = \frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \frac{3639 - \frac{345^2}{74}}{74-1} = 27.8158$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{27.8158} = 5.274$

The mean has decreased from 5.24 to 4.66, while the standard deviation decreased from 7.244 to 5.274.

2.117 a. Using MINITAB, the boxplot is:



From the boxplot, there appears to be 10 outliers: 69, 73, 74, 78, 83, 84, 84, 86, 86, and 86.

b. From Exercise 2.81, $\bar{x} = 95.699$ and $s = 4.963$. Since the data are skewed to the left, we will consider observations more than 2 standard deviations from the mean to be outliers. An observation with a z-score of 2 would have the value:

$$z = \frac{x - \bar{x}}{s} \Rightarrow 2 = \frac{x - 95.699}{4.963} \Rightarrow 2(4.963) = x - 95.699 \Rightarrow 9.926 = x - 95.699 \Rightarrow x = 105.625$$

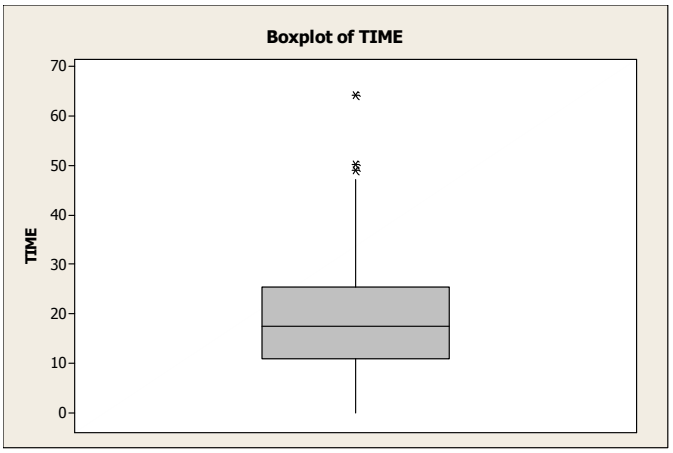
An observation with a z-score of -2 would have the value:

$$z = \frac{x - \bar{x}}{s} \Rightarrow -2 = \frac{x - 95.699}{4.963} \Rightarrow -2(4.963) = x - 95.699 \Rightarrow -9.926 = x - 95.699 \Rightarrow x = 85.773$$

Observations greater than 105.625 or less than 85.773 would be considered outliers. Using this criterion, the following observations would be outliers: 69, 73, 74, 78, 83, 84, and 84.

c. No, these methods do not agree exactly. Using the boxplot, 10 observations were identified as outliers. Using the z-score method, only 7 observations were identified as outliers. However, the 3 additional points that were not identified as outliers using the z-score method were very close to the cutoff value.

2.118 a. Using MINITAB, the box plot is:



The median is about 18. The data appear to be skewed to the right since there are 3 suspect outliers to the right and none to the left. The variability of the data is fairly small because the IQR is fairly small, approximately 26 – 10 = 16.

b. The customers associated with the suspected outliers are customers 268, 269, and 264.

c. In order to find the *z*-scores, we must first find the mean and standard deviation.

$$\bar{x} = \frac{\sum x}{n} = \frac{815}{40} = 20.375 \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{24129 - \frac{815^2}{40}}{40-1} = 192.90705$$

$$s = \sqrt{192.90705} = 13.89$$

The *z*-scores associated with the suspected outliers are:

Customer 268 $z = \dfrac{49 - 20.375}{13.89} = 2.06$

Customer 269 $z = \dfrac{50 - 20.375}{13.89} = 2.13$

Customer 264 $z = \dfrac{64 - 20.375}{13.89} = 3.14$

All the *z*-scores are greater than 2. These are unusual values.

2.119 From the stem-and-leaf display in Exercise 2.34, the data are fairly mound-shaped, but skewed somewhat to the right.

The sample mean is $\bar{x} = \dfrac{\sum x}{n} = \dfrac{1493}{25} = 59.72$.

The sample variance is $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{96{,}885 - \dfrac{1493^2}{25}}{25-1} = 321.7933$.

The sample standard deviation is $s = \sqrt{321.7933} = 17.9386$.

The z-score associated with the largest value is $z = \dfrac{x - \bar{x}}{s} = \dfrac{102 - 59.72}{17.9386} = 2.36$.

Since the data are not extremely skewed to the right, this observation is probably not an outlier.

The observations associated with the one-time customers are 5 of the largest 7 observations. Thus, repeat customers tend to have shorter delivery times than one-time customers.

2.120 For **Perturbed Intrinsics, but no Perturbed Projections**:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{8.1}{5} = 1.62 \qquad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{15.63 - \dfrac{8.1^2}{5}}{5-1} = \frac{2.508}{4} = .627 \qquad s = \sqrt{s^2} = \sqrt{.627} = .792$$

The z-score corresponding to a value of 4.5 is $z = \dfrac{x - \bar{x}}{s} = \dfrac{4.5 - 1.62}{.792} = 3.63$

Since this z-score is greater than 3, we would consider this an outlier for perturbed intrinsics, but no perturbed projections.

For **Perturbed Projections, but no Perturbed Intrinsics**:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{125.8}{5} = 25.16 \qquad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{3350.1 - \dfrac{125.8^2}{5}}{5-1} = \frac{184.972}{4} = 46.243$$

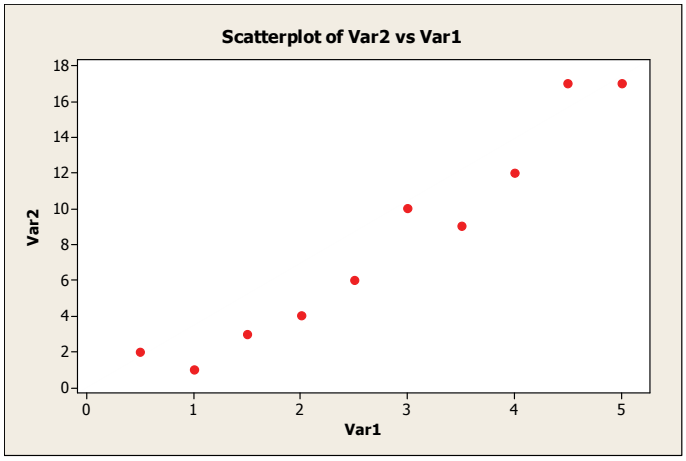
$$s = \sqrt{s^2} = \sqrt{46.243} = 6.800$$

The z-score corresponding to a value of 4.5 is $z = \dfrac{x - \bar{x}}{s} = \dfrac{4.5 - 25.16}{6.800} = -3.038$

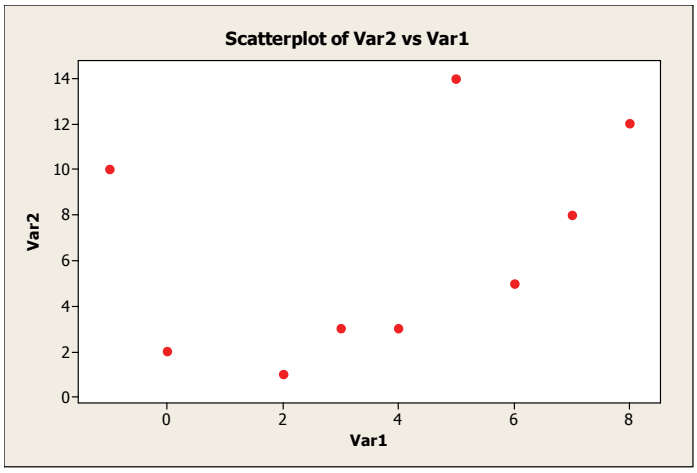
Since this z-score is less than -3, we would consider this an outlier for perturbed projections, but no perturbed intrinsics.

Since the z-score corresponding to 4.5 for the perturbed projections, but no perturbed intrinsics is smaller in absolute value than that for perturbed intrinsics, but no perturbed projections, it is more likely that the that the type of camera perturbation is perturbed projections, but no perturbed intrinsics.

2.121 Using MINITAB, the scatterplot is:

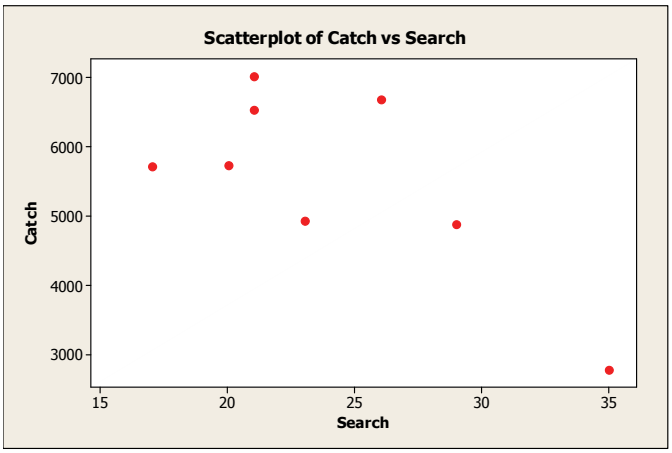


2.122 Using MINITAB, a scatterplot of the data is:



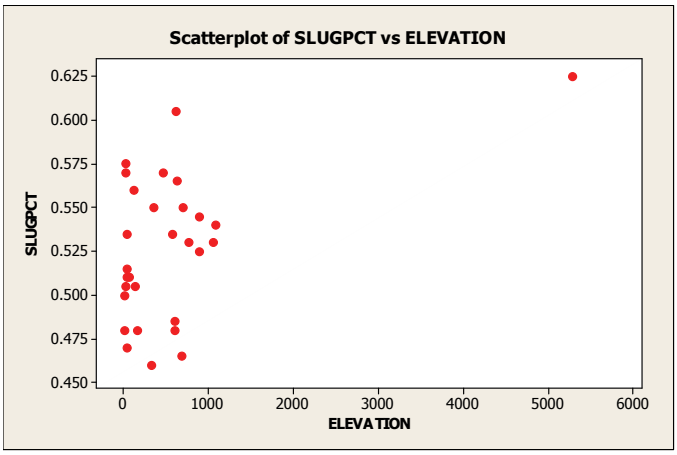
2.123. From the scatterplot of the data, it appears that as the number of punishments increases, the average payoff decreases. Thus, there appears to be a negative linear relationship between punishment use and average payoff. This supports the researchers conclusion that "winners" don't punish".

2.124 Using MINITAB, the scatterplot of the data is:



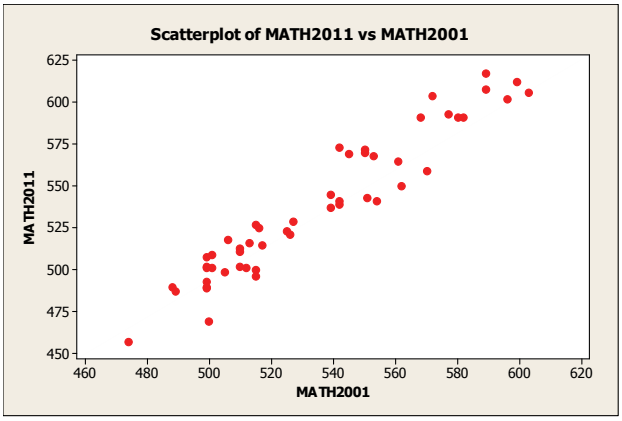
There is an apparent negative linear trend between the search frequency and the total catch. As the search frequency increases, the total catch tends to decrease.

2.125 Using MINITAB, a scattergram of the data is:



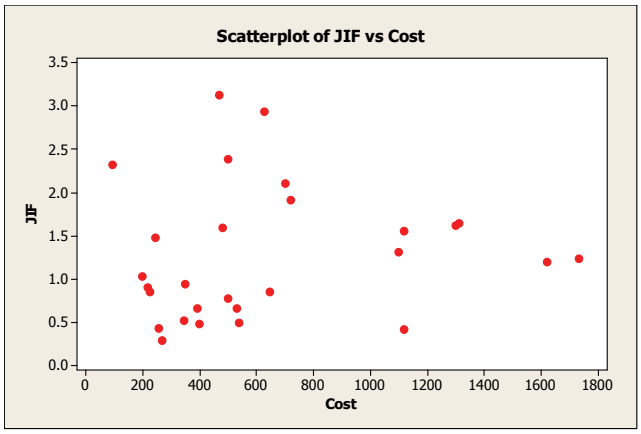
If we include the observation from Denver, then we would say there might be a linear relationship between slugging percentage and elevation. If we eliminated the observation from Denver, it appears that there might not be a relationship between slugging percentage and elevation.

2.126 Using MINITAB, the scatterplot of the data is:



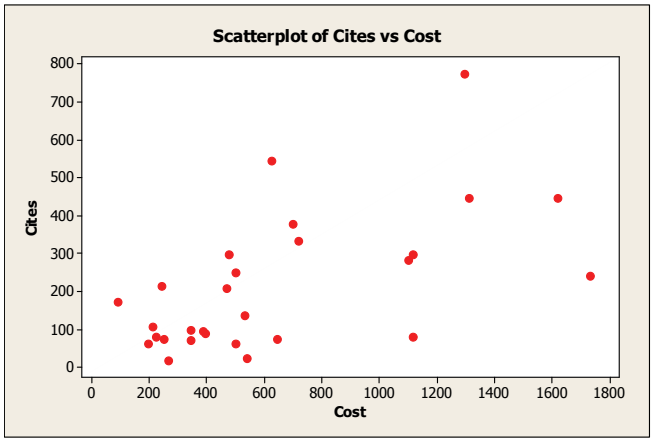
There appears to be a positive linear trend between the Math SAT scores in 2001 and the Math SAT scores in 2011. As the 2001 Math SAT scores increase, the 2011 Math SAT scores also tend to increase.

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



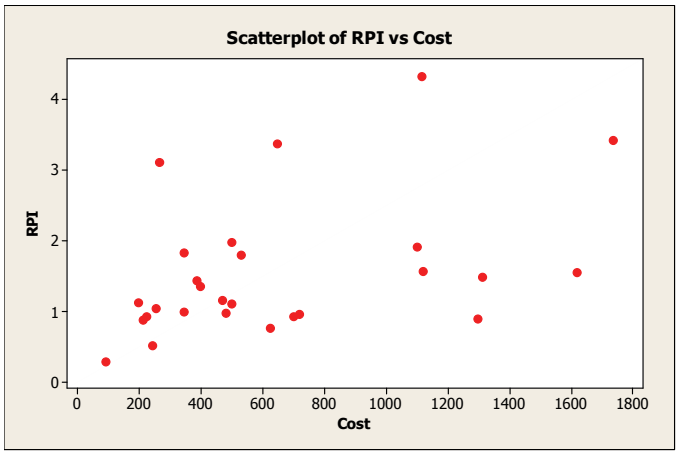
There is a slight negative linear trend to the data. As cost increases, JIF tends to decrease.

b. Using MINITAB, a scatterplot of the number of cities and cost is:



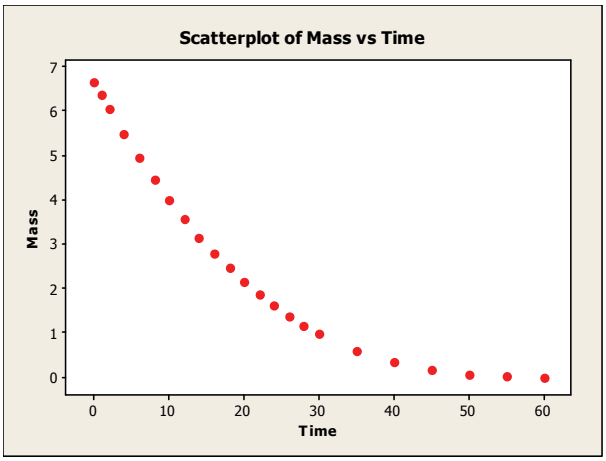
There is a moderate positive trend to the data. As cost increases, the number of cities tends to increase.

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



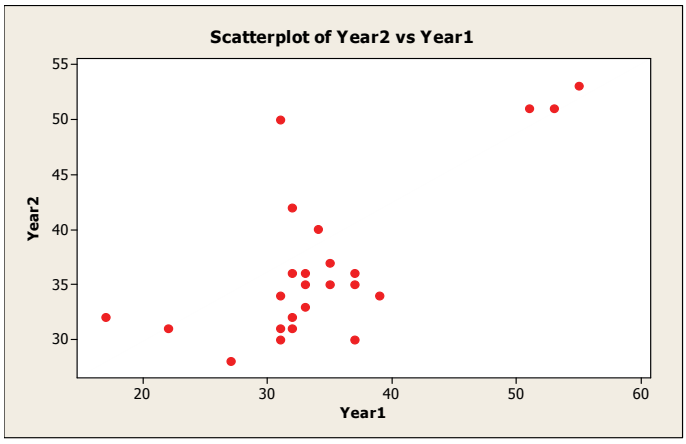
There is a slight positive trend to the data. As cost increases, RPI tends to increase.

2.128 Using MINITAB, the scatterplot of the data is:



There is evidence to indicate that the mass of the spill tends to diminish as time increases. As time is getting larger, the mass is decreasing.

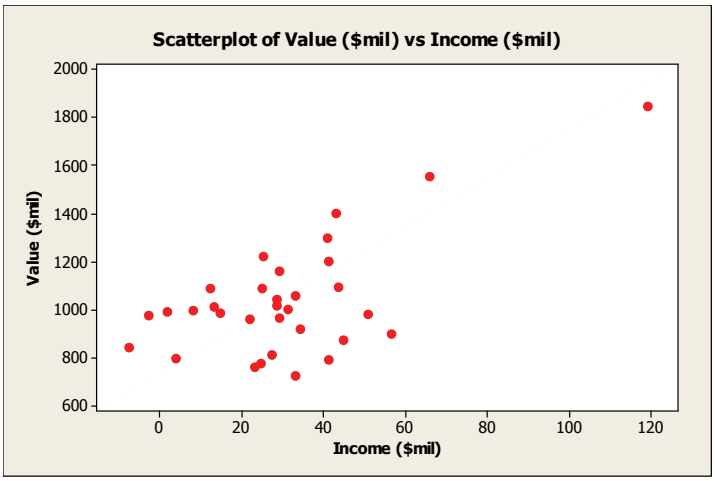
2.129 a. Using MINITAB, a scatterplot of the data is:



There is a moderate positive trend to the data. As the scores for Year1 increase, the scores for Year2 also tend to increase.

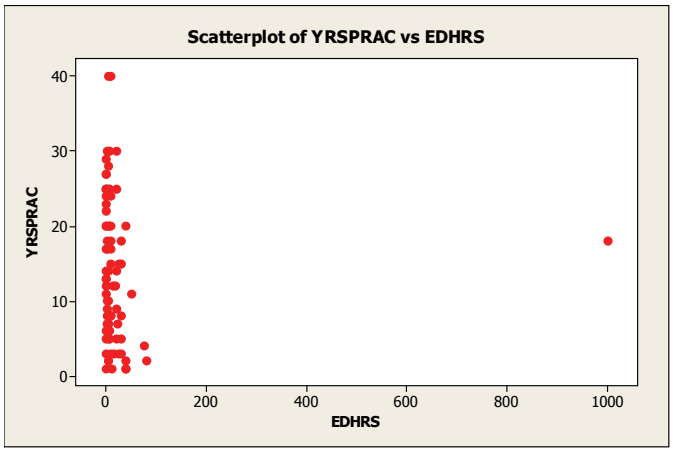
b. From the graph, two agencies that had greater than expected PARS evaluation scores for Year2 were USAID and State.

2.130 Using MINITAB, the scattergram of the data is:



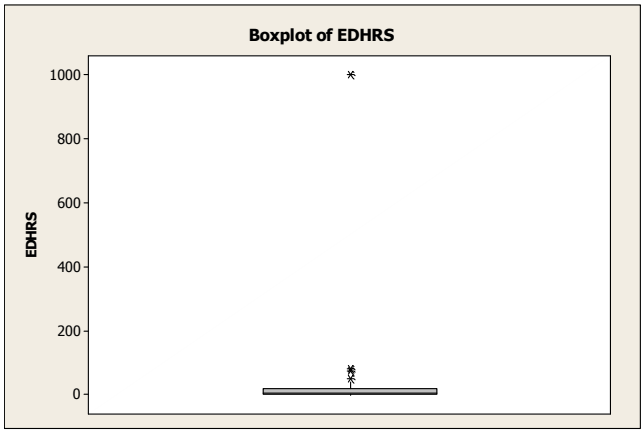
There is a moderate positive trend to the data. As operating income increases, the 2011 value also tends to increase. Since the trend is moderate, we would recommend that an NFL executive use operating income to predict a team's current value.

2.131 a. Using MINITAB, the scatterplot of the data is:



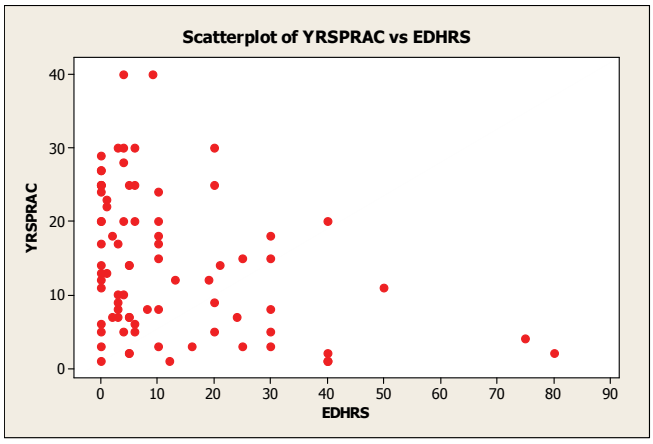
There does not appear to be much of a relationship between the years of experience and the amount of exposure to ethics in medical school.

b. Using MINITAB, a boxplot of the amount of exposure to ethics in medical school is:



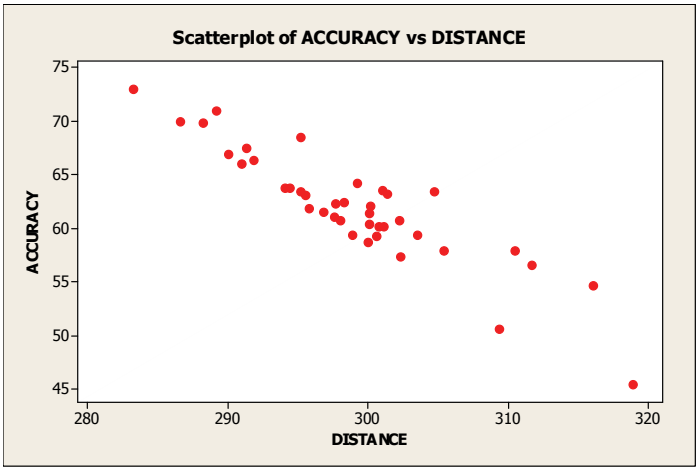
The one data point that is an extreme outlier is the value of 1000.

c. After removing this data point, the scatterplot of the data is:



With the data point removed, there now appears to be a negative trend to the data. As the amount of exposure to ethics in medical school increases, the years of experience decreases.

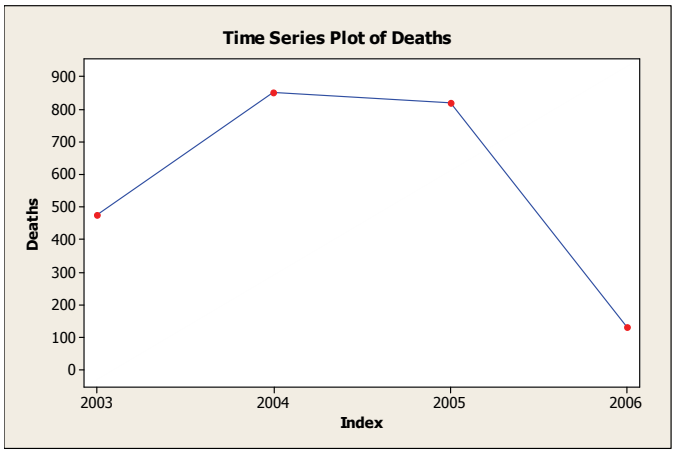
2.132 Using MINITAB, a scatterplot of the data is:



Yes, his concern is a valid one. From the scatterplot, there appears to be a fairly strong negative relationship between accuracy and driving distance. As driving distance increases, the driving accuracy tend to decrease.

2.133 One way the bar graph can mislead the viewer is that the vertical axis has been cut off. Instead of starting at 0, the vertical axis starts at 12. Another way the bar graph can mislead the viewer is that as the bars get taller, the widths of the bars also increase.

2.134 a. Using MINITAB, the time series plot is:

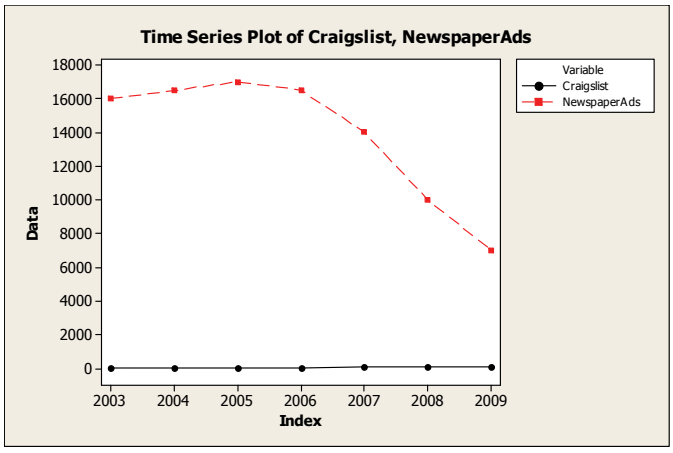


b. The time series plot is misleading because the information for 2006 is incomplete – it is based on only 2 months while all of the rest of the years are based on 12 months.

c. In order to construct a plot that accurately reflects the trend in American casualties from the Iraq War, we would want complete data for 2006 and information for the years 2007 through 2011.

2.135 a. The graph might be misleading because the scales on the vertical axes are different. The left vertical axis ranges from 0 to \$120 million. The right vertical axis ranges from 0 to \$20 billion.

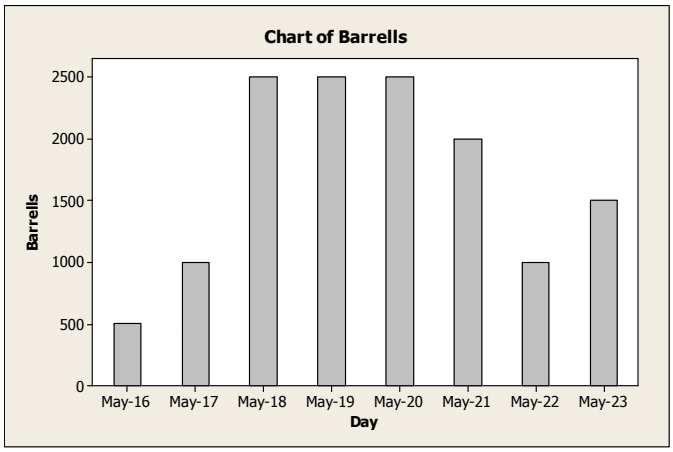
b. Using MINITAB, the redrawn graph is:



Although the amount of revenue produced by Craigslist has increased dramatically from 2003 to 2009, it is still much smaller than the revenue produced by newspaper ad sales.

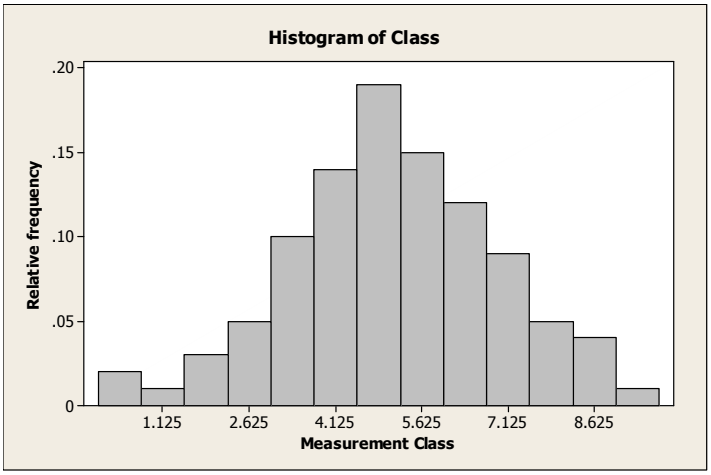
1.136 a. This graph is misleading because it looks like as the days are increasing, the number of barrels collected per day are also increasing. However, the bars are the cumulative number of barrels collected. The cumulative value can never decrease.

b. Using MINITAB, the graph of the daily collection of oil is:



From this graph, it shows that there has not been a steady improvement in the suctioning process. There was an increase for 3 days, then a leveling off for 3 days, then a decrease.

2.137 The relative frequency histogram is:



2.138 The mean is sensitive to extreme values in a data set. Therefore, the median is preferred to the mean when a data set is skewed in one direction or the other.

2.139 a. $z = \dfrac{x-\mu}{\sigma} = \dfrac{50-60}{10} = -1$ $\quad z = \dfrac{70-60}{10} = 1$ $\quad z = \dfrac{80-60}{10} = 2$

b. $z = \dfrac{x-\mu}{\sigma} = \dfrac{50-50}{5} = 0$ $\quad z = \dfrac{70-50}{5} = 4$ $\quad z = \dfrac{80-50}{5} = 6$

c $z = \dfrac{x-\mu}{\sigma} = \dfrac{50-40}{10} = 1$ $\quad z = \dfrac{70-40}{10} = 3$ $\quad z = \dfrac{80-40}{10} = 4$

d. $z = \dfrac{x-\mu}{\sigma} = \dfrac{50-40}{100} = .1$ $\quad z = \dfrac{70-40}{100} = .3$ $\quad z = \dfrac{80-40}{100} = .4$

2.140 a. If we assume that the data are about mound-shaped, then any observation with a z-score greater than 3 in absolute value would be considered an outlier. From Exercise 2.139, the z-score corresponding to 50 is −1, the z-score corresponding to 70 is 1, and the z-score corresponding to 80 is 2. Since none of these z-scores is greater than 3 in absolute value, none would be considered outliers.

b. From Exercise 2.139, the z-score corresponding to 50 is −2, the z-score corresponding to 70 is 2, and the z-score corresponding to 80 is 4. Since the z-score corresponding to 80 is greater than 3, 80 would be considered an outlier.

c. From Exercise 2.139, the z-score corresponding to 50 is 1, the z-score corresponding to 70 is 3, and the z-score corresponding to 80 is 4. Since the z-scores corresponding to 70 and 80 are greater than or equal to 3, 70 and 80 would be considered outliers.

d. From Exercise 2.139, the z-score corresponding to 50 is .1, the z-score corresponding to 70 is .3, and the z-score corresponding to 80 is .4. Since none of these z-scores is greater than 3 in absolute value, none would be considered outliers.

2.141 a. $\sum x = 13+1+10+3+3 = 30 \qquad \sum x^2 = 13^2+1^2+10^2+3^2+3^2 = 288$

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6 \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{288 - \frac{30^2}{5}}{5-1} = \frac{108}{4} = 27 \qquad s = \sqrt{27} = 5.20$$

b. $\sum x = 13+6+6+0 = 25 \qquad \sum x^2 = 13^2+6^2+6^2+0^2 = 241$

$$\bar{x} = \frac{\sum x}{n} = \frac{25}{4} = 6.25 \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{241 - \frac{25^2}{4}}{4-1} = \frac{84.75}{3} = 28.25 \qquad s = \sqrt{28.25} = 5.32$$

c. $\sum x = 1+0+1+10+11+11+15 = 49 \qquad \sum x^2 = 1^2+0^2+1^2+10^2+11^2+11^2+15^2 = 569.$

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

d. $\sum x = 3+3+3+3 = 12 \qquad \sum x^2 = 3^2+3^2+3^2+3^2 = 36$

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

2.142 a. $\sum x = 4+6+6+5+6+7 = 34 \qquad \sum x^2 = 4^2+6^2+6^2+5^2+6^2+7^2 = 198$

$$\bar{x} = \frac{\sum x}{n} = \frac{34}{6} = 5.67 \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{198 - \frac{34^2}{6}}{6-1} = \frac{5.3333}{5} = 1.0667 \qquad s = \sqrt{1.067} = 1.03$$

b. $\sum x = -1+4+(-3)+0+(-3)+(-6) = -9 \qquad \sum x^2 = (-1)^2+4^2+(-3)^2+0^2+(-3)^2+(-6)^2 = 71$

$$\bar{x} = \frac{\sum x}{n} = \frac{-9}{6} = -\$1.5 \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{71 - \frac{(-9)^2}{6}}{6-1} = \frac{57.5}{5} = 11.5 \text{ dollars squared}$$

$s = \sqrt{11.5} = \$3.39$

c. $\sum x = \frac{3}{5}+\frac{4}{5}+\frac{2}{5}+\frac{1}{5}+\frac{1}{16} = 2.0625 \qquad \sum x^2 = \left(\frac{3}{5}\right)^2+\left(\frac{4}{5}\right)^2+\left(\frac{2}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{16}\right)^2 = 1.2039$

$$\bar{x} = \frac{\sum x}{n} = \frac{2.0625}{5} = .4125\%$$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{1.2039 - \frac{2.0625^2}{5}}{5-1} = \frac{.3531}{4} = .0883\% \text{ squared}$$

$s = \sqrt{.0883} = .30\%$

d. (a) Range = 7 – 4 = 3

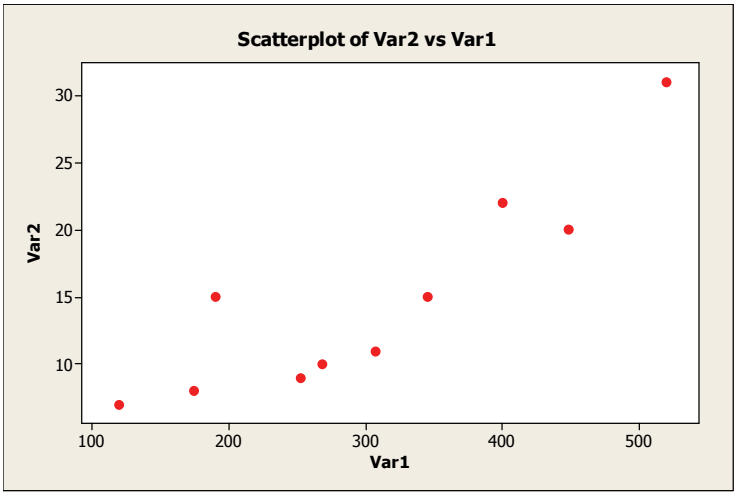
(b) Range = \$4 – (\$-6) = \$10

(c) Range = $\frac{4}{5}\% - \frac{1}{16}\% = \frac{64}{80}\% - \frac{5}{80}\% = \frac{59}{80}\% = .7375\%$

2.143 The range is found by taking the largest measurement in the data set and subtracting the smallest measurement. Therefore, it only uses two measurements from the whole data set. The standard deviation uses every measurement in the data set. Therefore, it takes every measurement into account—not just two. The range is affected by extreme values more than the standard deviation.

2.144 $\sigma \approx \frac{\text{range}}{4} = \frac{20}{4} = 5$

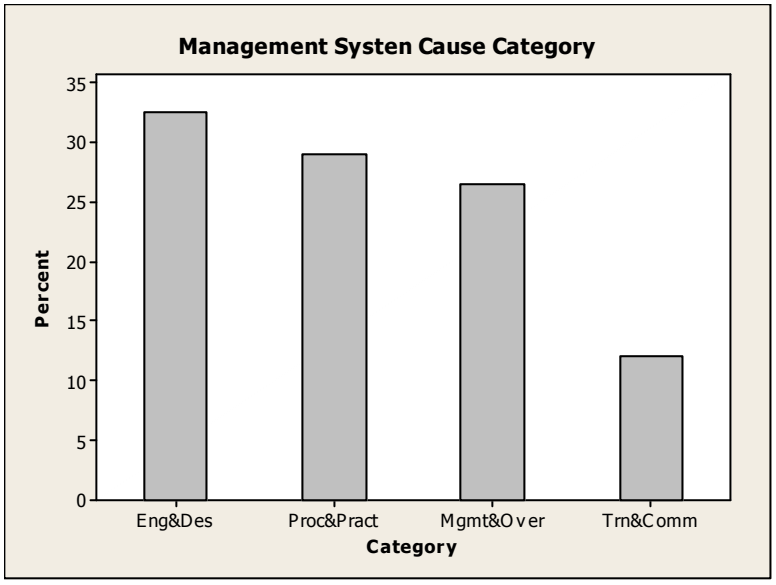
2.145 Using MINITAB, the scatterplot is:



2.146 a. To find relative frequencies, we divide the frequencies of each category by the total number of incidents. The relative frequencies of the number of incidents for each of the cause categories are:

Management System Cause Category	*Number of Incidents*	*Relative Frequencies*
Engineering & Design	27	27 / 83 = .325
Procedures & Practices	24	24 / 83 = .289
Management & Oversight	22	22 / 83 = .265
Training & Communication	10	10 / 83 = .120
TOTAL	83	1

b. The Pareto diagram is:



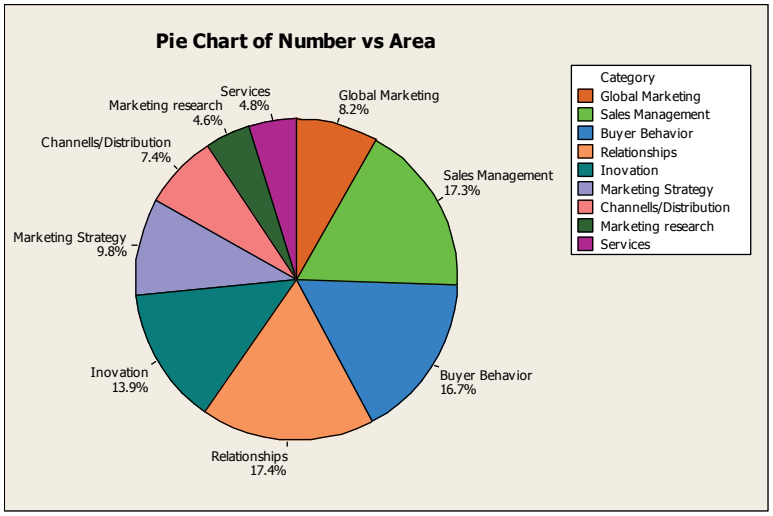
c. The category with the highest relative frequency of incidents is Engineering and Design. The category with the lowest relative frequency of incidents is Training and Communication.

2.147 a. The relative frequency for each response category is found by dividing the frequency by the total sample size. The relative frequency for the category "Global Marketing" is 235/2863 = .082. The rest of the relative frequencies are found in a similar manner and are reported in the table.

Area	***Number***	***Relative Frequencies***
Global Marketing	235	235/2863 = .082
Sales Management	494	494/2863 = .173
Buyer Behavior	478	478/2863 = .167
Relationships	498	498/2863 = .174
Innovation	398	398/2863 = .139
Marketing Strategy	280	280/2863 = .098
Channels/Distribution	213	213/2863 = .074
Marketing Research	131	131/2863 = .046
Services	136	136/2863 = .048
TOTAL	2,863	1.00

Relationships and sales management had the most articles published with 17.4% and 17.3%, respectively. Not far behind was Buyer Behavior with 16.7%. Of the rest of the areas, only innovation had more than 10%.

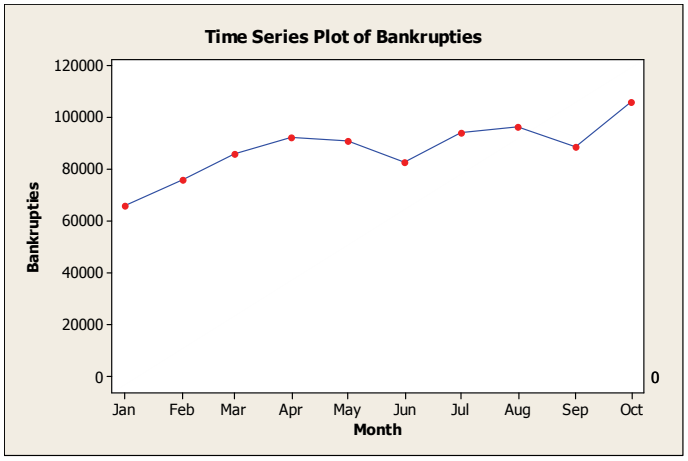
b. Using MINITAB, the pie chart of the data is:



The slice for Marketing Research is smaller than the slice for Sales Management because there were fewer articles on Marketing Research than for Sales Management.

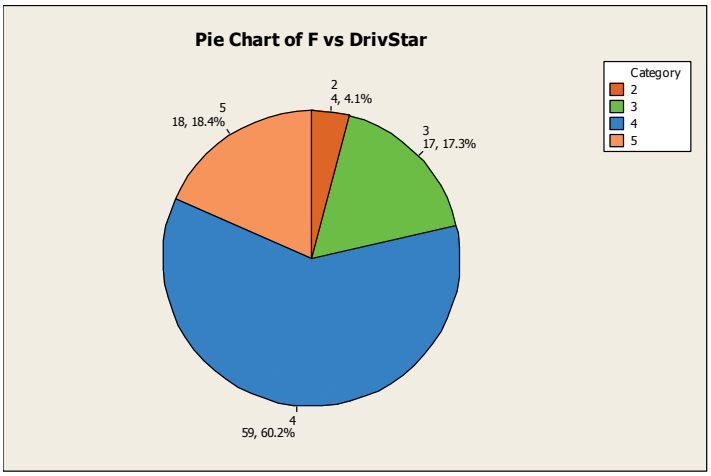
2.148 a. The data are time series data because the numbers of bankruptcies were collected over a period of 10 months.

b. Using MINITAB, the time series plot is:



c. There is a generally increasing trend in the number of bankruptcies as the months increase.

2.149 Using MINITAB, the pie chart is:



60% of cars have 4-star rating and only 4% have 2-star ratings.

2.150 a. The average driver's severity of head injury in head-on collisions is 603.7.

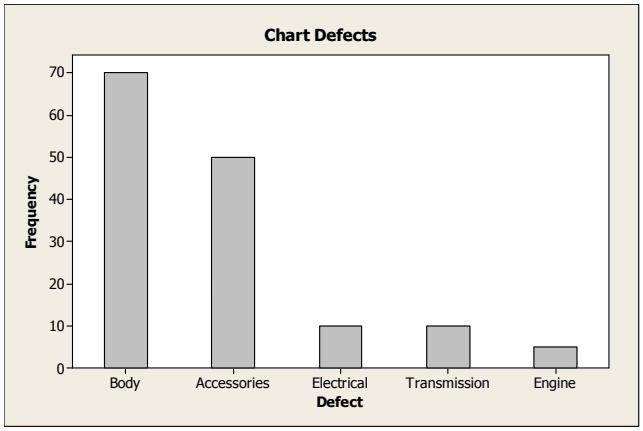
b. Since the mean and median are close in value, the data should be fairly symmetric. Thus, we can use the Empirical Rule. We know that about 95% of all observations will fall within 2 standard deviations of the mean. This interval is $\bar{x} \pm 2s \Rightarrow 603.7 \pm 2(185.4) \Rightarrow 603.7 \pm 370.8 \Rightarrow (232.9, 974.5)$

Most of the head-injury ratings will fall between 232.9 and 974.5.

c. The z-score would be: $z = \dfrac{x - \bar{x}}{s} = \dfrac{408 - 603.7}{185.4} = -1.06$

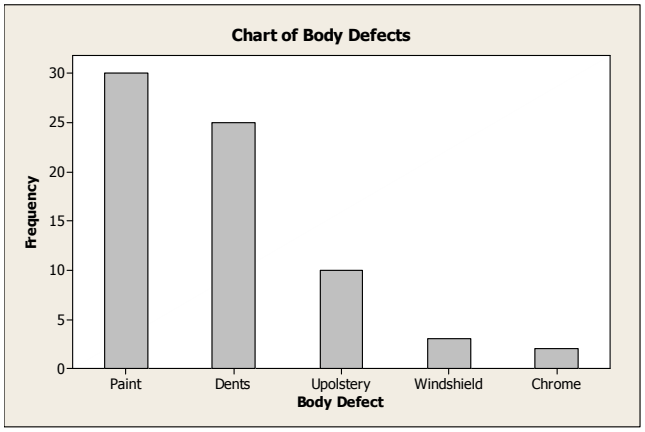
Since the absolute value is not very big, this is not an unusual value to observe.

2.151 a. Using MINITAB, a Pareto diagram for the data is:



The most frequently observed defect is a body defect.

b. Using MINITAB, a Pareto diagram for the Body Defect data is:

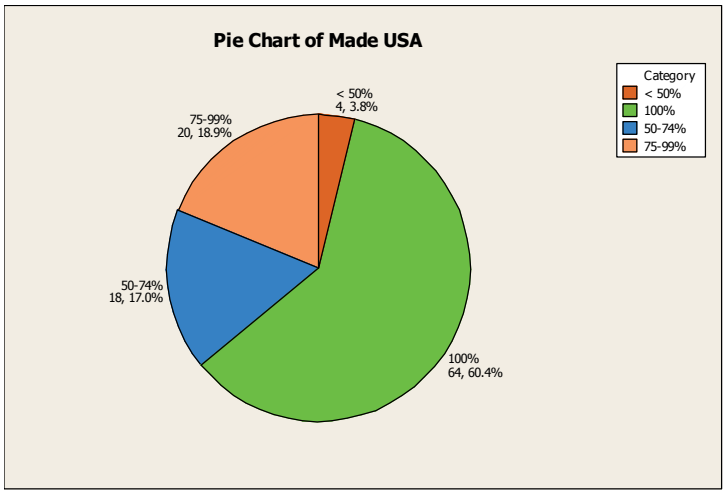


Most body defects are either paint or dents. These two categories account for $(30+25)/70 = 55/70 = .786$ of all body defects. Since these two categories account for so much of the body defects, it would seem appropriate to target these two types of body defects for special attention.

2.152 a. The data collection method was a survey.

b. Since the data were 4 different categories, the variable is qualitative.

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



About 60% of those surveyed believe that "Made in USA" means 100% US labor and materials.

2.153 a. From the information given, we have $\bar{x} = 375$ and $s = 25$. From Chebyshev's Rule, we know that at least three-fourths of the measurements are within the interval: $\bar{x} \pm 2s$, or (325, 425)

Thus, at most one-fourth of the measurements exceed 425. In other words, more than 425 vehicles used the intersection on at most 25% of the days.

b. According to the Empirical Rule, approximately 95% of the measurements are within the interval:

$\bar{x} \pm 2s$, or (325, 425)

This leaves approximately 5% of the measurements to lie outside the interval. Because of the symmetry of a mound-shaped distribution, approximately 2.5% of these will lie below 325, and the remaining 2.5% will lie above 425. Thus, on approximately 2.5% of the days, more than 425 vehicles used the intersection.

2.154 The percentile ranking of the age of 25 years would be 100% – 75% = 25%. Thus, an age of 25 would correspond to the 25th percentile.

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

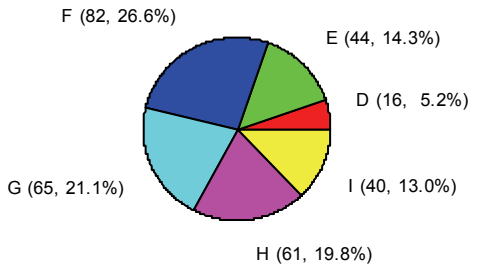
```
Stem-and-Leaf of PENALTY           N = 38
Leaf Unit = 10

(28)     0 0011111222222233333444489 9
 10      1 00239
  5      2
  5      3 0
  4      4 0
  3      5
  3      6
  3      7
  3      8 5
  2      9 3
  1     10 0
```

b. See the highlighted leaves in part **a**.

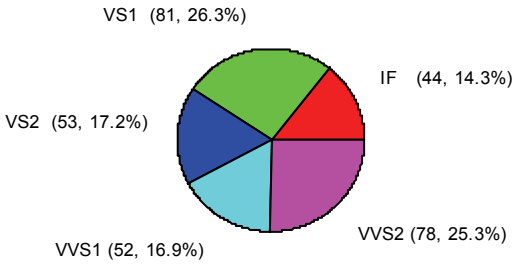
c. Most of the penalties imposed for Clean Air Act violations are relatively small compared to the penalties imposed for other violations. All but two of the penalties for Clean Air Act violations are below the median penalty imposed.

2.156 Using MINITAB, the pie charts are:

Color

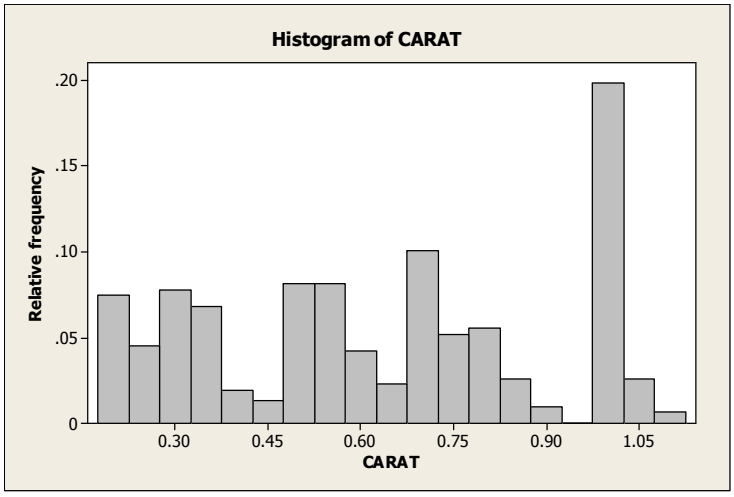


Clarity

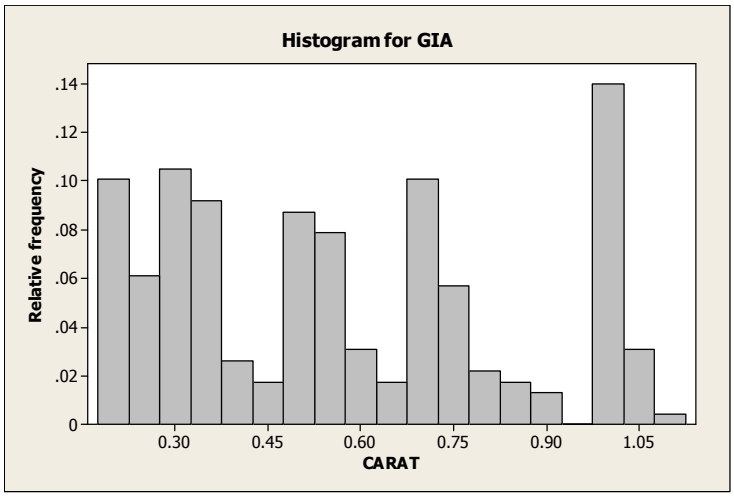


The F color occurs the most often with 26.6%. The clarity that occurs the most is VS1 with 26.3%. The D color occurs the least often with 5.2%. The clarity that occurs the least is IF with 14.3%.

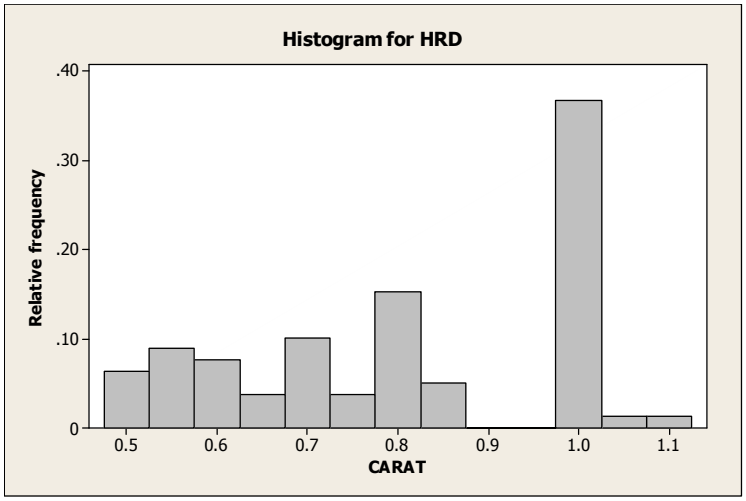
2.157 a. Using MINITAB, the relative frequency histogram is:

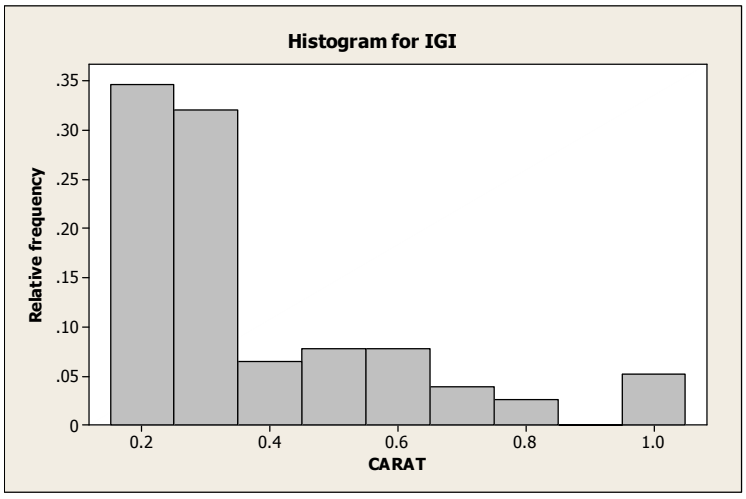


b. Using MINITAB, the relative frequency histogram for the GIA group is:



c. Using MINITAB, the relative frequency histograms for the HRD and IGI groups are:





d. The HRD group does not assess any diamonds less than .5 carats and almost 40% of the diamonds they assess are 1.0 carat or higher. The IGI group does not assess very many diamonds over .5 carats and more than half are .3 carats or less. More than half of the diamonds assessed by the GIA group are more than .5 carats, but the sizes are less than those of the HRD group.

e. The sample mean is: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{194.32}{308} = .631$

The average number of carats for the 308 diamonds is .631.

f. The median is the average of the middle two observations once they have been ordered. The 154th and 155th observations are .62 and .62. The average of these two observations is .62.

Half of the diamonds weigh less than .62 carats and half weigh more.

g The mode is 1.0. This observation occurred 32 times.

h. Since the mean and median are close in value, either could be a good descriptor of central tendency.

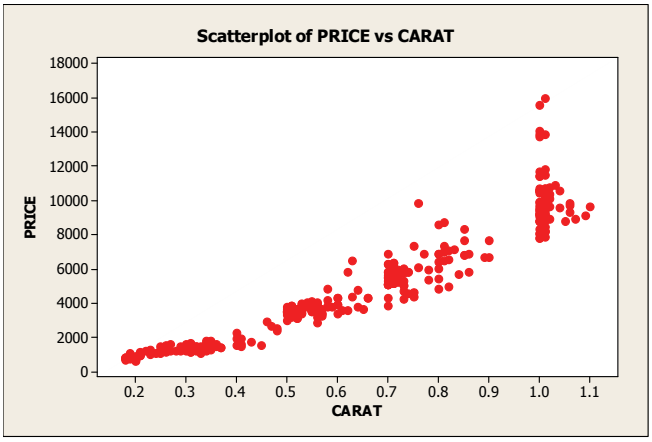
i. From Chebyshev's Theorem, we know that at least ¾ or 75% of all observations will fall within 2 standard deviations of the mean. From part e, $\bar{x} = .63$.

The variance is: $s^2 = \dfrac{\sum_i x_i^2 - \dfrac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \dfrac{146.19 - \dfrac{194.32^2}{308}}{308-1} = .0768$ square carats

The standard deviation is: $s = \sqrt{s^2} = \sqrt{.0768} = .277$ carats

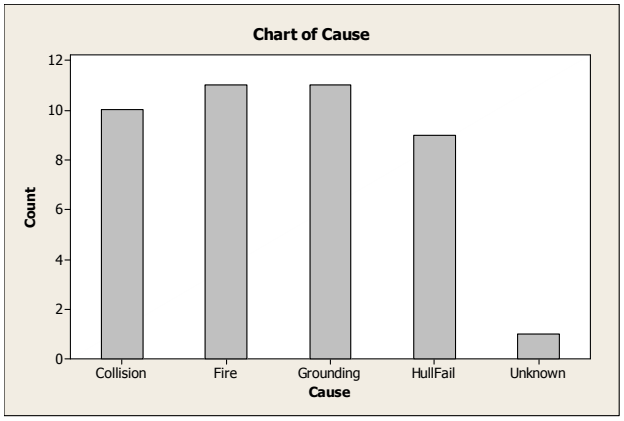
This interval is: $\bar{x} \pm 2s \Rightarrow .631 \pm 2(.277) \Rightarrow .631 \pm .554 \Rightarrow (.077, 1.185)$

2.158 Using MINITAB, the scatterplot is:



As the number of carats increases the price of the diamond tends to increase. There appears to be an upward trend.

2.159 a. Using MINITAB, a bar graph of the data is:



Fire and grounding are the two most likely causes of puncture.

b. Using MINITAB, the descriptive statistics are:

Descriptive Statistics: Spillage

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Spillage	42	66.19	56.05	25.00	32.00	43.00	77.50	257.00

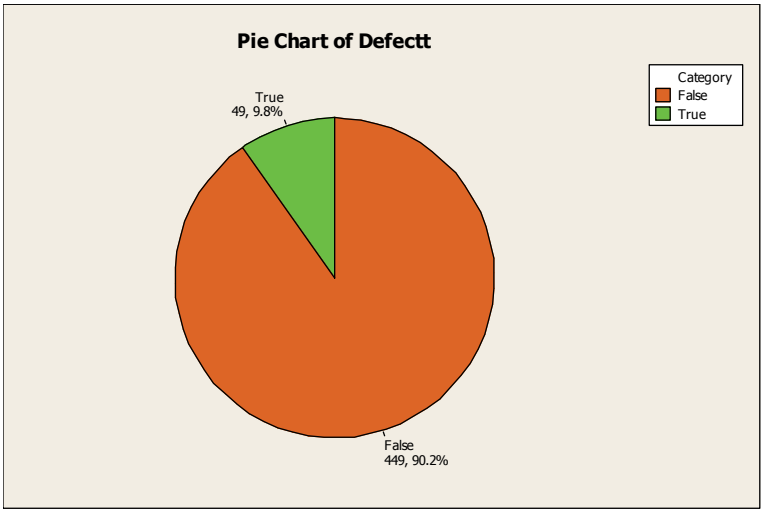
The mean spillage amount is 66.19 thousand metric tons, while the median is 43.00. Since the median is so much smaller than the mean, it indicates that the data are skewed to the right. The standard deviation is 56.05. Again, since this value is so close to the value of the mean, it indicates that the data are skewed to the right.

Since the data are skewed to the right, we cannot use the Empirical Rule to describe the data. Chebyshev's Rule can be used. Using Chebyshev's Rule, we know that at least 8/9 of the observations will fall within 3 standard deviations of the mean.

$\bar{x} \pm 3s \Rightarrow 66.19 \pm 3(56.05) \Rightarrow 66.19 \pm 168.15 \Rightarrow (-101.96,\ 234.34)$ or (0, 234.34) since we cannot have negative spillage.

Thus, at least 8/9 of all oil spills will be between 0 and 234.34 thousand metric tons.

2.160 Using MINITAB, a pie chart of the data is:



A response of 'true' means the software contained defective code. Thus, only 9.8% of the modules contained defective software code.

2.161 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:

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

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

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

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.

2.162 a. First, we must compute the total processing times by adding the processing times of the three departments. The total processing times are as follows:

Request	Total Processing Time	Request	Total Processing Time	Request	Total Processing Time
1	13.3	17	19.4*	33	23.4*
2	5.7	18	4.7	34	14.2
3	7.6	19	9.4	35	14.3
4	20.0*	20	30.2	36	24.0*
5	6.1	21	14.9	37	6.1
6	1.8	22	10.7	38	7.4
7	13.5	23	36.2*	39	17.7*
8	13.0	24	6.5	40	15.4
9	15.6	25	10.4	41	16.4
10	10.9	26	3.3	42	9.5
11	8.7	27	8.0	43	8.1
12	14.9	28	6.9	44	18.2*
13	3.4	29	17.2*	45	15.3
14	13.6	30	10.2	46	13.9
15	14.6	31	16.0	47	19.9*
16	14.4	32	11.5	48	15.4
				49	14.3*
				50	19.0

The stem-and-leaf displays with the appropriate leaves highlighted are as follows:

```
Stem-and-leaf of Mkt
  Leaf Unit = 0.10

  6  0  0112446
  7  1  3
 14  2  0024699
 16  3  25
 22  4  001577
(10) 5  0344556889
 18  6  00022247 99
  8  7  0038
  4  8  07
  2  9
  2 10  0
  1 11  0
```

```
Stem-and-leaf of Engr
  Leaf Unit = 0.10

  7  0  4466699
 14  1  3333788
 19  2  12246
 23  3  1568
(5)  4  24688
 22  5  233
 19  6  01239
 14  7  22379
  9  8
  9  9  66
  7 10  0
  6 11  3
  5 12  023
  2 13  0
  1 14  4
```

```
  Stem-and-leaf of Accnt       Stem-and-leaf of Total
     Leaf Unit = 0.10             Leaf Unit = 1.00

 19  0  111111111112 2333444      1   0  1
(8)  0  55556888                  3   0  33
 23  1  00                       5   0  45
 21  1  79                      11   0  666677
 19  2  0023                    17   0  888999
 15  2                          21   1  0000
 15  3  23                     (5)   1  33333
 13  3  78                      24   1  4444445555
 11  4                          14   1  6677
 11  4                          10   1  8999
 11  5                           6   2  0
 11  5  8                        5   2  3
 10  6  2                        4   2  44
  9  6                              HI 30, 36
  9  7  0
  8  7
  8  8  4
    HI  99, 105, 135, 144,
        182, 220, 300
```

Of the 50 requests, 10 were lost. For each of the three departments, the processing times for the lost requests are scattered throughout the distributions. The processing times for the departments do not appear to be related to whether the request was lost or not. However, the total processing times for the lost requests appear to be clustered towards the high side of the distribution. It appears that if the total processing time could be kept under 17 days, 76% of the data could be maintained, while reducing the number of lost requests to 1.

b. For the Marketing department, if the maximum processing time was set at 6.5 days, 78% of the requests would be processed, while reducing the number of lost requests by 4. For the Engineering department, if the maximum processing time was set at 7.0 days, 72% of the requests would be processed, while reducing the number of lost requests by 5. For the Accounting department, if the maximum processing time was set at 8.5 days, 86% of the requests would be processed, while reducing the number of lost requests by 5.

c. Using MINITAB, the summary statistics are:

Descriptive Statistics: REQUEST, MARKET, ENGINEER, ACCOUNT

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
MARKET	50	4.766	2.584	0.100	2.825	5.400	6.250	11.000
ENGINEER	50	5.044	3.835	0.400	1.775	4.500	7.225	14.400
ACCOUNT	50	3.652	6.256	0.100	0.200	0.800	3.725	30.000
TOTAL	50	13.462	6.820	1.800	8.075	13.750	16.600	36.200

d. The z-scores corresponding to the maximum time guidelines developed for each department and the total are as follows:

Marketing: $z = \dfrac{x - \bar{x}}{s} = \dfrac{6.5 - 4.77}{2.58} = .67$

Engineering: $z = \dfrac{x - \bar{x}}{s} = \dfrac{7.0 - 5.04}{3.84} = .51$

Accounting: $z = \dfrac{x - \bar{x}}{s} = \dfrac{8.5 - 3.65}{6.26} = .77$

Total: $z = \dfrac{x - \bar{x}}{s} = \dfrac{17 - 13.46}{6.82} = .52$

e. To find the maximum processing time corresponding to a z-score of 3, we substitute in the values of $z, \bar{x}$, and s into the z formula and solve for x.

$$z = \frac{x - \bar{x}}{s} \Rightarrow x - \bar{x} = zs \Rightarrow x = \bar{x} + zs$$

Marketing: $x = 4.77 + 3(2.58) = 4.77 + 7.74 = 12.51$
None of the orders exceed this time.

Engineering: $x = 5.04 + 3(3.84) = 5.04 + 11.52 = 16.56$
None of the orders exceed this time.

These both agree with both the Empirical Rule and Chebyshev's Rule.

Accounting: $x = 3.65 + 3(6.26) = 3.65 + 18.78 = 22.43$
One of the orders exceeds this time or 1/50 = .02.

Total: $x = 13.46 + 3(6.82) = 13.46 + 20.46 = 33.92$
One of the orders exceeds this time or 1/50 = .02.

These both agree with Chebyshev's Rule but not the Empirical Rule. Both of these last two distributions are skewed to the right.

f. Marketing: $x = 4.77 + 2(2.58) = 4.77 + 5.16 = 9.93$
Two of the orders exceed this time or 2/50 = .04.

Engineering: $x = 5.04 + 2(3.84) = 5.04 + 7.68 = 12.72$
Two of the orders exceed this time or 2/50 = .04.

Accounting: $x = 3.65 + 2(6.26) = 3.65 + 12.52 = 16.17$
Three of the orders exceed this time or 3/50 = .06.

Total: $x = 13.46 + 2(6.82) = 13.46 + 13.64 = 27.10$

Two of the orders exceed this time or 2/50 = .04.

All of these agree with Chebyshev's Rule but not the Empirical Rule.

g. No observations exceed the guideline of 3 standard deviations for both Marketing and Engineering. One observation exceeds the guideline of 3 standard deviations for both Accounting (#23, time = 30.0 days) and Total (#23, time = 36.2 days). Therefore, only (1/10) × 100% of the "lost" quotes have times exceeding at least one of the 3 standard deviation guidelines.

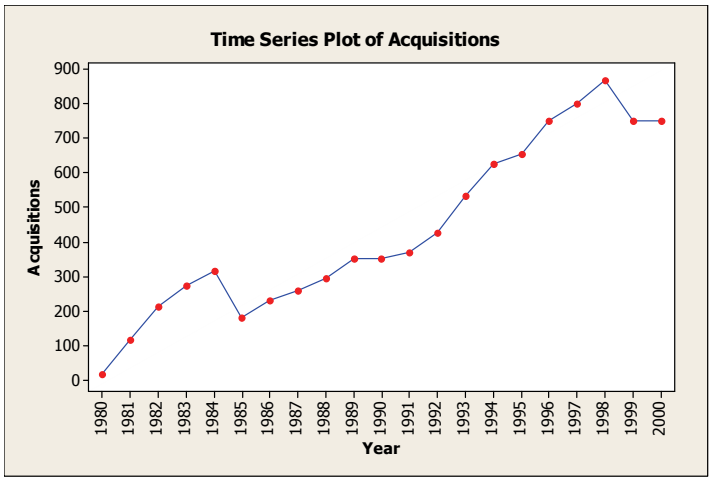
Two observations exceed the guideline of 2 standard deviations for both Marketing (#31, time = 11.0 days and #48, time = 10.0 days) and Engineering (#4, time = 13.0 days and #49, time = 14.4 days). Three observations exceed the guideline of 2 standard deviations for Accounting (#20, time = 22.0 days; #23, time = 30.0 days; and #36, time = 18.2 days). Two observations exceed the guideline of 2 standard deviations for Total (#20, time = 30.2 days and #23, time = 36.2 days). Therefore, (7/10) × 100% = 70% of the "lost" quotes have times exceeding at least one the 2 standard deviation guidelines.

We would recommend the 2 standard deviation guideline since it covers 70% of the lost quotes, while having very few other quotes exceed the guidelines.

2.163 a. One reason the plot may be interpreted differently is that no scale is given on the vertical axis. Also, since the plot almost reaches the horizontal axis at 3 years, it is obvious that the bottom of the plot has been cut off. Another important factor omitted is who responded to the survey.

b. A scale should be added to the vertical axis. Also, that scale should start at 0.

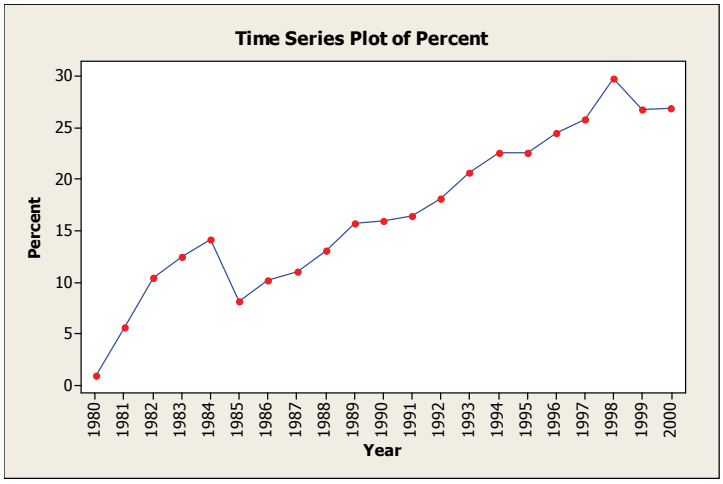
2.164 a. Using MINITAB, the time series plot of the data is:



b. To find the percentage of the sampled firms with at least one acquisition, we divide number with acquisitions by the total sampled and then multiply by 100%. For 1980, the percentage of firms with at least on acquisition is (18/1963)*100% = .92%. The rest of the percentages are found in the same manner and are listed in the following table:

Year	Number of firms	Number with Acquisitions	Percentage with Acquisitions
1980	1,963	18	.92%
1981	2,044	115	5.63%
1982	2,029	211	10.40%
1983	2,187	273	12.48%
1984	2,248	317	14.10%
1985	2,238	182	8.13%
1986	2,277	232	10.19%
1987	2,344	258	11.01%
1988	2,279	296	12.99%
1989	2,231	350	15.69%
1990	2,197	350	15.93%
1991	2,261	370	16.36%
1992	2,363	427	18.07%
1993	2,582	532	20.60%
1994	2,775	626	22.56%
1995	2,890	652	22.56%
1996	3,070	751	24.46%
1997	3,099	799	25.78%
1998	2,913	866	29.73%
1999	2,799	750	26.80%
2000	2,778	748	26.93%
TOTAL	51,567	9,123	

Using MINITAB, the time series plot is:



c. In this case, both plots are almost the same. In general, the time series plot of the percents would be more informative. By changing the observations to percents, one can compare time periods with different sample sizes on the same basis.

2.165 a. Since the mean is greater than the median, the distribution of the radiation levels is skewed to the right.

b. $\bar{x} \pm s \Rightarrow 10 \pm 3 \Rightarrow (7, 13)$; $\bar{x} \pm 2s \Rightarrow 10 \pm 2(3) \Rightarrow (4, 16)$; $\bar{x} \pm 3s \Rightarrow 10 \pm 3(3) \Rightarrow (1, 19)$

Interval	Chebyshev's	Empirical
(7, 13)	At least 0	≈68%
(4, 16)	At least 75%	≈95%
(1, 19)	At least 88.9%	≈100%

Since the data are skewed to the right, Chebyshev's Rule is probably more appropriate in this case.

c. The background level is 4. Using Chebyshev's Rule, at least 75% or .75(50) ≈ 38 homes are above the background level. Using the Empirical Rule, ≈ 97.5% or .975(50) ≈ 49 homes are above the background level.

d. $z = \dfrac{x - \bar{x}}{s} = \dfrac{20 - 10}{3} = 3.333$

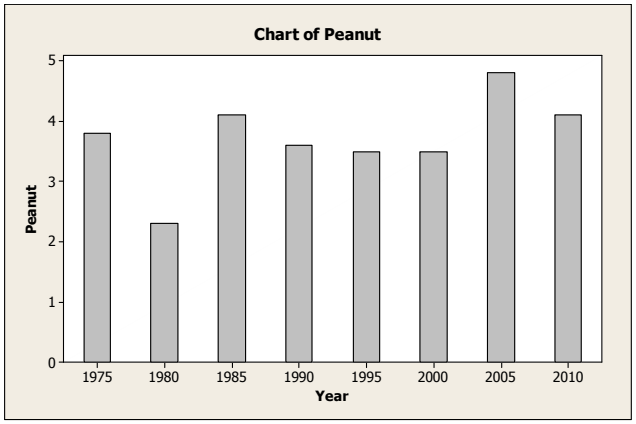
It is unlikely that this new measurement came from the same distribution as the other 50. Using either Chebyshev's Rule or the Empirical Rule, it is very unlikely to see any observations more than 3 standard deviations from the mean.

2.166 a. Since it is given that the distribution is mound-shaped, we can use the Empirical Rule. We know that 1.84% is 2 standard deviations below the mean. The Empirical Rule states that approximately 95% of the observations will fall within 2 standard deviations of the mean and, consequently, approximately 5% will lie outside that interval. Since a mound-shaped distribution is symmetric, then approximately 2.5% of the day's production of batches will fall below 1.84%.

b. If the data are actually mound-shaped, it would be extremely unusual (less than 2.5%) to observe a batch with 1.80% zinc phosphide if the true mean is 2.0%. Thus, if we did observe 1.8%, we would conclude that the mean percent of zinc phosphide in today's production is probably less than 2.0%.

2.167 a. Both the height and width of the bars (peanuts) change. Thus, some readers may tend to equate the area of the peanuts with the frequency for each year.

b. Using MINITAB, the frequency bar chart is:



2.168 a. Clinic A claims to have a *mean* weight loss of 15 during the first month and Clinic B claims to have a *median* weight loss of 10 pounds in the first month. With no other information, I would choose Clinic B. It is very likely that the distributions of weight losses will be skewed to the right – most people lose in the neighborhood of 10 pounds, but a couple might lose much more. If a few people lost much more than 10 pounds, then the mean will be pulled in that direction.

b. For Clinic A, the median is 10 and the standard deviation is 20. For Clinic B, the mean is 10 and the standard deviation is 5.

For Clinic A:

The mean is 15 and the median is 10. This would indicate that the data are skewed to the right. Thus, we will have to use Chebyshev's Rule to describe the distribution of weight losses.

$\bar{x} \pm 2s \Rightarrow 15 \pm 2(20) \Rightarrow 15 \pm 40 \Rightarrow (-25,\ 55)$

Using Chebyshev's Rule, we know that at least 75% of all weight losses will be between -25 and 55 pounds. This means that at least 75% of the people will have weight losses of between a loss of 55 pounds to a gain of 25 pounds. This is a very large range.

For Clinic B:

The mean is 10 and the median is 10. This would indicate that the data are symmetrical. Thus, the Empirical Rule can be used to describe the distribution of weight losses.

$\bar{x} \pm 2s \Rightarrow 10 \pm 2(5) \Rightarrow 10 \pm 10 \Rightarrow (0,\ 20)$

Using the Empirical Rule, we know that approximately 95% of all weight losses will be between 0 and 20 pounds. This is a much smaller range than in Clinic A.

I would still recommend Clinic B. Using Clinic A, a person has the potential to lose a large amount of weight, but also has the potential to gain a relatively large amount of weight. In Clinic B, a person would be very confident that he/she would lose weight.

c. One would want the clients selected for the samples in each clinic to be representative of all clients in that clinic. One would hope that the clinic would not choose those clients for the sample who lost the most weight just to promote their clinic.

2.169 First we make some preliminary calculations.

Of the 20 engineers at the time of the layoffs, 14 are 40 or older. Thus, the probability that a randomly selected engineer will be 40 or older is 14/20 = .70. A very high proportion of the engineers is 40 or over.

In order to determine if the company is vulnerable to a disparate impact claim, we will first find the median age of all the engineers. Ordering all the ages, we get:

29, 32, 34, 35, 38, 39, 40, 40, 40, 40, 40, 41, 42, 42, 44, 46, 47, 52, 55, 64

The median of all 20 engineers is $\frac{40+40}{2} = \frac{80}{2} = 40$

Now, we will compute the median age of those engineers who were not laid off. The ages underlined above correspond to the engineers who were not laid off. The median of these is $\frac{40+40}{2} = \frac{80}{2} = 40$.

The median age of all engineers is the same as the median age of those who were not laid off. The median age of those laid off is $\frac{40+41}{2} = \frac{81}{2} = 40.5$, which is not that much different from the median age of those not laid off. In addition, 70% of all the engineers are 40 or older. Thus, it appears that the company would not be vulnerable to a disparate impact claim.

2.170 Answers will vary. The graph is made to look like the amount of money spent on education has risen dramatically from 1980 to 2000, but the 4^{th} grade reading scores have not increased at all. The graph does not take into account that the number of school children has also increased dramatically in the last 20 years. A better portrayal would be to look at the per capita spending rather than total spending.

2.171 There is evidence to support this claim. The graph peaks at the interval above 1.002. The heights of the bars decrease in order as the intervals get further and further from the peak interval. This is true for all bars except the one above 1.000. This bar is greater than the bar to its right. This would indicate that there are more observations in this interval than one would expect, suggesting that some inspectors might be passing rods with diameters that were barely below the lower specification limit.