1.2 Descriptive statistics utilizes numerical and graphical methods to look for patterns, to summarize, and to present the information in a set of data. Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

1.7 A population is a set of existing units such as people, objects, transactions, or events. A variable is a characteristic or property of an individual population unit such as height of a person, time of a reflex, amount of a transaction, etc.

1.8 A population is a set of existing units such as people, objects, transactions, or events. A sample is a subset of the units of a population.

1.10 An inference without a measure of reliability is nothing more than a guess. A measure of reliability separates statistical inference from fortune telling or guessing. Reliability gives a measure of how confident one is that the inference is correct.

1.15 a. The population of interest is all citizens of the United States.

b. The variable of interest is the view of each citizen as to whether the president is doing a good or bad job. It is qualitative.

c. The sample is the 2000 individuals selected for the poll.

d. The inference of interest is to estimate the proportion of all citizens who believe the president is doing a good job.

e. The method of data collection is a survey.

f. It is not very likely that the sample will be representative of the population of all citizens of the United States. By selecting phone numbers at random, the sample will be limited to only those people who have telephones. Also, many people share the same phone number, so each person would not have an equal chance of being contacted. Another possible problem is the time of day the calls are made. If the calls are made in the evening, those people who work in the evening would not be represented.

1.27 a. The population from which the sample was selected is the set of all department store executives.

b. There are two variables measured by the authors. They are job-satisfaction and Machiavellian rating for each of the executives.

c. The sample is the set of 218 department store executives who completed the questionnaire.

d. The method of data collection is a survey.

e. The inference made by the authors is that those executives with higher job-satisfaction scores are likely to have a lower 'mach' rating.

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:

<table>
<thead>
<tr>
<th>Grade on Statistics Exam</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 90 – 100</td>
<td>16</td>
<td>.08</td>
</tr>
<tr>
<td>B: 80 – 89</td>
<td>36</td>
<td>.18</td>
</tr>
<tr>
<td>C: 65 – 79</td>
<td>90</td>
<td>.45</td>
</tr>
<tr>
<td>D: 50 – 64</td>
<td>30</td>
<td>.15</td>
</tr>
<tr>
<td>F: Below 50</td>
<td>28</td>
<td>.14</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>1.00</td>
</tr>
</tbody>
</table>

2.4 a. The variable summarized in the table is ‘Reason for requesting the installation of the passenger-side on-off switch.’ The values this variable could assume are: Infant, Child, Medical, Infant & Medical, Child & Medical, Infant & Child, and Infant & Child & Medical. Since the responses name something, the variable is qualitative.

b. The relative frequencies are found by dividing the number of requests for each category by the total number of requests. For the category ‘Infant’, the relative frequency is 1,852/30,337 = .061. The rest of the relative frequencies are found in the table below:

<table>
<thead>
<tr>
<th>Reason</th>
<th>Number of Requests</th>
<th>Relative frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infant</td>
<td>1,852</td>
<td>1,852/30,337 = .061</td>
</tr>
<tr>
<td>Child</td>
<td>17,148</td>
<td>17,148/30,337 = .565</td>
</tr>
<tr>
<td>Medical</td>
<td>8,377</td>
<td>8,377/30,337 = .276</td>
</tr>
<tr>
<td>Infant &amp; Medical</td>
<td>44</td>
<td>44/30,337 = .0014</td>
</tr>
<tr>
<td>Child &amp; Medical</td>
<td>903</td>
<td>903/30,337 = .030</td>
</tr>
<tr>
<td>Infant &amp; Child</td>
<td>1,878</td>
<td>1,878/30,337 = .062</td>
</tr>
<tr>
<td>Infant &amp; Child &amp; Medical</td>
<td>135</td>
<td>135/30,337 = .0045</td>
</tr>
<tr>
<td>TOTAL</td>
<td>30,337</td>
<td>.9999</td>
</tr>
</tbody>
</table>

c. Using MINITAB, a pie chart of the data is:
c. There are 4 categories where Medical is mentioned as a reason: Medical, Infant & Medical, Child & Medical, and Infant & Child & Medical. The sum of the frequencies for these 4 categories is $8,377 + 44 + 903 + 135 = 9,459$. The proportion listing Medical as one of the reasons is $9,459/30,337 = .312$.

2.5 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 \( \frac{30 + 25}{70} = \frac{55}{70} = 0.786 \) 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.

2.21 a. For male USGA golfers, there are about 28.5% with handicaps greater than 20.

b. For female USGA golfers, there are about 82% with handicaps greater than 20.

2.23 a. The stem-and-leaf display for the runs scored by St. Louis is:

```
Stem-and-leaf of St. Louis       N = 58
Leaf Unit = 0.10
  2  1 00
  7  2 00000
 15  3 00000000
 20  4 00000
 26  5 000000
 (11)  6 0000000000000000
 21  7 00000
 17  8 00000000
  9  9 00
  7 10 00
  5 11 00
  3 12
  3 13 0
  2 14 0
  1 15 0
```

b. The games where McGuire hit multiple home runs are circled above. The number of runs scored when McGuire hit multiple home runs tends to be at or above the median. Thus, the Cardinals tend to score more runs when McGuire hit multiple home runs.
a. Using MINITAB, the stem-and-leaf display is:

```
Stem-and-Leaf of PENALTY N = 38
Leaf Unit = 10000

(28)  0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 8
9 9
10 1 0 0 2 3 9
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 circled 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 one of the penalties for Clean Air Act violations are below the median penalty imposed.

2.37 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:

\[
\frac{2.5 + 2.8}{2} = \frac{5.3}{2} = 2.65
\]

2.39 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.

2.43 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.

2.49 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.
2.59  

a. \[ \sum x = 3 + 1 + 10 + 10 + 4 = 28 \]
\[ \sum x^2 = 3^2 + 1^2 + 10^2 + 10^2 + 4^2 = 226 \]
\[ \bar{x} = \frac{\sum x}{n} = \frac{28}{5} = 5.6 \]
\[ s^2 = \frac{\sum x^2 - \left( \frac{\sum x}{n} \right)^2}{n-1} = \frac{226 - \frac{28^2}{5}}{5-1} = \frac{69.2}{4} = 17.3 \]
\[ s = \sqrt{17.3} = 4.1593 \]

b. \[ \sum x = 8 + 10 + 32 + 5 = 55 \]
\[ \sum x^2 = 8^2 + 10^2 + 32^2 + 5^2 = 1213 \]
\[ \bar{x} = \frac{\sum x}{n} = \frac{55}{4} = 13.75 \text{ feet} \]
\[ s^2 = \frac{\sum x^2 - \left( \frac{\sum x}{n} \right)^2}{n-1} = \frac{1213 - \frac{55^2}{4}}{4-1} = \frac{456.75}{3} = 152.25 \text{ square feet} \]
\[ s = \sqrt{152.25} = 12.339 \text{ feet} \]

c. \[ \sum x = -1 + (-4) + (-3) + 1 + (-4) + (-4) = -15 \]
\[ \sum x^2 = (-1)^2 + (-4)^2 + (-3)^2 + 1^2 + (-4)^2 + (-4)^2 = 59 \]
\[ \bar{x} = \frac{\sum x}{n} = \frac{-15}{6} = -2.5 \]
\[ s^2 = \frac{\sum x^2 - \left( \frac{\sum x}{n} \right)^2}{n-1} = \frac{59 - \frac{-15^2}{6}}{6-1} = \frac{21.5}{5} = 4.3 \]
\[ s = \sqrt{4.3} = 2.0736 \]

d. \[ \sum x = \frac{1}{5} + \frac{1}{5} + \frac{2}{5} + \frac{1}{5} + \frac{4}{5} = \frac{10}{5} = 2 \]
\[ \sum x^2 = \left( \frac{1}{5} \right)^2 + \left( \frac{1}{5} \right)^2 + \left( \frac{2}{5} \right)^2 + \left( \frac{1}{5} \right)^2 + \left( \frac{4}{5} \right)^2 = \frac{24}{25} = .96 \]
\[ \bar{x} = \frac{\sum x}{n} = \frac{2}{6} = \frac{1}{3} = .33 \text{ ounce} \]
\[ s^2 = \frac{\sum x^2 - \left( \frac{\sum x}{n} \right)^2}{n-1} = \frac{24 - \frac{2^2}{6}}{6-1} = \frac{2933}{5} = .0587 \text{ square ounce} \]
\[ s = \sqrt{.0587} = .2422 \text{ ounce} \]
2.60 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, 5, 5, 5, 5, 5, 5

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

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

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

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

\[
s_2^2 = \frac{\sum x^2 - (\sum x)^2}{n-1} = \frac{110 - \frac{30^2}{10}}{9} = \frac{20}{9} = 4.4444
\]

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

2.70 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 \( \bar{x} + 3s \).

2.71 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 \( \bar{x} + 3s \).

2.73 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 range = 760 − 135 = 625.

The Range/6 = 625/6 = 104.17 and Range/4 = 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.

2.75  

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 \( \frac{3}{4} \) of the observations will fall within 2 standard deviations of the mean. Thus, at least \( \frac{3}{4} \) 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.