INSTRUCTOR'S SOLUTIONS MANUAL

NANCY S. BOUDREAU

Emeritus Associate Professor
Bowling Green State University

A FIRST COURSE IN STATISTICS

TWELFTH EDITION

James McClave

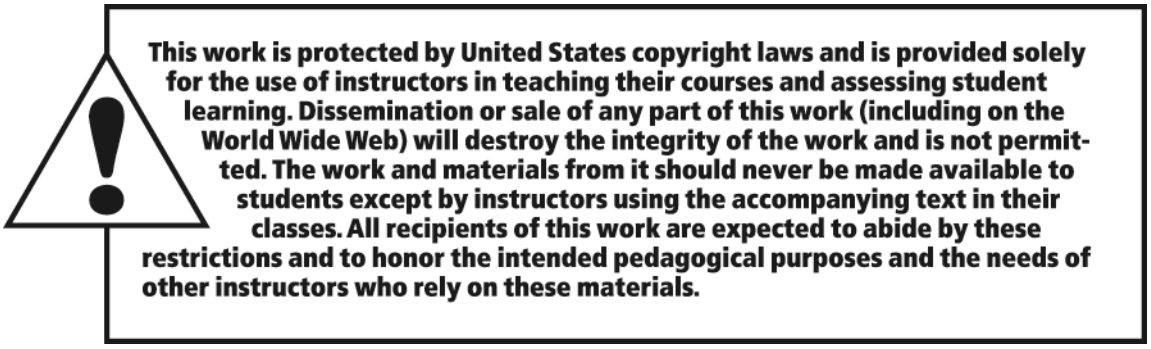
Info Tech, Inc.
University of Florida

Terry Sincich

University of South Florida

PEARSON

Boston Columbus Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo



Reproduced by Pearson from electronic files supplied by the author.

ISBN-13: 978-0-13-408110-6
ISBN-10: 0-13-408110-2

www.pearsonhighered.com

PEARSON

TABLE OF CONTENTS

Preface

This solutions manual is designed to accompany the text, *A First Course in Statistics*, Twelfth Edition, by James T. McClave and Terry Sincich. It provides solutions to most even-numbered exercises for each chapter in the text.

Other methods of solution may also be appropriate; however, the author has presented one that she believes to be the most instructive to the beginning statistics student.

The manual is provided to help instructors save time in preparing presentations of the solutions and possibly provide another point of view regarding their meaning.

Some of the exercises are subjective in nature. Subjective decisions regarding these exercises have been made and are explained by the author. Solutions based on these decisions are presented; the solution to this type of exercise is often most instructive. When an alternative interpretation of an exercise may occur, the author has often addressed it and given justification for the approach taken.

Nancy S. Boudreau, Associate Professor Emeritus
Bowling Green State University
Bowling Green, Ohio

Statistics, Data, and Statistical Thinking

Chapter **1**

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.

1.4 The first major method of collecting data is from a published source. These data have already been collected by someone else and is available in a published source. The second method of collecting data is from a designed experiment. These data are collected by a researcher who exerts strict control over the experimental units in a study. These data are measured directly from the experimental units. The third method of collecting data is from a survey. These data are collected by a researcher asking a group of people one or more questions. Again, these data are collected directly from the experimental units or people. The final method of collecting data is observationally. These data are collected directly from experimental units by simply observing the experimental units in their natural environment and recording the values of the desired characteristics.

1.6 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 representative sample is a sample that exhibits characteristics similar to those possessed by the target population. A representative sample is essential if inferential statistics is to be applied. If a sample does not possess the same characteristics as the target population, then any inferences made using the sample will be unreliable.

1.10 Statistical thinking involves applying rational thought processes to critically assess data and inferences made from the data. It involves not taking all data and inferences presented at face value, but rather making sure the inferences and data are valid.

1.12 a. High school GPA is a number usually between 0.0 and 4.0. Therefore, it is quantitative.

b. High school class rank is a number: 1st, 2nd, 3rd, etc. Therefore, it is quantitative.

c. The scores on the SAT's are numbers between 200 and 800. Therefore, it is quantitative.

d. Gender is either male or female. Therefore, it is qualitative.

e. Parent's income is a number: $25,000, $45,000, etc. Therefore, it is quantitative.

f. Age is a number: 17, 18, etc. Therefore, it is quantitative.

1.14 a. The experimental unit for this experiment is a drafted NFL quarterback.

b. Draft position is one of three categories. Therefore, it is a qualitative variable. NFL winning ratio is a number. Therefore, it is a quantitative variable. QB production score is a number. Therefore, it is a quantitative variable.

c. Because all quarterbacks drafted over a 38-year period were used, the application of this study is descriptive statistics.

1.16 a. The variable “difference between before and after sprint times” is measured in seconds. Thus, it is quantitative. The variable “improvement” is measured as one of three categories. Thus, it is qualitative.

b. The data set is a sample. It contains observations from only 14 of all high school football players.

1.18 a. The population of interest is all the students in the class. The variable of interest is the GPA of a student in the class.

b. Since GPA is measured on a numerical scale, it is quantitative.

c. Since the population of interest is all the students in the class and you obtained the GPA of every member of the class, this set of data would be a census.

d. Assuming the class had more than 10 students in it, the set of 10 GPAs would represent a sample. The set of ten students is only a subset of the entire class.

e. This average would have 100% reliability as an "estimate" of the class average, since it is the average of interest.

f. The average GPA of 10 members of the class will not necessarily be the same as the average GPA of the entire class. The reliability of the estimate will depend on how large the class is and how representative the sample is of the entire population.

g. In order for the sample to be a random sample, every member of the class must have an equal

1.20 a. Flight capability can have only 2 possible outcomes: volant or flightless. Thus, it is qualitative.

b. Habitat type can have only 3 possible outcomes: aquatic, ground terrestrial, or aerial terrestrial. Thus, it is qualitative.

c. Nesting site can have only 4 possible outcomes: ground, cavity within ground, tree, or cavity above ground. Thus, it is qualitative.

d. Nest density can have only 2 possible outcomes: high or low. Thus, it is qualitative.

e. Diet can have only 4 possible outcomes: fish, vertebrates, vegetables, or invertebrates. Thus, it is qualitative.

f. Body mass is measured in grams, a meaningful number. Thus, it is quantitative.

g. Egg length is measured in millimeters, a meaningful number. Thus, it is quantitative.

h. Extinct status can have only 3 possible outcomes: extinct, absent from island, or present. Thus, it is qualitative.

1.22 a. The population of interest to CSI is all computer security personnel at all U.S. corporations and government agencies.

b. The data collection method is a survey. A survey was sent to 5,412 firms with 351 firms responding.

c. The variable collected was whether or not the respondents admitted unauthorized use of computer systems at their firms during the year. Since the response to the questions was either "yes" or "no", this variable is qualitative.

d. In the sample 41% of the respondents admitted unauthorized use of computer systems at their firms during the year. If there is no nonresponse bias, then we can conclude that 41% of all firms would admit to unauthorized use of computer systems at their firms during the year.

1.24 The following variables would be qualitative because the response would be a category: country of operator/owner, primary use, and class of orbit. The following variables would be quantitative because the response would be a number: Longitudinal position, apogee, launch mass, usable electric power, and expected lifetime.

1.26 a. The population of interest is all senior managers at CPA firms.

b. The data collection method used is a survey.

c. Because only 992 of the 23,500 surveys sent out were returned and useable, there may be a problem with selection bias and/or nonresponse bias.

d. The validity of the inferences drawn from the study would be suspect. The inferences would only be valid if the 992 returned surveys were indeed, representative of the entire population. This is very unlikely.

1.28 a. The experimental unit for this study is a single-family residential property in Arlington, Texas.

b. The variables measured were the Zillow estimated value and the actual sale price. Both are quantitative variables.

c. If the population was described as all single-family residential properties in Arlington, Texas that sold within a given time period, then these 2,045 single-family residential properties could be the population if these were the only single-family residential property sales in Arlington, Texas in that time period.

d. The population could be all single-family residential properties sold in Arlington, Texas in a given time period and these 2,045 single-family residential properties did not include all the properties sold.

e. No. The single-family residential properties sold in Arlington, Texas probably are not similar to all single-family residential properties sold in the United States. Single-family residential properties sold in Arlington, Texas are not similar to single-family residential properties sold in places like New York City or San Francisco, California.

1.30 a. The population of interest is the set of all adults living in Tennessee. The sample of interest is the set of 575 people selected from Tennessee.

b. The data collection method used was a survey. A random-digit telephone dialing procedure was used to collect the sample. Since some people do not own phones, this would not be a random sample. Everyone in the state of Tennessee would not have an equal chance of being selected. Those without telephones would tend to be the undereducated. Thus, there could be potential biases in the data.

c. The two variables identified in this problem are the number of years of education and the insomnia status of each subject. The number of years of education is quantitative and the insomnia status is qualitative.

d. The researchers inferred that the fewer the years of education, the more likely the person was to have chronic insomnia.

1.32 a. The experimental units of this study were people who used a popular website for engaged couples.

b. The variables of interest are the engagement ring price and the level of appreciation of the recipient.

c. The population of interest were all those people on the popular website for engaged couples with "average" American names.

d. This sample of 33 respondents is probably not representative of the population. Those who decided to respond to the online survey self-selected themselves. They were not randomly selected. Generally speaking, those people who choose to respond to a survey have very strong feelings and are not representative of the entire population.

e. Answers will vary. Enter the numbers 1-50 in the first column of Minitab. Now, apply the random number generator of Minitab, requesting that 25 individuals be selected without replacement. The sample generated include individuals 1, 3, 6, 7, 9, 10, 11, 12, 13, 17, 18, 22, 24, 25, 27, 28, 30, 31, 32, 36, 37, 46, 47, 48, 50. These individuals would be placed in the gift-receiver role and the remaining individuals would be placed in the gift-giver role.

1.34 a. In Method 1, the researchers controlled which hot spots received the new program (through random assignment) and which did not. Therefore, a designed experiment was used to collect data in Method 1.

b. In Method 2, the researchers first divided the 56 hot spots into 4 groups based on the level of drug crimes. The researchers then controlled which hot spots received the new program (through random assignment) and which did not in each of the 4 groups. Therefore, a designed experiment was also used to collect data in Method 2.

c. This would be an application of inferential statistics because not all hot spots in Jersey City were used in the study. Only a sample of 56 hot spots was used.

d. Method 2 would be recommended. By creating 4 groups where the crime rate within each group is similar, we can control for a known source of variation. Within each group, we can then see how the new program compares to no program.

1.36 a. Although eating oat bran can reduce cholesterol, oat bran must be the only thing eaten. Reporting that eating oat bran is an easy and cheap way to lower your cholesterol implies that if you add oat bran to your diet, you can reduce your cholesterol, which may not be the case at all.

b. One would need to collect data on a sample of women who gave birth to babies with birth defects. Then, each woman in the study would also be asked whether she was a victim of domestic violence while pregnant or not. The data collection method for this study would be an observational study. Specifically, one would use a survey to collect the data.

c. In this study, only the results of the most positive response were reported. Not that many high school girls would 'Always' be happy with the way they were. To be more representative of those who are happy with the way they were, one should combine the results of those who responded 'Always true', 'Sort of true', and 'Sometimes true'.

Methods for Describing Sets of Data

Chapter **2**

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

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

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

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

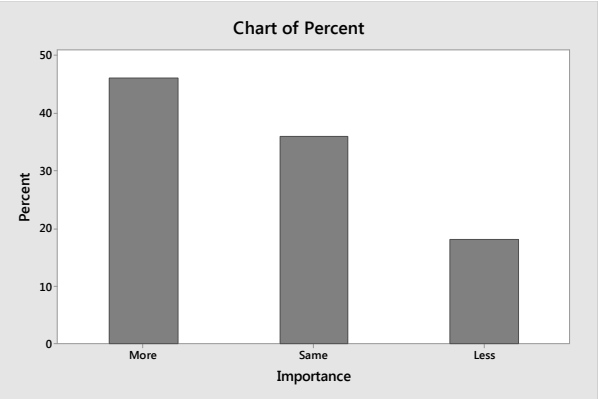
Grade on Statistics Exam	Frequency	Relative Frequency
A: 90–100	16	.08
B: 80– 89	36	.18
C: 65– 79	90	.45
D: 50– 64	30	.15
F: Below 50	28	.14
Total	200	1.00

2.6 a. The graph shown is a pie chart.

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

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

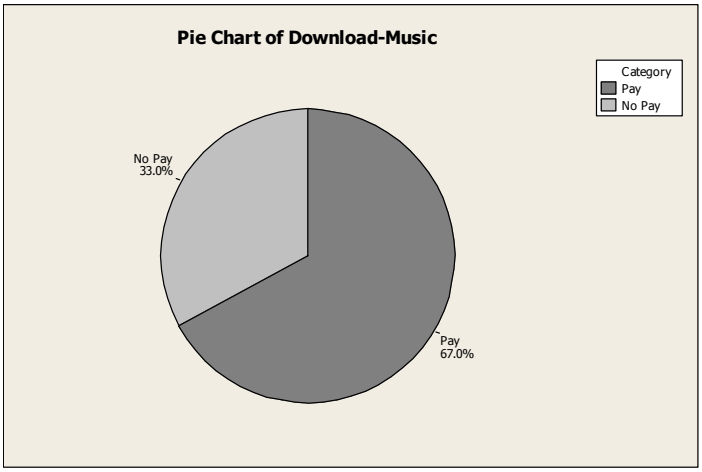
d. Using MINITAB, the Pareto diagram is:



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

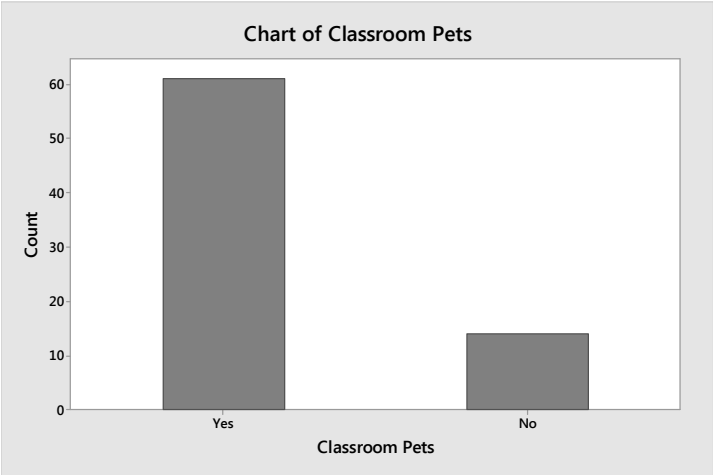
2.8 a. From the pie chart, 50.4% or 0.504 of adults living in the U.S. use the internet and pay to download music. From the data, 506 out of 1,003 adults or 506/1,003 = 0.504 of adults in the U.S. use the internet and pay to download music. These two results agree.

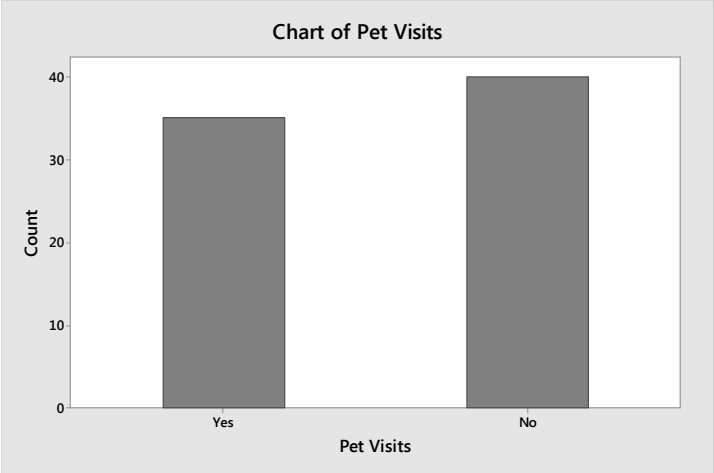
b. Using MINITAB, a pie chart of the data is:

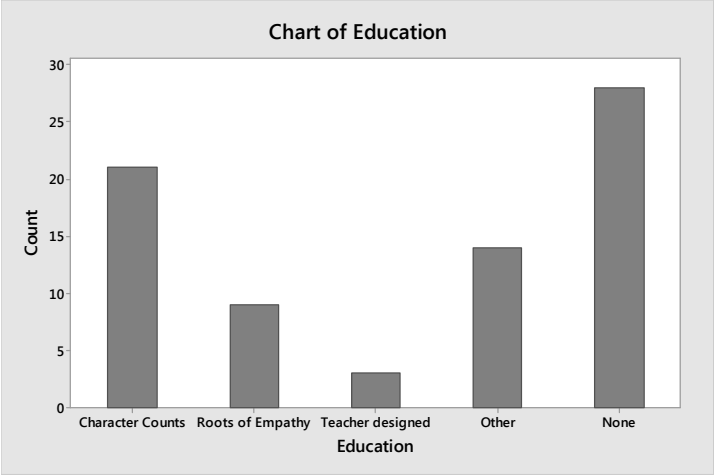


2.10 a. Data were collected on 3 questions. For questions 1 and 2, the responses were either 'yes' or 'no'. Since these are not numbers, the data are qualitative. For question 3, the responses include 'character counts', 'roots of empathy', 'teacher designed', other', and 'none'. Since these responses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

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

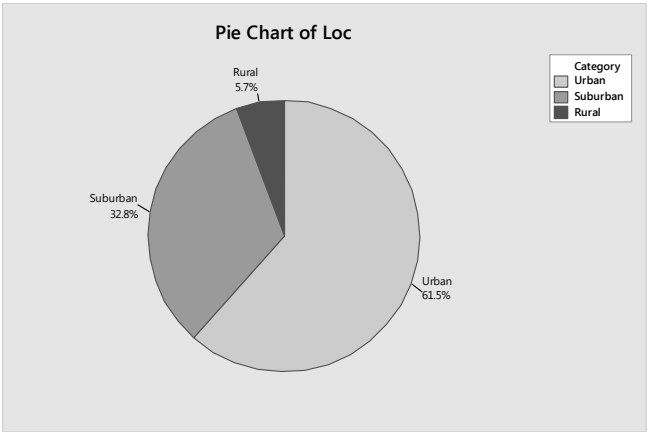






c. Many different things can be written. Possible answers might be: Most of the classroom teachers surveyed $(61/75 = 0.813)$ keep classroom pets. A little less than half of the surveyed classroom teachers $(35/75 = 0.467)$ allow visits by pets.

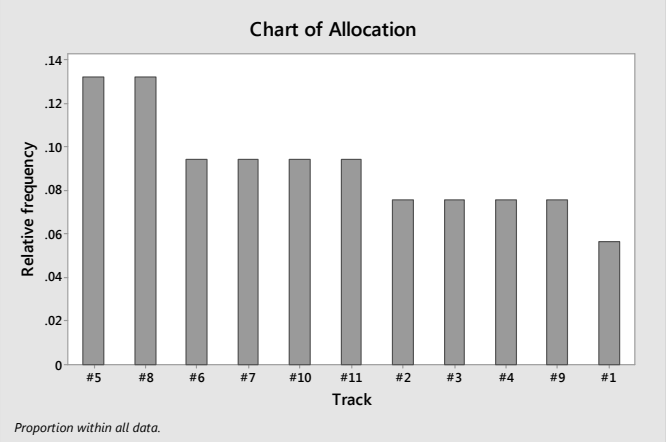
2.12 Using MINITAB, the pie chart is:



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

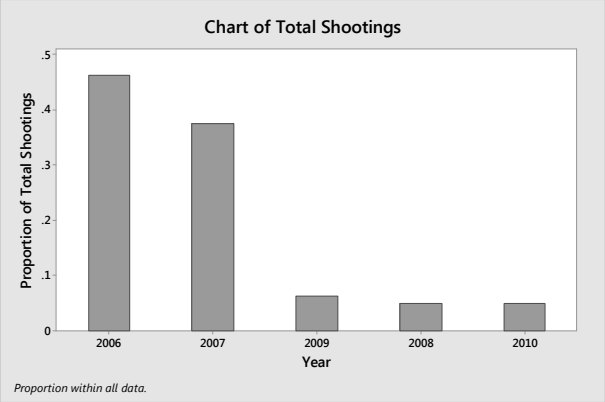
b. In the Continued Line, about 63% were in either the high or the middle grade. Only about 20% were in the nonofficial category. In the Dropout Line, only about 22% were in either the high or middle grade while about 64% were in the nonofficial category. The percentages in the low grade and provincial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

2.16 Using MINITAB, the Pareto chart is:



From the graph, it appears that tracks #5 and #8 were over-utilized and track #1 is under-utilized.

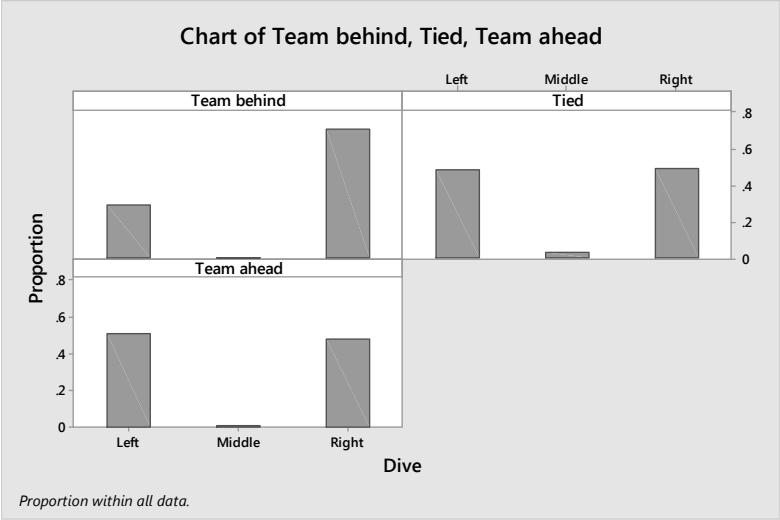
2.18 a. Using MINITAB, the Pareto chart of the total annual shootings involving the Boston street gang is:



b. Using MINITAB, the Pareto chart of the annual shootings of the Boston gang members is:

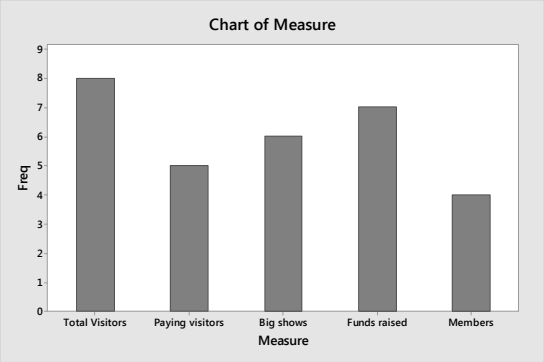
c. Because the proportion of shootings per year dropped drastically after 2007 for both the total annual shootings and annual shootings of the Boston street gang members, it appears that Operation Ceasefire was effective.

2.20 Using MINITAB, the side-by-side bar graphs showing the distribution of dives for the three match situations are:



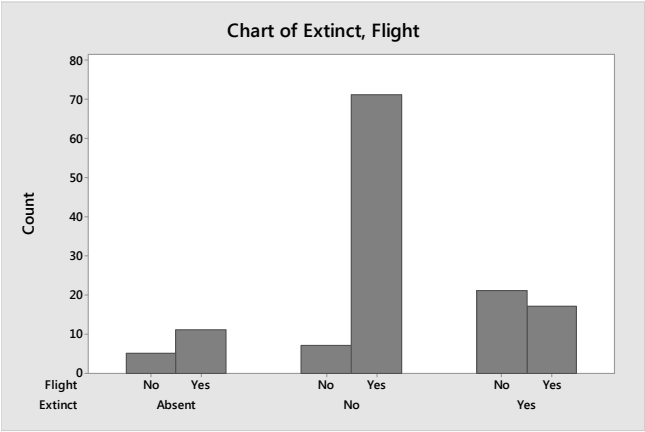
From the graphs, it appears that when a team is tied or ahead, there is no difference in the proportion of times the goal-keeper dives right or left. However, if the team is behind, the goal-keeper tends to dive right much more frequently than left.

2.22 Using MINITAB, a bar graph of the data is:



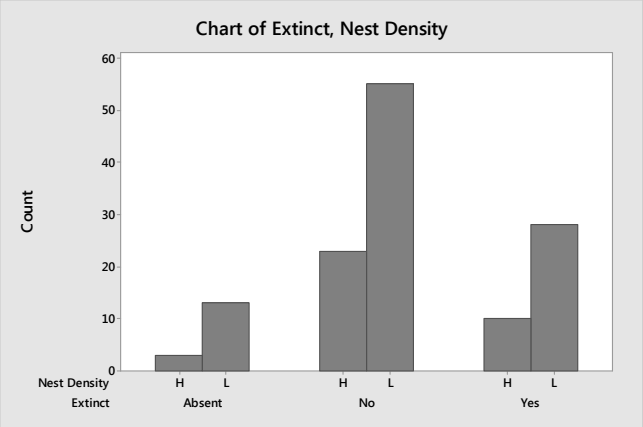
The researcher concluded that "there is a large amount of variation within the museum community with regards to . . . performance measurement and evaluation". From the data, there are only 5 different performance measures. I would not say that this is a large amount. Within these 5 categories, the number of times each is used does not vary that much. I would disagree with the researcher. There is not much variation.

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



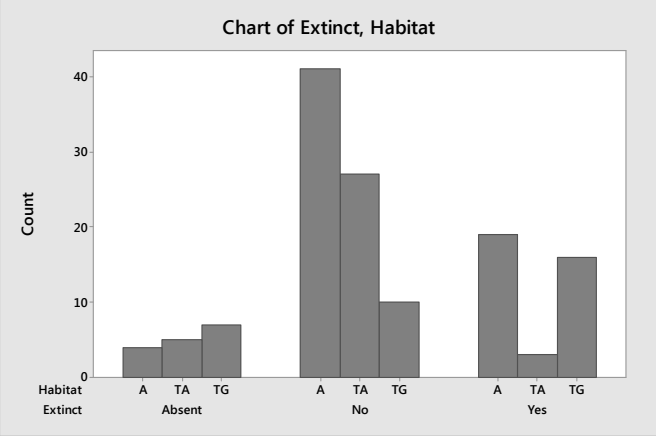
It appears that extinct status is related to flight capability. For birds that do have flight capability, most of them are present. For those birds that do not have flight capability, most are extinct.

The bar chart for Extinct status versus Nest Density is:



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

The bar chart for Extinct status versus Habitat is:



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

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

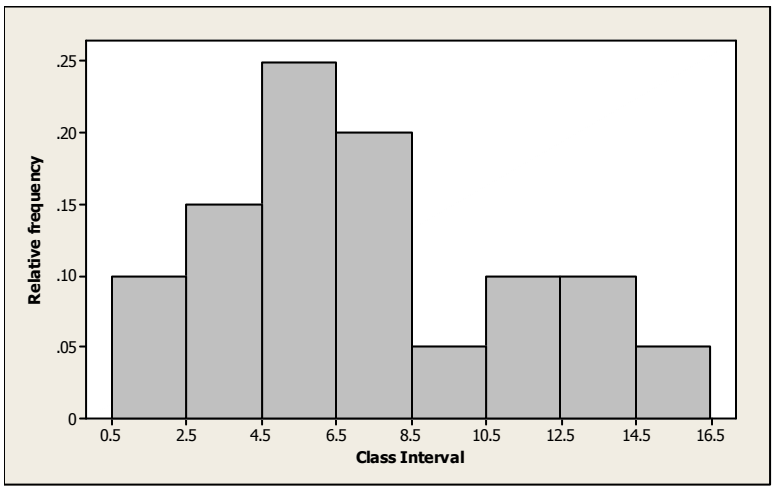
2.28 In a histogram, a class interval is a range of numbers above which the frequency of the measurements or relative frequency of the measurements is plotted.

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

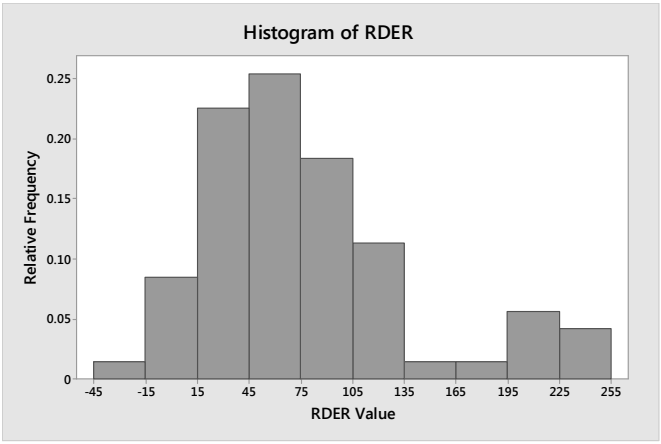
b. There are 14 class intervals used in this histogram.

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

2.32 Using MINITAB, the relative frequency histogram is:



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



b. From the graph, the proportion of subjects with RDER values between 75 and 105 is about 0.18. The exact proportion is $13/71 = 0.183$.

b. From the graph, the proportion of subjects with RDER values below 15 is about $0.01 + 0.08 = 0.09$. The exact proportion is $(1+6)/71 = 0.099$.

2.36 a. Because the label on the vertical axis is 'Percent' , this is a relative frequency histogram.

b. From the graph, the percentage of the 992 senior managers who reported a high level of support for corporate sustainability is about
$3.8 + 2.4 + 2.1 + 1.2 + 1.2 + 0.5 + 0.7 + 0.2 + 0.1 + 0 + 0.1 = 12.3\%$.

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

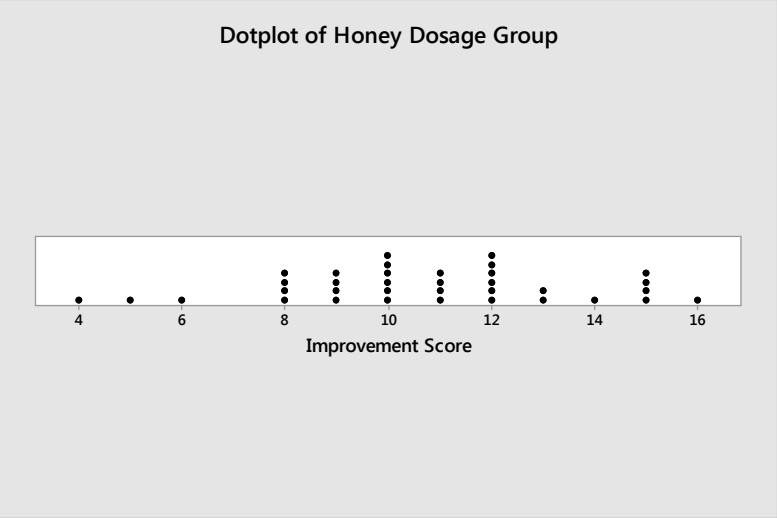
Stem-and-Leaf Display: Depth

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

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

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

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



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

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

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

Stem-and-Leaf Display: Spider

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

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

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

c. The sample proportion of spiders that a bird could detect is $3/10 = 0.3$. Thus, we could infer that a bird could detect a crab-spider sitting on the yellow central part of a daisy about 30% of the time.

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

Stem-and-Leaf Display: FNE

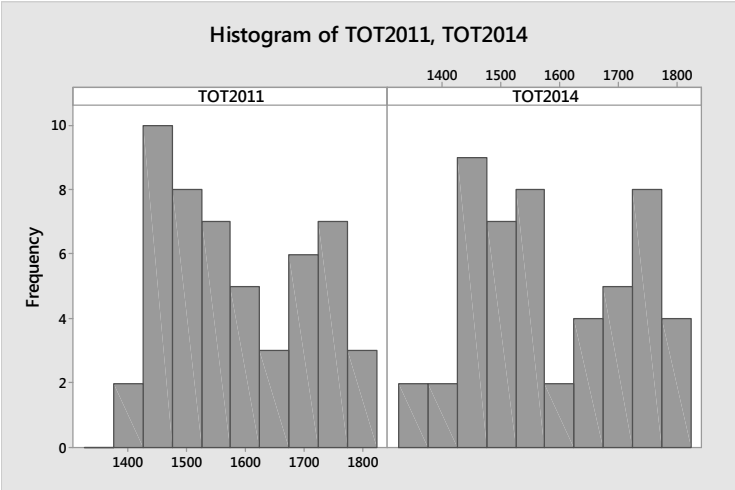
```
Stem-and-leaf of FNE       N  = 25
Leaf Unit = 1.0

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

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

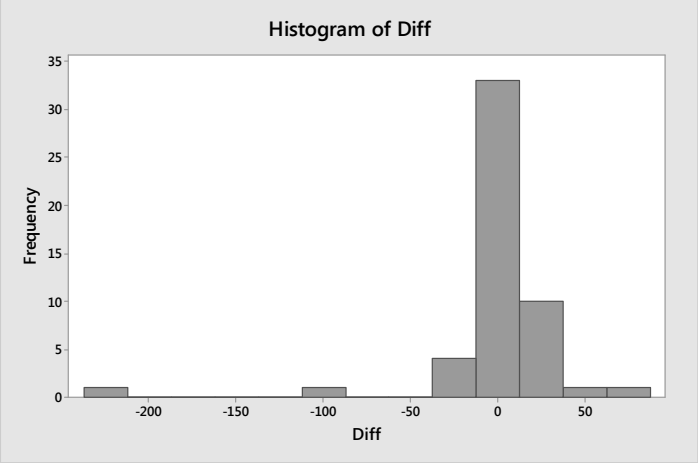
c. A measure of reliability indicates how certain one is that the conclusion drawn is correct. Without a measure of reliability, anyone could just guess at a conclusion.

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



It appears that the distributions of both sets of scores are somewhat skewed to the right. Although the distributions are not identical for the two years, they are similar.

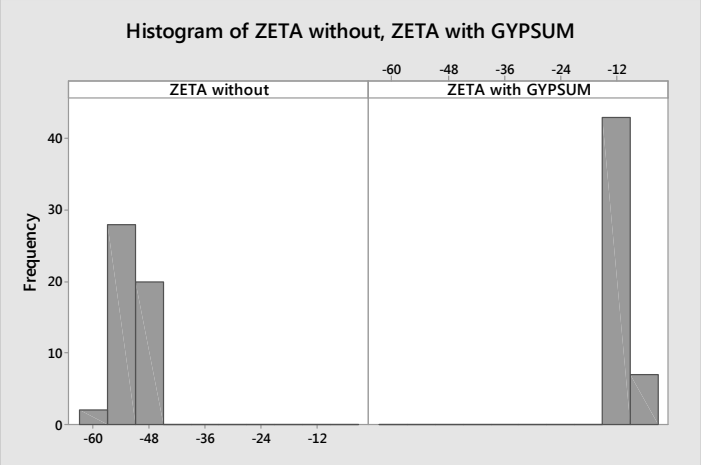
b. Using MINITAB, a histogram of the differences of the 2014 and 2011 SAT scores is:



c. It appears that there are more differences above 0 than below 0. Thus, it appears that in general, the 2014 SAT scores are higher than the 2011 SAT scores. However, there are many differences that are very close to 0.

d. Wyoming had the largest improvement in SAT scores from 2011 to 2014, with an increase of 65 points.

2.48 Using MINITAB, the side-by-side histograms are:



The addition of calcium/gypsum increases the values of the zeta potential of silica. All of the values of zeta potential for the specimens containing calcium/gypsum are greater than all of the values of zeta potential for the specimens without calcium/gypsum.

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

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

2.54 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.

2.56 Assume the data are a sample. The mode is the observation that occurs most frequently. For this sample, the mode is 15, which occurs 3 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.

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

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

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

2.60 a. The mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+...+9}{13} = \frac{42}{13} = 3.23$. This is the average number of sword shafts buried at each grave site.

b. To find the median, the data must be arranged in order. The data arranged in order are:

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

There are a total of 13 observations, which is an odd number. The median is the middle number which is 2. Half of the grave sites had 2 or fewer sword shafts buried and half had 2 or more.

c. The mode is the number that occurs most frequently. In this case, the mode is 2.

2.62 a. The mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{54+42+51+...+40}{8} = \frac{365}{8} = 45.625$. This is the Performance Anxiety Inventory scale for participants in 8 different studies.

b. To find the median, the data must be arranged in order. The data arranged in order are:

39 40 41 42 43 51 54 55

There are a total of 8 observations, which is an even number. The median is the average of the middle 2 numbers which are 42 and 43. The median is $\frac{42+43}{2} = \frac{85}{2} = 42.5$. Half of the studies had a PAI scale less than 42.5 and half had a value greater than 42.5.

c. If 39 were eliminated, the mean becomes

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{54+42+51+...+40}{7} = \frac{326}{7} = 46.571$. The data arranged in order are now: 40 41 42 43 51 54 55. The median is the middle number which is 43. The mean increased by 0.946 while the median only increased by 0.5.

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

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

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

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

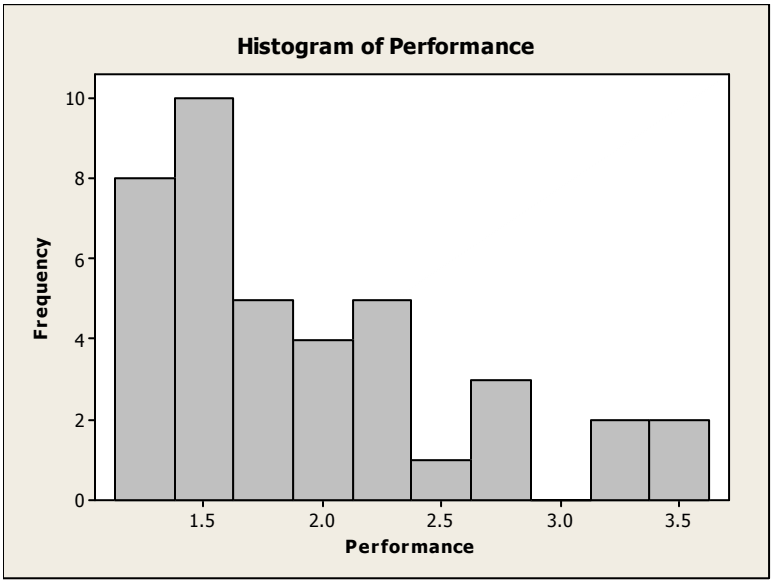
2.66 a. The mean of the driving performance index values is: $\bar{x} = \frac{\sum x}{n} = \frac{77.07}{40} = 1.927$

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

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

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

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



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

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

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

The median is 5.

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

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

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

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

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

The median is 40.

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

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

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

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

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

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

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

2.70 a. Using MINITAB, the simple statistics are:

Descriptive Statistics: ZETA without, ZETA with GYPSUM

Variable	N	Mean	Median	Mode	N for Mode
ZETA without	50	-52.070	-52.250	-50.2	3
ZETA with GYPSUM	50	-10.958	-11.300	-11.3	5

For the liquid solutions prepared without calcium/gypsum, the mean zeta potential measurement is -52.070, the median is -52.250 and the mode is -50.2. The average zeta potential measurement for liquid solutions prepared without calcium/gypsum is -52.070. Half of the zeta potential measurements for liquid solutions prepared without calcium/gypsum are less than or equal to -52.250 and half are greater than -52.250. The most common zeta potential measurement for liquid solutions prepared without calcium/gypsum is -50.2.

b. For the liquid solutions prepared with calcium/gypsum, the mean zeta potential measurement is -10.958, the median is -11.300 and the mode is -11.3. The average zeta potential measurement for liquid solutions prepared with calcium/gypsum is -10.958. Half of the zeta potential measurements for liquid solutions prepared with calcium/gypsum are less than or equal to -11.300 and half are greater than -11.300. The most common zeta potential measurement for liquid solutions prepared with calcium/gypsum is -11.3.

c. The interpretation remains the same as in Exercise 2.48. The addition of calcium/gypsum increases the values of the zeta potential of silica. The mean, median and mode of the values of zeta potential for the specimens containing calcium/gypsum are greater than the mean, median and mode of the values of zeta potential for the specimens without calcium/gypsum.

2.72 a. The mean number of power plants is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+3+4+\ldots+3}{20} = \frac{79}{20} = 3.95$$

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

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

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

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+3+4+\ldots+3}{19} = \frac{68}{19} = 3.58$$

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

The median is 3.

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

By dropping the largest measurement from the data set, the mean drops from 3.95 to 3.58. The median drops from 3.5 to 3 and the mode stays the same.

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

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

The new mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+1+1+...+7}{16} = \frac{57}{16} = 3.56$$

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

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

2.76 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 0.8, 0.7, 0.9, 0.5, and 0.3. The sample mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{0.8+0.7+0.9+0.5+0.3}{.5} = \frac{3.2}{5} = 0.64$$

The sample variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{2.28 - \frac{3.2^2}{5}}{5-1} = \frac{2.28 - 2.048}{4} = 0.058$$

The standard deviation is $s = \sqrt{0.058} = 0.241$

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

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

c. $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{18 - \dfrac{17^2}{20}}{20-1} = 0.1868$ $\qquad s = \sqrt{0.1868} = 0.432$

2.80 a. Range $= 4 - 0 = 4$

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

b. Range $= 6 - 0 = 6$

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

c. Range $= 8 - (-2) = 10$

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

d. Range $= 2 - (-3) = 5$

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

2.82 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

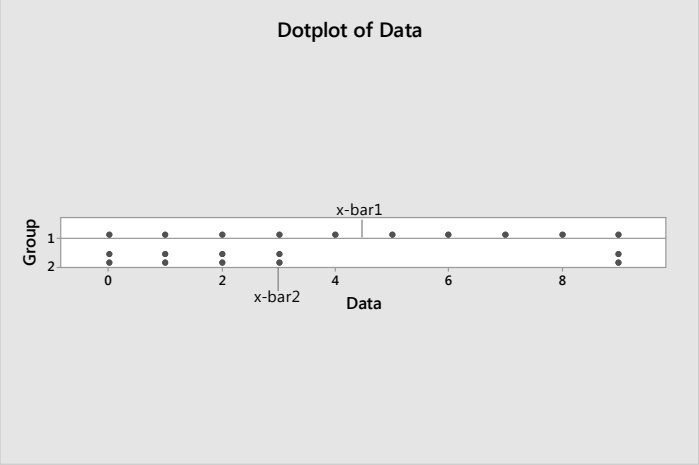
The two sets of data above have the same range = largest measurement − smallest measurement $= 9 - 0 = 9$.

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.



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

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

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

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

d. $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{\dfrac{24}{25} - \dfrac{2^2}{6}}{6-1} = \dfrac{0.2933}{5} = 0.0587$ square ounces

$s = \sqrt{0.0587} = 0.2422$ ounce

2.86 a. The range is the difference between the largest and smallest numbers. For this data set, the range is $9 - 1 = 8$.

b. The sample variance is $s^2 = \dfrac{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}{n-1} = \dfrac{202 - \dfrac{42^2}{13}}{13-1} = 5.526$.

c. The sample standard deviation is $s = \sqrt{5.526} = 2.351$.

d. Both the range and the standard deviation have the same units of measure as the original variable.

2.88 a. The range is the difference between the largest and smallest observations and is $17.83 - 4.90 = 12.93$ meters.

b. The variance is:

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

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

2.90 a. For Group A, the range is 67.20. From the printout, the range is $122.40 - 55.20 = 67.2$.

b. For Group A, the standard deviation is 14.48. From the printout, $s = \sqrt{s^2} = \sqrt{209.53} = 14.48$.

c. Group B has the more variable permeability data. Group B has the largest range, the largest variance and the largest standard deviation.

2.92 a. Range $= 11 - 1 = 10$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{455 - \frac{79^2}{20}}{20-1} = 7.524 \qquad s = \sqrt{s^2} = \sqrt{7.524} = 2.743$$

b. Dropping the largest measurement:
Range $= 9 - 1 = 8$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{334 - \frac{68^2}{19}}{19-1} = 5.035 \qquad s = \sqrt{s^2} = \sqrt{5.035} = 2.244$$

By dropping the largest observation from the data set, the range decreased from 10 to 8, the variance decreased from 7.524 to 5.035 and the standard deviation decreased from 2.743 to 2.244.

c. Dropping the largest and smallest measurements:
Range $= 9 - 1 = 8$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{333 - \frac{67^2}{18}}{18-1} = 4.918 \qquad s = \sqrt{s^2} = \sqrt{4.918} = 2.218$$

By dropping the largest and smallest observations from the data set, the range decreased from 10 to 8, the variance decreased from 7.524 to 4.918 and the standard deviation decreased from 2.743 to 2.218.

2.94 The Empirical Rule applies only to data sets that are mound-shaped—that are approximately symmetric, with a clustering of measurements about the midpoint of the distribution and that tail off as one moves away from the center of the distribution.

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

a. At least 0 of the measurements 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 $\bar{x} + 3s$.

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

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

b.

Interval	Number of Measurements in Interval	Percentage
$\bar{x} \pm s$, or (6.41, 10.17)	18	18 / 25 = 0.72 or 72%
$\bar{x} \pm 2s$, or (4.58, 11.90)	24	24 / 25 = 0.96 or 96%
$\bar{x} \pm 3s$, or (2.75, 13.73)	25	25 / 25 = 1 or 100%

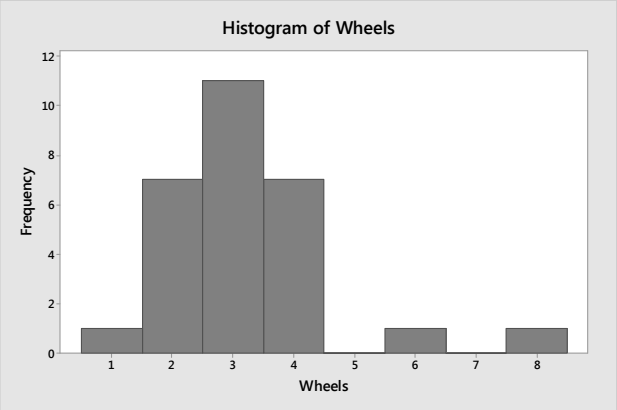
c. The percentages in part **b** are in agreement with Chebyshev's rule and agree fairly well with the percentages given by the Empirical Rule.

d. Range = 12 − 5 = 7

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

The range approximation provides a satisfactory estimate of s.

2.100 a. Using MINITAB, the histogram is:



Although the distribution is somewhat mound-shaped, the distribution is skewed to the right.

b. Using MINITAB, the mean and standard deviation are:

Descriptive Statistics: Wheels

```
Variable   N   Mean  StDev
Wheels    28  3.214  1.371
```

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

d. According to Chebyshev's rule, at least $\frac{3}{4}$ of the measurements will fall within 2 standard deviations of the mean.

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

f. Twenty-six of the twenty-eight observations fall within the interval. The proportion is $\frac{26}{28} = 0.929$. The Empirical Rule does provide a good estimate of the proportion. The actual percentage is 92.9% which is close to 95%.

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

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

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

2.104 a. Because the histogram in Exercise 2.34 is skewed to the right, Chebyshev's rule is more appropriate for describing the distribution of the RDER values.

b. The interval $\bar{x} \pm 2s$ is $78.19 \pm 2(63.24) \Rightarrow 78.19 \pm 126.48 \Rightarrow (-48.29, 204.67)$. At least ¾ or 75% of the observations will be between -48.29 and 204.67.

2.106 a. By Chebyshev's rule, at least 75% of the observations will fall within 2 standard deviations of the mean. This interval is $\bar{x} \pm 2s \Rightarrow 0.90 \pm 2(1.10) \Rightarrow 0.90 \pm 2.20 \Rightarrow (-1.30, 3.10)$.

b. No. A value of 7 would be $(7 - 0.90)/1.10 = 5.55$ standard deviations above the mean. This would be very unusual.

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

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

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

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

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

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

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

2.110 a. The mean and standard deviation for Group A are 73.62 and 14.48. The histogram of the data from Group A is skewed to the right so Chebyshev's rule is more appropriate. We know at least 8/9 or 88.9% of the observations will fall within 3 standard deviations of the mean. This interval is
$\bar{x} \pm 3s \Rightarrow 73.62 \pm 3(14.48) \Rightarrow 73.62 \pm 43.44 \Rightarrow (30.18, 117.06)$.

b. The mean and standard deviation for Group B are 128.54 and 21.97. The histogram of the data from Group B is skewed to the left so Chebyshev's rule is more appropriate. We know at least 8/9 or 88.9% of the observations will fall within 3 standard deviations of the mean. This interval is
$\bar{x} \pm 3s \Rightarrow 128.54 \pm 3(21.97) \Rightarrow 128.54 \pm 65.91 \Rightarrow (62.63, 194.45)$.

c. The mean and standard deviation for Group C are 83.07 and 20.05. The histogram of the data from Group C is skewed to the right so Chebyshev's rule is more appropriate. We know at least 8/9 or 88.9% of the observations will fall within 3 standard deviations of the mean. This interval is
$\bar{x} \pm 3s \Rightarrow 83.07 \pm 3(20.05) \Rightarrow 83.07 \pm 60.15 \Rightarrow (22.92, 143.22)$.

d. It appears that weathering type B results in faster decay.

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

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

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

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

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

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

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

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

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

2.118 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.

2.120 The mean score is 153. This is the arithmetic average score of U.S. twelfth graders on the mathematics assessment test. The 25th percentile score is 111. This indicates that 25% of the U.S. twelfth graders scored 111 or lower on the assessment test. The 75th percentile score is 177. This indicates that 75% of the U.S. twelfth graders scored 177 or lower on the

assessment test. The 90th percentile score is 197. This indicates that 90% of the U.S. twelfth graders scored 197 or lower on the assessment test.

2.122 a. From Exercise 2.35, the proportion of fup/fumic ratios that fall above 1 is 0.034. The percentile rank of 1 is $(1-0.034)100\% = 96.6^{\text{th}}$ percentile.

b. From Exercise 2.35, the proportion of fup/fumic ratios that fall below 0.4 is 0.695. The percentile rank of 0.4 is $(0.695)100\% = 69.5^{\text{th}}$ percentile.

2.124 a. The *z*-score associated with a score of 30 is $z = \dfrac{x-\bar{x}}{s} = \dfrac{30-39}{6} = -1.50$. This means that a score of 30 is 1.5 standard deviations below the mean.

b. The *z*-score associated with a score of 39 is $z = \dfrac{x-\bar{x}}{s} = \dfrac{39-39}{6} = 0$. Half or 0.5 of the observations are below a score of 39.

2.126 Since the 90th percentile of the study sample in the subdivision was 0.00372 mg/L, which is less than the USEPA level of 0.015 mg/L, the water customers in the subdivision are not at risk of drinking water with unhealthy lead levels.

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

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

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

2.130 Yes. From the graph in Exercise 2.129 c, we can see that there are 4 observations with *z*-scores greater than 3. There is then a gap down to 2.18. Those 4 observations are quite different from the rest of the data. After those 4 observations, the data are fairly similar. We know that by ranking the data, we can reduce the influence of outliers. But, by doing this, we lose valuable information.

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

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

2.136 The interquartile range is $\text{IQR} = Q_U - Q_L = 85 - 60 = 25$.

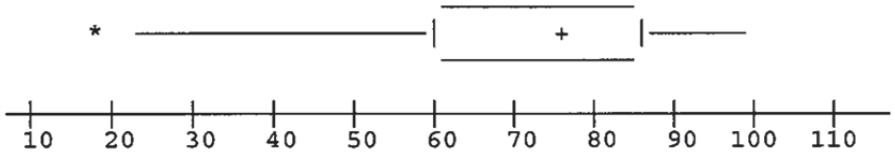
The lower inner fence $= Q_L - 1.5(\text{IQR}) = 60 - 1.5(25) = 22.5$.

The upper inner fence $= Q_U + 1.5(\text{IQR}) = 60 + 1.5(25) = 122.5$.

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

The upper outer fence $= Q_U + 3(\text{IQR}) = 60 + 3(25) = 160$.

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.

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

a. $z = \dfrac{x - \bar{x}}{s} = \dfrac{65 - 57}{11} = .727$

Since this *z*-score is less than 3 in magnitude, 65 is not an outlier.

b. $z = \dfrac{x - \bar{x}}{s} = \dfrac{21 - 57}{11} = -3.273$

Since this *z*-score is more than 3 in magnitude, 21 is an outlier.

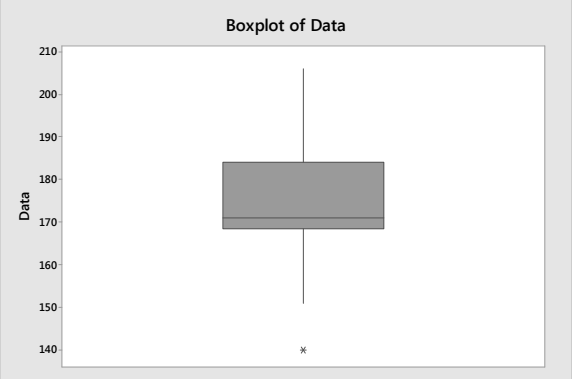
c. $z = \dfrac{x - \bar{x}}{s} = \dfrac{72 - 57}{11} = 1.364$

Since this *z*-score is less than 3 in magnitude, 72 is not an outlier.

d. $z = \dfrac{x - \bar{x}}{s} = \dfrac{98 - 57}{11} = 3.727$

Since this *z*-score is more than 3 in magnitude, 98 is an outlier.

2.140 a. Using MINITAB, the box plot for data is given below.



b. In this data set, there is one outlier. It corresponds to the value 140.

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

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

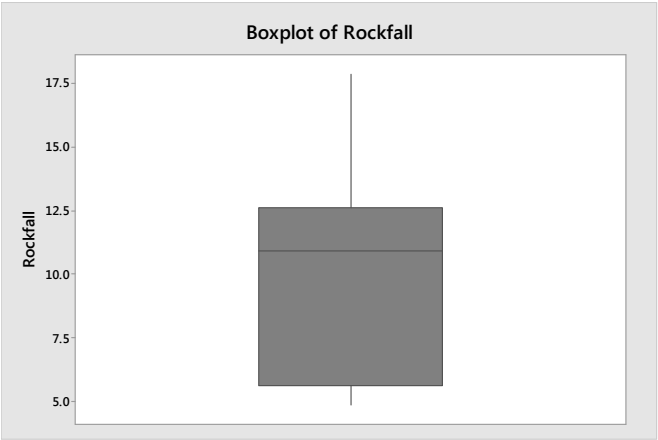
2.144 The z-score corresponding to 155 is: $z = \dfrac{x - \bar{x}}{s} = \dfrac{155 - 67.755}{26.871} = 3.25$. Because this observation is more than 3 standard deviations from the mean, it is considered a highly suspect outlier. It would not be considered typical of the study sample.

2.146 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 27.5.

b. The approximate 25th percentile PASI score after treatment is 3.5. 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.

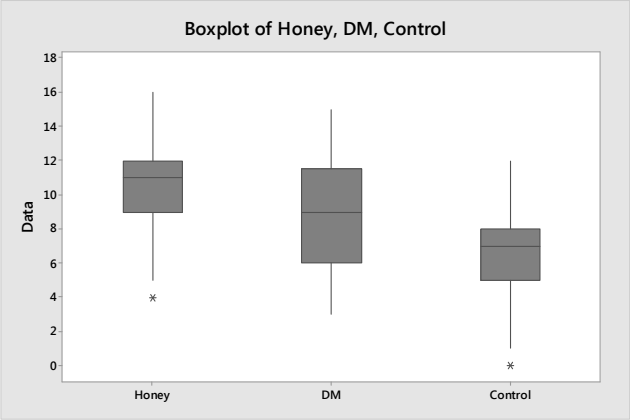
2.148 Using MINITAB, a boxplot of the data is:



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

We will now use the z-score criterion for determining outliers. From Exercises 2.61 and 2.88, $\bar{x} = 9.72$ and $s = 4.095$. The z-score associated with the minimum value is $z = \dfrac{x - \bar{x}}{s} = \dfrac{4.9 - 9.72}{4.095} = -1.18$ and the z-score associated with the maximum value is $z = \dfrac{x - \bar{x}}{s} = \dfrac{17.83 - 9.72}{4.095} = 1.98$. Neither of these indicates there are any outliers.

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



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

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

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

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

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

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

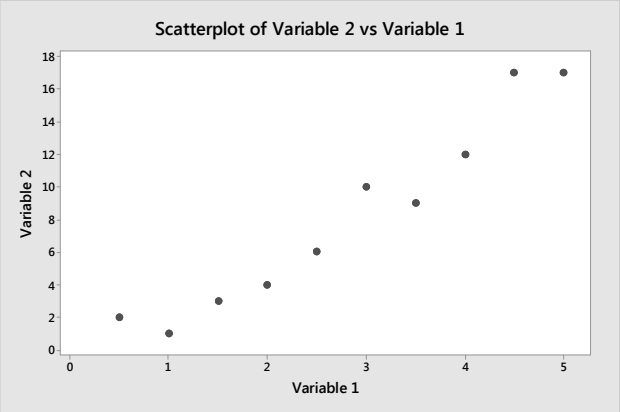
d. When $\sigma = 2$, $z = \dfrac{x - \mu}{\sigma} = \dfrac{4 - 7}{2} = -1.5$

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

2.154 A bivariate relationship is a relationship between 2 quantitative variables.

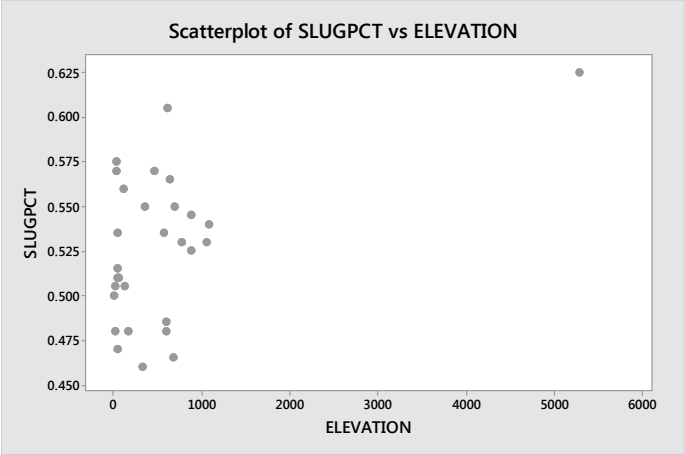
2.156 A positive association between two variables means that as one variable increases, the other variable tends to also increase. A negative association between two variables means that as one variable increases, the other variable tends to decrease.

2.158 Using MINITAB, the scatterplot is as follows:



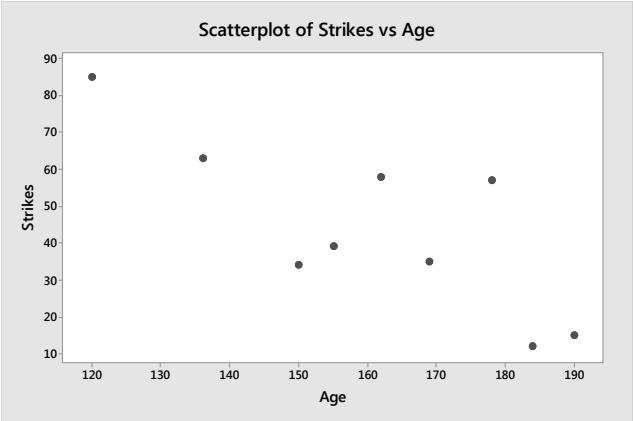
It appears that as variable 1 increases, variable 2 also increases.

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



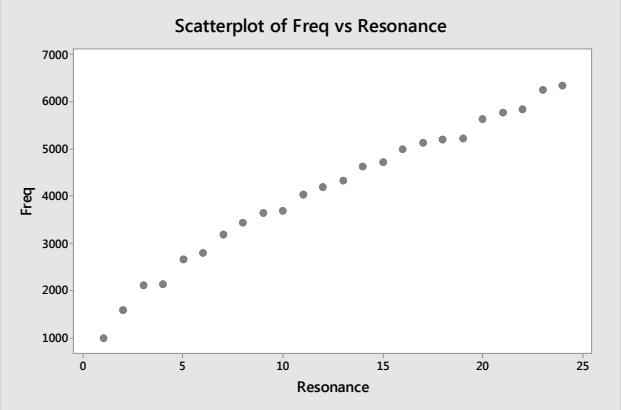
If one uses the one obvious outlier (Denver), then there does appear to be a trend in the data. As the elevation increases, the slugging percentage tends to increase. However, if the outlier is removed, then it does not look like there is a trend to the data.

2.162 a. A scattergram of the data is:



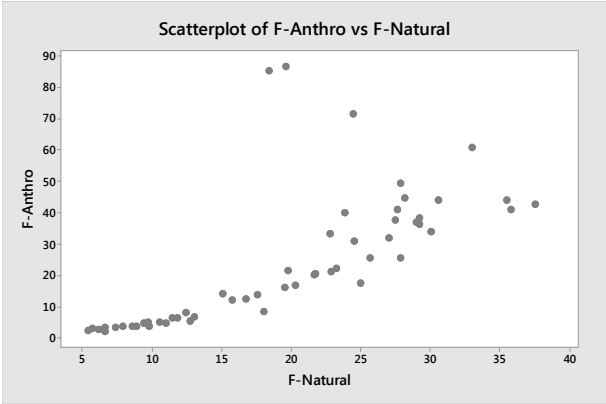
b. There appears to be a trend. As the age increases, the number of strikes tends to decrease.

2.164 Using MINITAB, a scatterplot of the data is:



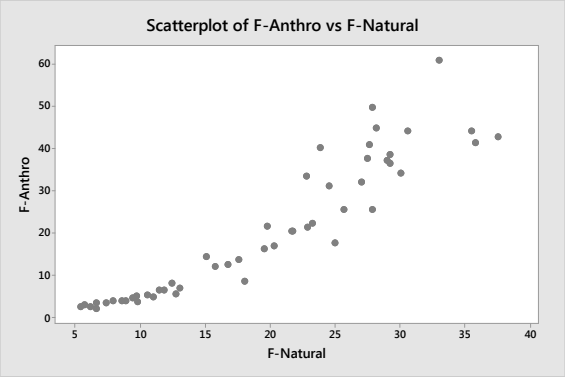
There is an increasing trend and there is very little variation in the plot. This supports the researcher's theory.

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



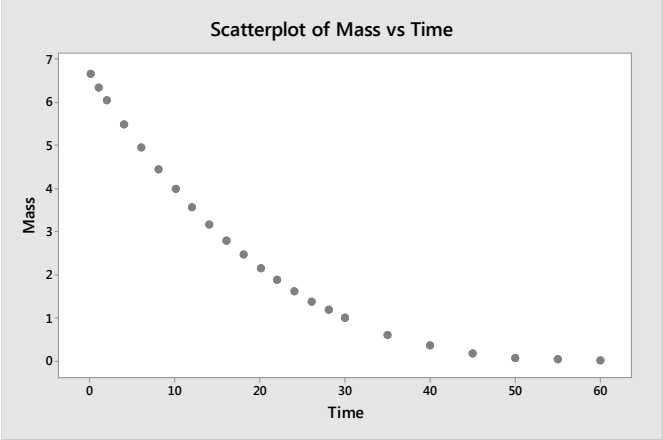
This graph does not support the theory that there is a straight-line relationship between the Anthropogenic Index against the Natural Origin Index. There are several points that do not lie on a straight line.

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



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

2.168 Using MINITAB, a scattergram of the data is:

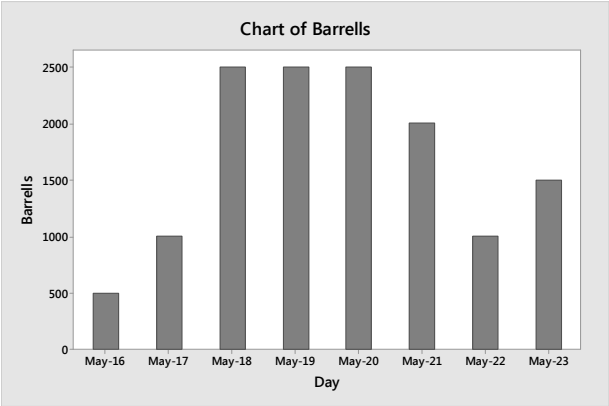


Yes, there appears to be a negative trend in this data. As time increases, the mass tends to decrease. There appears to be a curvilinear relationship. As time increases, mass decreases at a decreasing rate.

2.170 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.

2.172 a. This graph is misleading because it looks like as the days are increasing, the number of barrels collected per day is 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.

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

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

2.178 One technique for distorting information on a graph is by stretching the vertical axis by starting the vertical axis somewhere above 0.

2.180 From part a of Exercise 2.179, the 3 *z*-scores are −1, 1 and 2. Since none of these *z*-scores are greater than 2 in absolute value, none of them are outliers.

From part b of Exercise 2.179, the 3 *z*-scores are −2, 2 and 4. There is only one *z*-score greater than 2 in absolute value. The score of 80 (associated with the *z*-score of 4) would be an outlier. Very few observations are as far away from the mean as 4 standard deviations.

From part c of Exercise 2.179, the 3 *z*-scores are 1, 3, and 4. Two of these *z*-scores are greater than 2 in absolute value. The scores associated with the two *z*-scores 3 and 4 (70 and 80) would be considered outliers.

From part d of Exercise 2.179, the 3 *z*-scores are .1, .3, and .4. Since none of these *z*-scores are greater than 2 in absolute value, none of them are outliers.

2.182 $\sigma \approx \text{range} / 4 = 20 / 4 = 5$

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

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

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

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{288 - \frac{30^2}{5}}{5-1} = \frac{108}{4} = 27 \qquad s = \sqrt{27} = 5.196$$

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

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

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

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{241 - \frac{25^2}{4}}{4-1} = \frac{84.75}{3} = 28.25 \qquad s = \sqrt{28.25} = 5.315$$

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

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

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

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{569 - \frac{49^2}{7}}{7-1} = \frac{226}{6} = 37.667 \qquad s = \sqrt{37.667} = 6.137$$

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

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

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

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

e. a) $\bar{x} \pm 2s \Rightarrow 6 \pm 2(5.2) \Rightarrow 6 \pm 10.4 \Rightarrow (-4.4, 16.4)$
All or 100% of the observations are in this interval.

b) $\bar{x} \pm 2s \Rightarrow 6.25 \pm 2(5.32) \Rightarrow 6.25 \pm 10.64 \Rightarrow (-4.39, 16.89)$
All or 100% of the observations are in this interval.

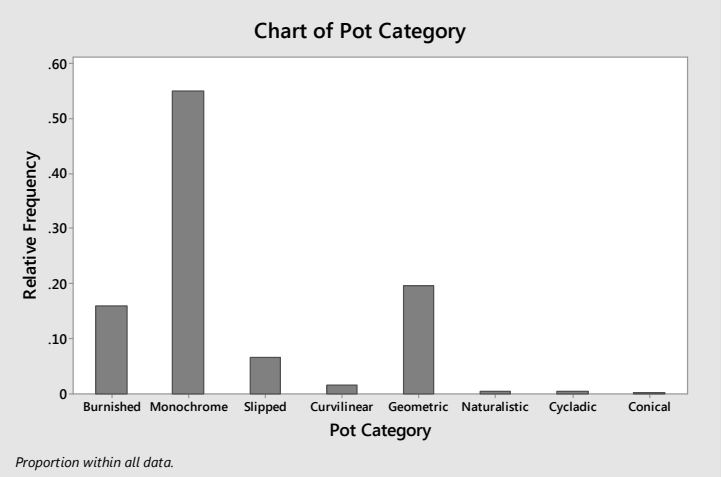
c) $\bar{x} \pm 2s \Rightarrow 7 \pm 2(6.14) \Rightarrow 7 \pm 12.28 \Rightarrow (-5.28, 19.28)$
All or 100% of the observations are in this interval.

d) $\bar{x} \pm 2s \Rightarrow 3 \pm 2(0) \Rightarrow 3 \pm 0 \Rightarrow (3, 3)$
All or 100% of the observations are in this interval.

2.186 Suppose we construct a relative frequency bar chart for this data. This will allow the archaeologists to compare the different categories easier. First, we must compute the relative frequencies for the categories. These are found by dividing the frequencies in each category by the total 837. For the burnished category, the relative frequency is $133 / 837 = 0.159$. The rest of the relative frequencies are found in a similar fashion and are listed in the table.

Pot Category	Number Found	Computation	Relative Frequency
Burnished	133	133 / 837	0.159
Monochrome	460	460 / 837	0.550
Slipped	55	55 / 837	0.066
Curvilinear Decoration	14	14 / 837	0.017
Geometric Decoration	165	165 / 837	0.197
Naturalistic Decoration	4	4 / 837	0.005
Cycladic White clay	4	4 / 837	0.005
Conical cup clay	2	2 / 837	0.002
Total	837		1.001

A relative frequency bar chart is:



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

2.188 a. Using MINITAB, the stem-and-leaf display is as follows.

Character Stem-and-Leaf Display

```
Stem-and-leaf of Books  N  = 14
Leaf Unit = 1.0

    1    1 6
    5    2 0124
    6    2 8
  (3)    3 044
    5    3 9
    4    4 002
    1    4
    1    5 3
```

b. The leaves that correspond to students who earned an "A" grade are highlighted in the graph above. Those students who earned A's tended to read the most books.

c. The mean is $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{53+42+40+...+16}{14} = \dfrac{443}{14} = 31.643$. This is the average number of books read per student.

To find the median, the data must be arranged in order. In this problem, the data are already arranged in order. There are a total of 14 observations, which is an even number. The median is the average of the middle 2 numbers which are 30 and 34. The median is $\dfrac{34+30}{2} = \dfrac{64}{2} = 32$. Half of the students read more than 32 books and half read fewer.

The mode is the observation appearing the most. In this data set, there are two modes - 34 and 40 because each appears 2 times in the data set. The most frequent number of books read is either 34 or 40.

d. Since the mean and the median are almost the same, the distribution of the data set is approximately symmetric. This can be verified by the stem-and-leaf display in part a.

e. For those students who earned A, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{53+42+40+...+24}{8} = \frac{296}{8} = 37.$$

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

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

f. For those students who earned a B or C, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{40 + 28 + 22 + \ldots + 16}{6} = \frac{147}{6} = 24.5$$

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

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

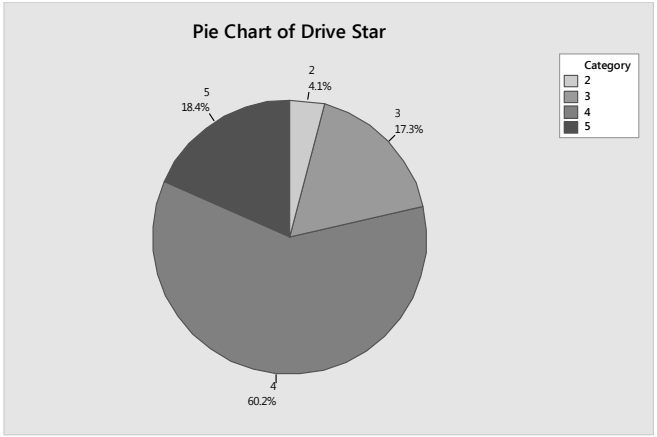
g. The students who received A's have a more variable distribution of the number of books read. The variance and standard deviation for this group are greater than the corresponding values for the B-C group.

h. The z-score for a score of 40 books is $z = \dfrac{x - \bar{x}}{s} = \dfrac{40 - 37}{8.701} = 0.345$. Thus, someone who read 40 books read more than the average number of books, but that number is not very unusual.

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

j. The group of students who earned A's is more likely to have read 40 books. For this group, the *z*-score corresponding to 40 books is 0.34. This is not unusual. For the B-C group, the *z*-score corresponding to 40 books is 1.82. This is close to 2 standard deviations from the mean. This would be fairly unusual.

2.190 A pie chart of the data is:



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

2.192 a. Using MINITAB, the descriptive statistics are:

Descriptive Statistics: Ratio

```
Variable   N   Mean  StDev  Minimum  Maximum
Ratio     26  3.507  0.634    2.250    5.060
```

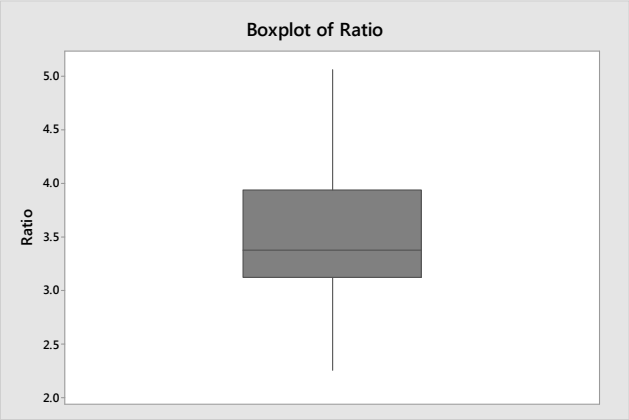
The z-score associated with the largest ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{5.06 - 3.507}{0.634} = 2.45$

The z-score associated with the smallest ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{2.25 - 3.507}{0.634} = -1.98$

The z-score associated with the mean ratio is $z = \dfrac{x - \bar{x}}{s} = \dfrac{3.507 - 3.507}{0.634} = 0$

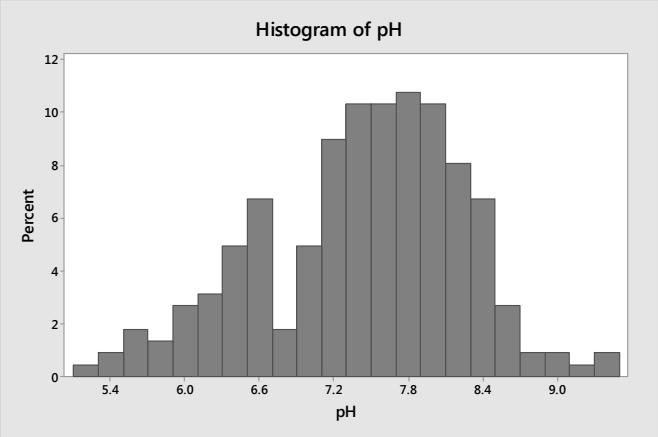
b. Yes, I would consider the z-score associated with the largest ratio to be unusually large. We know if the data are approximately mound-shaped that approximately 95% of the observations will be within 2 standard deviations of the mean. A z-score of 2.45 would indicate that less than 2.5% of all the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

c. Using MINITAB, the box plot is:



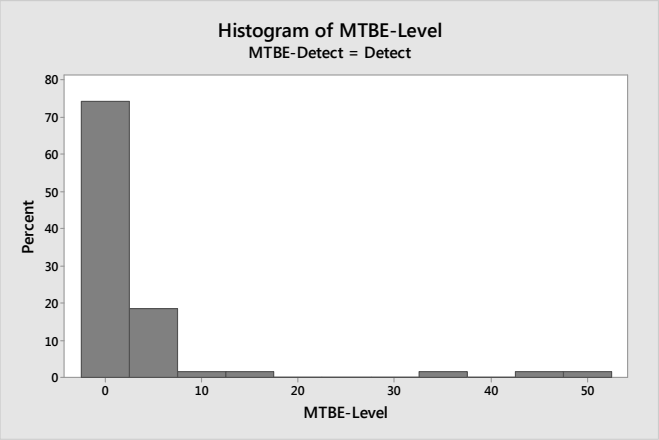
From this box plot, there are no observations marked as outliers.

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



From the graph, it looks like the proportion of wells with ph levels less than 7.0 is:
$0.005 + 0.01 + 0.02 + 0.015 + 0.027 + 0.031 + 0.05 + 0.07 + 0.017 + 0.05 = 0.295$

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



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

$0.03 + 0.01 + 0.01 + 0.01 + 0.01 + 0.01 + 0.01 = 0.09$

c. The sample mean is:

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

The variance is:

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{12{,}447.9812 - \frac{1{,}656.16^2}{223}}{223 - 1} = \frac{148.13391}{222} = 0.66727$$

The standard deviation is: $s = \sqrt{s^2} = \sqrt{0.66727} = 0.8169$

$\bar{x} \pm 2s \Rightarrow 7.427 \pm 2(0.8169) \Rightarrow 7.427 \pm 1.6338 \Rightarrow (5.7932,\ 9.0608)$.

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

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.23 + 0.24 + 0.24 + \cdots + 48.10}{70} = \frac{240.86}{70} = 3.441$$

The variance is:

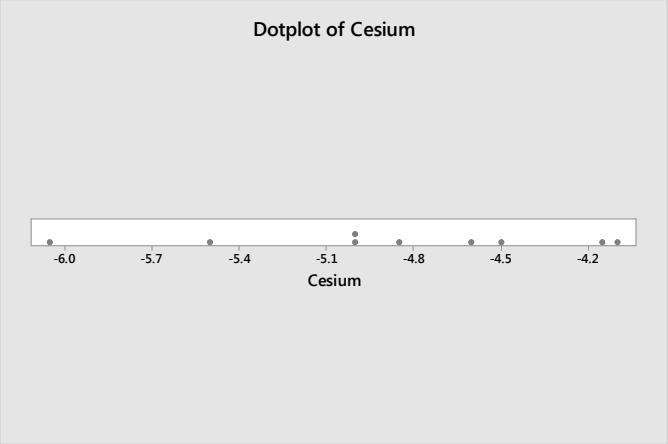
$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{6112.266 - \frac{240.86^2}{70}}{70\text{-}1} = \frac{5283.5011}{69} = 76.5725$$

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

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

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

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



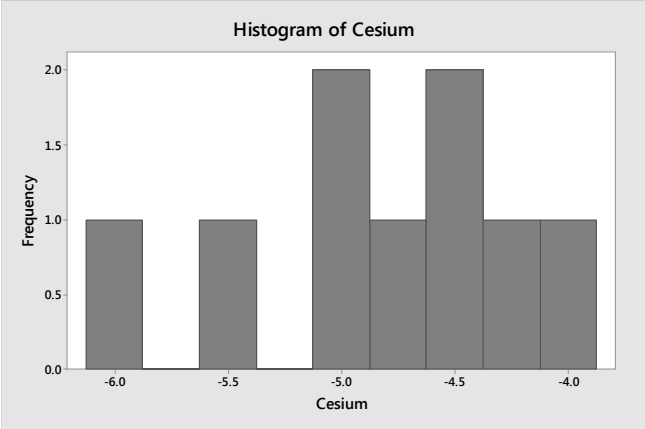
b. Using MINITAB, the stem-and-leaf display is:

Character Stem-and-Leaf Display

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

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

c. Using MINITAB, the histogram is:



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

e. There are 4 observations with radioactivity level of -5.00 or lower. The proportion of measurements with a radioactivity level of -5.0 or lower is $4/9 = 0.444$.

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

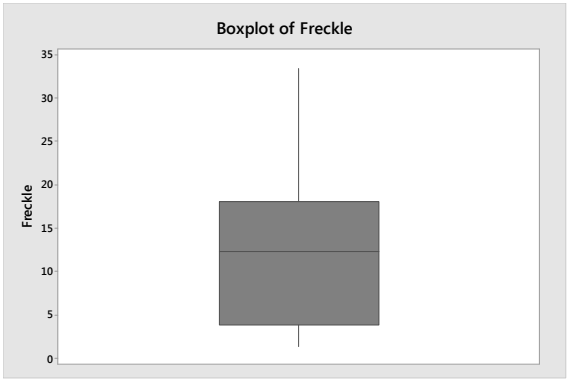
b. Using Chebyshev's rule, at least $1-1/4^2 = 1-1/16 = 15/16$ or 93.8% of the velocities will fall within 4 standard deviations of the mean. This interval is:

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

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

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

2.200 a. Using MINITAB, the boxplot is:



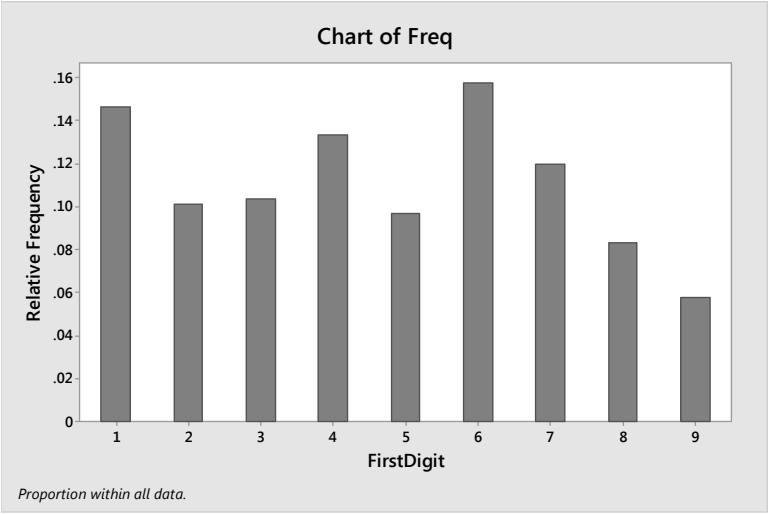
From the boxplot, there are no outliers detected.

b. Since there are no outliers, we cannot find the z-score for the points identified as outliers.

2.202 The relative frequency for each cell is found by dividing the frequency by the total sample size, $n = 743$. The relative frequency for the digit 1 is $109 / 743 = 0.147$. The rest of the relative frequencies are found in the same manner and are shown in the table.

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

Using MINITAB, the relative frequency bar chart is:



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

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

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

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

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

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

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

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

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

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

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

Even though 4,500 completed questionnaires were returned, the response rate was extremely small, only 4.5%. Also, it is very likely that this is not a representative sample of the population of American women. First, the questionnaires were not distributed to a random sample of women from the U.S. The questionnaires were sent specifically to certain groups and the directors of those groups were asked to distribute the surveys to members of their organizations. Thus, if a woman did not belong to one of the targeted groups, she had a very small chance of participating in the survey.

In these types of mail-in surveys, usually only those with very strong opinions respond to the surveys. Thus, those who actually responded to the survey were probably not the same as the population in general. The results are very likely not very reliable.