CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the “FlexBook,” CK-12 intends to pioneer the generation and distribution of high quality educational content that will serve both as core text as well as provide an adaptive environment for learning.

Copyright ©2009 CK-12 Foundation

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.
Contents

1 An Introduction to Analyzing Statistical Data - Solution Key 5
  1.1 Definitions of Statistical Terminology ........................................ 5

2 Visualizations of Data - Solution Key 15
  2.1 Review Questions ........................................................................ 15

3 An Introduction to Probability - Solution Key 39
  3.1 Review Questions ........................................................................ 39

4 Discrete Probability Distribution - Solution Key 47
  4.1 Probability Distribution for a Discrete Random Variable Review questions 47

5 Normal Distribution - Solution Key 53
  5.1 Review Questions ........................................................................ 53

6 Planning and conducting an Experiment or Study - Solution Key 65
  6.1 Surveys and Sampling Review Questions ................................. 65

7 Sampling Distributions - Solution Key 73
  7.1 Sampling Distributions .............................................................. 73

8 Hypothesis Testing - Solution Key 81
  8.1 Hypothesis Testing .................................................................... 81

9 Regression and Correlation - Solution Key 87
9.1 Regression and Correlation ......................................................... 87

10 Chi-Square - Solution Key ....................................................... 97
   10.1 Chi-Square ....................................................................... 97

11 Analysis of Variance and the F-Distribution - Solution Key ............ 101
   11.1 Analysis of Variance and the F-Distribution ............................ 101
Chapter 1

An Introduction to Analyzing Statistical Data - Solution Key

1.1 Definitions of Statistical Terminology

1. POPULATION: All candy bars made by the company
   UNITS: each individual candy bar
   VARIABLE: weight of the candy bars
   TYPE: Quantitative. It is continuous. The weights could be any weight reasonably close to
   the desired weight due to variation in the number and weight of individual candies. Note:
   if the worker decided to sort the candy bars as acceptable, too light, or too heavy, the same
   scenario could include a categorical variable.

b. POPULATION: All of Doris’ socks
   UNITS: each sock
   VARIABLE: color of socks
   TYPE: Categorical

c. POPULATION: All diabetes sufferers
   UNITS: each individual diabetes patient
   VARIABLE 1: change in sugar level (+ or −)
   TYPE: Quantitative, continuous
   VARIABLE 2: gender
   TYPE: Categorical
An Overview of Data

1. a. Ordinal
b. Nominal
c. Interval. Even though Celsius has a “0,” this is a completely arbitrary decision to set the freezing point of water and not the “absence” of temperature.
d. Ratio. The Kelvin scale is based on an absolute zero, the theoretical temperature at which molecules stop moving.

2. d. The levels of measurement theory is a useful tool to help categorize data, but like much of statistics, it is not an absolute “rule” that applies easily to every situation and several statisticians have pointed out some of the difficulties with the theory. See: http://en.wikipedia.org/wiki/Level_of_measurement

3. d. Population densities are certainly measured up to the interval level as there is meaning to the values and distance between two observations. To decide if it is measured at the ratio level, we need to establish a meaning for absolute zero. In this case, it would be 0 individuals per km². This is possible and indeed represents the extinct populations.

4. a. This is an experiment as each subject is drinking both waters (the imposed treatment). However, it will have to be designed properly. Students should not know which water is bottled and which is tap (this is called a “blind” experiment) and they should be randomly assigned the order in which they drink the water. Other conditions such as the appearance, amount, and temperature would also need to be tightly controlled.
b. Observational study.
c. Experiment. The research is imposing a treatment (different color rooms) on the mice.
d. Observational Study. It would be unacceptable to intentionally expose a baby to potentially harmful substances. The dangers of lead paint were discovered through years of careful observational studies.

Measures of Center

1. b. When one data point is a distance away from the others, it doesn’t change the median but the mean is pulled towards that data point. In this problem the data point (62") is larger than all the other points so the mean will be greater than the median.

2. To compute the mean you must add up all of Enrique’s scores and then divide by the number of scores – in this case 4. This will yield the mean. This problem states that grades are rounded so that to have a mean of 93 the calculated mean could be as low as 92.5
\[
\frac{91 + 87 + 95 + x}{4} = 92.5 \\
91 + 87 + 95 + x = 92.5 \times 4 \\
x = 370 - 273 \\
x = 97
\]

3. To calculate the 10% trimmed mean you must remove 10% of the data points. This would mean removing 5% of the data from each end. In this problem there are 300 data points. 10% of 300 = 30. So you must remove a total of 30 points, 15 from each end.

4. After removing these points there will be 270 points remaining. To calculate the mean you would add all remaining points and then divide by 270.

5. First put the data points in numerical order:

\[0 \ 0 \ 0 \ 40 \ 55 \ 210 \ 286 \ 357 \ 498 \ 552 \ 1293\]

a. The mode is the data point that occurs most often. In this case the mode is 0.

b. The median is the data point such that half of the data falls above the median and half falls below the median. There are 11 pieces of data. This means that the median will be the sixth data point, once the data is put in numerical order. In this problem the median is 210.

c. The mean is determined by adding all the data points and then dividing the total by 11.

\[
\frac{0 + 0 + 0 + 40 + 55 + 210 + 286 + 357 + 498 + 552 + 1293}{11} = \frac{3291}{11} = 299.18
\]

d. The 10% trimmed mean is calculated by first removing 1.1 pieces of data (10% of 11). This is not possible so we decide to remove 2 points, one from each end of the ordered data. In this case we would be removing 0 and 1293 and then calculate the mean.

\[
\frac{0 + 0 + 40 + 55 + 210 + 286 + 357 + 498 + 552}{9} = \frac{1998}{9} = 222
\]

e. The midrange is calculated by finding the mean of the minimum and maximum data points.
\[
\frac{0 + 1293}{2} = 646.5
\]

f. The lower quartile \( Q_1 \) is the median of the lower half of the data and the upper quartile \( Q_3 \) is the midpoint of the upper half of the data. The bottom half of the data is 0 0 0 40 55. The median of this is 0. So \( Q_1 = 0 \) The top half of the data is 286 357 498 552 1293. The median of this is 498. So \( Q_3 = 498 \)

g. There are 8 data points below Santiago (498) so the percentile is \( \frac{8}{11} = .127 = 72.7 \) percent.

6. There is one extreme point, 1293, which causes the mean to be greater than the median.

**Measures of Spread**

1. The sample mean is calculated by taking the sum of all observations and dividing by 10, the number of observations. The sample mean is 69.3

<table>
<thead>
<tr>
<th>Year</th>
<th>Rainfall (inches)</th>
<th>Deviation ( x - 69.3 )</th>
<th>Squared deviations ( (x - 69.3)^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>90</td>
<td>20.7</td>
<td>428.49</td>
</tr>
<tr>
<td>1999</td>
<td>56</td>
<td>-13.3</td>
<td>176.89</td>
</tr>
<tr>
<td>2000</td>
<td>60</td>
<td>-9.3</td>
<td>86.49</td>
</tr>
<tr>
<td>2001</td>
<td>59</td>
<td>-10.3</td>
<td>106.09</td>
</tr>
<tr>
<td>2002</td>
<td>74</td>
<td>4.7</td>
<td>22.09</td>
</tr>
<tr>
<td>2003</td>
<td>76</td>
<td>6.7</td>
<td>44.89</td>
</tr>
<tr>
<td>2004</td>
<td>81</td>
<td>11.7</td>
<td>136.89</td>
</tr>
<tr>
<td>2005</td>
<td>91</td>
<td>21.7</td>
<td>470.89</td>
</tr>
<tr>
<td>2006</td>
<td>47</td>
<td>-22.3</td>
<td>497.29</td>
</tr>
<tr>
<td>2007</td>
<td>59</td>
<td>-10.3</td>
<td>106.09</td>
</tr>
<tr>
<td></td>
<td>Sum ( \rightarrow )</td>
<td></td>
<td>2076.91</td>
</tr>
</tbody>
</table>

Variance: \( \frac{2070.91}{9} = 230.68 \)

Standard Deviation: \( \sqrt{230.86} = 15.19 \)

*Use the Galapagos Tortoise data below to answer questions 2 and 3.*
Table 1.2:

<table>
<thead>
<tr>
<th>Island or volcano</th>
<th>Number of individuals repatriated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wolf</td>
<td>40</td>
</tr>
<tr>
<td>Darwin</td>
<td>0</td>
</tr>
<tr>
<td>Alcedo</td>
<td>0</td>
</tr>
<tr>
<td>Sierra Negra</td>
<td>286</td>
</tr>
<tr>
<td>Cerro Azul</td>
<td>357</td>
</tr>
<tr>
<td>Santa Cruz</td>
<td>210</td>
</tr>
<tr>
<td>Española</td>
<td>1293</td>
</tr>
<tr>
<td>San Cristóbal</td>
<td>55</td>
</tr>
<tr>
<td>Santiago</td>
<td>498</td>
</tr>
<tr>
<td>Pinzón</td>
<td>552</td>
</tr>
<tr>
<td>Pinta</td>
<td>0</td>
</tr>
</tbody>
</table>

Use technology to solve problems 2 and 3. Enter the data into STAT – EDIT – L1 Then use STAT – CALC -1-Var Stats, L1. This will give you the mean, variance and standard deviation of the data. It will also give you the first and third quartiles and this will enable you to find the IQR.

2. The range = maximum - minimum: $1293 - 0 = 1293$. The IQR = $498 - 0 = 498$

3. The standard deviation for this data is 387.03 as determined through technology.

4. If $\sigma^2 = 9$ then the population standard deviation is $\sqrt{9} = 3$. So a) is the correct response.

5. The data set in b has the largest standard deviation since the data points are more spread out than the others.

**Chapter Review – Part 1: Multiple Choice**

1. b. The standard deviation is not resistant. When you change a data point the standard deviation changes.

2. The following shows the mean number of days of precipitation by month in Juneau Alaska:

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Number of days</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with precipitation</td>
<td>&gt; 0.1 inches</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>20</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

To determine the median number of days of rain first put the data points in order:
There are 12 data points. The median will split the data set into two pieces with half the data in one piece and the other half in the second piece. In this data set there will be six pieces of data in each half. 18 is the median.

3. Given the data set 2, 10, 14, 6. There is no mode since each data point occurs only once. To find the median put the data points in order: 2, 6, 10, 14. The median = 8 since half the data is below 8 and half is above 8. To find the mean add all the points and divide by 4: \( \frac{32}{4} = 8 \). Thus, the median is equal to the mean, \( \bar{x} \)

4.e. while the standard deviation MUST always be smaller than the range, the variance is not always smaller than the range. It is also true that the variance is the square of the standard deviation, but some standard deviations will get smaller when they are squared. For example

\[
\left( \frac{1}{2} \right)^2 = \frac{1}{4}
\]

5. There are two quantitative variables in this example: age and homeroom number.

6. The highest level of measurement when you are looking at shoe sizes is interval. The correct response is c.

7. d is the correct response. In this situation a treatment is introduced and a placebo is used. Because a treatment is introduced this is an experiment.

8. d is the correct response. The mean of the data is 164.8 with a standard deviation of 6.4. 152 is two standard deviations below the mean and 177.6 is two standard deviations above the mean. The empirical rule tells us that 95% of the data is within two standard deviations of the mean (when the data is from a normal distribution). Height is considered to follow a normal distribution so we expect to see 95% of the Chinese men between 152 and 177.6 cm. So at least 75% of Chinese men between 30 and 65 are between 152 and 177.6 = cm.

9. b is the correct response. Sampling error is best described as the natural variation that is present when you do not get data from the entire population.

10. a is the correct response. If the sum of the squared deviations for a sample of 20 individuals is 277 then the variance will be \( \frac{277}{19} = 14.576 \) and the standard deviation will be \( \sqrt{14.576} = 3.818 \approx 3.82 \)

Part Two: Open-Ended Questions

11. a. quizzes count once, labs count twice as much as a quiz and test count three times. The data then looks like the following: 62, 88, 82, 89, 89, 96, 96, 87, 87, 87, 99, 99, 99
i. the mode is 99 and 87 since these each appear three times, more than the other scores.

ii. The mean is determined by adding all the scores and dividing by 13:

\[
\frac{62 + 88 + 82 + 2 \times 89 + 2 \times 96 + 3 \times 87 + 3 \times 99}{13} = \frac{1160}{13} = 89.23
\]

iii. To find the median first order the data:

62 82 87 87 87 88 89 89 96 96 99 99 99

There are 13 pieces of data so the median is the 7 data point which is 89 (there are 6 pieces of data below and 6 pieces of data above 89).

iv. The lower quartile, \( Q_1 \), is the median of the lower half of the data. There are 6 pieces of data in the lower half. The lower quartile is 89. The upper quartile falls between 96 and 99. We find the upper quartile by taking the average of these two numbers:

\[
\frac{99 + 96}{2} = \frac{195}{2} = 97.5
\]

v. The midrange is found by taking the mean of the minimum and the maximum data values:

\[
\frac{62 + 99}{2} = \frac{161}{2} = 80.5
\]

vi. The range is maximum value – minimum value. In this case it is 99 – 62 = 37

b. If the quiz score 62 were removed from the data the mode would not change, the mean would become larger (since 62, the minimum value would no longer be there), the range would become smaller and \( Q_1 \) would increase.

12. a. The sample mean is

\[
\frac{13.81 + 12.65 + 12.87 + 13.32 + 12.93}{5} = \frac{64.95}{5} = 12.99
\]
b. Complete the chart below to calculate the standard deviation of Spud’s sample. Observed data Deviations

<table>
<thead>
<tr>
<th>Observed data</th>
<th>Deviations $x - 12.99$</th>
<th>$(x - 12.99)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.18</td>
<td>.19</td>
<td>.0361</td>
</tr>
<tr>
<td>12.65</td>
<td>-.34</td>
<td>.1156</td>
</tr>
<tr>
<td>12.87</td>
<td>-.12</td>
<td>.0144</td>
</tr>
<tr>
<td>13.32</td>
<td>.33</td>
<td>.1089</td>
</tr>
<tr>
<td>12.93</td>
<td>-.06</td>
<td>.0036</td>
</tr>
<tr>
<td><strong>Sum of the squared deviations</strong> →</td>
<td></td>
<td><strong>.2786</strong></td>
</tr>
</tbody>
</table>

c. Calculate the variance $\frac{.2786}{4} = .06965$

d. The standard deviation $= \sqrt{.06965} = .2639$

e. the standard deviation tells you that the “typical” or “average” bag of chips in this sample is within .264 grams of the mean weight. Based on our sample, we would not have reason to believe that the company is selling unusually light or heavy bags of chips. Their quality control department appears to be doing a good job.

13.

Table 1.4:

<table>
<thead>
<tr>
<th>Island</th>
<th>Approximate area (sq. km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baltra</td>
<td>8</td>
</tr>
<tr>
<td>Darwin</td>
<td>1.1</td>
</tr>
<tr>
<td>Española</td>
<td>60</td>
</tr>
<tr>
<td>Fernandina</td>
<td>642</td>
</tr>
<tr>
<td>Floreana</td>
<td>173</td>
</tr>
<tr>
<td>Genovesa</td>
<td>14</td>
</tr>
<tr>
<td>Isabela</td>
<td>4640</td>
</tr>
<tr>
<td>Marchena</td>
<td>130</td>
</tr>
<tr>
<td>North Seymour</td>
<td>1.9</td>
</tr>
<tr>
<td>Pinta</td>
<td>60</td>
</tr>
<tr>
<td>Pinzón</td>
<td>18</td>
</tr>
<tr>
<td>Rabida</td>
<td>4.9</td>
</tr>
<tr>
<td>San Cristóbal</td>
<td>558</td>
</tr>
<tr>
<td>Santa Cruz</td>
<td>986</td>
</tr>
<tr>
<td>Santa Fe</td>
<td>24</td>
</tr>
<tr>
<td>Santiago</td>
<td>585</td>
</tr>
</tbody>
</table>
Table 1.4: (continued)

<table>
<thead>
<tr>
<th>Island</th>
<th>Approximate area (sq. km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Plaza</td>
<td>0.13</td>
</tr>
<tr>
<td>Wolf</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Source: [http://en.wikipedia.org/wiki/Gal%C3%A1pagos_Islands](http://en.wikipedia.org/wiki/Gal%C3%A1pagos_Islands)

a. Calculate the mode, mean, median, quartiles, range, and standard deviation for this data. Using technology: STAT - EDIT – put the data in L1. Use STAT – CALC – 1-Var Statistics, L1 This will give you the following information

Mean $-439.3 \text{ km}^2$
Median $-42 \text{ km}^2$
Upper Quartile $-558 \text{ km}^2$
Lower quartile $-4.9 \text{ km}^2$
Range $4640 - .13 = 4639.87 \text{ km}^2$
Standard deviation $1088.69 \text{ km}^2$
Mode $60 \text{ km}^2$

b. There is one very extreme outlier. Isabela is by far the largest island. In addition to that, there are many points in the lower half of the data that are very closely grouped together. Many of these islands are volcanic rock that barely poke above the surface of the ocean. The upper 50% of the data is much more spread out. This creates a situation in which the median stays very small, but the mean will be strongly pulled towards the larger numbers because it is not resistant.

c. The standard deviation is a statistic that is based on the mean. Therefore, if the mean is not resistant, the standard deviation is not, and it will also be influenced by the larger numbers. If it is a measure of the “typical” distance from the mean, then the larger points will have a disproportionate influence on the calculation. On a more intuitive level, if the upper 50% of the data is very widely spread, the standard deviation reflects that extreme variation.

14. a. Answers will vary
b. answers will vary
c. **Mean**: the average salary of the players on this team in 2007

**Median**: the salary at which half the players on the team make more than that, and half the players make less than that.

**Mode**: the salary that more players make than any other individual salary. Usually, this is
a league minimum salary that many players make.

**Midrange:** The mean of just the highest paid and lowest paid players.

**Lower quartile:** The salary at which only 25% of the players on the team make less.

**Upper quartile:** The salary at which 75% of the players make less, or the salary at which only one quarter of the team makes more.

**IQR:** The middle 50% of the players varies by this amount.

a. Answers will vary.

b. **Range:** the gap in salary between the highest and lowest paid players.

**Standard deviation:** the amount by which a typical player’s salary varies from the mean salary.

c. Answers will vary, but students should comment on spread in one sentence and center in the other. Since many baseball teams have a few star players who make much higher salaries, most examples should give the students an opportunity to comment on the presence of outliers and their affect on the statistical measures of center and spread.
Chapter 2

Visualizations of Data - Solution Key

2.1 Review Questions

1.

Table 2.1:

<table>
<thead>
<tr>
<th>Number of Plastic Beverage Bottles per week</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

To answer this question complete the frequency column. Add these numbers and subtract the sum from 30, the total number of students in the class. The sum is 24 and when subtracted from 30 the result is 6. So 6 should be in the second line of the frequency table.

2. a is the correct response. The only possible positive multiples of 10 for this problem are 10, 20, and 30. There are no data points in the intervals containing 20 and 30. So the only possible interval is the third one in the table, the one containing the number 10.

3. a:
Table 2.2:

<table>
<thead>
<tr>
<th>Liters of Water Per Person</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[60 – 70)</td>
<td>4</td>
</tr>
<tr>
<td>[70 – 80)</td>
<td>3</td>
</tr>
<tr>
<td>[80 – 90)</td>
<td>0</td>
</tr>
<tr>
<td>[90 – 100)</td>
<td>1</td>
</tr>
<tr>
<td>[100 – 110)</td>
<td>3</td>
</tr>
<tr>
<td>[110 – 120)</td>
<td>2</td>
</tr>
<tr>
<td>[120 – 130)</td>
<td>1</td>
</tr>
<tr>
<td>[130 – 140)</td>
<td>0</td>
</tr>
<tr>
<td>[140 – 150)</td>
<td>0</td>
</tr>
<tr>
<td>[150 – 160)</td>
<td>1</td>
</tr>
</tbody>
</table>

Completed Frequency Table for World Bottled Water Consumption Data (1999)

b: 

[Image of histogram]

Figure: Histogram of 1999 World Bottled Water Consumption Data

Student answers may vary if they choose a different bin width for their histogram.

c: This data set does appear to be have some characteristics of being skewed right. There also appears to be two distinct mounds. This shape is called “bimodal”.

4. a.
Table 2.3:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency (%)</th>
<th>Cumulative Frequency</th>
<th>Relative Cumulative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0 – 25)</td>
<td>7</td>
<td>50</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>(25 – 50)</td>
<td>1</td>
<td>7.1</td>
<td>8</td>
<td>57.1</td>
</tr>
<tr>
<td>(50 – 75)</td>
<td>3</td>
<td>21.4</td>
<td>11</td>
<td>78.6</td>
</tr>
<tr>
<td>(75 – 100)</td>
<td>1</td>
<td>7.1</td>
<td>12</td>
<td>85.7</td>
</tr>
<tr>
<td>(100 – 125)</td>
<td>1</td>
<td>7.1</td>
<td>13</td>
<td>92.9</td>
</tr>
<tr>
<td>(125 – 150)</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>92.9</td>
</tr>
<tr>
<td>(150 – 175)</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>92.9</td>
</tr>
<tr>
<td>(175 – 200)</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>92.9</td>
</tr>
<tr>
<td>(200 – 225)</td>
<td>1</td>
<td>7.1</td>
<td>14</td>
<td>100</td>
</tr>
</tbody>
</table>

b.
e. This distribution is skewed to the right, which means that most of the materials are concentrated in the area of saving up to 75 million BTU’s by using recycled materials and there are just a few materials (copper wire, carpet, and aluminum cans) that use inordinately large amounts of energy to create from raw materials.

f. 99.8% > The total should be all of the data, or 100%. The reason for the difference is rounding error.

g. The horizontal portion of the ogive is where there is no data present, so the amount of accumulated data does not change.

h. Because the ogive shows the increase in the percentage of data, the steepest section (in this case between 0 and 50%) is where most of the data is located and the accumulation of data is therefore changing at the most rapid pace.
1. a.

b. Divide each weight by 27.

Table 2.4:

<table>
<thead>
<tr>
<th>Material</th>
<th>Kilograms</th>
<th>Approximate Percentage of total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plastics</td>
<td>6.21</td>
<td>23</td>
</tr>
<tr>
<td>Lead</td>
<td>1.71</td>
<td>6.3</td>
</tr>
<tr>
<td>Aluminum</td>
<td>3.83</td>
<td>14.2</td>
</tr>
<tr>
<td>Iron</td>
<td>5.54</td>
<td>20.5</td>
</tr>
<tr>
<td>Copper</td>
<td>2.12</td>
<td>7.8</td>
</tr>
<tr>
<td>Tin</td>
<td>0.27</td>
<td>1</td>
</tr>
<tr>
<td>Zinc</td>
<td>0.60</td>
<td>2.2</td>
</tr>
<tr>
<td>Nickel</td>
<td>0.23</td>
<td>0.8</td>
</tr>
<tr>
<td>Barium</td>
<td>0.05</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Table 2.4: (continued)

<table>
<thead>
<tr>
<th>Material</th>
<th>Kilograms</th>
<th>Approximate percentage of total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other elements and chemicals</td>
<td>6.44</td>
<td>23.9</td>
</tr>
</tbody>
</table>

The data from the table has been adjusted a bit from the original source to simplify the problem and to account for rounding errors.

c.

![Materials in a Typical Desktop Computer](image)

2. a.

![Percentage of Municipal Waste Recycled by US States, 1998](image)

**Figure:** Percentage of Municipal Waste Recycled by US States, 1998

b. This data is fairly symmetric, is centered around 25 or 26 percent, and has a spread of 43 percentage points (the range).

```
0 | 5 5 7 8
1 | 1 1 2 2 3 4 5 8 8 9 9
2 | 0 0 1 3 3 3 5 5 5 6 6 6 6 7 8 8 8 8 9
3 | 0 1 2 3 3 4 5 6 9
4 | 0 0 1 2 2 5 8
```
3 means 13 percent

In this example, we chose the stems to represent every 10 percentage points, which compresses the data and forfeits some of the information about the shape of the distribution. Instead, we can split the stems in half (10 – 14, and 15 – 19). This plot is much more informative about the true shape of the data.

\[
\begin{array}{c|cccc}
0 & 5 & 5 & 7 & 8 \\
1 & 2 & 2 & 3 & 4 \\
1 & 5 & 8 & 8 & 9 \\
2 & 0 & 1 & 3 & 3 \\
2 & 5 & 5 & 6 & 6 \\
3 & 0 & 1 & 2 & 3 \\
3 & 5 & 6 & 9 \\
4 & 0 & 0 & 1 & 2 \\
4 & 5 & 8 \\
\end{array}
\]

d. There are 49 data points, so the 25\textsuperscript{th} value, counted from either end, is the median. In this case, that is 26.

3. a. This data set is mound-shaped and skewed left.

b. This data set has one obvious outlier. The remainder of the data is mound-shaped and fairly symmetric.

c. This data appears to be bimodal.

d. This data set appears mound-shaped and skewed right.

4. a. mound-shaped and symmetric

b. mound-shaped and symmetric

c. bimodal

d. uniform

5. All four distributions have almost the same center, which appears to be around 52.

6. a. The distribution is symmetric, so the center should be very close, if not equal to the mean. Standard deviation is a measure of the typical distance away from the mean. Most of the data points in this distribution are concentrated very close to the center.

7. c. The bimodal shape means that most of the points are located at the extreme values making their distance from the mean greater.

8. a. total amount of municipal waste: explanatory variable

percentage of total waste recycled: response variable
In this situation, we would most likely be interested in showing if states that have more or less total waste would have different recycling performance, i.e. does municipal waste explain a response in the recycling rate.

b. 13,386,000 tons. The data is indicated to be recorded in thousands of tons.

c.

d. There does appear to be at least one obvious outlier. California creates by far the most waste, most likely due to its large population. If we ignore California, there appears to be only a weak positive association between the amount of waste and the percentage recycled. If we removed two other points that appear atypical, New York and Florida, there is almost no association between the two. States with low average waste creation have recycling rates varying from the lowest, to the highest rates. If we remove the three potential outliers and rescale the axes, the data cloud is almost a circle, showing virtually no association.
9. a. HDPE plastic recycling showed a dramatic growth in 1996 and a slight growth the following year, but decline for all other years. PET bottle recycling has declined steadily through the entire time range.

b. The recycling rate declined by approximately 20% from 1995 to 2001.

c. One contributing cause has been the dramatic growth of immediate use personal size containers like bottled water that are typically consumed away from home and are not as likely to end up in curbside recycling programs.

d. The decline was the greatest during 1995 and 1996.

Box and Whisker Plots

1. a. To find the five number summary first put the data in numerical order. An easy way to do this is to draw a box and whisker plot of the data:
The median will be the data point that separates the data into two parts with an equal number of data points in each part. There are 32 data points so the median will fall between the 16th and 17th data points: 67 and 68. You take the average of these two points so the median is 67.5.

The lower quartile is the median of the first half of the data: in this case the lower quartile is between the 8th and 9th data points: 51 and 55. Take the average of these two points. The lower quartile is 53.

The upper quartile is between the data points 75 and 76. Take the average of these. The upper quartile is 75.5.

The maximum value is 95 and the minimum value is 35.

b) The IQR is $75.5 - 53 = 22.5$. A data point will be an outlier if it is less than $53 - 1.5 \times IQR$ or if it is greater than $75.5 + 1.5 \times IQR$.

\[
\begin{align*}
53 - 1.5 \times IQR &= 53 - 1.5 \times 22.5 \\
&= 53 - 33.75 \\
&= 19.25 \\
75.5 + 1.5 \times IQR &= 75.5 + 33.75 \\
&= 109.2
\end{align*}
\]

Since there are no data points less than 19.25 or greater than 109.25 there are no outliers.

c) c.

www.ck12.org 24
d. The distribution of Idaho reservoir capacities is roughly symmetric and is centered somewhere in the middle 60 percents. The capacities range from 35 up to 95 and the middle 50% of the data is between 53 and 75.5%.

e. The data between the median and the upper quartile is slightly more compressed which causes the median to be slightly larger than the median. The mean is approximately 65.7.

2.

a. To find the five number summary use technology. STAT, EDIT and put the data into L1. Then STAT, CALC, One-var stat, L1. The five number summary will be displayed: 
\{0, 72, 82, 89, 105\}

b. IQR = 17. 1.5 * IQR = 25.5

Upper bound for outliers > 25.5 + 89 = 114.5. There is no data above 105

Lower bound for outliers < 72 − 25.5 = 49.5. 63 is the last value that is above this point, so 0, 4, and 46 are all outliers 35.
c.

![Box plot of Reservoir Capacity in Utah](image)

**Figure:** Reservoir Capacity in Utah

d. The distribution of Utah reservoir capacities has three outliers. If those points were removed, the distribution is roughly symmetric and is centered somewhere in the low 80 percents. The capacities range from 0 up to 105 including the outliers and the middle 50% of the data is between 72 and 89%.

e. There are three extreme outliers, the mean is not resistant to the pull of outliers and there will be significantly lower than the median. The mean is approximately 75.4.

3.

![Box plot of Comparison of Idaho and Utah Reservoir Capacities](image)

**Figure:** Comparison of Idaho and Utah Reservoir Capacities

If we disregard the three outliers, the distribution of water capacities in Utah is higher than that of Idaho at every point in the five number summary. From this we might conclude it is centered higher and that the reservoir system is overall at safer levels in Utah, than in Idaho. Again eliminating the outliers, both distributions are roughly symmetric and their spreads are also fairly similar. The middle 50% of the data for Utah is just slightly more closely grouped.

e. If the mean is greater than the median, then it has been pulled to the right either by
an outlier, or by a skewed right shape. The median will not be affected by either of those things.

5. a. Find the five number summary using the technology as explained in the solution to problem 2. The five number summary is: \(\{3.12, 3.22, 3.282, 3.393, 3.528\}\)

b. IQR = 0.173

\[1.5 \times \text{IQR} = 0.2595\]

Upper bound for outliers > \(0.2595 + 3.393 = 3.6525\). There is no data above 3.528

Lower bound for outliers < \(3.22 - 0.2595 = 2.9605\). There is no data below 3.12, so there are no outliers.

c.

d. By dividing the data by 3.7854, we will obtain the average cost per liter. The mean, median will be decrease, being divided by 3.7854. The same is true for the measures of spread (range, IQR, and standard deviation), which will result in the data being compressed into a smaller area if we were to graph both distributions on the same scale. The shape of the distributions will remain the same.

e.

<table>
<thead>
<tr>
<th>State</th>
<th>Average Price of a Gallon of Gasoline (US $)</th>
<th>Average Price of a liter of Gasoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>3.458</td>
<td>0.914</td>
</tr>
<tr>
<td>Washington</td>
<td>3.528</td>
<td>0.932</td>
</tr>
<tr>
<td>Idaho</td>
<td>3.26</td>
<td>0.861</td>
</tr>
<tr>
<td>Montana</td>
<td>3.22</td>
<td>0.851</td>
</tr>
<tr>
<td>North Dakota</td>
<td>3.282</td>
<td>0.867</td>
</tr>
<tr>
<td>Minnesota</td>
<td>3.12</td>
<td>0.824</td>
</tr>
<tr>
<td>Michigan</td>
<td>3.352</td>
<td>0.886</td>
</tr>
</tbody>
</table>
Table 2.5: (continued)

<table>
<thead>
<tr>
<th>State</th>
<th>Average Price of a Gallon of Gasoline (US $)</th>
<th>Average Price of a liter of Gasoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>3.393</td>
<td>0.896</td>
</tr>
<tr>
<td>Vermont</td>
<td>3.252</td>
<td>0.859</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>3.152</td>
<td>0.833</td>
</tr>
<tr>
<td>Maine</td>
<td>3.309</td>
<td>0.874</td>
</tr>
</tbody>
</table>

Chapter Review

Part One Questions

1. c The mean is greater than the median because the graph is skewed right and the larger values draw the mean towards them.

2. \textit{(omitted because I don’t believe the question makes sense as it is stated)}

3. Her graph is a frequency polygon because an ogive is always increasing and in this graph there is a portion that is decreasing.

4. b

5. d

6. a

7. c

8. a

9. a
10. d

**Part Two – Open Ended Questions**

11. 11. a.

Table 2.6:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency(%)</th>
<th>Cumulative Frequency</th>
<th>Relative Cumulative Frequency(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1100 – 1150)</td>
<td>1</td>
<td>7.1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[1150 – 1200)</td>
<td>1</td>
<td>7.1</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>[1200 – 1250)</td>
<td>3</td>
<td>21.4</td>
<td>5</td>
<td>14.3</td>
</tr>
<tr>
<td>[1250 – 1300)</td>
<td>3</td>
<td>21.4</td>
<td>8</td>
<td>35.7</td>
</tr>
<tr>
<td>[1300 – 1350)</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>57.1</td>
</tr>
<tr>
<td>[1350 – 1400)</td>
<td>2</td>
<td>14.3</td>
<td>10</td>
<td>57.1</td>
</tr>
<tr>
<td>[1400 – 1450)</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>71.4</td>
</tr>
<tr>
<td>[1450 – 1500)</td>
<td>2</td>
<td>14.3</td>
<td>12</td>
<td>85.7</td>
</tr>
<tr>
<td>[1500 – 1550)</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>85.7</td>
</tr>
<tr>
<td>[1550 – 1600)</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>85.7</td>
</tr>
<tr>
<td>[1600 – 1650)</td>
<td>1</td>
<td>7.1</td>
<td>13</td>
<td>92.9</td>
</tr>
<tr>
<td>[1650 – 1700)</td>
<td>1</td>
<td>7.1</td>
<td>14</td>
<td>100</td>
</tr>
</tbody>
</table>
The median would correspond to the 50th percentile. Locate 50% on the vertical axis and trace across to the intersection width the ogive line. Follow that line down to the height axis and approximate the value.

Approximately 1280 ft
d. approximately 1625 ft

12. a. There isn’t necessarily a wrong way or right way to create this graph and to interpret the different time intervals, but a year should be the same distance apart for the entire graph so that the rate of change of the lines means the same thing across the entire plot. In this case, we plotted the average as a point in the middle of the five-year interval. It is possible that a student could devise a better representation, as long as the relationship in the data is clearly and correctly represented.
b. Answers will vary, but comments should focus on features of the plot that are placed in the context of the actual situation. For example, the plot of adult salmon increases dramatically after 1995 to a peak in 2002. This could be due to many factors, one of which was the inclusion of the Chinook salmon under the endangered species act. The plot for the Jack salmon stays relatively horizontal, indicating that the Jack population remained relatively constant until the most recent downturn. Other comments could be made and interested students might be encouraged to research things such as climate conditions or changes in the management of the salmon populations that may have led to the increases or decreases.

13. a. The various plots are shown below:
The only plot that does not seem to be a good fit is a stem-and-leaf plot. There is an extremely wide spread with the outlier, and creating meaningful stems would be difficult.

b. The plot is spread very widely, extending from a group of islands with almost no significant area, to the largest island, Isabela, which is so large at 4600 mi² that it is an extreme outlier. Even without the outlier, there is still a significant variation in the remaining islands. Ignoring Isabela, the distribution is still significantly skewed right. You can see this in all three graphs and it shows that most of the islands in the archipelago are smaller. The box plot does not appear to have a left whisker, but it is in fact, so small in relation to the scale of the graph, that it is indistinguishable. Here is a box-and-whisker plot without the outlier that has been rescaled.
The center would most appropriately be measured by the median because the extreme skewing and outliers will raise the mean substantially. The median island size is approximately 42 square kilometers.
3.1 Review Questions

1. a. The sample space is the following:
\[1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T, 3T, 4T, 5T, 6T\]
Each of these simple events has a probability of \(\frac{1}{12}\).

b) Find the probability of each of the following events: 
   A: 2 on die, H on coin. The probability of this event is \(\frac{1}{12}\). It can happen only one way out of the 12 possible outcomes for the experiment.
   B: even number on the die and a T on the coin. This can happen in 3 ways \(2T, 4T, 6T\) out of the 12 possible outcomes. So the probability is \(\frac{3}{12}\) or \(\frac{1}{4}\).
   C: Even number on the die: This can happen 6 ways out of the possible 12 outcomes: \(2H, 4H, 6H, 2T, 4T, 6T\). The probability is \(\frac{6}{12}\) or \(\frac{1}{2}\).
   D: a T on the coin. This can happen 6 ways so the probability is \(\frac{6}{12}\).

2. The Venn diagram below shows an experiment with six simple events. Events A and B are also shown. The probabilities of the simple events are:

\[
P(1) = P(2) = P(4) = \frac{2}{9}\\
P(3) = P(5) = P(6) = \frac{1}{9}
\]
a. \( P(A) = P(2) + P(3) + P(5) = \frac{2}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9} \)

b. \( P(B) = P(4) + P(6) = \frac{2}{9} + \frac{1}{9} = \frac{3}{9} = \frac{1}{3} \)

2. a. \{B_1B_2, B_1R_1, B_1R_2, B_1R_3, B_2B_1, B_2R_2, B_2R_3, R_1B_1, R_1B_2, R_1R_2, R_1R_3, \}
\{R_2B_1, R_2B_2, R_2R_1, R_2R_3, R_3B_1, R_3B_2, R_3R_1, R_3R_2\}

b. \( P(A) = P(2 \text{ blue marbles are drawn}) = \frac{2}{20} = \frac{1}{10} \) because you can get \( B_1B_2 \) or \( B_2B_1 \).
\( P(B) = P(1 \text{ red and } 1 \text{ blue}) = \frac{12}{20} = \frac{3}{5} \) You can see this by looking at the listing of the sample space in part a and counting the number of times you get one red and one blue. \( P(C) = P(2 \text{ red marbles are drawn}) = \frac{6}{20} = \frac{3}{10} \). You can get \( \{R_1R_2, R_1R_3, R_2R_1, R_2R_3, R_3R_1\} \).

Complement of an Event Review Questions

1. a. The sample space for tossing a coin three times: HHH, HHT, HTT, TTT, HTH, THT, THH, TTH

b. \( A = \) at least one head is observed. The outcomes for \( A \) are \{HHH, HHT, HTT, HTH, THT, THH, TTH\}

c. \( B = \) the number of heads observed is odd. The outcomes for \( B \) are \{HHH, HTT, THT, TTH\}

d. The outcomes for :

\[ A \cup B = \{HHH, HHT, HTT, HTH, THT, THH, TTH\} \]
\[ A' = TTT \]
\[ A \cap B = \{HHH, HTT, THT, TTH\} \]
e.

\[ P(A) = \frac{7}{8} \]
\[ P(B) = \frac{4}{8} = \frac{1}{2} \]

\[ P(A \cup B) = \frac{7}{8} \]
\[ P(A') = \frac{1}{8} \]
\[ P(A \cap B) = \frac{4}{8} = \frac{1}{2} \]

2. The Venn diagram below shows an experiment with five simple events. The two events A and B are shown.

The probabilities of the simple events are:

\[ P(1) = \frac{1}{10}, P(2) = \frac{2}{10}, P(3) = \frac{3}{10}, P(4) = \frac{1}{10}, P(5) = \frac{3}{10}. \]

Find \( P(A'), P(B'), P(A' \cap B), P(A \cap B), P(A \cup B'), P(A \cup B), P[(A \cap B)'] \) and \( P[(A \cup B)'] \)
\[
P(A') = P(5) + P(4) = \frac{4}{10}
\]
\[
P(B') = P(1) + P(4) = \frac{2}{10}
\]
\[
P(A' \cap B) = P(5) = \frac{3}{10} \text{ (everything in } B \text{ but not in } A)
\]
\[
P(A \cap B) = P(2) + P(3) = \frac{5}{10} = \frac{1}{2}
\]
\[
P(A \cup B') = P(1) + P(2) + P(3) + P(4) = \frac{7}{10}
\]
\[
P(A \cup B) = P(1) + P(2) + P(3) + P(5) = \frac{9}{10}
\]
\[
P((A \cap B)') = P(1) + P(5) + P(4) = \frac{5}{10}
\]
\[
P((A \cup B)') = P(4) = \frac{1}{10}
\]

**Conditional Probability Review Questions**

1. 

\[
P(A \mid B) = \frac{(A \cap B)}{P(B)}
\]

\[
= \frac{.15}{.7} = .214
\]

\[
P(B \mid A) = \frac{P(B \cap A)}{P(A)}
\]

\[
= \frac{.15}{.3} = .5
\]

2. i. the outcomes when two fair coins are tossed: \{HH, HT, TH, TT\}

ii.

www.ck12.org
\[ P(A) = P \{ HH, HT, TH \} = \frac{3}{4} \]
\[ P(B) = P \{ HT, TH \} = \frac{2}{4} \]
\[ P(A \cap B) = P \{ HT, TH \} = P(B) = \frac{2}{4} \]
\[ P(A \mid B) = \frac{(A \cap B)}{P(B)} \]
\[ = \frac{.5}{.5} \]
\[ = 1 \]
\[ P(B \mid A) = \frac{(B \cap A)}{P(A)} \]
\[ = \frac{.5}{.75} \]
\[ = \frac{2}{3} \]

3. i. \{ww, wr, wb, rw, rr, rb, bw, br, bb\}
ii. A: both marbles have same color
B: both marbles are red
C: At least one marble is red or white

\[ P(B \mid A) = \frac{(B \cap A)}{P(A)} \]
\[ = \frac{P(B)}{P(A)} \]
\[ = \frac{1}{15} \]
\[ = \frac{1}{5} \]
\[ = \frac{1}{3} \]
\[ P(B \mid A') = \frac{(B \cap A')}{P(A')} \]

\[ = \frac{0}{\frac{15}{44}} \]

\[ = 0 \]

\[ P(B \mid C) = \frac{(B \cap C)}{P(C)} \]

\[ = \frac{P(B)}{P(C)} \]

\[ = \frac{\frac{15}{14}}{\frac{15}{14}} \]

\[ = \frac{1}{14} \]

\[ P(A \mid C) = \frac{(A \cap C)}{P(C)} \]

\[ = \frac{P(rr, ww)}{P(C)} \]

\[ = \frac{\frac{1}{15} + \frac{1}{15}}{\frac{15}{15}} \]

\[ = \frac{2}{15} \]

\[ = \frac{1}{7} \]

\[ P(C \mid A') = \frac{(A' \cap C)}{P(A')} \]

\[ = \frac{P(wr, wb, rw, rb, bw, br)}{P(A')} \]

\[ = \frac{6 \times \frac{2}{6} \times \frac{2}{5}}{\frac{4}{5}} \]

\[ = 1 \]
Additive and Multiplicative Rules Review Questions

1. a. Events $A$ and $B$ are not independent. Consider the following table of outcomes when two fair dice are tossed and the variable of interest is the sum (shown in bold).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

$P(A) = P(\text{sum is odd}) = \frac{18}{36}$

$P(B) = P(\text{sum is 9, 11 or 12}) = \frac{7}{36}$

$$P(A \mid B) = \frac{(A \cap B)}{P(B)}$$

$$= \frac{P(9, 11)}{P(B)}$$

$$= \frac{6}{36} \div \frac{7}{36}$$

$$= \frac{6}{7}$$

Therefore, $A$ and $B$ are not independent.

c. The events $A$ and $B$ are not mutually exclusive because $A \cap B \neq \emptyset$

2. The probability that a new television of the same brand will last 1 year equal the probability that the tv doesn’t fail when it is first used (.9) and the tv will work properly for 1 year(.99). The probability that a new television of the same brand will last 1 year is $9 \times .99 = .891$.

Basic Counting Rules Review Problems

1. a. 4
b. 8

c. 32

d. \(2^n\)

2. You have \(3 \times 2 = 6\) travel options. You can choose airline 1, first class or airline 1, economy or airline 2 first class or airline 2 economy or airline 3 first class or airline 3 economy.

3. This would be \(52\) choose \(5\) which is

\[
\begin{align*}
\frac{52!}{5! \times (52 - 5)!} &= \frac{52 \times 51 \times 50 \times 49 \times 48 \times 47!}{5 \times 4 \times 3 \times 2 \times 1 \times 47!} \\
&= 2,598,960
\end{align*}
\]

4. You would have \(26 \times 10 \times 10 \times 10 \times 10 \times \times\) different license plates. This is equal to \(2,600,000\) license plates.
Chapter 4

Discrete Probability Distribution - Solution Key

4.1 Probability Distribution for a Discrete Random Variable Review questions

1. a. \([-4,0,1,3]\)
b. The value of \(x\) most likely to occur is 1 because it has the highest probability associated with it.
c. \(P(x > 0) = .4 + .2 = .6\) d. \(P(x = -2) = 0\)

2.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

b. \(P(x \geq 8) = \frac{15}{36} = \frac{5}{12}\)
c.
\[ P(X < 8) = 1 - P(X \geq 8) \]
\[ = 1 - \frac{15}{36} \]
\[ = \frac{21}{36} \]

d.
\[ P(x \text{ is odd}) = \frac{18}{36} \]
\[ P(x \text{ is even}) = \frac{18}{36} \]
e. \( P(x = 7) = \frac{6}{36} \)

2. Suppose a couple has three children. Using a tree diagram you can find that the possible outcomes are \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}. The probability of having at least one boy is the probability of having one, two or three boys. This is equal to \(1-\) probability of having all girls.

\[ P(GGG) = \frac{1}{8} \]
\[ 1 - P(GGG) = \frac{7}{8} \]

Mean and Standard Deviation of Discrete Random Variable

Review problems

1. a.
\[ \mu_x = \sum xP(x) \]
\[ = 0(.1) + 1(.4) + 2(.3) + 3(.1) + 4(.1) \]
\[ = 1.7 \]

b.
\[ \sigma_x^2 = \sum (x_i - \mu_x)^2 p(x) \]
\[ = (0 - 1.7)^2(.1) + (1 - 1.7)^2(.4) + (2 - 1.7)^2(.3) + (3 - 1.7)^2(.1) + (4 - 1.7)^2(.1) \]
\[ = 1.21 \]
c. 

\[ \sigma = \sqrt{\sigma_x^2} = \sqrt{1.21} = 1.1 \]

2. The expected value of the number of times of previous convictions of an inmate is the mean of the distribution.

\[ E(x) = \mu_x = 0(.16) + 1(.53) + 2(.2) + 3(.08) + 4(.03) = 1.29 \]

The Binomial Probability Distribution Review Problems

1. 

\[ X \times B(4, .2) \]

\[ P(X = x) = \binom{4}{x} (0.2)^x (0.8)^{4-x} \]

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>.4096</td>
<td>.4096</td>
<td>.1536</td>
<td>.0256</td>
<td>.0016</td>
</tr>
</tbody>
</table>

2. a. 

\[ X \times B(5, 0.5) \]

\[ P(X = x) = \binom{5}{x} (0.5)^x (0.5)^{5-x} \]

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>.03125</td>
<td>.15625</td>
<td>.3125</td>
<td>.3125</td>
<td>.15625</td>
<td>.03125</td>
</tr>
</tbody>
</table>

b. 

\[ \mu_x = np = 5(.5) = 2.5 \]

\[ \sigma_x^2 = np(1 - p) = 5(.5)(.5) = 1.25 \]
3. The probability of developing antibodies against the hormone if you are a diabetic patient is .1. This can be modeled with a binomial distribution:

a.

\[ X \times B(5, .1) \]
\[ P(X = 0) = 5C_0(.1)^0(.9)^5 \]
\[ = .590 \]

b.

\[ P(\text{at least one will develop antibodies}) = 1 - P(X = 0) \]
\[ = 1 - .590 \]
\[ = .40951 \]

The Geometric Probability and Poisson Distribution Review problems

1. \( X \times \text{Poisson} \lambda = 1.5 \)

a.

\[ P(X = 3) = \frac{(1.5)^3e^{-15}}{3!} \]
\[ = \frac{3.357(.223)}{6} \]
\[ = .1255 \]

b.

\[ P(X = 1) = \frac{(1.5)e^{-15}}{1} \]
\[ = .3345 \]

2. \( X \times \text{Poisson} \lambda = 2.5 \)

\[ P(X = x) = \frac{2.5^x e^{-25}}{x!} \]

The hospital will not have enough beds of the number of patients is greater than 4. For the Poisson distribution:
\[
\begin{array}{cccccc}
X & 0 & 1 & 2 & 3 & 4 \\
P(X = x) & .082 & .205 & .2565 & .2138 & .1336 \\
\end{array}
\]

\[
P(x > 4) = 1 - P(X \leq 4) \\
= 1 - (.82 + .205 + .2565 + .2138 + .1336) \\
= 1 - .89117 \\
= .1088
\]

3.

\[
X \times \text{Geom}(.2) \\
P(X = 5) = (.8)^4(.2) \\
= .08192
\]

For the geometric probability distribution if the first success is on the fifth one then the first four had to be failures. The probability of failure is .8 since the probability of success is .2.
Chapter 5

Normal Distribution - Solution Key

5.1 Review Questions

1. a. You would expect this situation to vary normally with most students’ hand spans centering around a particular value and a few students having much larger or much smaller hand spans.

b. Most employees could be hourly laborers and drivers and their salaries might be normally distributed, but the few management and corporate salaries would most likely be much higher, giving a skewed right distribution.

c. Many studies have been published detailing the shrinking, but still prevalent income gap between male and female workers. This distribution would most likely be bi-modal, with each gender distribution by itself possibly being normal.

d. You might expect most of the pennies to be this year or last year, fewer still in the previous few years, and the occasional penny that is even older. The distribution would most likely be skewed left.

2. 

a. \[ z = \frac{(65 - 81)}{6.3} \]

\[ z \approx -2.54 \]

b. \[ z = \frac{(83 - 81)}{6.3} \]
\[ z \approx 0.32 \]

c. \[ z = \frac{(93-81)}{6.3} \]

\[ z = +1.90 \]

d. \[ z = \frac{(100-81)}{6.3} \]

\[ z \approx 3.02 \]

3. Because the data is normally distributed, students should use the 68 – 95 – 99.7 rule to answer these questions.

a. about 16% (less than one standard deviation below the mean)

b. about 95% (within 2 standard deviations)
c. about 0.15% (more than 3 standard deviations above the mean)

4. Using the empirical rule (68 − 95 − 99.7) with

\[ X \sim N(0, 1) \]

\[ P(X < 1) = .34 + .5 = .84 \]

\[ P(X < -1) = .5 - .34 = .16 \]

\[ \mu = 0 \]

\[ \sigma = 1 \]

\[ P(X > 2) \approx 0.25 \]

So the proper order from smallest to largest is: \( \mu \), percentage of data above 2, percentage of data less than −1, percentage of data less than 1, \( \sigma \)

5. The correct response is c

\[ x = 5 \]

\[ z = \frac{5 - 2.80}{1.341} \]

\[ = 1.64 \]

6. 0.025.157.7 is exactly 2 standard deviations above the mean height. According to the empirical rule, the probability of a randomly chosen value being within 2 standard deviations is about 0.95, which leaves 0.05 in the tails. We are interested in the upper tail only as we are looking for the probability of being above this value.

7.

a. Here are the possible plots showing a symmetric, mound shaped distribution.
b. $\mu = 262.222 \quad \sigma = 67.837$
<table>
<thead>
<tr>
<th>Number of Sprinkles</th>
<th>deviations</th>
<th>Z – scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>241</td>
<td>−21.2222</td>
<td>−0.313</td>
</tr>
<tr>
<td>282</td>
<td>19.7778</td>
<td>0.292</td>
</tr>
<tr>
<td>258</td>
<td>−4.22222</td>
<td>−0.062</td>
</tr>
<tr>
<td>223</td>
<td>−38.2222</td>
<td>−0.563</td>
</tr>
<tr>
<td>133</td>
<td>−129.222</td>
<td>−1.905</td>
</tr>
<tr>
<td>335</td>
<td>72.7778</td>
<td>1.073</td>
</tr>
<tr>
<td>322</td>
<td>59.7778</td>
<td>0.881</td>
</tr>
<tr>
<td>323</td>
<td>60.7778</td>
<td>0.896</td>
</tr>
<tr>
<td>354</td>
<td>91.7778</td>
<td>1.353</td>
</tr>
<tr>
<td>194</td>
<td>−68.2222</td>
<td>−1.006</td>
</tr>
<tr>
<td>332</td>
<td>69.7778</td>
<td>1.029</td>
</tr>
<tr>
<td>274</td>
<td>11.7778</td>
<td>0.174</td>
</tr>
<tr>
<td>233</td>
<td>−29.2222</td>
<td>−0.431</td>
</tr>
<tr>
<td>147</td>
<td>−115.222</td>
<td>−1.699</td>
</tr>
<tr>
<td>213</td>
<td>−49.2222</td>
<td>−0.726</td>
</tr>
<tr>
<td>262</td>
<td>−0.222222</td>
<td>−0.003</td>
</tr>
<tr>
<td>227</td>
<td>−35.2222</td>
<td>−0.519</td>
</tr>
<tr>
<td>366</td>
<td>103.778</td>
<td>1.530</td>
</tr>
</tbody>
</table>

c. The normal probability plot shows a fairly linear pattern which is an indication that there are no obvious departures from normality in this distribution.

d. The normal probability plot shows a fairly linear pattern which is an indication that there are no obvious departures from normality in this distribution.

**Density Curve of Normal Distribution Review Problems**

1. Here is the distribution with a density curve drawn and the inflection points estimated.
The distribution appears to be normal. The inflection points appear to be a little more than two units from the mean of 84, therefore we would estimate the standard deviation to be a little more than two. Using the frequencies, the middle three bins contain 42 of the 50 values. Approximately 34 of the values should be within one standard deviation, which is consistent with our estimate.

2.

a. \( P(Z \geq -0.79) = 1 - P(Z < -0.79) = 1 - 0.2148 = .7852 \)

b. \( P(Z \leq -1) = 0.1587, P(Z \leq 1) = 0.8413, 0.8413 - 0.1587 = 0.6826 \)

c. \( P(Z \leq -1.56) = 0.0594, P(Z \leq 0.32) = 0.6255, 0.6255 - 0.0594 = 0.5661 \)

3. a. Brielle did not enter the mean and standard deviation. The calculator defaults to the standard normal curve, so the calculation she performed is actually explaining the percentage of data between the \( z \)-scores of 80 and 90. There is virtually 0 probability of experiencing data that is over 80 standard deviations away from the mean, especially given a test grade presumably out of 100.

b. 0.525 or about 53%

4.

\[
\begin{align*}
x & \times N(72, 6.5) \\
X & = 78 \\
z & = \frac{78 - 72}{6.5} = .92 \\
X & \times N(77, 8.4) \\
X & = 83 \\
z & = \frac{83 - 77.8}{8.4} = .071
\end{align*}
\]

In both cases the scores are above the mean. The \( z \) score represents the number of standard deviations above the mean each test score is. The test score of 78 is more standard deviations above the mean so it is the better grade.
5.

\[ A \times N(72, 10) \]
\[ a = 90 \]
\[ z = \frac{90 - 72}{10} = 1.8 \]

\[ B \times N(82, 5) \]
\[ b = 90 \]
\[ z = \frac{90 - 82}{5} = 1.6 \]

A 90 is a better score with teacher A.

The \( p \)-values can be found using the technology:

\( 2^{\text{nd}} \) VARS DISTR normalcdf\((-1000000, 90, 72, 10) = .964 \)

\( 2^{\text{nd}} \) VARS DISTR normalcdf\((-1000000, 90, 82, 5) = .945 \)
b. A 60 is by far a much lower score with teacher B.

2\text{nd} \text{ VARS DISTR normalCDF}(-1000000, 60, 72, 10) = .115
2\text{nd} \text{ VARS DISTR normalCDF}(-1000000, 60, 82, 5) = 5.4 \times 10^{-6}

\textbf{Applications of the Normal Distribution Review Problems}

1. If we are looking for the interval containing 95\% of the data we are looking for the interval which is within two standard deviations of the mean (by the empirical rule). This is $-1.96 \leq z \leq 1.96$

$X \sim N(85, 4.5)$

$P(X < x) = .68$

Use technology 2\text{nd} \text{ VARS invNorm}(.68, 85, 4.5) = 87.1

a. 87.1

b.
\[ P(X < 16) = .05 \]
\[ P\left( \frac{X - \mu}{1} < \frac{16 - \mu}{1} \right) = .05 \]
\[ P\left( z < \frac{16 - \mu}{1} = .05 \right) \]

2nd Vars InvNorm(.05) = -1.645

\[ 16 - \mu = -1.645 \]
\[ \mu = 17.645 \]

c. 
\[ P \left( z < \frac{85 - 13}{\sigma} \right) = .91 \]

InvNorm(.91) = 1.34
\[ \frac{85 - 73}{\sigma} = 1.34 \]
\[ \frac{12}{1.34} = \sigma \]
\[ 8.96 = \sigma \]
d.

\[ P \left( z < \frac{x - 93}{5} \right) = .90 \]

InvNorm(.90, 93, 5) = 1.28

\[ \frac{x - 93}{5} = 1.28 \]

\[ x - 93 = 1.28 \times 5 \]

\[ x - 93 = 6.4077 \]

\[ x = 99.41 \]

3. The \( z \)-score for the lower quartile in a standard normal distribution is that score that has 25% of the data below it. You can use technology to find this:

\( \text{2}^{\text{nd}} \text{ VARS InvNorm(.25, 0, 1)} = -.6745 \)

4. The manufacturer wants to be .999 sure that the mean diameters are between 1.397 and 1.4035. He/she wants only .001 area in the tails. This means there will be .9995 area to the left of 1.4035.

\[ P(X < 1.4035) = .9995 \]

\[ P \left( Z < \frac{1.4035 - 1.4}{\sigma} \right) = .9995. \]

Use technology : \( \text{2}^{\text{nd}} \text{ VARS InvNorm(.9995)} = 3.29 \)
\[
\frac{1.4035 - 1.4}{\sigma} = 3.29 \\
\frac{.0035}{\sigma} = 3.29 \\
\sigma = \frac{.0035}{3.29} \\
\sigma = .00106
\]

5. a.

\[X \times N(2.2, .04)\]

\[a) P(X < 2.13)\]

\[= P\left(\frac{X - 2.2}{.04} < \frac{2.13 - 2.2}{.04}\right)\]

\[= P\left(Z < -\frac{.07}{.04}\right)\]

\[(Z < -1.75)\]

Use technology 2\(^{nd}\) VARS normalcdf\((-1000000, -1.75)\) to find this probability: .04

b. To find the proportion of candy bars between 2.2 and 2.3 ounces use technology:

2\(^{nd}\) VARS normalcdf\((2.2, 2.3, 2.2, .04)\) to find the probability: .494

c. To find the weight that would be heavier than all but 1% of the candy bars use technology:

2\(^{nd}\) VARS InvNorm\((.99, 2.2, .04)\) to find the probability 2.29 oz.

d. The manufacturer wants the probability of the a bar weighing less that 2.13 to be .001.

\[P(X < 2.13) = .001\]

\[P\left(Z < \frac{2.13 - \mu}{.04}\right) = .001\]

Use technology to find the z score associated with this probability:

2\(^{nd}\) VARS InvNorm (.001) = −3.09
\[
\begin{align*}
\frac{2.13 - \mu}{.04} &= -3.09 \\
2.13 - \mu &= -3.09 \times .04 \\
2.13 - (-3.09 \times .04) &= \mu \\
\mu &= 2.254
\end{align*}
\]
e. The manufacturer now wants the probability of a candy bar weighing less than the advertised weight of 2.13 ounces is .001. The mean remains at 2.2 ounces.

\[
P(X < 2.13) = .001 \\
P \left( Z < \frac{2.13 - 2.2}{\sigma} \right) = .001 \\
\frac{2.13 - 2.2}{\sigma} = -3.09 \\
\frac{2.13 - 2.2}{-3.09} = \sigma \\
.023 = \sigma
\]
Chapter 6

Planning and conducting an Experiment or Study - Solution Key

6.1 Surveys and Sampling Review Questions

1. a. All high school soccer players.
b. Each individual high school soccer players.
c. A census.
d. Boys, students from other areas of the country of different socio-economic or cultural backgrounds, if she is on a varsity team, perhaps JV or freshman soccer players might have different preferences.
e. There are multiple answers, which is why the explanation is very important. The two most obvious sources are:
   Convenience bias, she asked the group that was most easily accessible to her, her own teammates.
   Incorrect Sampling frame, boys or some of the other undercovered groups mentioned in d, have no chance of being included in the sample.
f. The sampling frame.
g. Stratification.

2. This is incorrect response bias. You are intentionally answering the question incorrectly so as to not antagonize the giant talking snake!

3. The biologist is using her knowledge of moose behavior to choose an area and a time in which to estimate the population, this is judgment sampling. She has also selected one
particular lake to estimate the entire region, which could be considered a form of cluster sampling.

4. Systematic sampling. The customer is selected based on a fixed interval.

5. Convenience bias. The first 30 riders is an easy group to access. Incorrect Sampling Frame. The first riders of the day are likely to be those who are most excited by high-thrill rides and may not have the same opinions as those who are less enthusiastic about riding.

6. Cluster sampling. A group is chosen because of a natural relationship that does not necessarily have any similarity of response, i.e. we have no reason to believe that people wearing a certain color would respond similarly, or differently, from anyone else in the population.

7. Voluntary response bias. Participants will self-select. Non-response bias. A large percentage of potential participants are not going to want to be bothered participating in a survey at the end of a long day at an amusement park.

8. There are several potential answers. Incorrect Response Bias. The chosen participants might not want to admit to being scared in front of the young lady. Questionnaire bias. The question is definitely worded in a manner that would encourage participants to answer in a particular way. This is also systematic sampling and someone used their judgment that only boys should be surveyed. A case could also be made for incorrect sampling frame as no girls or other age groups have a chance of being represented. All of these examples also eliminate the opinions of those in the park who do not choose to ride.

9. Stratification. It could be that people who ride at different times during the day have different opinions about thrill rides or are from different age groups. In this case, each hour is a stratum. For example, it could be that those riding early in the morning are more of the thrill seeker types, and the more hesitant folks might take some time to muster the courage to ride.

10. To make it easier to keep track of repeated choices, we have generated 100 numbers and stored them in L1.

The chosen students are:

16, 20, 9, 31, 30, 29, 8, 10, 23, 33

In this example there were no repeated digits.

**Experimental Design Review Questions**

1.

a. The population is all fruit flies of this species. The treatment is breeding for intelligence.
The other treatment is really a control group. The second group of flies were not bred for any special quality.

b. By the strict definition, this is an observational study as the subjects (fruit flies) are not randomly assigned to the treatment. A group of fruit flies was selectively bred for intelligence.

c. Because the treatments were not randomly assigned the results are susceptible to lurking variables. It is possible that some other trait not observed in the population of intelligent fruit flies led to their lower survival rate. It is also questionable to generalize the behavior of fruit flies to the larger population of all animals. We have no guarantee that other animals will not behave differently than fruit flies. Without reading the study completely, it is difficult to determine how many of these concerns were addressed by the scientists performing the study. You can read more at:

http://www.nytimes.com/2008/05/06/science/06dumb.html?ref=science

2. a. This is a repeated measures design. Each student becomes their own matched pair as they are sampling both colas.

b. Students may have a preconceived idea of which cola they prefer for many possible reasons. You could have the colas already poured into identical unmarked cups, or hide the label of the bottle. This would be an example of a blind experiment.

c. It is possible that the taste of the first cola might affect the taste of the second. In general, the order in which they taste the colas could affect their perception in a number of ways. To control for this, we could randomly assign the order in which the colas are sampled. Assign one of the colas to be 1 and the other to be 2, then use your calculator to choose 1 or 2 randomly for each subject. If the student is given the two cups and given the option of choosing which one to drink first, we could randomly assign the position of each cup (right or left).

3. a. Because students with lower pulses may react differently than students with higher pulses, we will block by pulse rate. Place the students in order from lowest to highest pulse rate, then take them two at a time.

Table 6.1:

<table>
<thead>
<tr>
<th>Pair Number</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6, 1</td>
</tr>
<tr>
<td>2</td>
<td>11, 7</td>
</tr>
<tr>
<td>3</td>
<td>9, 10</td>
</tr>
<tr>
<td>4</td>
<td>3, 4</td>
</tr>
<tr>
<td>5</td>
<td>2, 8</td>
</tr>
<tr>
<td>6</td>
<td>12, 5</td>
</tr>
</tbody>
</table>
b. The calculator would generate the following 6 random ones and twos.

Using the order in which the students appear in the table as their number, the students could be assigned by placing the chosen student for each pair into treatment 1, and the remaining student to treatment 2:

<table>
<thead>
<tr>
<th>Treatment 1 (black ink)</th>
<th>Treatment 2 (red ink)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6, 11, 9, 3, 8, 5</td>
<td>1, 7, 10, 4, 2, 12</td>
</tr>
</tbody>
</table>

c. It is possible that different students react to testing taking and other situations differently and it may not affect their pulse directly. Some students might be better test takers than others. The level of mathematics ability or previous success on the subject matter being tested could also affect the stress level. Perhaps amount of sleep, diet, and amount of exercise may also be lurking variables.

d. A repeated measures design would help control for individual differences in pulse rate. Each student would have to take both a black ink and red ink test. A second test would have to be carefully designed that was similar to the first, but with different color ink. If you just gave the students the same test twice, their stress level might be significantly lower when they take it the second time.

4. a. The population is children with epilepsy who have not responded to other traditional medications.

b. We need assurances that the children were randomly assigned to the treatment and control groups.

c. The treatment is starting on the high fat diet immediately, the control group is the group who started the diet 3 months later. Notice in this case, researchers did not completely withhold the treatment from the control group for ethical reasons. This treatment has already shown some effectiveness in non-clinical trials.

d. We would conclude that the high fat diet is effective in treating seizures among children with epilepsy who do not respond to traditional medication.

5. We will need at least 6 blocks to impose the various treatments, which are:

Organic fertilizer, 1 inch
Chemical fertilizer, 1 inch
Organic fertilizer, 2 inches
Chemical fertilizer, 2 inches
Organic fertilizer, 4 inches
Chemical fertilizer, 4 inches
Assign the plots numbers from 1 to 6.

```
  1   2   3
  4   5   6
```

Then randomly generate a number from 1 to 6, without replacement, until all six treatments are assigned to a plot.

In this example, the random number generator was seeded with 625, repeated digits were ignored, and the assignments were as follows:

Organic fertilizer, 1 inch PLOT 6
Chemical fertilizer, 1 inch PLOT 2
Organic fertilizer, 2 inches PLOT 1
Chemical fertilizer, 2 inches PLOT 5
Organic fertilizer, 4 inches PLOT 4
Chemical fertilizer, 4 inches PLOT 3

### Chapter Review Questions

#### Multiple Choice:
1. c
2. b
4. e
5. c

#### Open-Ended Questions
1. Incorrect response bias. The main focus of the piece, and an issue in exit polling in general is that there is no guarantee that, for many possible reasons, subjects in an exit poll will answer truthfully. The pollsters also ask the questions in a variety of rude, unethical, and inappropriate ways that would manipulate the responses. Even though a real pollster would never actually engage in this type of behavior, it could be considered questionnaire bias.

2. a. Randomly assign each tortoise a number from 1 – 9 using a random number generator, then incubate the eggs from tortoises 1 – 3 at 25.50\(^\circ\)C, 4 – 6 at 29.50\(^\circ\)C, and 7 – 9 at 33.50\(^\circ\)C.
When the tortoises hatch, observe and compare the ratio of female and male tortoises (which is not easy to do) at the various temperatures. The results of this study did confirm that the ratio of females is higher found that 29.50°C is the optimum temperature for a higher female ratio and good survival rate, and 280°C is the best temperature to insure more males (source: Restoring the Tortoise Dynasty, Godfrey Merlin, Charles Darwin Foundation, 1999.)

b. This would be a blocking design. We would block on species and temperature, so there would be 9 blocks, 3 of each species, and three at each incubation temperature. There really would not be any randomization in this design.

3. a. There are two treatments, the new medication, and the existing medication. All the subjects could be told that they were receiving a new treatment, and then only some would be given the new treatment and the rest would be given their original medication. The resulting differences in skin condition between the two groups would be compared.

b. You could assign the subjects a different numbering from 1 to 12, but this time generating the assignments at random. Then subjects 1 – 6 would be given the new treatment, and subjects 7 – 12 would be given the original medication. Compare the results.

c. In blocking for condition, each block should be homogeneous for that trait. So, you would create three blocks: all 4 mild subjects, all 4 moderate subjects, and all 4 severe subjects. Then, within each block, randomly assign two subjects to receive the new treatment, and two to receive the original. Compare the results.

d.

Table 6.3:

<table>
<thead>
<tr>
<th>Pair Number</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>F</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>18</td>
</tr>
</tbody>
</table>

Place the females in chronological order, then group the two youngest, the next two, and the last two. Repeat the same procedure with the males. This way we have pairs that are similar in both age and gender. One of the subjects would be chosen at random for the new
treatment and the other would receive the traditional medication.

e. Repeated measures are not a good idea with medication studies as it would be hard to distinguish if the effects from the repeated treatment are not in fact from still occurring from the presence of the first medication that was given.
Chapter 7

Sampling Distributions - Solution Key

7.1 Sampling Distributions

1. Many students may guess normal, but in reality the distribution is likely to be skewed toward the older pennies. (Remember that this means there are more newer pennies.)
2. The histogram will probably show the distribution skewed toward the older ages.
3. Answers will vary
4. The mean of the sampling distribution should be the same as the mean of the population.
6. The shape of the sampling distribution becomes approximately normal as the sample size increases.

Note: This activity would work very well done with an entire class. Each student could use 20 coins and the sample means could be an accumulation of sample means from each student.

Z Score and Central Limit Theorem Review Questions

1.
   a.

\[
\begin{align*}
z &= \frac{x - \mu}{\sigma} \\
z &= \frac{360 - 400}{50} \\
z &= -0.8 \\
z &= \frac{460 - 400}{50} \\
z &= 1.2
\end{align*}
\]

and

\[
\begin{align*}
z &= \frac{x - \mu}{\sigma} \\
z &= \frac{460 - 400}{50} \\
z &= 1.2
\end{align*}
\]
Using the graphing calculator the area for \( z = -0.8 \) is 0.2881 and for \( z = 1.2 \) is 0.3849

\[
\text{Area is: } 0.2881 + 0.3849 = 0.6730
\]

\[
(.6730)(6000) = 4038
\]

This means that 67.3\% of the 6000 batteries lasted between 360 and 460 days.

Note: For \( z = -0.8 \), the area is to the left of the mean. However, the curve is symmetrical about the mean and the value of the area for \( z = 0.8 \) is used and added to the area of \( z = 1.2 \).

a. 4038 batteries will last between 360 and 460 days.

b. 
\[
z = \frac{x - \mu}{\sigma}
\]
\[
z = \frac{320 - 400}{50}
\]
\[
z = -1.6
\]

The area for \( z = 1.6 \) is 0.4452.

\[
0.4452 + 0.5000 = 0.9452
\]

\[
(.9452)(6000) = 5671
\]

This means that 94.52\% of the 6000 batteries lasted more than 320 days.

Note: For \( z = -1.6 \), the total area to the right of the mean is needed. Since the total area under the curve is one, the total area on either side of the mean is 0.5000. This area must be added to the area 0.4452

b. 5671 batteries will last more than 320 days

c. 
\[
z = \frac{x - \mu}{\sigma}
\]
\[
z = \frac{280 - 400}{50}
\]
\[
z = -2.4
\]

The area for \( z = 2.4 \) is 0.4918.

\[
0.5000 - 0.4918 = 0.0082
\]

\[
(.0082)(6000) = 49
\]
This means that 0.82% of the 6000 batteries lasted less than 280 days.

Note: Since the total area to the left of \( z = -2.4 \) is required, the area for \( z = 2.4 \) is subtracted from 0.5000

c. 49 batteries will last less than 280 days

2.

a. \( \mu = 25 \) the population mean of 25 was given in the question.

b. \( \bar{x} = 25.5 \) The sample mean is 25.5 and is determined by using 1 – VarsStat on the TI - 83.

c. \( \sigma = 4 \) The population standard deviation of 4 was given in the question.

d. The sample standard deviation is 3.47 and is determined by using 1 – VarsStat on the TI - 83.

e. \( \mu_x = 25 \) A property of the Central Limit Theorem.

f. \[
\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{40}} = 0.63 \text{ Central Limit Theorem}
\]

Binomial Distribution and Binomial Experiments

a. The operative word in this problem is ‘median’; if the median income is $56,400, then this indicates that one-half of the income falls below $56,400 and one-half of the income lies above it. Therefore, the chance of a randomly selected income being below the median income is 0.5.

Let \( X \) represent the number of households with incomes below the median in a random sample of size 24. \( X \) has a binomial distribution with \( n = 24 \) and \( p = 0.5 \).

\[
E(X) = np = \mu_x
\]

\[
E(X) = (24)(0.5) = 12
\]

b. The standard deviation of households with incomes less than $56,400 is
\[
\sigma_x = \sqrt{np(1-p)} = \sqrt{12(1-0.5)} = 2.45
\]

c. \( P(X \geq 18) \approx \)

\[
\begin{align*}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
\binom{24}{18} &= \frac{24!}{18!(24-18)!} = 134596 \\
\binom{24}{19} &= \frac{24!}{19!(24-19)!} = 42504 \\
\binom{24}{20} &= \frac{24!}{20!(24-20)!} = 10626 \\
\end{align*}
\]

\[
\begin{align*}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
\binom{24}{21} &= \frac{24!}{21!(24-21)!} = 2024 \\
\binom{24}{22} &= \frac{24!}{22!(24-22)!} = 276 \\
\binom{24}{23} &= \frac{24!}{23!(24-23)!} = 24 \\
\end{align*}
\]

\[
\begin{align*}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
\binom{24}{24} &= \frac{24!}{24!(24-24)!} = 1 \\
\end{align*}
\]

\[P(X = 18) = 134596(0.5)^{24} \approx 0.0080\]

\[P(X = 19) = 42504(0.5)^{24} \approx 0.0025\]

\[P(X = 20) = 10626(0.5)^{24} \approx 0.0006\]

\[P(X = 21) = 2024(0.5)^{24} \approx 0.0001\]

\[P(X = 22) = 276(0.5)^{24} \approx 0.000016\]

\[P(X = 23) = 24(0.5)^{24} \approx 0.0000014\]

\[P(X = 24) = 1(0.5)^{24} \approx 0.00000006\]

The sum of these probabilities gives the answer to the question:

0.01121746.
Confidence Intervals

1. 
   a. 
   \[
   \hat{p} = \frac{x}{n} \quad \hat{p} = \frac{142}{250} \quad \hat{p} = 0.568 \\
   \hat{p} = 0.568 
   \]
   
   The interval is from 0.488 to 0.648 \text{ OR from } 48.8\% \text{ to } 64.8\% 

   b. The 99\% confidence interval could be narrowed by increasing the sample size from 250 to a larger number.

2. 
   a. 
   \[
   \hat{p} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \\
   0.61 \pm 2.57 \sqrt{\frac{0.61(1 - 0.61)}{300}} \quad \text{from 0.555 to 0.665} \\
   \]
   \[
   \hat{p} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \\
   0.61 \pm 1.96 \sqrt{\frac{0.61(1 - 0.61)}{300}} \quad \text{from 0.538 to 0.682} 
   \]

   b. The confidence interval got wider as the confidence level increased.

   \[
   \begin{align*}
   99\% \text{ CI} & : 0.61 \pm 2.57(0.028) \\
   95\% \text{ CI} & : 0.61 \pm 1.97(0.028) \\
   95\% \text{ CI} & : 0.61 \pm 1.645(0.028) 
   \end{align*}
   \]

   To increase the probability of enclosing the population proportion a wider confidence interval must be chosen.

   c. Yes, all three confidence intervals would capture the population proportion if it were 0.58.
Sums and Differences of Independent Random Variables

1. Outcomes and their probabilities are:  
   \[ P(\text{none take the bus}) = (0.3)(0.3)(0.3) = 0.027 \]
   \[ P(\text{only the first student takes the bus}) = (0.7)(0.3)(0.3) = 0.063 \]
   \[ P(\text{only the second student takes the bus}) = (0.3)(0.7)(0.3) = 0.063 \]
   \[ P(\text{only the third student takes the bus}) = (0.3)(0.3)(0.7) = 0.063 \]
   \[ P(\text{the first and second students take the bus}) = (0.7)(0.7)(0.3) = 0.147 \]
   \[ P(\text{the second and third students take the bus}) = (0.3)(0.7)(0.7) = 0.147 \]
   \[ P(\text{the first and third students take the bus}) = (0.7)(0.3)(0.7) = 0.147 \]
   \[ P(\text{all three students take the bus}) = (0.7)(0.7)(0.7) = 0.343 \]

<table>
<thead>
<tr>
<th>Number of Students who ride the bus, (x)</th>
<th>Probability (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.027</td>
</tr>
<tr>
<td>1</td>
<td>(3)(0.063) = 0.189</td>
</tr>
<tr>
<td>2</td>
<td>(3)(0.147) = 0.441</td>
</tr>
<tr>
<td>3</td>
<td>0.343</td>
</tr>
</tbody>
</table>

2. The value of the prize is $200. If they sell 500 tickets expected value of the ticket is $2.40 (\(\frac{12}{00/500}\)).

If the students decide to sell tickets on three monetary prizes \(\text{(1500, 500, 500)}\) the expected value of the ticket will be \(1500 + 1000 = 2500\) and then \(\frac{2500}{500}\) which gives an expected value of \$5.00. 

3. 

| Outcome \(\), \(N\) \(|S\) \(|N\) \(N\) | Probability |
|--------------------------------|-------------|
| \(NNN\)                       | (0.76)(0.76)(0.76) = .438976 |
| \(NN\)                        | (0.76)(0.76)(0.24) = .138624 |
| \(NSN\)                       | (0.76)(0.24)(0.76) = .138624 |
| \(SNN\)                       | (0.24)(0.76)(0.76) = .138624 |
| \(NSS\)                       | (0.76)(0.24)(0.24) = .043776 |
| \(SNS\)                       | (0.24)(0.76)(0.24) = .043776 |
| \(SSN\)                       | (0.24)(0.24)(0.76) = .043776 |
| \(SSS\)                       | (0.24)(0.24)(0.24) = .013824 |
Table 7.3:

<table>
<thead>
<tr>
<th>Number of Drivers Using a Cell Phone, x</th>
<th>Probability P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.438976</td>
</tr>
<tr>
<td>1</td>
<td>0.415872</td>
</tr>
<tr>
<td>2</td>
<td>0.131328</td>
</tr>
<tr>
<td>3</td>
<td>0.013824</td>
</tr>
</tbody>
</table>

\[
E(X) = np = 3(0.76) = 2.28\\
\text{Var}(X) = npq = 3(0.76)(1 - 0.76) = 3(0.76)(0.24) = 0.547
\]

**Student’s t Distribution**

1. a. 8 degrees of freedom mean that the sample size is \((8 + 1)\) or 9. The graphing calculator will be used in order to randomly select scores from a normally distributed population.

   Math \[\rightarrow\] PRB \[6. \text{randNorm}(110, 20, 9) \rightarrow\] STO \[2^\text{nd}1\]

   This command will select 9 scores from the population and store the values in List One.

   ![Graphing calculator screen showing random numbers generated.]

   These are values selected by the calculator.

   b. To calculate \(\frac{\bar{x} - \mu}{s/\sqrt{n}}\), enter on the calculator

   \text{randNorm}(110, 20, 9) L_1 : (\text{mean } L_1 - 110)/\text{(stdDev } L_1)/\sqrt{9} \text{ Press enter 4 times to generate the } t \text{ values.}
You can find “mean” and “stdDev” in the catalog of the calculator. The catalog is located above the numeral 0.
Chapter 8

Hypothesis Testing - Solution Key

8.1 Hypothesis Testing

1. do not accept, fail to reject

2. \( H_0 : \mu = 19 \)
\( H_A : \mu \neq 19 \)
3. fail to reject, reject

4. For a single tailed test at the .01 level of significance the critical value= 2.325 if the alternate hypothesis is the mean is greater than some specified value and is −2.325 if the alternate hypothesis is that the mean is less than some specified value.

5. \( H_0 : \mu = 1020 \)
\( H_a : \mu > 1020 \)

In this problem we are told that \( n = 144 \) and that the average SAT score in the sample is 1100 with a standard deviation of 144. The test statistic is:

\[
\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1100 - 1020}{\frac{144}{12}} = 6.67
\]

6. probability

\[
\mu = 500 \quad \frac{s}{\sqrt{n}} = 100
\]

7.
\[ P(X > 520) = P \left( Z > \frac{520 - 500}{100} \right) \]
\[ = P(Z > .2) \]
\[ = .4207 \]

You can calculate the probability by using the technology:

2nd VARS normcdf(.2, 1000000, 0, 1)

8. Type I
9. Type II
10. Power of the test, power

With an \( \alpha \) level of .05, we have a critical value of 1.64 for the single tailed test which would have a value of:

\[ z = \frac{\bar{x} - \mu}{\sigma_x} \]
\[ 1.64 = \frac{\bar{x} - 24,500}{\frac{4800}{\sqrt{144}}} \]
\[ \bar{x} = 1.64 \left( \frac{4800}{12} \right) + 24,500 \approx 25,156 \]

Now, with a new mean set at the alternative hypothesis (\( H_a : \mu = 25,100 \)) we want to find the value of the critical score (25,156) when centered around this score. Therefore, we can figure that:

\[ z = \frac{25156 - 25100}{400} \approx 14 \]

We are concerned with making a type II error only if the average salary is less than 25,100. So we are concerned with the area to the left of .14. Using technology: 2nd VARS normalcdf(−1000000, .14, 0, 1) we find this area to be approximately .56 This means that since we assumed the alternative hypothesis to be true, there is a 56% chance of rejecting the null hypothesis. The power of this test is about .4.

Hypothesis Testing About Population Proportions Review Questions

1. The magnitude of the difference between the observed sample mean and the hypothesized population mean.
2. True

3. False. A 95% confidence interval would say that if you found 100 confidence intervals, 95 of them would contain the mean. A single confidence interval either contains the mean or does not contain the mean.

4. $H_0 : p = .60$  
   $H_a : p > .60$

5. The observed value of the sample proportion is $p = \frac{495}{750} = .66$

6. The standard error of the proportion is $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.66 \times .34}{750}} = .0173$

7. The test statistic is $z = \frac{\hat{p} - p_0}{s_p} = \frac{.66 - .60}{.0173} \approx 3.47$

8. For $\alpha = .01$ the critical value for a one sided test is 2.33. That is, we would reject the null hypothesis if our test statistic is larger than the critical value. Since 3.47 > 2.33, our test statistic falls in the rejection region and we do not accept the null hypothesis. We conclude that the probability is less than .01 that a sample proportion of .66 would appear due to sampling error if in fact the population proportion was equal to .60.

   $$p \pm z_a \times S_p = .66 \pm 2.57(.0173)$$

9. 

   The confidence interval is = $.66 \pm .044$

   (.616, .704)

10. We are 99% confident that the interval (.616 < $p < .704$) contains the proportion mean. In other words, this confidence interval show perhaps as many as 70 percent of the voters favor the bill, but it is very unlikely that less than 61 percent favor the bill.

**Testing a Mean Hypothesis Review Problems**

1. When working with large samples (typically samples of more than 30) we use the normal distribution. When working with small samples (typically samples under 30) we use the $t$ distribution.

2. True

3. $H_0 : \mu = 2.75$  
   $H_a : \mu \neq 2.75$

4. standard error is: $\frac{s}{\sqrt{n}} = \frac{.65}{\sqrt{250}} \approx .041$
5. When setting the critical regions for this hypothesis, it is important to consider the repercussions of the decision. Since there does not appear to be major financial or health repercussions of this decision, a more conservative alpha level need not be chosen. With an alpha level of .05 and a sample size of 256, we find the area under the curve associated in the z distribution and set the critical regions accordingly. With this alpha level and sample size, the critical regions are at 1.96 standard deviations above and below the mean.

6. \[ z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.85 - 2.75}{0.6406} = 2.46 \] Since 2.46 > 1.96 our test statistic falls in the critical region and we decide to reject the null hypothesis. This means that the probability that the observed sample mean would have occurred by chance if the null hypothesis is true is less than 5%.

7. Actually, at 30 students one could still use the normal distribution but below 30 one should use the t distribution.

8. \[ t = \frac{2.85 - 2.75}{s/\sqrt{n}} = 0.8425 \] One determines the critical value by using a t with 29 degrees of freedom (sample size −1). The critical t value is 2.045. since .8425 < 2.045 our test statistic does NOT fall in the critical region and we fail to reject the null hypothesis. Therefore, we can conclude that the probability that the observed sample mean could have occurred by chance if the null hypothesis is true is greater than 5%.

9. You would need a larger difference because the standard error of the mean would be greater with a sample size of 30 than with a sample size of 256.

10. a. A one tailed test because a two tailed test splits the rejection region into two pieces. The rejection region for a one tailed test at the same level of significance is larger.

    b. A .05 level of significance has a larger rejection region than a .01 level of significance.

    c. \( n = 144 \) since a smaller sample size leads to a larger standard error (since the sample size is in the denominator of the standard error) and this leads to a smaller value of the test statistic (since the standard error is in the denominator of the test statistic).

**Dependent and Independent Samples Review Problems**

1. Answers will vary. An example of dependent samples might be the pre and post test scores on the same person. An example of independent samples might be SAT scores for males and females.

2. True

3. The scenario involves independent samples since it is assumed that the results of one sample do not affect the other.

4. \( H_0 : \mu_a = \mu_b \)

   \( H_a : \mu_a \neq \mu_b \)
5. The pooled estimate for the population variance is

\[
\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{17 \times 12.25 + 23 \times 10.24}{18 + 24 - 2} = 11.09
\]

6. The estimated standard error for this scenario is

\[
\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{11.09 \left( \frac{1}{18} + \frac{1}{24} \right)} \approx 1.04
\]

7. The test statistic is

\[
\frac{\bar{x}_a - \bar{x}_b - 0}{sd} = \frac{75.6 - 72.8}{1.04} = 2.697.
\]

This exceeds the critical value of \(t = 2.021\) scores above or below the mean. Therefore, we would reject the null hypothesis and conclude that it is highly unlikely that the difference between the means of the two samples occurred by chance.

8.\[
\begin{align*}
H_0 &: \mu_h = \mu_w \quad \text{or} \quad H_0 &: \mu_h - \mu_w = \mu_d = 0 \\
H_a &: \mu_h \neq \mu_w \\
H_a &: \mu_d \neq 0
\end{align*}
\]

9. Put the data into \(L1\) and \(L2\) in the calculator. In \(L3\) put \(L1 - L2\). Then calculate \(1\) Var stats on \(L3\). This will give you the mean of the differences (1.25) and the standard deviation of the differences (3.15).

10. The test statistic is

\[
\frac{\bar{x}_a - \bar{x}_b}{sd} = \frac{1.25}{1.04} \approx 1.12.
\]

The critical value at alpha = .05 and the \(t\) distribution with 7 degrees of freedom is \(t = 2.021\). 1.12 < 2.021 and so we fail to reject the null hypothesis. Therefore, we can conclude that the attitudes towards drinking for married couples are dependent or related to each other.
Chapter 9

Regression and Correlation - Solution

Key

9.1 Regression and Correlation

1. Answers will vary. Two examples of bivariate situations are:
   a. Are height and weight of high school students related?
   b. Are SAT Math and SAT Verbal scores related?

2. a. positive correlation

Figure: A scatterplot of two variables with a positive correlation.

b. negative correlation
Figure: A scatterplot of two variables with a negative correlation.

c.

Figure: A scatterplot of two variables with a perfect (negative) correlation since they all fall on the same line.

d. Zero correlation
Figure: A scatterplot of two variables with near-zero correlation.

3. a. weak correlation

Figure: A scatterplot of two variables with a weak correlation.

b. strong correlation

Figure: A scatterplot of two variables with a strong correlation.

4. The correlation coefficient measures the strength and direction of a linear association between two variables.

5. The correlation for the following scatterplot is perfect. All the data points fall on a straight line.
6.

\[ r_{xy} = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left(n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right) \left(n \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}} \]

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>XY</th>
<th>X^2</th>
<th>Y^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>84</td>
<td>36</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>60</td>
<td>25</td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>40</td>
<td>16</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>18</td>
<td>46</td>
<td>200</td>
<td>82</td>
<td>492</td>
</tr>
</tbody>
</table>

\[ r_{xy} = \frac{5 \times 200 - 18 \times 46}{\sqrt{(5 \times 82 - 324)(5 \times 492 - 2116)}} \]
\[ = \frac{1000 - 828}{\sqrt{86 \times 344}} \]
\[ = \frac{172}{172} \]
\[ = 1 \]

7.
Table 9.2:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>X</th>
<th>Y</th>
<th>XY</th>
<th>X^2</th>
<th>Y^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>300</td>
<td>225</td>
<td>400</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>180</td>
<td>144</td>
<td>225</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>120</td>
<td>100</td>
<td>144</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>10</td>
<td>18</td>
<td>252</td>
<td>196</td>
<td>324</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>8</td>
<td>13</td>
<td>104</td>
<td>64</td>
<td>169</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6</td>
<td>12</td>
<td>72</td>
<td>36</td>
<td>144</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>150</td>
<td>225</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>288</td>
<td>256</td>
<td>324</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>195</td>
<td>169</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td></td>
<td>total</td>
<td></td>
<td>1761</td>
<td>1515</td>
<td>2155</td>
</tr>
</tbody>
</table>

\[
r_{xy} = \frac{10 \times 1761 - 119 \times 143}{\sqrt{(10 \times 1615 - 119^2)(10 \times 2155 - 143^2)}}
\]

\[
= \frac{17710 - 17017}{\sqrt{989 \times 1101}}
\]

\[
= \frac{693}{1043.5}
\]

\[
= .568
\]

8. \( r^2 = .568 = .323 \)

9. The correlation between the two quizzes is positive and is moderately strong. Only a small proportion of the variance is shared by the two variables (32.3%).

10. The three factors that we should be aware of that affect the size and accuracy of the Pearson correlation coefficient are curvilinear relationships, homogeneity of the group and small group size.

**Least Squares Regression Review Problems**

1. 

91

www.ck12.org
2. Yes. The data appear to be negatively correlated and in a linear pattern.

3. \( Y = bX + a \)

4. 

\[
b = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}
\]

\[
b = \frac{15 \times 166 - 26 \times 128}{15 \times 64 - 26^2}
\]

\[
= \frac{2490 - 3328}{960 - 675}
\]

\[
= -2.94
\]

The regression coefficient meant that every -2.94 percent change in memory test score is associated with a one percent change in exercise per week.

5. To determine the regression equation first find \( a \).

\[
a = \frac{\Sigma Y - b \Sigma X}{n}
\]

\[
= \frac{128 + 2.94 \times 26}{15}
\]

\[
= 13.63
\]

The regression equation is \( Y = -2.94X + 13.63 \)

6. See scatterplot in previous problem
7. 

\[ Y(predicted) = -2.95(3) + 13.63 \]

\[ = 4.78 \]

This means that if a student exercised 3 times per week, you would expect that they would have a memory test score of 4.78.

8. No data transformation is necessary because the relationship between the two variables is linear.

9. 

<table>
<thead>
<tr>
<th>Student</th>
<th>Exercise Per Week</th>
<th>Memory Score</th>
<th>Test Score</th>
<th>Predicated Value</th>
<th>Residual Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>15</td>
<td>13.7</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7.7</td>
<td>-4.7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>12</td>
<td>7.7</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>11</td>
<td>10.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4.8</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>8</td>
<td>10.7</td>
<td>-2.7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>15</td>
<td>7.7</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>13</td>
<td>13.7</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2</td>
<td>4.8</td>
<td>-2.8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>4.8</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2</td>
<td>1.8</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>8</td>
<td>10.7</td>
<td>-2.7</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>10</td>
<td>10.7</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>12</td>
<td>10.7</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>8</td>
<td>7.7</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

10. Upon first glance, a transformation of the data is not necessary since the residual values are relatively evenly distributed on the scatterplot. However, it is worth considering dropping
several of the outliers (namely observation 7) if they are over three standard deviations from the mean.

Inferences about Regression Review Problems

1. The predictor variable is the number of years of college. The reasoning behind this decision is that we are trying to determine and predict the financial benefits of further education (as measured by annual salary) by using the number of years of formal education (the predictor, or the \( X \) variable.

2. Yes. With an \( r \) value of 0.67, these two variables appear to be moderately strongly correlated. The nature of the relationship is a relatively strong, positive correlation.

3. 

\[
Y = bX + a \\
b = r \frac{s_y}{s_x} \\
a = \bar{y} - b\bar{X} \\
b = 0.67 \times \frac{4.2}{3.6} \approx 0.782 \\
a = 15.45 - 0.782(14.85) \approx 3.84 \\
Y(\text{predicted}) = 0.782X + 3.84
\]

4. \( Y(\text{predicted}) = 0.782(13.5) + 3.84 \approx 14.4 \) thousand dollars

5. 

\[
H_0 : \beta = 0 \\
H_a : \beta \neq 0 \\
t = \frac{b - \beta_0}{s_b} \text{ where } s_b = \frac{s_{xy}}{ss_x} \\
t = \frac{0.782 - 0}{\frac{3.12}{\sqrt{1542}}} = 9.84
\]

The critical value at the .05 level of significance is 1.98. This is a \( t \) with 118 degrees of freedom.

The value of the test statistic exceeds the critical value so we reject the null hypothesis and conclude that if the null hypothesis was true, we would observe a regression coefficient of .792 by chance less than 5% of the time.
6. \( X = 15Y \text{(predicted)} = .782(15) + 3.84 = 15.57 \)

\[
\text{For } Z = \frac{Y - Y \text{(predicted)}}{s_{yy} \text{(predicted)}} = \frac{18500 - 15570}{3.12} = \frac{2930}{3.12} = .939
\]

\[ P(z > .939) \approx .174 \]

This was calculated using technology 2nd VARS normalcdf(.939, 10000, 0, 1) This means that with 15 years of formal education an estimated 17.4% of parents will have incomes greater than $18,500.

7. \( X = 12Y \text{(predicted)} = .782(15) + 3.84 = 13.22 \)

\[
\text{For } Z = \frac{Y - Y \text{(predicted)}}{s_{yy} \text{(predicted)}} = \frac{18500 - 15570}{3.12} = \frac{5.28}{3.12} = 1.69
\]

\[ P(z > 1.69) \approx .0455 \]

This was calculated using technology 2nd VARS normalcdf(1.69, 1000, 0, 1) This means with 12 years of formal education an estimated 4.6% of parents will have incomes greater than $18,500.

8. \( s_y = S_Y \cdot X \sqrt{1 + \frac{1}{n} + \frac{(X - X)^2}{SS_x}} = 3.12 \sqrt{1 + \frac{1}{120} + \frac{(16 - 14.85)^2}{1542}} = 3.14 \)

Using the general formula for the confidence interval, \( Y \text{(predicted)} \pm tS_y \text{(predicted)} \) where \( Y \text{(predicted)} = .782(16) + 3.84 = 16.35 \) and we find that

\[ CI_{95} = 16.35 \pm (1.98)(3.14) = 16.35 \pm 6.22 \]

\[ CI_{95} = (10.13, 22.57) \]

9. Answers here will vary though one hopes the students will realize that the more formal education one the higher their salary will likely be.

**Multiple Regression Review Problems**

1. There are 3 predictor values – Test 1, Test 2 and Test 3.

2. The regression coefficient of 0.5560 tells us that every 0.5560 percent change in Test 2 is associated with a 1.000 percent change in the final exam when everything else is held constant.

3. From the data given, the regression equation is \( Y = 0.0506X_1 + 0.5560X_2 + 0.2128X_3 + 10.7592. \)
4. The $R^2$ value is 0.4707 and indicates that 47% of the variance in the final exam can be attributed to the variance of the combined predictor variables.

5. Using the print out, we see that the $F$ statistic is 13.621 and has a corresponding $p$ value of 0.000. This means that the probability that the observed $R$ value would have occurred by chance if it was not significant is very small (slightly greater than 0.000).

6. Test 2. Upon closer examination, we find that on the Test 2 predictor variable is significant since the t value of 3.885 exceed the critical value (as evidenced by the low $p$ value of .003).

7. No. It is not necessary to include Test 1 and Test 3 in the multiple regression model since these two variables do not have a significant test statistic that exceeds the critical value.
Chapter 10

Chi-Square - Solution Key

10.1 Chi-Square

Goodness of Fit Test Review Questions

1. c. The name of the statistical test used to analyze the patterns between two categorical variables is the chi-square test.

2. a. The type of chi-square test that estimates how closely a sample matches an expected distribution is the Goodness of Fit test.

3. b. Gender is considered a categorical variable.

4. a. If there were 250 observations in a data set and 2 uniformly distributed categories that were being measured you would expect to see 125 \( \left( \frac{250}{2} \right) \) observations in each category.

5. The formula for calculating the chi-square statistic is: \( \chi^2 = \frac{\sum(O-E)^2}{E} \) where \( O \) stands for the observed frequency and \( E \) stands for the expected frequency.

6. The observed frequency value for the Science Museum category is 29.

7. The expected frequency value for the Sporting Event category is \( \frac{100}{3} = 33.3 \)

8. a. The null hypothesis for the situation would be there is no preference e between the types of field trips that students prefer.
9. 

\[
\chi^2 = \frac{(53 - 33.3)^2}{33.3} + \frac{(18 - 33.3)^2}{33.3} + \frac{(29 - 33.3)^2}{33.3} \\
= \frac{388.09 + 234.09 + 18.49}{33.3} \\
= \frac{640.67}{33.3} \\
\approx 19.24
\]

10. If the critical chi-square level of significance was 5.99, you would reject the null hypothesis since 19.24 > 5.99.

**Test of Independence Review Problems**

1. The chi-square Test of Independence is used to examine if two variables are related.
2. True. In the test of independence, you can test if two variables are related but you cannot test the nature of the relationship itself.
3. a. When calculating the expected frequency for a cell in a contingency table, you use the formula

   Expected Frequency = (row total)(column total)/Total Number observations.

4. d. The total number of females in the study is $322 + 460 = 782$
5. c. The total number of observations in the study is $322 + 460 + 128 + 152 = 1062$
6. a. The expected frequency for males that did not study abroad is $\frac{(280)(612)}{1062} = 161.36$
7. a. The degrees of freedom is $(number\ of\ rows - 1)\ (number\ of\ columns - 1)$.

   In this problem that is $(2 - 1)(2 - 1) = 1$.
8. True. The null hypothesis is females are as likely to study abroad as males.
9. a. The chi-square statistic for this example is

   \[
   \chi^2 = \frac{(322 - 331)^2}{331} + \frac{(460 - 451)^2}{451} + \frac{(128 - 119)^2}{119} + \frac{(152 - 161)^2}{161} \\
   = .245 + .180 + .681 + .503 \\
   = 1.60
   \]

10 b. If the chi-square critical value at .05 and 1 degree of freedom is 3.81 and we have a calculated chi-square value of 2.22 we reject because 2.22 < 3.81.
11 True. The Test of Homogeneity is used to evaluate the equality of several samples of certain variables.

12 b. The Test of homogeneity is carried out the exact same way as the Test of Independence.

**Testing One Variance Review Problems**

1. d. We use the chi-square for testing goodness of fit, independence and for testing a hypothesis of single variance.

2. False. The chi-square distribution can be used for testing a hypothesis about a single variance for a normal population.

3. a. In testing variance, our null hypothesis states that the two population means that we are testing are equal with respect to variance.

4. c. In the formula for calculating the chi-square statistic for the single variance, \( \sigma^2 \) is the hypothesized population variance.

5. d. There is no other information needed to solve for the chi-square statistic if we know the number of observations in the sample, the standard deviation of the sample and the hypothesized variance of the population.

6. Given \( n = 30 \), \( s = 1.1 \) and \( \sigma^2 \) is the hypothesized population variance, the null hypothesis would be \( H_0 \) : the sample comes from a population with \( \sigma^2 \leq 3.22 \)

7.

\[
\chi^2 = \frac{df \times s^2}{\sigma^2} = \frac{29 \times 1.1^2}{3.22} = 10.897
\]

8. At \( \alpha = 0.5 \) and a critical chi-square value of 42.557, 10.897 < 42.557 we fail to reject the null hypothesis.

9.

\[
\chi_{.025}^2 \leq \frac{df \times s^2}{\sigma^2} \leq \chi_{1-.025}^2 \\
\frac{df \times s^2}{\chi_{.05}^2} \leq \sigma^2 \leq \frac{df \times s^2}{\chi_{.05}^2} \\
\frac{29(1.1)^2}{45.52} \leq \sigma^2 \leq \frac{29(1.1)^2}{17.71} \\
.824 \leq \sigma^2 \leq 1.98
\]
10. We are 90% confident that the overall population variance is between .824 and 1.98.
Chapter 11

Analysis of Variance and the F-Distribution - Solution Key

11.1 Analysis of Variance and the F-Distribution

The F Distribution and Testing Two Variances Review Questions

1. We use the $F$ Max test to examine the differences in the variances between two independent samples.

2. Answers may vary but could include:
   a. We use the $t$ distribution when testing the difference between the means of two independent samples and the $F$ distribution when testing the difference between the variances of two independent samples.
   b. The $t$ distribution is based off of one degree of freedom and the $F$ distribution is based off of two.
   c. $F$ distributions are not symmetrical, $t$ distributions are.
   d. $T$ values range from $-\infty$ to $-\infty$ while $F$ ratios range from zero to $\infty$

3. When we test the differences between the variance of two independent samples, we calculate the $F$ ratio.

4. When calculating the $F$ ration, it is recommended that the sample with the larger sample variance be placed in the numerator and the sample with the smaller sample variance be placed in the denominator.

5. $H_0 : \sigma_1^2 = \sigma_2^2$
\(H_a : \sigma_1^2 \neq \sigma_2^2\). For \(\alpha = .10\) the critical value is 2.04

7. The \(F\) ratio is \(\frac{42.30}{18.80} = 2.25\). This is an \(F\) with degrees of freedom 30, 20.

8. We would reject the null hypothesis because the calculated \(F\) ratio (2.25) exceeds the critical value (2.04).

9. We can conclude that the variance of the student achievement scores for the second sample is less than the variance for the students in the first sample. Since the achievement test means are practically equal, the variance in student achievement scores may help the guidance counselor in her selection of a preparatory program.

10. True The test of the null hypothesis \(H_0 : \sigma_1^2 = \sigma_2^2\) using the \(F\) distribution is only appropriate when it can be safely assumed that the population is normally distributed.

### One Way ANOVA Test Review Questions

1. ANOVA stands for Analysis of Variance.

2. If we are testing whether pairs of ample means differ by more than we would expect due to chance using multiple \(t\) tests, the probability of making a Type I error would increase.

3. c. In the ANOVA method, we use the \(F\) distribution.

4. In the ANOVA method, we complete a series of steps to evaluate our hypothesis. The steps are done in the following order:
   a. State the null hypothesis
   b. Calculate the mean squares between groups and the mean squares within groups.
   c. Calculate the test statistic
   d. Determine the critical values in the \(F\) distribution
   e. Evaluate the hypothesis

5. \(H_0 : \mu_1 = \mu_2 = \mu_3\)

6. Table 11.1:

<table>
<thead>
<tr>
<th></th>
<th>Ms. Jones</th>
<th>Mr. Smith</th>
<th>Mrs. White</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number ((n_k))</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>Total ((T_k))</td>
<td>109</td>
<td>131</td>
<td>137</td>
<td>377</td>
</tr>
<tr>
<td>Mean ((X))</td>
<td>10.9</td>
<td>14.6</td>
<td>17.1</td>
<td>13.96</td>
</tr>
<tr>
<td>Sum of Squared Obs. ((\sum_{i=1}^{n_k} X_{ik}^2))</td>
<td>1,339</td>
<td>2,113</td>
<td>2,529</td>
<td>5,981</td>
</tr>
</tbody>
</table>

www.ck12.org 102
Table 11.1: (continued)

<table>
<thead>
<tr>
<th></th>
<th>Ms. Jones</th>
<th>Mr. Smith</th>
<th>Mrs. White</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Obs.</td>
<td>1,188</td>
<td>1,907</td>
<td>2,346</td>
<td>5,441</td>
</tr>
<tr>
<td>Squared/Number of Obs. ( \frac{T^2_k}{n_k} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>907</td>
<td>346</td>
<td>441</td>
</tr>
</tbody>
</table>

7.

\[
MS_B = \frac{SS_B}{k-1} = \frac{\left( \sum \frac{T^2_k}{n_k} - \frac{T^2}{N} \right)}{2} = \frac{\left( 1188 + 1907 + 2346 - \frac{377^2}{27} \right)}{2} = \frac{176.96}{2} = 88.48
\]

8.

\[
MS_w = \frac{SS_w}{N-k} = \frac{\left( \sum X^2_{ik} - \sum \frac{T^2_k}{n_k} \right)}{N-k} = \frac{5981 - 5441}{27 - 3} = 22.5
\]

9. The \( F \) ratio of these two values is \( \frac{88.48}{22.5} = 3.95 \)

10. The critical value for the \( F_{2.24} \) with \( \alpha = 0.05 \) is 3.40

11. The calculated test statistic exceeds the critical value so we would reject the null hypothesis. Therefore, we conclude that not all the population means are equal.

**The Two-Way ANOVA Test Review Questions**

1. In two-way ANOVA, we study not only the effect of two independent variables on the dependent variable, but also the interaction between the variables.
2. d. We could conduct multiple t tests between pairs of hypotheses but there are several advantages when we conduct a two way ANOVA. These include efficiency, control over additional variables and the study of interaction between variables.

3. d. Calculating the total variation in two-way ANOVA includes calculating 4 types of variation – variation with the group, variation in the dependent variable attributed to one independent variable, variation in the dependent variable attributed to the other independent variable, and variation between the independent variables.

4. The three hypotheses are:
   a. The mean for genders are the same.
   b. The mean for the three doses are the same.
   c. All effects are zero.

5. Answers may vary. They could include (1) $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, $H_0 : \text{all effects} = 0$ or (2) written hypotheses that the means of the independent variable in the rows are equal to each other, the means of the independent variable in the rows columns are equal to each other and there is no interaction.

6. The three critical values are 4.07, 3.23 and 3.23. These values are derived from the $F$ distribution. If the calculated $F$ statistic exceeds these values, we will reject the null hypothesis.

7. Since 14.94 > 4.07 and 8.62 > 3.23 we reject that gender and dosage are the same. Since 1.30 < 3.23 we fail to reject the third hypothesis.

8. We can conclude that not all means in the populations are equal with regard to gender and drug dosage. Because the $F$ ratio for the interaction effect (gender x drug dosage) was not statistically significant, the conclusion is that there is no difference in the performance of the male and female rats across the levels of drug dosage.