Section 3.1 Part 1: Explanatory/Response Variables, Scatterplots, & DOFS pg. 142-149

Today, you will learn how to:
- Distinguish between explanatory and response variables for quantitative data (also called Bivariate data)
- Make a scatterplot to display the relationship between two quantitative variables
- Describe the direction, unusual features/OUTLIERS, form, and strength in a scatterplot (DOFS or DUFS)

Example 1: The Endangered Manatee
Manatees are large, gentle, slow-moving creatures found along the coast of Florida. Many manatees are injured or killed by boats. The table at right contains data on the number of Manatees killed by boats for the years 1977 to 2010.

1. Identify the explanatory and response variables.

   Explanatory variable = boats registered
   Response variables = manatees killed

2. How many variables do we have? Are they categorical or quantitative?

   2 quantitative variables = bivariate data

3. Enter the data into L1 and L2. Make a scatterplot of the relationship between boat registrations and manatees killed. (STAT PLOT > Select Scatter plot option > ZOOM STAT>Adjust window as you see fit).

4. Describe the relationship formed in the scatterplot.

   Direction: positive association because as the # of boat registrations increases, the # of manatees killed also increases.
   Outliers/Unusual Features: there are no outliers.
   Form: linear association
   Strength: moderately strong

<table>
<thead>
<tr>
<th>YEAR</th>
<th>BOATS</th>
<th>MANATEES</th>
<th>YEAR</th>
<th>BOATS</th>
<th>MANATEES</th>
<th>YEAR</th>
<th>BOATS</th>
<th>MANATEES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977</td>
<td>447</td>
<td>13</td>
<td>1989</td>
<td>711</td>
<td>50</td>
<td>2001</td>
<td>944</td>
<td>81</td>
</tr>
<tr>
<td>1978</td>
<td>490</td>
<td>21</td>
<td>1990</td>
<td>719</td>
<td>47</td>
<td>2002</td>
<td>962</td>
<td>95</td>
</tr>
<tr>
<td>1980</td>
<td>470</td>
<td>16</td>
<td>1992</td>
<td>679</td>
<td>38</td>
<td>2004</td>
<td>983</td>
<td>69</td>
</tr>
<tr>
<td>1981</td>
<td>513</td>
<td>24</td>
<td>1993</td>
<td>678</td>
<td>35</td>
<td>2005</td>
<td>1010</td>
<td>79</td>
</tr>
<tr>
<td>1982</td>
<td>512</td>
<td>20</td>
<td>1994</td>
<td>666</td>
<td>49</td>
<td>2006</td>
<td>1024</td>
<td>92</td>
</tr>
<tr>
<td>1983</td>
<td>526</td>
<td>15</td>
<td>1995</td>
<td>713</td>
<td>42</td>
<td>2007</td>
<td>1027</td>
<td>73</td>
</tr>
<tr>
<td>1984</td>
<td>559</td>
<td>34</td>
<td>1996</td>
<td>732</td>
<td>60</td>
<td>2008</td>
<td>1010</td>
<td>90</td>
</tr>
<tr>
<td>1985</td>
<td>585</td>
<td>33</td>
<td>1997</td>
<td>755</td>
<td>54</td>
<td>2009</td>
<td>962</td>
<td>97</td>
</tr>
<tr>
<td>1986</td>
<td>614</td>
<td>33</td>
<td>1998</td>
<td>809</td>
<td>66</td>
<td>2010</td>
<td>942</td>
<td>83</td>
</tr>
<tr>
<td>1987</td>
<td>845</td>
<td>39</td>
<td>1999</td>
<td>650</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>675</td>
<td>43</td>
<td>2000</td>
<td>880</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Big Ideas from Section 3.1 Part 1: Variables, Making Scatterplots, and Describing Scatterplots

Explanatory and Response Variables
- Explanatory: used to predict
- Response: outcome; responds to explanatory

DOFS
- Directions: positive, negative, none
- Outliers: look for unusual departures from the overall pattern
- Form: linear or nonlinear
- Strength: how close together (strong) or far apart (weak) the points are
(fairly or moderately strong/weak)

Now You Try: Track and Field Day

Each member of a small statistics class ran a 40-yard sprint and then did a long jump (with a running start). The table below shows the sprint time (in seconds) and the long-jump distance (in inches).

<table>
<thead>
<tr>
<th>Sprint time (s)</th>
<th>5.41</th>
<th>5.05</th>
<th>7.01</th>
<th>7.17</th>
<th>6.73</th>
<th>5.68</th>
<th>5.78</th>
<th>6.31</th>
<th>6.44</th>
<th>6.50</th>
<th>6.80</th>
<th>7.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-jump distance (in)</td>
<td>171</td>
<td>184</td>
<td>90</td>
<td>65</td>
<td>78</td>
<td>130</td>
<td>173</td>
<td>143</td>
<td>92</td>
<td>139</td>
<td>120</td>
<td>110</td>
</tr>
</tbody>
</table>

Make a scatterplot of the relationship between sprint time and long-jump distance. Describe what the scatterplot reveals:

D.O.F.S: (AP-style)
- There is a fairly strong, negative, linear association between sprint time and long distance jumps.
- There do not appear to be any outliers or unusual features.

Check Your Understanding: pg. 144 #1-2 and pg. 149 #1-5

*optional practice*
Section 3.1 Part 2: Correlation Coefficient $r$

Today, you will learn how to:
- Determine and interpret the correlation coefficient $r$ and understand the effect of outliers on $r$

The Correlation Coefficient $r$ BIG IDEAS

A **scatterplot** describes the direction, form, and strength of the relationship between two quantitative variables.

1. A linear relationship is **strong** if points are close to linear pattern & $r$ close to $\pm 1$
2. A linear relationship is **weak** if points are far apart from linear pattern & $r$ is __
3. The Correlation $r$ measures the direction (+ or -) and strength of the linear relationship between two quantitative variables

4. The Correlation $r$ is always a number between $-1$ and $1$.
   - $r > 0$ indicates **positive association**
   - $r < 0$ indicates **negative association**
   - The extreme values $r = 1$ and $r = -1$ only occur in a perfect linear relationship, when the points lie directly on the line.

**HOW TO CALCULATE THE CORRELATION $r$**

Suppose that we have data on variables $x$ and $y$ for $n$ individuals. The values for the first individual are $x_1$ and $y_1$, the values for the second individual are $x_2$ and $y_2$, and so on. The means and standard deviations of the two variables are $\bar{x}$ and $s_x$ for the $x$-values, and $\bar{y}$ and $s_y$ for the $y$-values. The correlation $r$ between $x$ and $y$ is

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

or, more compactly,

$$r = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

**Important Facts about Correlation (KNOW THESE FOR TESTS):**

1. Correlation makes no distinction between explanatory and response variables!!
2. $r$ does not change when we change the units of measure. Changing the units does not change the way the scatterplot looks either.
3. The correlation $r$ has no units. It's just a number to help you interpret strength!
4. **CORRELATION DOES NOT IMPLY CAUSATION.**
5. $r$ can only be calculated from quantitative variables, never categorical variables.
6. Correlation does not describe curved relationships, only linear relationships. That's why you should always plot the data and analyze the graph to ensure that you're interpreting $r$ correctly.
7. The correlation is **NOT RESISTANT** because it is strongly affected by outliers or extreme values.
   - Suggestion: play around with [http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html](http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html)
8. Always give the mean and standard deviation first because correlation is not a complete summary of the two-variable data.
**Example 1: Track and Field Day**

<table>
<thead>
<tr>
<th>Sprint time (s)</th>
<th>5.41</th>
<th>5.05</th>
<th>7.01</th>
<th>7.17</th>
<th>6.73</th>
<th>5.68</th>
<th>5.78</th>
<th>6.31</th>
<th>6.44</th>
<th>6.50</th>
<th>6.80</th>
<th>7.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-jump distance (in)</td>
<td>171</td>
<td>184</td>
<td>90</td>
<td>65</td>
<td>78</td>
<td>130</td>
<td>173</td>
<td>143</td>
<td>92</td>
<td>139</td>
<td>120</td>
<td>110</td>
</tr>
</tbody>
</table>

Here is a scatterplot of the sprint time and long-jump distance data from earlier. Make sure the sprint data is still entered into L1 and the long distance jump data is entered into L2.

a) **Determine** the correlation coefficient $r$ using your graphing calculator. 

\[ r = \boxed{-0.839} \]

* negative
* closer to $r = -1$ than $r=0$ - $strong$?

b) **Interpret** the correlation coefficient $r$ in context.

* There is a negative and fairly strong association between sprint times and long jumps.

**c) What effect** does the bold point have on the correlation? **Explain.**

The point (7.25, 110) would weaken the correlation because it is further away from the rest of the linear pattern. The $r$-value $r = -0.839$ is closer to 0 with the point included, than it would be without the point.

**Know the Facts about $r$:**

d) Is it correct to say that your $r$-value's units are meters per second?

No. $r$ doesn't have units, it is just a number.

e) What would happen to $r$ if we changed sprint time to minutes and long jump to feet?

It would stay the same.

f) What would happen to $r$ if we switched the $x$ and $y$ axes?

It would stay the same.

g) Would we be able to use $r$ to determine the strength of the data if the pattern in the scatterplot was curved?

No. $r$ is only used for linear relationships.

h) Can we conclude that longer sprint times cause long-distance runners to have shorter long jumps? **Explain.**

No! Correlation does not imply causation. Sprint times do not necessarily cause a decrease in long jump distances. Perhaps runner's weight, height, fitness levels, and experience impacted the results.
Big Ideas from Section 3.1 Part 2: Correlation Coefficient $r$

Interpreting $r$: sentence frame
- There is a (direction) and (strength) association between explanatory and response context.

Facts about $r$:
- no units!
- $-1 \leq r \leq 1$
- nonresistant to outliers
- points outside the pattern weaken $r$.
- points inside the pattern strengthen $r$.
- $r \neq$ slope!

Causal relationship to make a better story—beware!

Always remember....

Example 2: Salads at McDonalds

McDonald’s restaurants offer a variety of salads. The table below lists 10 different salads, along with the amount of sodium (in mg) and the amount of fat (in grams). Source: http://nutrition.mcdonalds.com/getnutrition/nutritionfacts.pdf

<table>
<thead>
<tr>
<th>Salad</th>
<th>Sodium</th>
<th>Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwest Salad with Grilled Chicken</td>
<td>650</td>
<td>8</td>
</tr>
<tr>
<td>Southwest Salad with Crispy Chicken</td>
<td>820</td>
<td>21</td>
</tr>
<tr>
<td>Southwest Salad without chicken</td>
<td>150</td>
<td>4.5</td>
</tr>
<tr>
<td>Bacon Ranch Salad with Grilled Chicken</td>
<td>700</td>
<td>9</td>
</tr>
<tr>
<td>Bacon Ranch Salad with Crispy Chicken</td>
<td>870</td>
<td>22</td>
</tr>
<tr>
<td>Bacon Ranch Salad without chicken</td>
<td>300</td>
<td>7</td>
</tr>
<tr>
<td>Caesar Salad with Grilled Chicken</td>
<td>580</td>
<td>5</td>
</tr>
<tr>
<td>Caesar Salad with Crispy Chicken</td>
<td>740</td>
<td>18</td>
</tr>
<tr>
<td>Caesar Salad without chicken</td>
<td>180</td>
<td>4</td>
</tr>
<tr>
<td>Side Salad</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

DOES!

a) Describe what the scatterplot reveals about the relationship between the amount of sodium and the amount of fats in McDonalds salads.

There is a positive, fairly strong, non-linear association between sodium & fat content. There are 3 distinctive clusters formed by salads with crispy chicken, grilled chicken, & no chicken.

b) Determine and interpret $r$.

$\text{r} = 0.849$

There is a positive and fairly strong association between sodium and fat content.

c) Does higher sodium content mean a burger contains higher fat content?

No! Correlation does not imply causation.

d) What would happen to $r$ if a new burger with 400 mg of Sodium and 10 grams of fat were added to the data set?

The $r$-value would be closer to 1 and the association is stronger.
Check Your Understanding #1 - 3: Chocolate and Nobel Prizes

Most people love chocolate for its great taste. But does it also make you smarter? A scatterplot like this one recently appeared in the New England Journal of Medicine. The explanatory variable is the chocolate consumption per person for a sample of countries. The response variable is the number of Nobel Prizes per 10 million residents of that country.

1. **Interpret** the correlation of $r = 0.791$.
   
   There is a **moderately strong**, positive association between chocolate consumption and Nobel prizes.

2. What effect does Switzerland have on the correlation? **Explain**.
   
   Switzerland strengthens the correlation since it is inside the linear pattern. The $r$-value moves move closer to 1.

3. If people in the United States started eating more chocolate, can we expect more Nobel prizes to be awarded to residents of the United States? **Explain**.
   
   No, correlation does not imply causation. Chocolate consumption does not necessarily cause an increase in Nobel Prizes. There is probably something else impacting both variables, like the economy.