Chapter 3: Describing Relationships

Section 3-1: Scatterplots
3-1 Scatterplots

In Chapters 1 and 2, we studied data for 1 variable.
   Example: Chapter 1 Test Scores

In Chapter 3, we investigate the relationship between 2 variables.
   Example: Ch. 1 Test Scores vs. Student GPA

Sometimes one variable explains or even causes changes in the other variable. If this is the case:
   The independent variable, or the explanatory variable attempts to explain the observed outcomes.
   The dependent variable, or the response variable measures an outcome of a study.
   
   Explanatory = green   Response = blue

Which is the explanatory variable and which is the response variable?
(1) For a student, hours studied and test scores.
(2) For a given high school, number of college bound students and teacher pay.
(3) For a high school student, SAT math score and SAT verbal score ??

Depends on your intentions. What do you want to predict?
Check Your Understanding Pg 144 #1-2

1) Explanatory: number of cans of beer
   Response: blood alcohol level

2) Explanatory: amount of debt AND income
   Response: stress caused by college debt
We want to think about two questions:

1) What does the data suggest?
   Is there a relationship between GPA & ACT?

2) Could we use one variable to predict the other?
   Does a high GPA imply a high ACT Score?
General Approach to 2-variable data:

1. Start with a graph. (Use a calculator to make scatterplot)
2. Look for an overall pattern and deviations from the pattern.
3. Add numerical summaries. (tomorrow)
4. Describe the overall pattern.

These are the same steps as the 1-variable data from Ch. 2

Strategy for exploring data:

1. Start with a __________.
2. Look for overall __________ and striking deviations such as outliers.
3. Choose a numerical summary to describe __________ and __________.
4. Describe the pattern using __________ curve.
A **scatterplot** shows the relationship between two **quantitative** variables measured on the same individuals.

Each individual appears on the scatterplot as a **point**.

The explanatory variable, if there is one, always goes on the **x** axis.

The response variable, if there is one, always goes on the **y** axis.

**Scatterplots are the only choice for displaying the relationship between 2 quantitative variables!**

**GRAPHING SCATTERPLOTS:**
1) Decide which variable should go on each axis.
2) **Label** and scale your axes.
3) Plot individual data values.

Graph the scatterplot for the GPA and ACT scores by hand first and then use a calculator.

<table>
<thead>
<tr>
<th>GPA</th>
<th>3.0</th>
<th>3.5</th>
<th>3.1</th>
<th>4.0</th>
<th>3.7</th>
<th>3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td>31</td>
<td>26</td>
<td>19</td>
</tr>
</tbody>
</table>
When examining a scatterplot, look for an overall pattern showing (1) **direction**, (2) **form**, and (3) **strength**.

(1) **Direction**
- Two variables are **positively** associated when above-average values of one go with above-average values of the other. Example: *Hours studied vs test grade*.
- Two variables are **negatively** associated when above-average values of one go with below-average values of the other. Example: *Temperature vs hot chocolate sales*.

(2) **Form**
- The most common forms of a scatterplot are **linear**, **curved**, or **clusters**. Not all scatterplots fit into one of these forms.

(3) **Strength**
- Determined by how close the points in the scatterplot resemble a simple form, such as a **line**.

Just as with 1-variable statistics, we must continue to watch out for **outliers**, or individual observations that fall outside the overall pattern of the graph.

When describing the relationship in a scatterplot, just remember: **DOFS + context**.
Let’s describe the relationship between GPA and ACT scores.

Remember **DOFS + context**.

There is a ***moderately strong positive linear*** relationship between GPA and ACT scores, with ***no obvious outliers***.

**Side notes:**

If there are 2 individuals with the same data point, use a different symbol. 

To add a categorical variable to a scatterplot, such as gender, simply use 2 different symbols.

For example, use X for men and O for women.
The relationship is positive. The longer the duration of the eruption, the longer the wait between eruptions. One reason for this may be that if the geyser erupted for longer, it expended more energy and it will take longer to build up the energy needed to erupt again.

The form is roughly linear with two clusters. The clusters indicate that in general there are two types of eruptions: one shorter, the other somewhat longer.

The relationship is fairly strong. Two points define a line, and in this case we could think of each cluster as a point, so the two clusters seem to define a line.

There are a few outliers around the clusters, but not many and not very distant from the main grouping of points.

The Starnes family needs to know how long the last eruption lasted in order to predict how long until the next one.
3-1 Correlation

From yesterday we know that a scatterplot allows us to look for an overall pattern by investigating its (1) **direction**, (2) **form**, and (3) **strength**.

Today, we will evaluate **strength**. Start by looking at the scatterplots on page 150. Which is stronger?

What do you notice?
The **correlation** measures the **strength** and **direction** of the **LINEAR** relationship between two **quantitative** variables.

Correlation: \( r \)

Two variables: \( x, y \)

Two means: \( \bar{x}, \bar{y} \)

Two standard deviations: \( s_x, s_y \)

Number of individuals: \( n \)

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

Also called the “**correlation coefficient**”

Example: Find the correlation, \( r \), for the following two variables \( x \) and \( y \):

**This is the GPA & ACT scores data**

<table>
<thead>
<tr>
<th>Individual #</th>
<th>x-value</th>
<th>y-value</th>
<th>( \frac{x_i - \bar{x}}{s_x} )</th>
<th>( \frac{y_i - \bar{y}}{s_y} )</th>
<th>( \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.0</td>
<td>19</td>
<td>-1</td>
<td>-1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>27</td>
<td>0.25</td>
<td>0.62</td>
<td>0.155</td>
</tr>
<tr>
<td>3</td>
<td>3.1</td>
<td>22</td>
<td>-0.75</td>
<td>-0.41</td>
<td>0.3075</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>31</td>
<td>1.5</td>
<td>1.45</td>
<td>2.175</td>
</tr>
<tr>
<td>5</td>
<td>3.7</td>
<td>26</td>
<td>0.75</td>
<td>0.41</td>
<td>0.3075</td>
</tr>
<tr>
<td>6</td>
<td>3.1</td>
<td>19</td>
<td>-0.75</td>
<td>-1.04</td>
<td>0.78</td>
</tr>
</tbody>
</table>

\[
\text{Sum} = 4.765
\]
\[ r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\[ r = \frac{1}{6 - 1} (4.765) \]

\[ r = 0.953 \]

Do #20b from page 161.

<table>
<thead>
<tr>
<th>Individual #</th>
<th>x-value</th>
<th>y-value</th>
<th>( \frac{x_i - \bar{x}}{s_x} )</th>
<th>( \frac{y_i - \bar{y}}{s_y} )</th>
<th>( \frac{x_i - \bar{x}}{s_x} \left( \frac{y_i - \bar{y}}{s_y} \right) )</th>
</tr>
</thead>
</table>

\[ r \approx 0.564 \]
Guess what!! Of course we can find $r$ with the calculator!

<table>
<thead>
<tr>
<th>GPA</th>
<th>3.0</th>
<th>3.5</th>
<th>3.1</th>
<th>4.0</th>
<th>3.7</th>
<th>3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td>31</td>
<td>26</td>
<td>19</td>
</tr>
</tbody>
</table>

1) Create two lists.

2) Turn DiagnosticOn [mode $\rightarrow$ STAT DIAGNOSTIC $\rightarrow$ ON]

3) STAT $\rightarrow$ CALC $\rightarrow$ 4:LinReg(ax+b) OR 8:LinReg(a+bx)

4) Input Xlist and Ylist $\rightarrow$ Calculate

5) Look for $r =$

6) Calculator: $r = 0.955$ By Hand: $r = 0.953$
Write the formula for the correlation \( r \) from memory: 

\[
\frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

**Facts about \( r \):**

- Positive \( r \) indicates a **positive** association.
- Negative \( r \) indicates a **negative** association.
- As one variable increases, the other decreases OR As one variable decreases, the other increases
- \(-1 \leq r \leq 1\) always.
  - \( r = 0 \) indicates **no linear relationship**.
  - \( r = 0.3 \) indicates **weak (+) linear relationship**.
  - \( r = -0.3 \) indicates **weak (-) linear relationship**.
  - \( r = 0.7 \) indicates **fairly strong (+) linear relationship**
  - \( r = 1.0 \) indicates **very strong (+) linear relationship**
  - \( r = -1.0 \) indicates **very strong (-) linear relationship**

-- Look at scatterplots on page 151.
Correlation makes no distinction between explanatory and response variables.

The units for correlation, $r$, are nothing.

The correlation value does not change if the units for either variable are changed.

The correlation value, $r$, describes the strength of linear relationships only.

The correlation value, $r$, is nonresistant, meaning vulnerable to outliers.

Correlation coefficient $r$ is unitless since formula is standardizing $x$ and $y$

Use your calculator to plot the scatter plot for this data. Describe the relationship.

<table>
<thead>
<tr>
<th>L₁</th>
<th>L₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
</tbody>
</table>

Nonlinear relationship because it’s curved data. The relationship is $y = x^2$.

 ALWAYS plot the data on your calculator to see the relationship.
1) Estimates of $r$ will vary.
   (a) The correlation is about 0.9. There is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed.
   (b) The correlation is about 0.5. There is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms.
   (c) The correlation is about 0.3. There is a weak, positive linear relationship between the healing rate of the two front limbs of the newts.
   (d) The correlation is about −0.1. There is a weak, negative linear relationship between last year’s percent return and this year’s percent return in the stock market.

2) The correlation would decrease. This point has the effect of strengthening the observed linear relationship that we see.
Correlation Coefficient Guessing Game:

http://www.rossmanchance.com/applets/GuessCorrelation.html