3.1: Scatterplots & Correlation

I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped. Well, maybe.
Section 3.1
Scatterplots and Correlation

After this section, you should be able to...

✓ IDENTIFY explanatory and response variables
✓ CONSTRUCT scatterplots to display relationships
✓ INTERPRET scatterplots
✓ MEASURE linear association using correlation
✓ INTERPRET correlation
## Explanatory & Response Variables

<table>
<thead>
<tr>
<th>Explanatory Variables (Independent Variables)</th>
<th>Response Variables (Dependent Variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car weight</td>
<td>Accident death rate</td>
</tr>
<tr>
<td>Number of cigarettes smoked</td>
<td>Life expectancy</td>
</tr>
<tr>
<td>Number of hours studied</td>
<td>SAT scores</td>
</tr>
</tbody>
</table>
Scatterplots

- A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals.
- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.
- Each individual in the data appears as a point on the graph.
Scatterplots

1. Decide which variable should go on each axis.
   
   • *Remember, the explanatory variable goes on the X-axis!*

2. Label and scale your axes.

3. Plot individual data values.
Scatterplots

Make a scatterplot of the relationship between body weight and pack weight. *Body weight is our explanatory variable.*

<table>
<thead>
<tr>
<th>Body weight (lb)</th>
<th>120</th>
<th>187</th>
<th>109</th>
<th>103</th>
<th>131</th>
<th>165</th>
<th>158</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backpack weight (lb)</td>
<td>26</td>
<td>30</td>
<td>26</td>
<td>24</td>
<td>29</td>
<td>35</td>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>
Social progress and economic development
Social Progress Index, 100=most progressive

Select to filter:
19 overperforming countries  35 underperforming countries

Asia  Europe  Latin America & Caribbean
M East & N Africa  N America  Sub-Saharan Africa

GDP per person, $’000 at PPP*, 2014

Source: Social Progress Imperative
* Purchasing-power parity at 2011 prices

http://www.economist.com/blogs/graphicdetail/2016/06/daily-chart-20
Constructing a Scatterplot: TI-Nspire

1. Enter x values into list 1 and enter y values into list 2.

2. Label each column. Label column x: weight and column y: bpack.

3. Press HOME/On, click Add Data & Statistics
Constructing a Scatterplot: TI-Nspire

4. Move the cursor to the bottom of the screen and “click to add variable”. Select weight.

5. Move the cursor to the left of the screen and “click to add variable”. Select bpack.
Constructing a Scatterplot

<table>
<thead>
<tr>
<th>Body weight (lb):</th>
<th>120</th>
<th>187</th>
<th>109</th>
<th>103</th>
<th>131</th>
<th>165</th>
<th>158</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backpack weight (lb):</td>
<td>26</td>
<td>30</td>
<td>26</td>
<td>24</td>
<td>29</td>
<td>35</td>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>
Describing Scatterplots

As in any graph of data, look for the overall pattern and for striking departures from that pattern.

• You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship.

• An important kind of departure is an outlier, an individual value that falls outside the overall pattern of the relationship.

• Also, clustering.
Words That Describe...

• **Direction (slope)**
  — Positive or Negative

• **Form**
  — Linear, quadratic, cubic, exponential, curved, non-linear, etc.

• **Strength**
  — Strong, weak, somewhat strong, very weak, moderately strong, etc.
More on Strength...

• Strength refers to how tightly grouped the points are in a particular pattern.
• Later on we use describe strength as “correlation”
Describe this Scatterplot
Describe this Scatterplot
Describe this Scatterplot
Interpreting a Scatterplot

• **Interpret**....tell what the data suggests in real world terms.

• **Example**: The data suggests that the more hours a student studied for Mrs. Daniel’s AP Stats test the higher grade the student earned. There is a positive relationship between hours studied and grade earned.
Describe and interpret the scatterplot below. The y-axis refers to backpack weight in pounds and the x-axis refers to body weight in pounds.
Describe and interpret the scatterplot below. The y-axis refers to backpack weight in pounds and the x-axis refers to body weight in pounds.

**Sample Answer:**
There is a moderately strong, positive, linear relationship between body weight and pack weight. There is one possible outlier, the hiker with the body weight of 187 pounds seems to be carrying relatively less weight than are the other group members. It appears that lighter students are carrying lighter backpacks.
Describe and interpret the scatterplot below. The y-axis refers to a school’s mean SAT math score. The x-axis refers to the percentage of students at a school taking the SAT.
Sample Answer:
There is a moderately strong, negative, curved relationship between the percent of students in a state who take the SAT and the mean SAT math score. Further, there are two distinct clusters of states and at least one possible outlier that falls outside the overall pattern.
What is Correlation?

• A mathematical value that describes the strength of a linear relationship between two quantitative variables.
• Correlation values are between -1 and 1.
• Correlation is abbreviated: r
• The strength of the linear relationship increases as r moves away from 0 towards -1 or 1.
What does “r” tell us?!

• Correlation describes what percent of variation in y is ‘explained’ by x.

• Notice that the formula is the sum of the z-scores of x multiplied by the z-scores of y.

\[ r = \frac{1}{n-1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right) \]
What does “r” mean?

<table>
<thead>
<tr>
<th>R Value</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>Perfectly linear; negative</td>
</tr>
<tr>
<td>-0.75</td>
<td>Strong negative relationship</td>
</tr>
<tr>
<td>-0.50</td>
<td>Moderately strong negative relationship</td>
</tr>
<tr>
<td>-0.25</td>
<td>Weak negative relationship</td>
</tr>
<tr>
<td>0</td>
<td>nonexistent</td>
</tr>
<tr>
<td>0.25</td>
<td>Weak positive relationship</td>
</tr>
<tr>
<td>0.50</td>
<td>Moderately strong positive relationship</td>
</tr>
<tr>
<td>0.75</td>
<td>Strong positive relationship</td>
</tr>
<tr>
<td>1</td>
<td>Perfectly linear; positive</td>
</tr>
</tbody>
</table>
How strong is the correlation? Is it positive or negative?

- 0.235
- -0.456
- 0.975
- -0.784
Describe and interpret the scatterplot below. Be sure to estimate the correlation.
Sample Answer:
As the number of boats registered in Florida increases so does the number of manatees killed by boats. This relationship is evidenced in the scatterplot by a strong, positive linear relationship. The estimated correlation is approximately $r = 0.85$.

**Answers between 0.75-0.95 would be acceptable.**
Describe and interpret the scatterplot below. Be sure to estimate the correlation.
Sample Answer:
As the number of predicted storms increases, so does the number of observed storms, but the relationship is somewhat weak. The relationship evidenced in the scatterplot is a fairly weak positive linear relationship. The estimated correlation is approximately $r = 0.35$. There is one outlier at (15, 28); very active storm year.

**Answers between 0.15 and 0.55 would be acceptable.**
Estimate the Correlation Coefficient
Estimate the Correlation Coefficient
Calculate Correlation: TI-Nspire

1. Enter x values in list 1 and y values in list 2.
2. Press MENU, then 4: Statistics
3. Option 1: Stat Calculations
4. Option 3: Linear Regression \( mx + b \)
5. \( X: a[], \ Y: b[], \) ENTER

6. Correlation = \( r \)

<table>
<thead>
<tr>
<th>Body weight (lb)</th>
<th>120</th>
<th>187</th>
<th>109</th>
<th>103</th>
<th>131</th>
<th>165</th>
<th>158</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backpack weight (lb)</td>
<td>26</td>
<td>30</td>
<td>26</td>
<td>24</td>
<td>29</td>
<td>35</td>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>

Correlation should be 0.79
Find the Correlation

| Height in Feet | 5.5 | 6.0 | 5.25 | 6.25 | 5.75 | 6.0 | 5.75 | 5.5 | 5.75 |
| Weight in pounds | 150 | 180 | 138 | 191 | 172 | 181 | 168 | 148 | 172 |

See Chapter 3: Correlation Transformation Investigation WS.
Facts about Correlation

1. Correlation requires that both variables be quantitative.
2. Correlation does not describe curved relationships between variables, no matter how strong the relationship is.
3. Correlation is not resistant. $r$ is strongly affected by a few outlying observations.
4. Correlation makes no distinction between explanatory and response variables.
5. $r$ does **not change** when we change the units of measurement of $x$, $y$, or both.
6. $r$ does not change when we add or subtract a constant to either $x$, $y$ or both.
7. The correlation $r$ itself has no unit of measurement.
R: Ignores distinctions between X & Y
R: Highly Effected By Outliers

- Left: $r = 0.4$
  - Outlier

- Right: $r = 0.7$
  - Outlier removed
Why?!

• Since r is calculated using standardized values (z-scores), the correlation value will not change if the units of measure are changed (feet to inches, etc.)

• Adding a constant to either x or y or both will not change the correlation because neither the standard deviation nor distance from the mean will be impacted.
Correlation Formula:

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:

\[
r = \frac{1}{n-1} \left[ \left( \frac{x_1 - \bar{x}}{s_x} \right) \left( \frac{y_1 - \bar{y}}{s_y} \right) + \left( \frac{x_2 - \bar{x}}{s_x} \right) \left( \frac{y_2 - \bar{y}}{s_y} \right) + \cdots + \left( \frac{x_n - \bar{x}}{s_x} \right) \left( \frac{y_n - \bar{y}}{s_y} \right) \right]
\]

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

The following scatterplot shows the arm span (in centimeters) and height (in centimeters) for a random sample of 19 twelfth graders. The correlation is $r = 0.902$.

(a) Explain why it is incorrect to say that the correlation is 0.902 centimeter.

(b) What would happen to the correlation if height was measured in inches instead of centimeters? Explain your answer.

(c) What would happen to the correlation if height was used as the explanatory variable and arm span was used as the response variable?
Check Your Understanding:

(a) Explain why it is incorrect to say that the correlation is 0.902 centimeter.

Correlation is a standardized measurement and therefore has no units.

(b) What would happen to the correlation if height was measured in inches instead of centimeters? Explain your answer.

There would be no change in the correlation because it is a standardized measurement. Changes in units do not change the relative relationship.

(c) What would happen to the correlation if height was used as the explanatory variable and arm span was used as the response variable?

There would be no change. The correlation formula multiplies the $x$ values by the $y$ values, so flipping the $x$ values and the $y$ values will not change the product/the correlation.