Chapter 6

Scatterplots, Association, and Correlation
6.1 Scatterplots

The resolution with the greatest chance of success was "to enjoy life more." 44% of people who bristle to compactness identify themselves as "natural salespeople" are more than 50% more likely to think they'd survive the year if it were a_REPEAT. You might think that the most common street name in the United States would be "First," but according to the U.S. Census Bureau, the second most common street name is "Second," with 10,131 occurrences. Strangely enough, the third place is "Third," with 10,131 occurrences, while "First" is in third place with 9,888 occurrences. 3% of all restaurant meals include French fries. 15% of people who dislike mayonnaise are good dancers, compared to 2% of people in general. In 2001, the American diet was 13% vegetables, 20% fruits, 15% whole grains, and 52% fat. In 2005, the American diet was 13% vegetables, 20% fruits, 15% whole grains, and 52% fat. In 2009, the American diet was 13% vegetables, 20% fruits, 15% whole grains, and 52% fat. In 2010, the American diet was 13% vegetables, 20% fruits, 15% whole grains, and 52% fat.

De VEAUx, VEELLEMAN, BOCK
Scatterplot of Hurricane Predictions

- **Scatterplots** exhibit the relationship between two variables.
- Used for detecting patterns, trends, relationships, and extraordinary values.
The Direction of the Association

• **Negative Direction:** As one goes up, the other goes down.

• **Positive Direction:** As one goes up, the other goes up also.

• **No Direction:**
Form

• **Linear:** The points cluster near a line.

• **Gently curves in a direction.** May be able to straighten with a transformation.

• **Curves up and down.** Difficult to straighten
Strength of the Relationship

• Strong Linear Relationship:

• Moderate Linear Relationship:

• No Linear Relationship:
Outliers

• An outlier is a point on a scatterplot that stands away from the overall pattern of the scatterplot.

• Outliers are almost always interesting and always deserves special attention.
Example: Comparing Prices Worldwide

- Describe the patterns.
  - *Cost of Women’s Clothes* and *Food Costs* are positively associated. The association is straight.
  
  - Higher clothes costs correspond to higher food costs.
Example: Comparing Prices Worldwide

• Describe the patterns.
  • *Average Hourly Wage* and *Hours to Earn an iPod* are negatively associated. The association is not straight.

• Higher average hourly wages correspond to fewer hours to earn an iPod.
Example: Comparing Prices Worldwide

- Describe the patterns.
  - There seems to be no association between Vacation Days and Food Costs.

- Knowing the vacation days per year tell us nothing about what food will cost.
Roles of Variables

• **Response Variable** \((y)\): The variable of interest. It is what we want to predict.

• **Explanatory or Predictor Variable** \((x)\): The variable that we use to provide information or a prediction of the response variable.

• Choosing the response variable and the explanatory variable depends on how we think about the problem.
Which is Response and Which is Explanatory?

• Do baseball teams that score more runs also sell more tickets?
  • Tickets = Response \((y)\), \quad \text{Runs} = \text{Explanatory} \((x)\)

• Do students with higher SAT scores get better grades?
  • Grades = Response \((y)\), \quad \text{SAT score} = \text{Explanatory} \((x)\)

• Can we estimate a person’s BMI by measuring their wrist size?
  • BMI = Response \((y)\), \quad \text{Wrist Size} = \text{Explanatory} \((x)\)
Correlation
Height and Weight

• How strong is the association between height and weight?

• It looks positive. How do we measure it?

• Positive association means above average height predicts above average weight.

• Green → +, Red → -, Blue → No Association
Correlation

• For the green dots: $z$-scores have the same sign, so multiplying the $z$-scores produces a positive value.

• For the red dots: $z$-scores have opposite signs, so multiplying the $z$-scores produces a negative value.

• Define the correlation coefficient by an almost average product of the $z$-scores:

$$ r = \frac{\sum z_xz_y}{n-1} $$
Finding the Correlation Using StatCrunch

• Enter the data in two columns. Then go to Stat → Regression → Simple Linear.
• Choose \( x \) and \( y \) and hit Calculate.
StatCrunch gives a lot of information. For now we just want $r = 0.7452$. 
Assumptions and Conditions for Correlation

• To use $r$, there must be a true underlying linear relationship between the two variables.

• The variables must be quantitative.

• The pattern for the points of the scatterplot must be reasonably straight.

• Outliers can strongly affect the correlation. Look at the scatterplot to make sure that there are no strong outliers.
Example: Clothes and Food Revisited

- The scatterplot indicates a straight-line pattern. The variables are both quantitative ($), and there are no strong outliers away from the linear pattern.

- The correlation of $r = 0.774$ represents a strong positive association.
Example: Vacation and Food Revisited

• The scatterplot indicates that there may be no underlying linear relationship between vacation days and food costs.

• The correlation is $r = -0.022$ supports that there may be no linear association between the two.
Systolic and Diastolic Blood Pressure

- **Think →**

- **Plan:** Examine the relationship between the two types of blood pressure.

- **Variables:** Systolic (SBP) and Diastolic (DBP) blood pressure measured for each of 1406 people from Framingham, MA.
Blood Pressure Continued

- **Plot:**
  - Both variables are quantitative.
  - The scatterplot is quite straight.
  - There are no extreme outliers.

- Therefore, correlation is a suitable measure of association.
Blood Pressure Continued

• Show →
  • **Mechanics**: Use StatCrunch to find $r = 0.792$.

• Tell →
  • **Conclusion**: The correlation of 0.792 indicates that there is a positive association between systolic and diastolic blood pressure. There is a tendency for DBP to be high when SBP is high.
Properties of Correlation

• $r > 0 \rightarrow$ positive association

• $r < 0 \rightarrow$ negative association

• $-1 \leq r \leq 1$, with $r = -1$ only if the points all lie exactly on a negatively sloped line and $r = 1$ only if the points all lie exactly on a positively sloped line.

• Interchanging $x$ and $y$ does not change the correlation.

• $r$ has no units.
Properties of Correlation Continued

- Changing the units of $x$ or $y$ does not affect $r$.
- Measuring in dollars, cents, or Euros will all produce the same correlation.

- Correlation measures the strength of the linear association between the two variables.

- Correlation is sensitive to outliers. An extreme outlier can cause a dramatic change in $r$.

- The adjectives *weak*, *moderate*, and *strong* can describe correlation, but there are no agreed upon boundaries.
6.3

Warning: Correlation ≠ Causation
Storks and Babies

• There is a clear positive association between the number of storks and the population.

• This does not prove that an increase in storks has caused an increase in babies being born.

• Causation is in reverse. Storks nest on house chimneys, so the increased population has increased nesting sites.
Reasons for Correlation

- Causation is a possibility, but more must be done to prove causation.

- The causation could be in reverse ($y$ causes $x$)

- A lurking variable may cause both.
  - Number of gray hairs and number of wrinkles are strongly correlated, but dyeing hair black does not undo wrinkles. Age is the lurking variable that causes both to increase.
How to Report Correlation

• **Bad:** Raising salaries increases productivity.
• **Good:** Employees with higher salaries tend to be more productive.

• **Bad:** $r = -0.99$. This proves that drinking more red wine lowers cholesterol.
• **Good:** There is a strong negative association between red wine consumption and cholesterol level.

• **Bad:** A child that has two educated parents will graduate from college.
• **Good:** Children whose parents are educated are more likely to graduate from college.
Correlation Tables

• Often surveys or experiments contain many quantitative variables. A table can be used to show the correlation for each pair of variables.

• Why do all the diagonals have 1.000?

<table>
<thead>
<tr>
<th></th>
<th>Assets</th>
<th>Sales</th>
<th>Market Value</th>
<th>Profits</th>
<th>Cash Flow</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales</td>
<td>0.746</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market Value</td>
<td>0.682</td>
<td>0.879</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profits</td>
<td>0.602</td>
<td>0.814</td>
<td>0.968</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cash Flow</td>
<td>0.641</td>
<td>0.855</td>
<td>0.970</td>
<td>0.989</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Employees</td>
<td>0.594</td>
<td>0.924</td>
<td>0.818</td>
<td>0.762</td>
<td>0.787</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Cautions about Correlation Tables

- Don’t blindly read a correlation table without first looking at each of the many scatterplots to check for linearity.

- You must also check for outliers.
*6.4

Straightening Scatterplots
Shutter Speed and f/stop

• What’s wrong with concluding that since $r = 0.979$ that shutter speed and f/stop are strongly correlated?

• The scatter plot indicates that the two variables are **not linearly related**.

• **Solution:** Use a transformation, much like we did for histograms.
Re-expressing with log?

- How about the log transformation?
- This is even less linear!
- Can you think of a transformation that will straighten out the curve?
Re-Expressing

- Clearly the shutter speed and the square of the f/stop are linearly related.

- Now we can conclude that there is a very strong correlation between shutter speed and the square of the f/stop.
Guidelines for Re-Expressions

- Scatterplot bends *downwards*.  
  ➢ $y^2$

- Scatterplot is *linear*.  
  ➢ No Change

- For data that is a *count*  
  ➢ $y^{1/2}$

- For data that is *always positive*  
  ➢ $\log y$

- If nothing else seems to work try  
  ➢ $y^{-1/2}$

- For *ratios* such as miles per gallon  
  ➢ $1/y$
What Can Go Wrong?

• Don’t say “correlation” when you mean “association.”
  • Correlation implies a **linear** relationship. Association means any relationship.

• Don’t correlate categorical variables.
  • It makes no sense to say *car model* and *personality type* are correlated.

• Don’t confuse correlation with causation.
  • Correlation only implies general tendencies.
What Can Go Wrong?

- Make sure the association is linear.
  - Always look at the scatterplot to check.

- Don’t assume the association is linear just because the correlation coefficient is high.
  - Always look at the scatterplot to check.

- Beware of outliers!
  - $r = 0.5$, but there is no correlation between shoe size and IQ.