Correlation

Course: Statistics 1
Lecturer: Dr. Courtney Pindling
Measures of Association

*Between Two Variables*

Measures of Linear Associations:

- Scatter Plots
- Covariance
- Correlation Coefficient
- Coefficient of Determination
Scatter Plot 1

- Scatter Plot
- Positive linear association
- As the Ability Index increase so does value of the y-axis
- Positive correlation
Scatter Plot 2

- Scatter Plot
- Negative linear association
- As the *Ability Index* increase so the value on the y-axis decreases
- Negative correlation
Scatter Plot 3

- Scatter Plot
- No linear association
- As the Ability Index increase there seem to be no trend in the value on the y-axis
- No correlation
Covariance

- A measure of the linear association between variables
  - *Positive* indicates positive linear relationship
  - *Negative* indicates a negative linear relationship
  - *Values close to zero* indicates no linear relationship

- It is dependent upon the units of measurement for x and y variables
  - Height in inches would give a larger covariance than height in feet; even with same degree of association
  - So the magnitude of the covariance is not significant
Covariance Formula

\[ s_{xy} = \frac{\sum (X_i - M_x)(Y_i - M_y)}{n - 1} \]

Where

- \( X_i \) is data point \( i \) for \( X \) variable
- \( Y_i \) is data point \( i \) for \( Y \) variable
- \( M_x \) is mean for \( X \) variable
- \( M_y \) is mean for \( Y \) variable
- \( n \) is sample size
## Covariance Example

The formula for covariance is given by:

\[
s_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}
\]

where

- \(X_i\) and \(Y_i\) are individual data points,
- \(\bar{X}\) and \(\bar{Y}\) are the means of \(X\) and \(Y\) respectively,
- \(n\) is the number of data points.

Given data points:

<table>
<thead>
<tr>
<th>(X_i)</th>
<th>(Y_i)</th>
<th>(X_i - \bar{X})</th>
<th>(Y_i - \bar{Y})</th>
<th>((X_i - \bar{X})(Y_i - \bar{Y}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>2</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>-2</td>
<td>-10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>37</td>
<td>-2</td>
<td>-13</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>2</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>-1</td>
<td>-5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\bar{X} = 3, \quad \bar{Y} = 50
\]

Sum of \(X_i\) = 0, Sum of \(Y_i\) = 0, Sum of \((X_i - \bar{X})(Y_i - \bar{Y})\) = 99

\[
s_{xy} = \frac{99}{9} = 11
\]
Correlation Coefficient

- A measure of the linear association between variables
  - *Positive* indicates positive linear relationship
  - *Negative* indicates a negative linear relationship
  - Values close to *zero* indicates no linear relationship

- It not affected by the units of measurement for x and y variables
  - Pearson product moment correlation coefficient or
  - Sample correlation coefficient, $r$
Correlation Coefficient Formula 1

\[ r_{xy} = r = \frac{s_{xy}}{s_x s_y} \]

Where

- \( r_{xy} \) = sample correlation coefficient,
- \( s_{xy} \) = sample covariance,
- \( s_x \) = sample standard deviation of \( x \), and
- \( s_y \) = sample standard deviation of \( y \)
Knowing the covariance and the standard deviations of each variable we can compute the sample correlation coefficient, $r$

- Covariance = 11, $SD_x = 1.49$, $SD_y = 7.93$
- So Pearson $r = 11/(1.49 \times 7.93) = 0.93$

**Descriptive Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>50.0000</td>
<td>7.93025</td>
<td>10</td>
</tr>
<tr>
<td>X</td>
<td>3.0000</td>
<td>1.49071</td>
<td>10</td>
</tr>
</tbody>
</table>
Correlation Coefficient Formula 2

Computational Formula

$$r = \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}}$$
## Correlation Example

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X^2$</th>
<th>$Y^2$</th>
<th>$XY$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49</td>
<td>4</td>
<td>2401</td>
<td>98</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>25</td>
<td>3136</td>
<td>280</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>1</td>
<td>1600</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>9</td>
<td>2809</td>
<td>159</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>16</td>
<td>2809</td>
<td>212</td>
</tr>
<tr>
<td>1</td>
<td>37</td>
<td>1</td>
<td>1369</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>25</td>
<td>3844</td>
<td>310</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>9</td>
<td>2209</td>
<td>141</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>16</td>
<td>3364</td>
<td>232</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>4</td>
<td>2025</td>
<td>90</td>
</tr>
</tbody>
</table>

| $\sum X = 30$ | $\sum Y = 500$ | $\sum X^2 = 110$ | $\sum Y^2 = 25566$ | $\sum XY = 1599$ |
Correlation Example cont.

\[
\begin{align*}
\Sigma X &= 30 \\
\Sigma Y &= 500 \\
\Sigma X^2 &= 110 \\
\Sigma Y^2 &= 25566 \\
\Sigma XY &= 1599
\end{align*}
\]

\[
r = \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} = \frac{10(1599) - (30)(500)}{\sqrt{10(110) - 30^2} \cdot \sqrt{10(25566) - 500^2}}
\]

\[
r = \frac{990}{1063.96} = 0.93
\]
Coefficient of Determination

- Tells us how much of the variation in the dependent variable, $Y$, is due to change in the independent variable, $X$
- Coefficient of Determination is $r^2$
- For example, $r^2 = 0.8649$
  - Therefore, 86.49% of the variation in $Y$ is associated with the change in $X$ or
  - 13.51% of variation in $Y$ is due to other factors
Properties of $r$

- Required Interval or Ratio Scales
- Relationship between $X$ and $Y$ must be linear
- Requires pairs of values for $X$ and $Y$
- The standard deviation about $Y$ for a given value of $X$ is about the same (homogeneity)
- The sample size, $N$, has little effect on $r$, but is used to make compute the significance of $r$
Limitations of $r$

- Correlation does not mean causality
  - *Patients’ height may correlate with their blood pressure, but it does not mean that their height is the cause for their blood pressure*

- When $r$ is based on sample data, you may get a strong positive or negative correlation purely by chance, even though there is no relationship between the two variables
  - *Patients’ shoe size in the hospital may correlates with their blood pressure at time of admission, but there may be no relationship between the two*
Other Correlational Methods

- Pearson $r$ is computed on interval and ratio scales
- Spearman $r$, is Pearson $r$ computed for ordinal scale
- Other correlational methods based on modified Pearson $r$ or probability functions for specific applications
# Correlation Methods

<table>
<thead>
<tr>
<th>Correlation Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point-biserial r</strong></td>
<td>One dichotomous variable (yes/no; male/female) and one interval or ratio variable</td>
</tr>
<tr>
<td><strong>Biserial r</strong></td>
<td>One variable forced into a dichotomy (grade distribution dichotomized to “pass” and “fail”) and one interval or ratio variable</td>
</tr>
<tr>
<td><strong>Phi coefficient</strong></td>
<td>Both variables are dichotomous on a nominal scale (male/female vs. high school graduate/dropout)</td>
</tr>
<tr>
<td><strong>Tetrachoric r</strong></td>
<td>Both variables are dichotomous with underlying normal distributions (pass/fail on a test vs. tall/short in height)</td>
</tr>
<tr>
<td><strong>Correlation ratio</strong></td>
<td>There is a curvilinear rather than linear relationship between the variables (also called the eta coefficient)</td>
</tr>
<tr>
<td><strong>Partial correlation</strong></td>
<td>The relationship between two variables is influenced by a third variable (e.g., mental age and height, which is influenced by chronological age)</td>
</tr>
<tr>
<td><strong>Multiple R</strong></td>
<td>The maximum correlation between a dependent variable and a combination of independent variables (a college freshman’s GPA as predicted by his high school grades in Math, chemistry, history, and English)</td>
</tr>
</tbody>
</table>