AMS7: WEEK 8. CLASS 1

Correlation
Monday May 18th, 2015
Type of Data and objectives of the analysis

- Paired sample data (Bivariate data)
- Determine whether there is an association between two variables
- This association is called **CORRELATION**
- We will consider mostly **linear associations** (scatterplots approximate a straight-line pattern).
Anscombe’s Scatterplots: Are these relationships linear?
Linear Correlation coefficient

- **Notation:**
  
r = Linear correlation coefficient calculated for the sample data

  \[ \rho = \text{Linear correlation coefficient for the population values.} \]

- **r** measures the strength of the linear association between the paired \( x \) and \( y \) values in the sample.

- This coefficient is sometimes referred as the Pearson correlation coefficient in honor to Karl Pearson (1857-1936) who originally developed it.
Requirements when making inferences about $r$

- Sample of paired data $(x,y)$ is a random sample of quantitative data
- Scatterplot must have an approximate straight-line pattern
- Outliers should be removed from data
- **Formal requirement**: $(x,y)$ data must have a bivariate normal distribution distribution.
Notation

- \( n \): Number of pairs of data
- \( \sum x \): sum of all \( x \) values
- \( \sum x^2 \): each \( x \) value is squared and then added
- \( (\sum x)^2 \): \( x \) values should be added and the total squared
- \( \sum xy \): Each \( x \) value is multiplied by its corresponding \( y \) value and the sum must be calculated.
- \( r \): linear correlation coefficient for the sample
- \( \rho \): linear correlation coefficient for the population
Formula for calculation

\[ r = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \]
Example of calculation

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x.y</th>
<th>x²</th>
<th>y²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>18</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>20</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>Σx=10</td>
<td>Σy=20</td>
<td>Σxy=48</td>
<td>Σx²=36</td>
<td>Σy²=120</td>
</tr>
</tbody>
</table>
Scatterplot of data
Applying the formula

- The formula can be always applied even though the scatterplot does not look like a straight-line

\[
r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}
\]

\[
= \frac{4(48) - (10)(20)}{\sqrt{4(36) - (10)^2} \sqrt{4(120) - (20)^2}} = \frac{-8}{\sqrt{44 \cdot 80}} = -0.135
\]
Properties of the linear correlation coefficient

• Values of r are always between -1 and +1
• The value of r does not change if all values of either variable are converted to a different scale (for example: if you covert inches to cms.)
• Exchanging the values of x and y does not change the value of r
• r measures the strength of a linear association. It is not good to measure associations that are non-linear.
Back to Anscombe’s Scatterplots: Are these relationships linear?
Comments on the graph

- The four $y$ variables have the same mean (7.5), standard deviation (4.12) and correlation (0.816).
- The first one (top left) seems to follow the assumption of normality.
- The second one (top right) is not distributed normally; there is a relationship between the two variables but this relationship is non-linear.
- In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- The fourth example (bottom right) shows another case when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.
Comments on the Graph (Cont.)

• These examples indicate that the correlation coefficient, as a summary statistic, cannot replace visual examination of the data.

• If we find that the coefficient of linear correlation between x and y is significant, we can find a linear equation that expresses y in terms of x. This equation can be used to predict the values of y given values of x. This equation is called the **regression equation**.

• The value $r^2$ is the proportion of the variation in y that is explained by the linear association between x and y.
Common errors involving correlation

• Correlation does not imply causality (lurking variables)
• When data averages are used correlation can be inflated
• A low correlation values doe not imply a lack of relationship between two variables. The relationship can be nonlinear as in the graph shown before.
Hypothesis testing for correlation: Method 1

- Determine whether there is a significant linear correlation between two variables using a formal hypothesis test.

**Hypothesis test:**
- $H_0: \rho = 0$
- $H_1: \rho \neq 0$
  (two-tailed test, although a one-tailed test is possible)
- **Test statistic:** Use a t Student distribution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

with n-2 degrees of freedom
Hypothesis testing for correlation: Method 1 (cont.)

**Conclusion:**

- If $|t| >$ critical value from Table A-3, reject the null hypothesis and conclude that **there is a significant linear correlation.**
- If $|t| \leq$ critical value, fail to reject the null hypothesis. **There is no sufficient evidence to conclude that there is a significant linear correlation.**
Hypothesis testing for correlation: Method 2

- Comparison with critical values from table A-6
- The test statistic is $r$

**Conclusion:**

- If $|r| >$ critical value from Table A-6, reject the null hypothesis and conclude that there is a significant linear correlation.
- If $|r| \leq$ critical value, fail to reject the null hypothesis. There is no sufficient evidence to conclude that there is a significant linear correlation.
Example: heights of fathers and sons

• Sect. 9.2, #8.

• Data about heights of fathers and sons:

<table>
<thead>
<tr>
<th>Height of father</th>
<th>70</th>
<th>69</th>
<th>64</th>
<th>71</th>
<th>68</th>
<th>66</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height of Son</td>
<td>62.5</td>
<td>64.6</td>
<td>69.1</td>
<td>73.9</td>
<td>67.1</td>
<td>64.4</td>
<td>71.1</td>
</tr>
</tbody>
</table>

• Construct a scatterplot, find the value of $r$ and use a significance level of $\alpha=0.05$ to determine whether there is a significant linear correlation between the two variables.
Scatterplot: Sons’ heights vs. fathers’ heights
Example (Cont.)

- The scatterplot does not approximate a straight-line pattern.
- A hypothesis test is not indicated.
- We will, however, perform the test in order to demonstrate the procedure.
- We now calculate the needed sums and the linear correlation coefficient.
Example (Cont.)

• $\sum x = 482$
• $\sum y = 472.7$
• $\sum x^2 = 33254$
• $\sum x \cdot y = 32576.3$
• $\sum y^2 = 32020.41$

\[ r = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \]
\[ = \frac{7(32576.3) - (482)(472.7)}{\sqrt{7(33254) - (482)^2} \sqrt{7(32020.4) - (472.7)^2}} = 0.342 \]
Hypothesis testing: Method 2

• The critical value from Table A-6 for \( n = 7 \) and with \( \alpha = 0.05 \) is 0.754.

• Since \( r = 0.342 \), there is not a significant linear correlation between heights of fathers and their sons.
Method 1

- The formal hypothesis test follows (alpha=0.025):
  - $H_0: \rho=0$
  - $H_1: \rho \neq 0$

- The test statistic is ($t$ Student with n-2 d.f.)
  
  $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.342}{\sqrt{\frac{1-0.342^2}{7-2}}} = 0.815$$

- The critical values are found in Table A-3 using df = 7 – 2=5, and are $t_{0.025}=\pm 2.571$. 
Conclusion

- The critical values are found in Table A-3 using degrees of freedom = n – 2, and α/2 = 0.025, are ±2.571.
- Since the test statistic is between -2.571 and 2.571, we fail to reject the null hypothesis.
- There is not sufficient sample evidence to conclude that there is a significant linear correlation between father’s height and their son’s height. The data does not suggest that taller fathers tend to have taller sons.