Section 10-2 Correlation

A simple relationship involves two variables,
the independent variable (can be controlled or manipulated; the explanatory or predictor variable),
the dependent variable (cannot be controlled, called the outcome or response variable).

Correlation is a method used to determine whether a relationship exists between two variables.

A Scatter plot is a graph of the ordered pairs, \((x,y)\), of the independent and dependent variable. This is a visual way to represent the nature of the relationship between two variables. The scales for the graph are often truncated (do not start at zero) and may be different for each variable. It is important to label such graphs clearly to avoid misrepresenting the data.

The correlation coefficient is a statistical value calculated from sample data to measure the strength and direction of a linear relationship between two variables. In this class we will use the “Pearson product moment correlation coefficient”.

\[ r = \text{the sample correlation coefficient}, \quad \rho = \text{the population correlation coefficient} \]

We will use EXCEL to calculate \( r \) using the regression option in the data analysis tool pack. The correlation coefficient is “Multiple \( R \)”. ROUND to 3 decimal places.

The range of the correlation coefficient is from -1 to 1.

Type of linear relationship:

<table>
<thead>
<tr>
<th>Strong negative</th>
<th>none</th>
<th>strong positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
</tr>
</tbody>
</table>

When testing the significance of the correlation coefficient, the hypotheses are:

\( H_0 : \rho = 0 \) (there is no linear relationship) \( H_1 : \rho \neq 0 \) (there is a linear relationship)

We will use critical values from Table I to test the significance of the correlation coefficient. When using this table the test value is \( r \).

Correlation does not imply causation:

There could be a direct cause and effect relationship, \( x \) causes \( y \).
There could be a reverse cause and effect relationship \( y \) causes \( x \).
The relationship could be caused by a third variable.
There could be interrelationships among many variables.
The relationship could be coincidental.
Example 1
For the following data draw a scatter plot and then test the significance of \( r = 0.933 \) at the 0.05 level.

<table>
<thead>
<tr>
<th>1st Test</th>
<th>73</th>
<th>86</th>
<th>93</th>
<th>92</th>
<th>72</th>
<th>65</th>
<th>58</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd Test</td>
<td>70</td>
<td>80</td>
<td>96</td>
<td>85</td>
<td>68</td>
<td>68</td>
<td>62</td>
<td>78</td>
</tr>
</tbody>
</table>

\[ H_0: \rho = 0 \quad H_1: \rho \neq 0 \]

- C.V. FROM \( t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \)
- TEST VALUE IS \( r = 0.933 \)
- REJECT \( H_0: \rho = 0 \)
- (NO LIN. REL.)
- There is enough ev. to reject claim \( \rho = 0 \)
- (So there is a linear relation between test 1 and test 2 scores)

Example 2
For the following data draw a scatter plot and then test the significance of \( r = -0.912 \) at the 0.01 level.

<table>
<thead>
<tr>
<th>Absences</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Grade</td>
<td>96</td>
<td>91</td>
<td>78</td>
<td>83</td>
<td>75</td>
<td>68</td>
<td>56</td>
</tr>
</tbody>
</table>

\[ H_0: \rho = 0 \quad H_1: \rho \neq 0 \]

- C.V. FROM \( t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \)
- TEST VALUE IS \( r = -0.912 \)
- REJECT \( H_0 \)
- There is a linear relation between absences and final grades.
Section 10-3 Regression

If the value of the correlation coefficient **is significant**, the next step is to determine the line of regression (line of best fit).

"Best fit" means that the sum of the squares of the vertical distance each point is from the line is as small as possible.

The regression line is used to note trends and make predictions. To make valid predictions using regression:

1) For a specific x value, the y values must be normally distributed.
2) The standard deviation of each y value must be the same.

The line of regression: \( a = \) the y intercept, \( b = \) the slope

\[
y' = a + bx
\]

This corresponds to slope intercept form, \( y = mx + b \);
in either form the slope is the coefficient of \( x \).

We will use EXCEL to calculate \( a \) and \( b \) using the **regression** option in the data analysis tool pack. The y intercept, \( a \), is the "intercept coefficient", the slope, \( b \), is the "X variable 1 coefficient". ROUND to 3 decimal places.

**Notes:**
1) Regression graphs are often truncated, do not use y intercept to graph.
2) **The sign of the slope is the sign of \( r \).**
3) The regression line will always pass through \((\bar{x}, \bar{y})\).
4) Predictions made beyond the bounds of the data must assume present trends and conditions will continue.

If \( r \) is **NOT** significant, then the line of regression does not apply, and the best estimate of \( x \) is the mean of the \( x \) values, and the best estimate of \( y \) is the mean of the \( y \) values.
Exercise #18 from text
A football fan wishes to see how the number of pass attempts (not completions) relates to the number of yards gained for quarterbacks in past NFL season playoff games. The data are shown for five quarterbacks. Describe the relationship.

<table>
<thead>
<tr>
<th>Pass attempts (x)</th>
<th>116</th>
<th>90</th>
<th>82</th>
<th>108</th>
<th>92</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yards gained (y)</td>
<td>1001</td>
<td>823</td>
<td>851</td>
<td>873</td>
<td>839</td>
</tr>
</tbody>
</table>

A) Graph the scatter plot:

[Graph image]

b) Given r = 0.819, test the significance of the correlation coefficient at \( \alpha = 0.01 \), using table I and explain the relationship.

1. \( H_0: \rho = 0 \) \hspace{1cm} \( H_1: \rho \neq 0 \)
   NO RELATION \hspace{1cm} IS A LIN. RELATIONSHIP

2. C.V. FROM \( \chi^2 \)

\[
- \frac{959}{9.59} = -99.9 \\
\text{DNR} \quad \text{C.V.} \quad 9.59
\]

3. TEST VALUE
\( r = 0.819 \)

4. DNR \( H_0: \rho = 0 \)

5. THERE IS NO LINEAR REL. BETWEEN PASSES & YDS GAINED

c) Since the linear relationship is not significant, the best estimate for either variable is that variable’s mean. If a quarterback makes 100 pass attempts, predict the yards gained.

IF \( x = 100 \), PREDICT \( y \).

FIND THE MEAN OF YARDS GAINED = 877

NOTE THAT THIS WOULD BE THE SAME PREDICTION FOR 85 PASS ATTEMPTS, OR 95 ETC.
The table below shows the IQ of 8 students and the number of hours of TV each student views per week.

<table>
<thead>
<tr>
<th>IQ (x)</th>
<th>105</th>
<th>125</th>
<th>135</th>
<th>100</th>
<th>115</th>
<th>130</th>
<th>140</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours of TV (y)</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>13</td>
<td>15</td>
<td>8</td>
<td>2</td>
<td>14</td>
</tr>
</tbody>
</table>

d) Graph the scatter plot:

![Scatter plot with regression line]

This regression line equation is from Part C

b) Given $r = -0.873$, test the significance of the correlation coefficient at $\alpha = 0.01$, using table I and explain the relationship.

1. $H_0: \rho = 0$ (No relationship)
2. $H_1: \rho \neq 0$ (Linear relationship)
3. $r = -0.873$
4. Reject $H_0$
5. There is a significant linear relationship (Negative)

\[ y' = a + bx \]

From Excel: $a = 38.731$, $b = -0.246$. The line of regression is: $y' = 38.731 - 0.246x$

Plot the regression line on the scatter plot. Choose two $x$ values in the range of the given data and find the corresponding $y$ values to generate two ordered pairs and graph.

- $x = 107$: $y' = 38.731 - 0.246(107) = 12.409$ (107, 12)
- $x = 138$: $y' = 38.731 - 0.246(138) = 4.783$ (138, 5)

d) How many hours of TV per week would a student with an IQ of 120 be predicted to watch?

\[ x = 120 \]
\[ y' = 9.3 \]
Factoid: Crickets make their chirping sounds by rapidly sliding one wing over the other. The faster they move their wings, the higher the chirping sound that is produced. Scientists have noticed that crickets move their wings faster in warm temperatures than in cold temperatures. Therefore, by listening to the pitch of the chirp of crickets, it is possible to tell the temperature of the ground. The table below gives the recorded pitch (in vibrations per second) of a cricket chirping recorded on 15 different occasions.

<table>
<thead>
<tr>
<th># of chirps (x)</th>
<th>20</th>
<th>16</th>
<th>20</th>
<th>18</th>
<th>17</th>
<th>16</th>
<th>15</th>
<th>17</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>16</th>
<th>17</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp. (y)</td>
<td>89</td>
<td>72</td>
<td>93</td>
<td>84</td>
<td>81</td>
<td>75</td>
<td>70</td>
<td>82</td>
<td>69</td>
<td>83</td>
<td>80</td>
<td>83</td>
<td>81</td>
<td>84</td>
</tr>
</tbody>
</table>

a) Graph the scatter plot:

b) Find the correlation coefficient. Using Excel this is “Multiple R” with the sign of the slope as found below. Test the significance of the correlation coefficient at \( \alpha = 0.01 \), using table I and explain the relationship. \( r = 0.825 \)

1. \( H_0: \rho = 0 \)  
2. \( H_1: \rho \neq 0 \)  
3. TEST \( r = 0.825 \)  
4. REJECT \( H_0: \rho = 0 \)  
5. THERE IS A LINEAR RELATION BETWEEN # OF CHIRPS AND TEMP.

c) Find the linear regression equation for this data if appropriate. Round the slope and the y-intercepts to the nearest thousandth. Using Excel in the coefficients column, the slope is “x variable 1”, the intercept is “intercept”. Plot the regression line on the scatter plot. Choose two x values in the range of the given data and find the corresponding y values to generate two ordered pairs and graph.

\[
\alpha = 26.742 \\
\beta = 3.216 \\
y' = \alpha + \beta x \\
y' = 26.742 + 3.216x
\]

d) If you had a listening device and used it in the morning when you woke up and measured a cricket chirping at a rate of 18 chirps per second, how warm would you say the temperature is?

\[
y' = 26.742 + 3.216x \\
y' = 26.742 + 3.216(18) = 85 \text{ (85, 85)}
\]

e) If the temperature reached 95°F, at what rate would you expect those little guys to be chirping?

\[
x' = 26.742 + 3.216x \\
95 = 26.742 + 3.216x \\
68.258 = 3.216x \\
\frac{68.258}{3.216} = x \\
x = 21 \text{ CHIRPS PER SECOND}
\]