Correlation

Are two numeric measurements related to each other? How related are they?

Outcome given as correlation coefficient $r$

$r$ varies from $-1$ to $+1$ (positive versus negative relationship)

(Correlation .79)
What does the correlation coefficient “$R$” mean?

$R$ indicates the strength and direction of the correlation. Look at the above graph of “income and years of education”. The coefficient is +.79. Since it is positive in means that higher incomes are related to higher education, and lower incomes to lower education. The correlation is positive because they both move in the same direction which results in an upwardly sloping line. If there were a perfect +1 correlation all of the dots would fall exactly on the line. This would mean that every year of education would be related to exactly a certain dollar amount in salary. Exact correlations are not often found in behavioral studies, but are in nature. For example, if you filled a container with water and weighed it, the weight of the water is perfectly correlated with the weight of the water; all of the volume and weight dots would be on the line.

A negative correlation occurs on the “GPA and hours of TV graph” (-.63). It is negative because more TV is correlated with a lower GPA, and a higher GPA is correlated with less TV; the two move in opposite directions. A perfect negative correlation of -1 would have all of the dots on the line and slope downward.

Coefficients close to zero indicate that there is no (or very little) relationship between the variables. The line would have no slope and the dots would be scattered all over the chart.
Click on this chart to look at an interactive graphic representation of what the coefficients represent. Click on the chart to add data points and see how the correlation changes.

**What does “p” mean?**

*P* is not specific to correlation, but is used for statistics in general. It represents the probability that the data that was analyzed could have occurred by random chance. You are interested in results that are NOT random. Imagine that you teach English to two classes using two different methods. You then give the students a test to see if one method leads to higher test scores that the other. You want to be able to show that the test scores in one class are significantly better than the other. If the difference in scores could be obtained by chance then you can’t conclude that one method is better than the other.

Significance is defined as a smaller than 1 in 20 probability of occurring by chance. The statistics programs make this calculation. In sum, if the *p* is 0.05 or SMALLER, then means the results are significant, that is they is a small probability of getting the results by chance. The SMALLER the *p* value the BETTER (more significant, less likely due to chance) the results.

**What is a regression line?**

It is the line that can be drawn that is closest to all the points in the graph.

**What is degrees of freedom?**

It is a number used in the statistical calculation. For correlations is is the number of pairs of numbers that are correlated minus 2.

**What are one- and two-tailed correlations?**

Generally use two-tailed, unless you are positive of the direction of the correlation beforehand.

**How is r related to variance?**

Example: Height vs. weight correlation isn't *r*=+1. Besides height, there may be something that also correlates with weight (e.g. body fat).

So if correlation between height and weight is *r*=.85 the variance is *r* squared (.72). This means that 72% of the variance in height and weight is accounted. But, about 28% of the variance is not accounted for by height and weight and must be due to another variable or variables.
How do you report the results of a correlation?

If \( r \) is -0.81, the degrees of freedom are 74, and the significance is less than 0.001 then:

\[
    r(74) = -0.81, \ p < 0.001, \ \text{two-tailed}
\]

Some possible linguistic correlations:

- Number of months in a foreign country and linguistic abilities in the country's language (positive or negative?)

  What would this mean? \( R = 0.56, \ p < .03 \)

  What would this mean? \( R = 0.56, \ p < .07 \)

- Number of native dialectal usages and time spent living outside of native dialect area (negative or positive?)

  What would this mean? \( R = -.23, \ p < .0001 \)

  What would this mean? \( R =-.67, \ p < .0001 \)

- Score on language aptitude test and language fluency achieved after 18 months (positive or negative? What if there is no correlation?)

- Percentage of a particular past tense of a nonce word and the percentage of responses given by people in an experiment.

<table>
<thead>
<tr>
<th>What is the past test of <em>spling</em>?</th>
<th>What is the past tense of <em>creeze</em>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>People</td>
</tr>
<tr>
<td>splung 35%</td>
<td>splung 22%</td>
</tr>
<tr>
<td>People</td>
<td>Computer</td>
</tr>
<tr>
<td>splung 22%</td>
<td>croze 12%</td>
</tr>
<tr>
<td>People</td>
<td>People</td>
</tr>
<tr>
<td>croze 6%</td>
<td></td>
</tr>
</tbody>
</table>
Correlation and Causation

- Drowning and ice-cream consumption

Depression and being LDS ([Read this article](#))
- What is the correlation it is based on?
- What are the conclusions reached?
  - What is the problem with the conclusions?
  - What are other explanations?
- Score on language aptitude test and language fluency achieved after 18 months (Problem: prior classes taken)
- **Bad oral health causes Alzheimers**
  - What other explanation is there?
Download and open this file in SPSS.

Table 3.1. GPA, TOEFL scores, and number of hours spent gaming by student.

<table>
<thead>
<tr>
<th>Student</th>
<th>GPA</th>
<th>TOEFL</th>
<th>Gaming Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eun Mi</td>
<td>2.7</td>
<td>39</td>
<td>4</td>
</tr>
<tr>
<td>Ji</td>
<td>3.1</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>Marina</td>
<td>2.9</td>
<td>51</td>
<td>3</td>
</tr>
<tr>
<td>Andrea</td>
<td>3.6</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>Hyo</td>
<td>2.6</td>
<td>57</td>
<td>3</td>
</tr>
<tr>
<td>Marcos</td>
<td>2.8</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>Luis</td>
<td>3.3</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>Zafir</td>
<td>3.8</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>Rashid</td>
<td>3.0</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>Hu Do</td>
<td>2.9</td>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>Jung</td>
<td>3.6</td>
<td>81</td>
<td>1</td>
</tr>
<tr>
<td>Bao Yu</td>
<td>2.8</td>
<td>81</td>
<td>0</td>
</tr>
<tr>
<td>Hassan</td>
<td>3.5</td>
<td>87</td>
<td>3</td>
</tr>
<tr>
<td>Ivan</td>
<td>3.4</td>
<td>90</td>
<td>2</td>
</tr>
<tr>
<td>Kun</td>
<td>3.6</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>Jorge</td>
<td>3.3</td>
<td>105</td>
<td>1</td>
</tr>
<tr>
<td>Roderick</td>
<td>3.7</td>
<td>114</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 3.1. Scatterplot of GPA and TOEFL scores.

Which point represents EunMi's and Roderick's scores?
Using SPSS to generate scatterplots

The GPAs and TOEFL scores are in the correlation.sav file. Once you have opened that in SPSS you will see that the student’s name, GPA, TOEFL score, and number of hours spend gaming appear in the first four columns. The data in the GPA_new and Gaming-Hours_new columns will be used later. To create the scatterplot in Figure 3.1:

Click on Graphs > Legacy Dialog > Scatter/Dot (Figure 3.2) > Simple Scatter > Define.

Figure 3.2. Creating a scatterplot.

Click on TEOFL and move it to the Y-Axis by clicking on the arrow to the left of the Y-Axis box.

Move GPA to the X-Axis box in the same way > OK (Figure 3.3).
Figure 3.3. *Scatterplot* dialog box.
Think back on the idea of weighing different volumes of water and correlating their weight against their volume. I told you I'd come back to it. It would result in a perfect correlation of $r = 1$. Is there any variable, other than volume of the water, that influences its weight (assuming it is pure water at the same temperature)? No, volume explains 100% of the differences in weight between the samples that were measured. In statistical terminology weight explains 100% of the VARIANCE, or in other words, 100% of the dispersion of the scores around the regression line. We can get this 100% figure by taking the $r$, which is 1, and squaring it which yields an $r^2$ value of 1. In this case, 1 means everything or 100%. An $r^2$ of .61 indicates that the independent variable accounts for 61% of the variance in the dependent variable.
Now do a correlation between GPA and gaming hours.

1-What is the correlation (r)?

2- Is it positive or negative? What does that tell you?

3-Is the correlation significant? at what level?

4-What does r square tell you?

Now do a scatterplot between gaming hours and GPA.

Click on Graphs > Legacy Dialog > Scatter/Dot > Simple Scatter > Define. Click on Gaming_hours and move it to the Y-Axis by clicking on the arrow to the left of the Y-Axis box. Move GPA to the X-Axis box in the same way > OK.

To place a regression line on the scatterplots:

Double click on the scatterplot itself to open the Chart Editor. Click on Elements > Fit Line at Total. The Properties box will open. Choose Linear > Apply.
Assumptions of correlation:

1 the data are continuous.
2 the relationship between the variables is linear.
3 the data are normally distributed.
4 the observations are independent.
5 the data are homoscedastic.

Figure 3.6. Scatterplot of GPA and gaming hours. Loess line at 70% of points to fit.
Using SPSS to generate graphs for visualizing linearity in the data

To test for linearity first inspect the scatterplot:

Click on Graphs > Legacy Dialog > Scatter/Dot > Simple Scatter > Define. Click on
Gaming_Hours_new and move it to the Y-Axis by clicking on the arrow to the left of the Y-Axis box.
Move GPA_new to the X-Axis box in the same way > OK.

To place a regression line on the scatterplots:

Double click on the scatterplot itself to open the Chart Editor. Click on Elements > Fit Line at Total.
The Properties box will open. Choose Linear > Apply > Close.

To place a Loess line on the scatterplots

Double click on the scatterplot itself to open the Chart Editor. Click on Elements > Fit Line at Total.
The Properties box will open. Choose Loess > Apply. Adjust % of points to fit to smooth Loess line if desired (I used 70%) > Apply > Close.

What to do if the data aren't linear

1-look for outliers and deal with them
2-transform the data into different scale
3-use Kendall's tau b instead of Pearson correlation

Using SPSS to generate graphs and measures of normal distribution

Click on Analyze > Descriptive Statistics > Explore. Click on the variable name, Gaming_hours_new, and move it to the Dependent List by clicking on the arrow between the boxes. Click on Statistics > check the box next to Descriptives > Continue. Click on Plots, check the boxes next to Histogram and Normality plots with tests > Continue > OK. In order to superimpose the normal distribution curve on the histogram double click on the histogram itself then in the Chart Editor choose Elements > Show Distribution Curve.

This generates the boxplot in Figure 3.10, the histogram in Figure 3.11, the Q-Q plot in Figure 3.12, and the results of the Shapiro-Wilk and Kolmogorov-Smirnov tests in Table 3.7.
What to do if the distribution is not normal?

1-Transform the variables

2-Use Kendall's tau b instead

3-Maybe they are not normally distributed because there is no relationship