Exercises for Workshop: Understanding correlations

For all these exercises you will need to download and then open the file “understanding\_correlations” from the DLE. This has multiple tabs (along the bottom).

This is an activity to complete in groups.

**Exercise 1**

**Make sure you click on the Exercise 1 tab at the bottom of the spreadsheet**

The data in the columns represent the hypothetical data from 40 participants on two scores. In fact, the numbers are randomly generated as if they came from a normal distribution (a bell-curve)., separately for Score 1 and Score 2.

They are each a random set of numbers from a distribution with a mean of 100, and a standard-deviation of 10. (We will discuss these terms in class). For now, we will focus on the means.

For each set of numbers you can see the following things:

1. The average values for each Score.
2. The Pearson correlation between the scores (and the significance level, known as the p value).
3. A scatterplot showing the relationship between the scores.
4. Distribution plots showing the range of scores on each variable (Score 1 in Blue, Score 2 in red).

Activity:

Regenerate sets of random numbers 10 times. Each time you generate it, write down the following things:

Mean Score 1

Mean Score 2

Pearson correlation

Whether or not the distributions looked like bell curves

P value

After 10 iterations, look at what you have found and discuss answers to the following:

1. Why do the means of Scores 1 and 2 vary?
2. Why does the correlation value vary?
3. Did you ever get a p value less than 0.05? If so, what does this mean?

**Exercise 2**

**Make sure you click on the Exercise 2 tab at the bottom of the spreadsheet**

This spreadsheet is identical to the previous one, except for one pair of values (participant number 40). You can make this person more of an outlier by adding values in the two boxes F43 and G43 (both in red). If the values here are zero, then this person is not an outlier. However, if you put values in these two boxes these are added to that person’s scores to make them become outliers.

Try adding 100 in boxes F43 and G43, and see what happens to the scores in boxes C43 and D43. Notice that they won’t be exactly 100 higher than they were before, because Excel has re-randomised the whole set. But they will be around 100 higher than the average of the rest. You can see this in the scatterplot.

Try adding various positive numbers to F43 and G43, and see what happens to the following:

Mean Score 1

Mean Score 2

Pearson correlation

Whether or not the distributions looked like bell curves

P value

Discussion point: what have you learned from this? Do the values depend upon how much of an outlier there is?

Second discussion point: What do you expect will happen if you put in negative values to F43 and G43. Discuss this before you try it.

**Exercises 3 and 4**

**Make sure you click on the Exercise 3 / 4 tabs at the bottom of the spreadsheet to access the correct spreadsheet.**

This exercise is based on the same random numbers, but in these two exercises, we do something to ALL the numbers.

In Exercise 3, entering values in the boxes F43 ADDS that value to \*\*all\*\* Score 1 values, and the box G43 does the same for Score 2 values.

In Exercise 4, entering values in the box G43 MULTIPLIES all Score 2 values.

Before you start, discuss what you think will happen to each of the measures you have been discussing:

Mean Score 1

Mean Score 2

Pearson correlation

Whether or not the distributions looked like bell curves

P value

Now try it. Remember that each time you enter a new value, the random numbers are updated, so things will vary (randomly). Have a few goes with different values to get a sense of what is happening.

Can you explain the pattern you observe?

**Exercises 5 and 6**

**Make sure you click on the Exercise 5 / 6 tabs at the bottom of the spreadsheet to access the correct spreadsheet.**

In these two examples I have created two variables that are correlated. To do this I have randomly generated Score 1, then made Score 2 equal the average of Score 1, and another random number. (This is like having a “true score” and adding “noise”, as discussed in the lecture).

In Exercise 5 I have created Score 3, which is simply Score 2 plus a fixed value (that you can enter, as in Exercise 3).

In Exercise 6 Score 3 is Score 2 multiplied by a fixed value (that you can enter, as in Exercise 3).

For both exercises, I have plotted the values of Score 1 and Score 2, and Score 3 in separate scatterplots.

Additionally, I have got Excel to plot the “best fitting” line on each set of points. This line can be described with two key values: the slope (how much Score 2 goes up as you increase Score 1), and the intercept (the value of Score 2 expected when Score 1 is zero). I have highlighted these values in green boxes next to each scatterplot.

What do you expect will happen to the scatterplots and correlations if you add something to Score 2 (Exercise 5) or multiply Score 2 by a value. In each case you can see by comparing the two scatterplots. What do you expect will happen with the values of the slope and intercept? Discuss these before you try it.

Now try it out: what happens, and why?

Now try it by using negative numbers (i.e. subtracting a score, or multiplying by a negative number).

What do you expect will happen?

What does happen, and why?

**Exercise 7 (if you have time)**

**Make sure you click on the Exercise 7 tab at the bottom of the spreadsheet to access the correct spreadsheet.**

In this example, I have created a “True” value, plus 8 different measures (numbered 1-8). These are all generated by adding random noise to the true value, and so are all positively correlated with the true score.

In the table in green, you can see the means and correlation values for the true score, and then averages of 1,2,4,6 or 8 measures.

What happens to the mean value in each case? Why?

What happens to the correlations in each case? Why?

Discuss in your group.