Introduction to correlation

Introduction to correlation

We often want to investigate the relationship between pairs of variables.

- Vegetation height and mean annual temperature
- Animal body size and metabolic rate
- Number of automobiles in a location and carbon emissions

Introduction to correlation

We often want to investigate the relationship between pairs of variables.

- Vegetation height and mean annual temperature
- Animal body size and metabolic rate
- Number of automobiles in a location and carbon emissions

The **correlation** between pairs of variables, such as those listed above, describes how the variation of each variable is related to the other variable.

Visualising the correlation between two variables

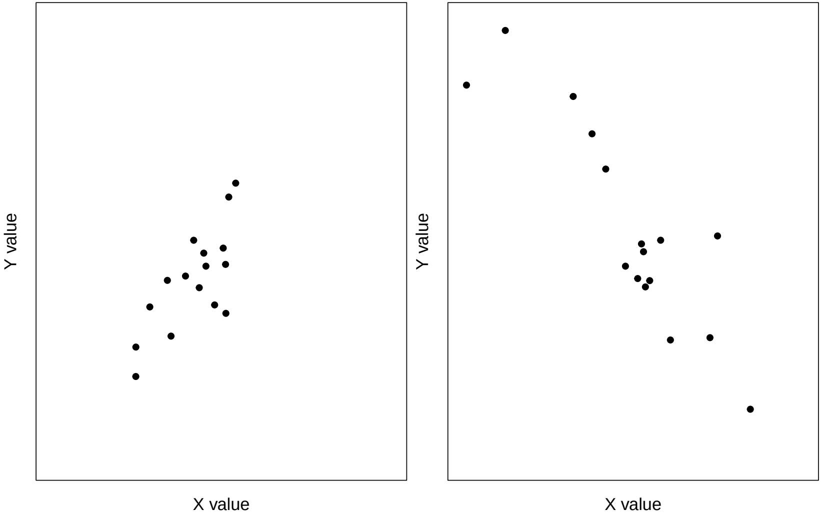


Figure 1: Variables illustrating a positive (left) and negative correlation

Visualising two variables that are not correlated

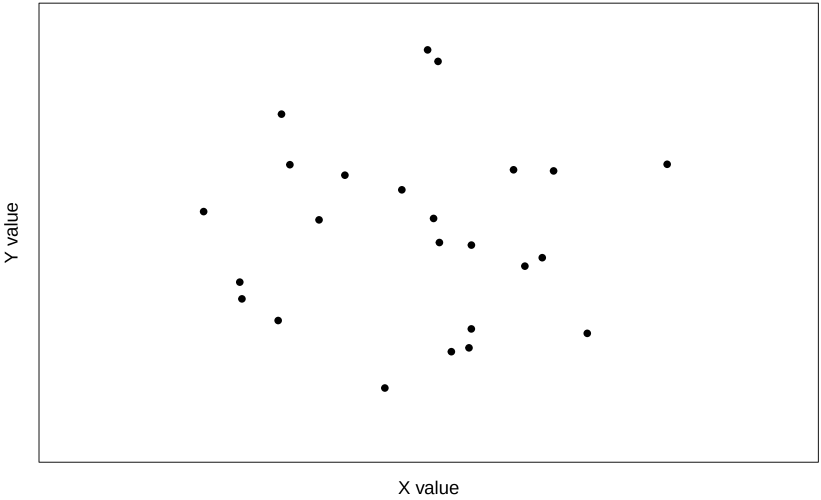


Figure 2: A plot of two hypothetical variables that are not correlated.

Getting a more intuitive sense of correlation

Formalised with the **correlation coefficient** (r)

- Provides a statistical measure of strength and direction of correlation
- Only describes association between variables (**not** cause and effect)

Getting a more intuitive sense of correlation

Formalised with the **correlation coefficient** (*r*)

- Provides a statistical measure of strength and direction of correlation
- Only describes association between variables (**not** cause and effect)

The value *r* ranges between -1 and 1

- Negative numbers indicate a negative correlation
- Positive numbers indicate a positive correlation
- Value of zero indicates no correlation

Getting a more intuitive sense of correlation

Formalised with the **correlation coefficient** (r)

- Provides a statistical measure of strength and direction of correlation
- Only describes association between variables (**not** cause and effect)

The value r ranges between -1 and 1

- Negative numbers indicate a negative correlation
- Positive numbers indicate a positive correlation
- Value of zero indicates no correlation

We can get a more intuitive understanding of the correlation coefficient with **[this application]**.

Introduction the correlation coefficient equation

Here we will consider the Pearson product-moment correlation coefficient

- Several equations are to follow
- Will walk through them step by step
- Explain each step verbally
- Relate the equation back to previous material

Introduction the correlation coefficient equation

Here we will consider the Pearson product-moment correlation coefficient

- Several equations are to follow
- Will walk through them step by step
- Explain each step verbally
- Relate the equation back to previous material

You will not need to memorise any of the equations you see here, but you should be able to recognise and understand the equation for the correlation coefficient.

How is the correlation coefficient defined?

The covariance between two variables X and Y, Cov(X, Y), divided by the standard deviation of X times the standard deviation of Y,

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}.$$

How is the correlation coefficient defined?

The covariance between two variables X and Y, Cov(X, Y), divided by the standard deviation of X times the standard deviation of Y,

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}.$$

But what is Cov(X, Y), exactly?

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

How is the correlation coefficient defined?

The covariance between two variables X and Y, Cov(X, Y), divided by the standard deviation of X times the standard deviation of Y,

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}.$$

But what is Cov(X, Y), exactly?

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Let's break this down a bit further!

What is the covariance?

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Assume 'x' is bird length and 'y' is bird mass. How would we calculate this, explained verbally?

What is the covariance?

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Assume ‘x’ is bird length and ‘y’ is bird mass. How would we calculate this, explained verbally?

1. For each bird, calculate its length minus mean bird length
2. For each bird, calculate its mass minus mean bird mass
3. Multiply the two steps above together
4. Do the above for all birds, and add up all the values
5. Divide what you added up by the total number of birds

This is the numerator part of the correlation coefficient

Back to the definition of the correlation

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}$$

Back to the definition of the correlation

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}$$

Now we can expand that covariance,

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{StDev(X) \times StDev(Y)}.$$

Back to the definition of the correlation

$$Correlation = \frac{Cov(X, Y)}{StDev(X) \times StDev(Y)}$$

Now we can expand that covariance,

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{StDev(X) \times StDev(Y)}.$$

We already know the formula for standard deviation

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

The Pearson's correlation coefficient,

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Testing whether or not a correlation is significant

Null Hypothesis: There is no correlation between two variables X and Y

Testing whether or not a correlation is significant

Null Hypothesis: There is no correlation between two variables X and Y

Alternative Hypothesis: There is a correlation between two variables X and Y

Testing whether or not a correlation is significant

Null Hypothesis: There is no correlation between two variables X and Y

Alternative Hypothesis: There is a correlation between two variables X and Y

Degrees of Freedom: Number of data points minus two (lose a degree of freedom for calculating each mean)

To test whether or not to reject the null hypothesis, we can obtain a p-value in jamovi or use a table of critical values to look up the value of r for a specific degrees of freedom

Testing whether or not a correlation is significant

We often want to test whether or not the correlation between two variables is significant

- Test of Pearson product moment correlation assumes variables are normally distributed
- Test of Spearman's rank correlation coefficient (i.e., correlation of ranks) does not assume normality

To test whether or not two variables are correlated, we first must test the null hypothesis that the two variables are normally distributed.

A data set of soil depths and root densities

Sample number	Soil depth (m)	Root density (g per m^3)
1	0.8	13
2	2.0	8
3	2.3	4
4	2.7	6
5	0.5	18
6	1.8	7
7	1.5	9
8	2.1	3
9	1.2	7
10	1.1	10

In the above table, soil depth is measured in metres and root density is measured in grams per cubic metre.

Testing for normality in jamovi

Select 'Exploration > Descriptives', then move both variables to the Variables list. Select 'Shapiro-Wilk' from the statistics pulldown

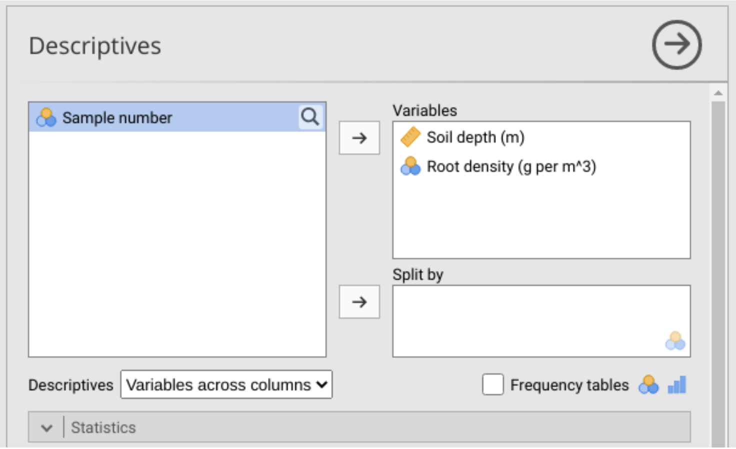


Figure 3: Jamovi input for testing normality.

Testing for normality in jamovi

Below shows the output of the tests for normality in jamovi.

Descriptives

Descriptives

	Soil depth (m)	Root density (g per m^3)
N	10	10
Missing	0	0
Mean	1.60000	8.50000
Median	1.65000	7.50000
Standard deviation	0.70079	4.40328
Minimum	0.50000	3
Maximum	2.70000	18
Shapiro-Wilk W	0.97905	0.92568
Shapiro-Wilk p	0.95985	0.40680

Figure 4: Jamovi output for tests of normality on two variables.

Plotting soil depth versus root density in jamovi

How to make a scatterplot in jamovi

- Select ‘Exploration > Scatterplot’
- A panel will pop up with the dataset variables
- Move ‘Soil depth (m)’ to ‘X-Axis’
- Move Root density (g per m^3) to ‘Y-Axis’

Plotting soil depth versus root density in Jamovi

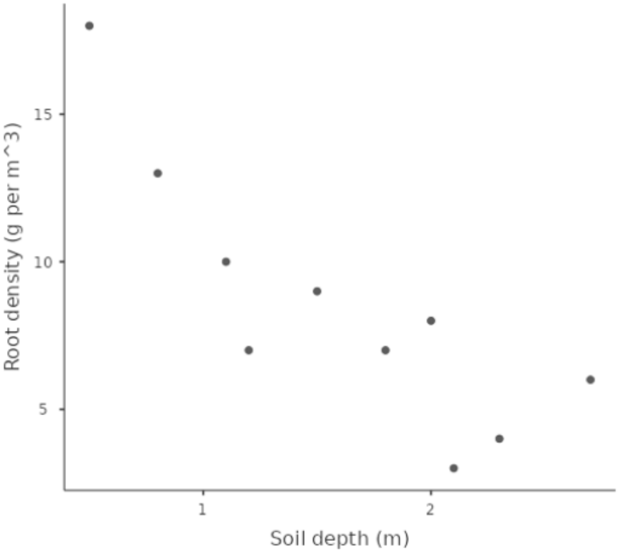
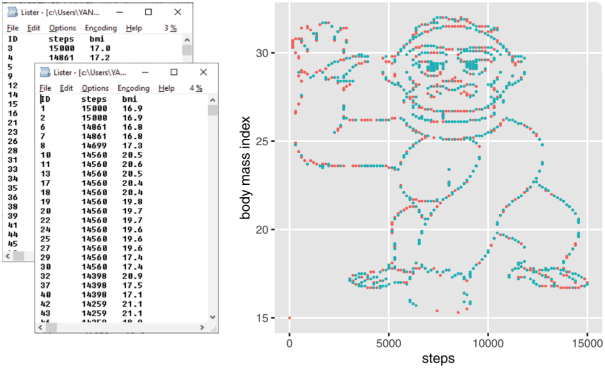


Figure 5: Jamovi output of a scatterplot for Soil Depth vs Root Density

Plotting soil depth versus root density in jamovi

a b



c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Testing whether soil depth and root density are correlated

Test whether or not our variables 'soil_depth' and 'root_density' are correlated.

Hypothesis for Pearson's correlation coefficient

Testing whether soil depth and root density are correlated

Test whether or not our variables 'soil_depth' and 'root_density' are correlated.

Hypothesis for Pearson's correlation coefficient

- **Null:** There is no correlation between root density and soil depth
- **Alternative:** There is a significant correlation between root density and soil depth

Testing whether soil depth and root density are correlated

Test whether or not our variables ‘soil_depth’ and ‘root_density’ are correlated.

Hypothesis for Pearson’s correlation coefficient

- **Null:** There is no correlation between root density and soil depth
- **Alternative:** There is a significant correlation between root density and soil depth

We will reject the null hypothesis if, assuming that the null hypothesis is true, the probability of getting an r value as or more extreme than the one we obtained from our sample (i.e., the p-value) is less than or equal to 0.05.

Testing whether soil depth and root density are correlated

Test the null hypothesis that this correlation is not significant

- Selecting ‘Regression > Correlation Matrix’
- Move both variables into the box to the right
- Make sure ‘Pearson’ selected for Correlation Coefficients
- Test ‘Correlation’ (two-tailed test for statistical significance

Testing whether soil depth and root density are correlated

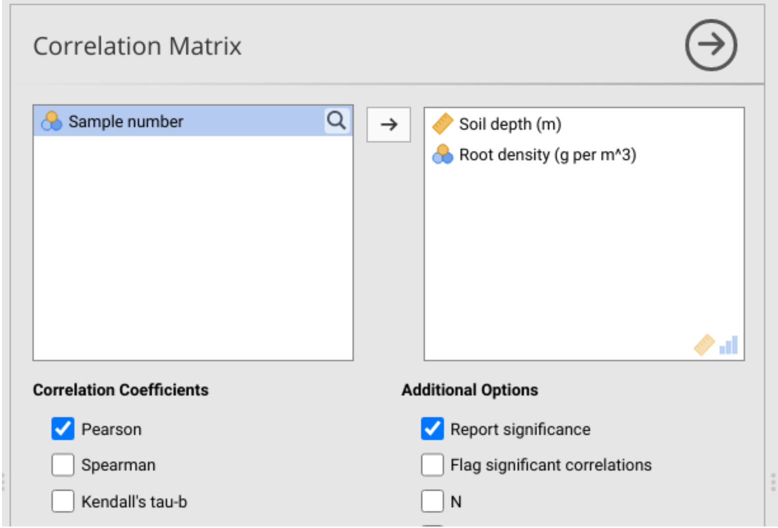


Figure 6: Jamovi box on how to run a test of the correlation coefficient.

Testing whether soil depth and root density are correlated

A table of output that looks like the one below.

Correlation Matrix

Correlation Matrix

		Soil depth (m)	Root density (g per m^3)
Soil depth (m)	Pearson's r	—	
	p-value	—	
Root density (g per m^3)	Pearson's r	-0.84257	—
	p-value	0.00221	—

Figure 7: Jamovi table showing output of a parametric test of the significance of a correlation coefficient.

Spearman rank correlation coefficient

If either variable is not normally distributed, we need a non-parametric test

- The Spearman rank correlation coefficient is a non-parametric alternative.
- Calculate the correlation of the **ranks** of the values
- Test whether this Spearman rank correlation coefficient is significant

Consider some measurements of per cent dissolved oxygen and ammonia concentration (in mg per litre) from eight locations in Scotland.

Spearman rank correlation coefficient

Sample	%O2	Rank %O2	NH3 Conc.	Rank NH3 Conc.
1	95.9	8	0.080	2
2	81.9	3	0.100	3
3	80.9	2	0.210	6
4	77.9	1	0.579	8
5	90.7	6	0.250	7
6	88.2	4	0.130	5
7	93.6	7	0.070	1
8	89.1	5	0.121	4

Testing whether per cent O_2 and Ammonia are correlated

Test whether or not our variables per cent O_2 and Ammonia concentration are correlated.

Testing whether per cent O_2 and Ammonia are correlated

Test whether or not our variables per cent O_2 and Ammonia concentration are correlated.

Hypothesis for Spearman's correlation coefficient

- **Null:** There is no correlation between per cent dissolved oxygen and ammonia concentration
- **Alternative:** There is a significant correlation between per cent dissolved oxygen and ammonia concentration

Testing whether per cent O_2 and Ammonia are correlated

Test whether or not our variables per cent O_2 and Ammonia concentration are correlated.

Hypothesis for Spearman's correlation coefficient

- **Null:** There is no correlation between per cent dissolved oxygen and ammonia concentration
- **Alternative:** There is a significant correlation between per cent dissolved oxygen and ammonia concentration

We will reject the null hypothesis if, assuming that the null hypothesis is true, the probability of getting an r value as or more extreme than the one we obtained from our sample (i.e., the p-value) is less than or equal to 0.05.

Testing whether per cent O_2 and Ammonia are correlated

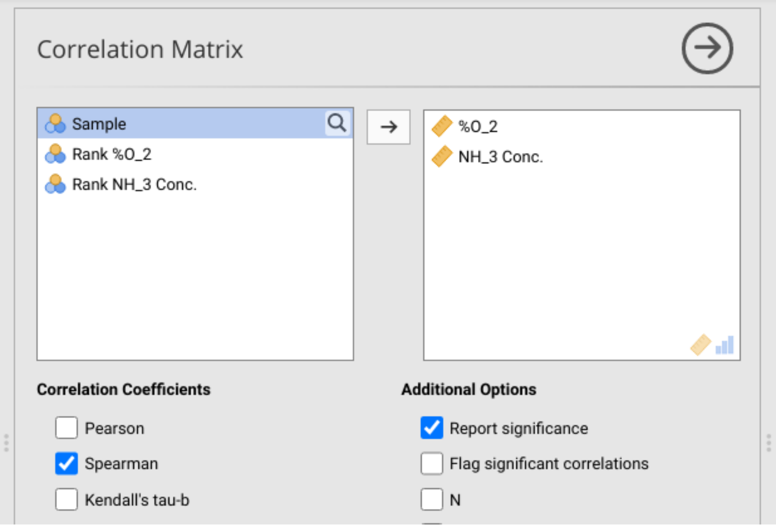


Figure 8: Jamovi box showing how to run a test of the Spearman rank correlation coefficient.

Testing whether per cent O2 and Ammonia are correlated

Spearman rank correlation coefficient between the variables per cent dissolved oxygen and ammonia concentration is -0.667, and the p-value for this test is 0.083, meaning that we cannot reject our null hypothesis that the two variables are uncorrelated.

Correlation Matrix

Correlation Matrix

		%O_2	NH_3 Conc.
%O_2	Spearman's rho	—	
	p-value	—	
NH_3 Conc.	Spearman's rho	−0.66667	—
	p-value	0.08309	—

Figure 9: Jamovi output table showing output of a test of the significance of a Spearman rank correlation coefficient.