

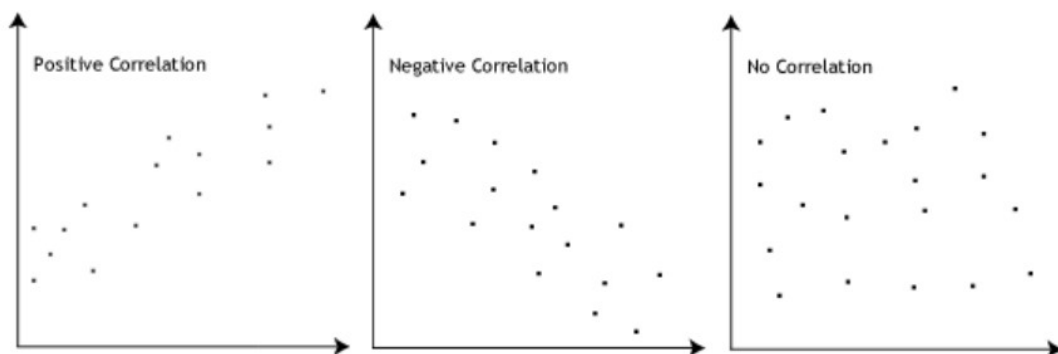
Pearson's Correlation

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r . Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r , indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

For example, you could use a Pearson's correlation to understand whether there is an association between exam performance and time spent revising. You could also use a Pearson's correlation to understand whether there is an association between depression and length of unemployment.

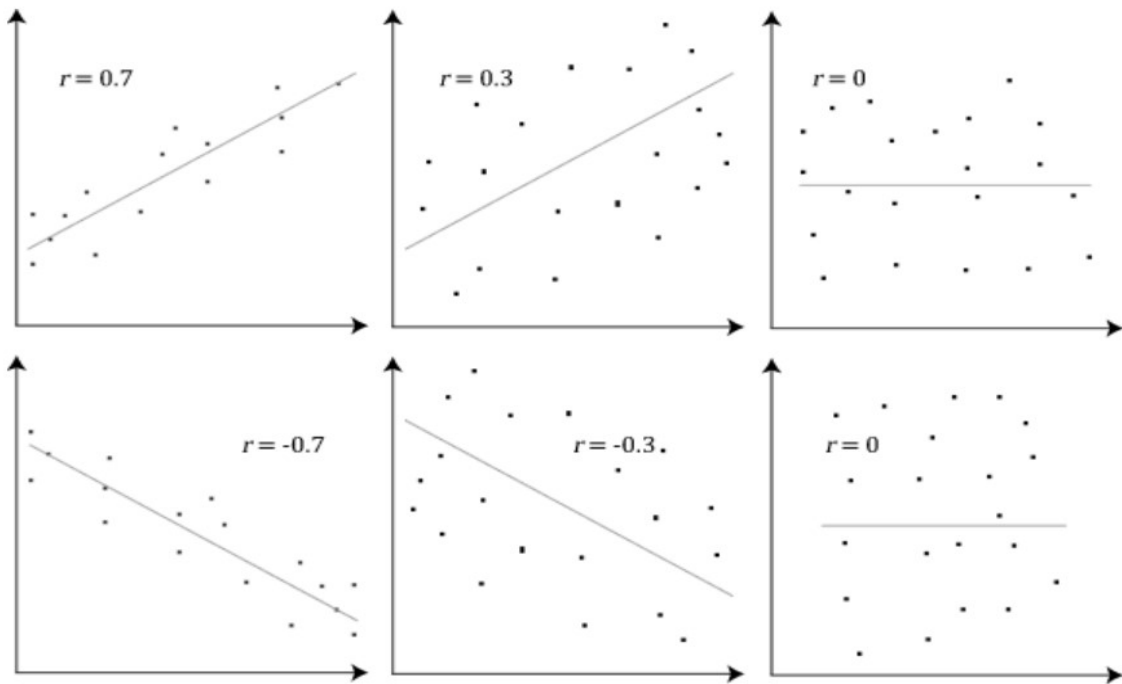
What values can the Pearson correlation coefficient take?

The Pearson correlation coefficient, r , can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



How can we determine the strength of association based on the Pearson correlation coefficient?

The stronger the association of the two variables, the closer the Pearson correlation coefficient, r , will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, $r = 0.8$ or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Can you use any type of variable for Pearson's correlation coefficient?

No, the two variables have to be measured on either an interval or ratio scale. However, both variables do not need to be measured on the same scale (e.g., one variable can be ratio and one can be interval).

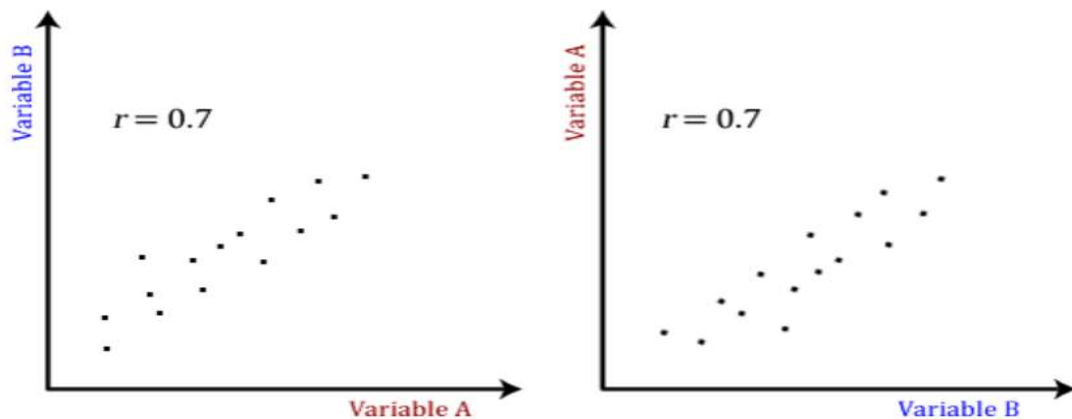
Do the two variables have to be measured in the same units?

No, the two variables can be measured in entirely different units. For example, you could correlate a person's age with their blood sugar levels. Here, the units are completely different; age is measured in years and blood sugar level measured in mmol/L (a measure of concentration). Indeed, the calculations for Pearson's correlation coefficient were designed such that the units of measurement do not affect the calculation. This allows the correlation coefficient to be comparable and not influenced by the units of the variables used.

What about dependent and independent variables?

The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally. For example, you might want to find out whether basketball performance is correlated to a person's height. You might, therefore, plot a graph of performance against height and calculate the Pearson correlation coefficient. Let's say, for example, that $r = .67$. That is, as height increases so does basketball performance. This makes sense. However, if we plotted the variables the other way around and wanted to determine whether a person's height was determined by their basketball performance (which makes no sense), we would still get $r = .67$. This is because the

Pearson correlation coefficient makes no account of any theory behind why you chose the two variables to compare. This is illustrated below:

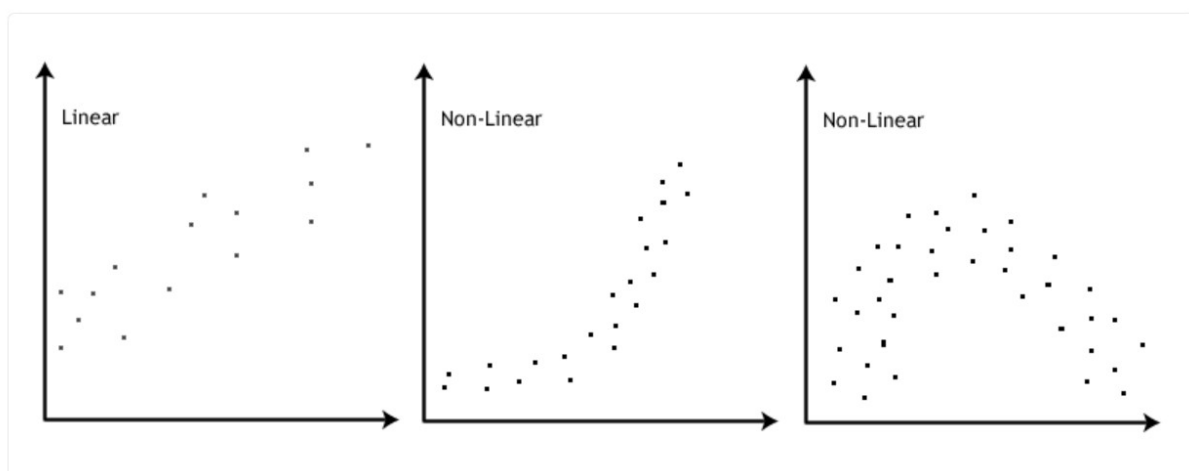


Assumptions

When you choose to analyse your data using Pearson's correlation, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using Pearson's correlation. You need to do this because it is only appropriate to use Pearson's correlation if your data "passes" four assumptions that are required for Pearson's correlation to give you a valid result.

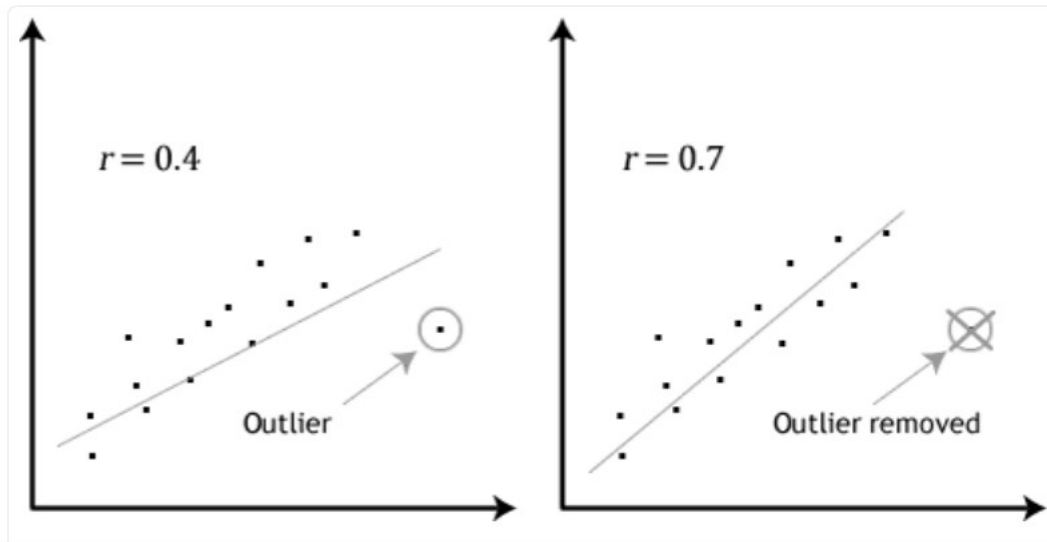
Assumption #1: Your two variables should be measured at the **interval** or **ratio level** (i.e., they are **continuous**). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

Assumption #2: There is a **linear relationship** between your two variables. Whilst there are a number of ways to check whether a linear relationship exists between your two variables, we suggest creating a scatterplot using SPSS Statistics, where you can plot the one variable against the other variable, and then visually inspect the scatterplot to check for linearity. Your scatterplot may look something like one of the following:



If the relationship displayed in your scatterplot is not linear, you will have to either run a nonparametric equivalent to Pearson's correlation or transform your data.

Assumption #3: There should be **no significant outliers**. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:



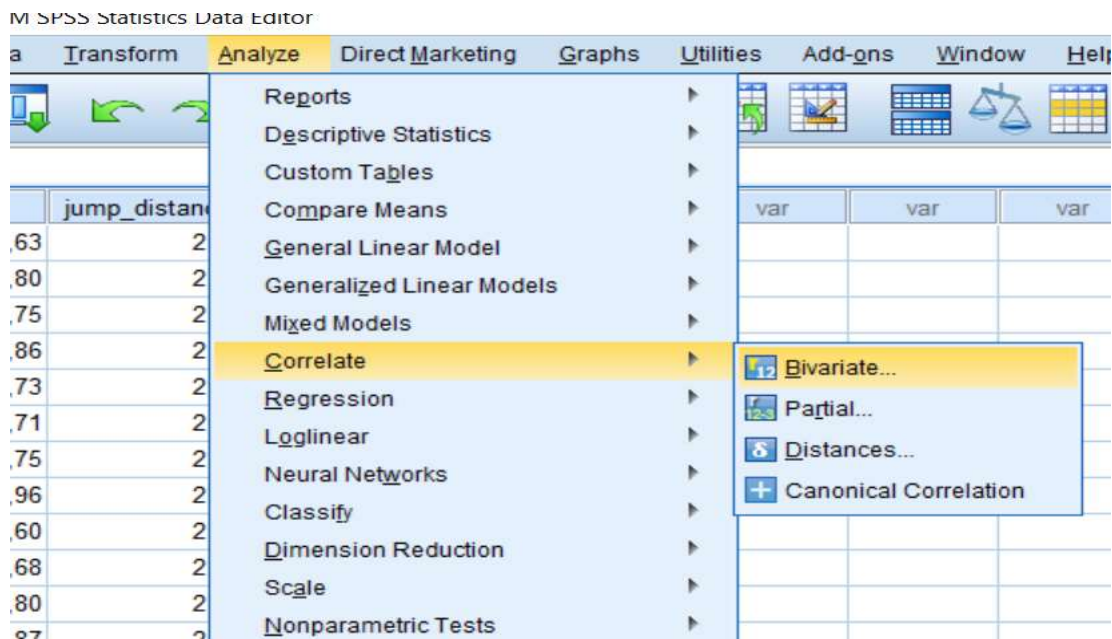
Pearson's correlation coefficient, r , is sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. Therefore, in some cases, including outliers in your analysis can lead to misleading results. Therefore, it is best if there are no outliers or they are kept to a minimum.

Assumption #4: Your variables should be **approximately normally distributed**. In order to assess the statistical significance of the Pearson correlation, you need to have bivariate normality.

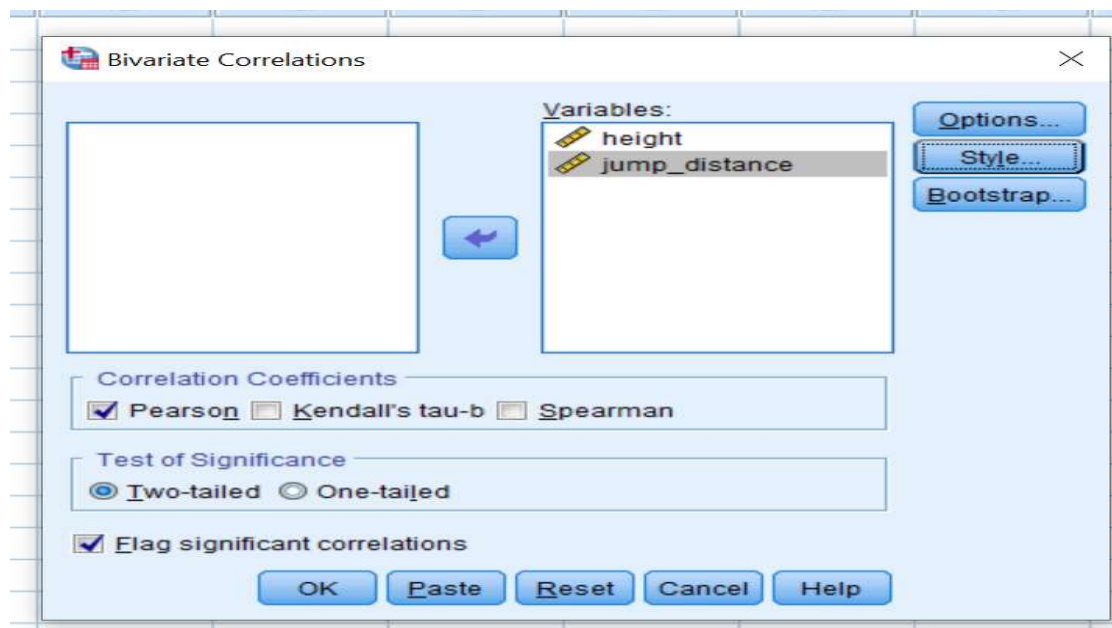
Example

A researcher wants to know whether a person's height is related to how well they perform in a long jump. The researcher recruited untrained individuals from the general population, measured their height and had them perform a long jump. The researcher then investigated whether there was an association between height and long jump performance by running a Pearson's correlation.

| Height | Jump_Distance | Height | Jump_Distance |
|--------|---------------|--------|---------------|
| 1,63 | 2,34 | 1,96 | 2,67 |
| 1,80 | 2,48 | 1,60 | 2,39 |
| 1,75 | 2,29 | 1,68 | 2,47 |
| 1,86 | 2,62 | 1,80 | 2,60 |
| 1,73 | 2,64 | 1,87 | 2,75 |
| 1,71 | 2,30 | 1,74 | 2,40 |
| 1,75 | 2,44 | 1,67 | 2,46 |



You will be presented with the **Bivariate Correlations** dialogue box:



Click on the **Options...** button and you will be presented with the **Bivariate Correlations: Options** dialogue box. If you wish to generate some descriptives, you can do it here by clicking on the relevant checkbox in the –Statistics– area.

Descriptive Statistics

| | Mean | Std. Deviation | N |
|---------------|--------|----------------|----|
| height | 1,7536 | ,09834 | 14 |
| jump_distance | 2,4893 | ,14467 | 14 |

| Correlations | | | |
|---------------|---------------------|--------|---------------|
| | | height | jump_distance |
| height | Pearson Correlation | 1 | ,706** |
| | Sig. (2-tailed) | | ,005 |
| | N | 14 | 14 |
| jump_distance | Pearson Correlation | ,706** | 1 |
| | Sig. (2-tailed) | ,005 | |
| | N | 14 | 14 |

** . Correlation is significant at the 0.01 level (2-tailed).

we can see that the Pearson correlation coefficient, r , is 0.706, and that it is statistically significant ($p = 0.005$). For interpreting multiple correlations.

A Pearson product-moment correlation was run to determine the relationship between height and distance jumped in a long jump. There was a strong, positive correlation between height and distance jumped, which was statistically significant ($r = .706$, $n = 14$, $p = .005$).

Spearman's Rank-Order Correlation

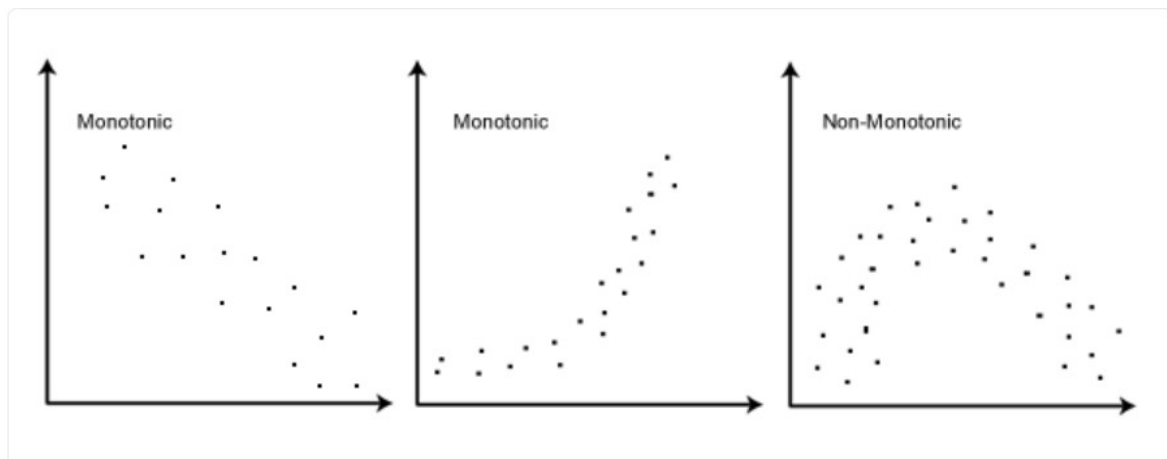
The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ , also signified by r_s) measures the strength and direction of association between two ranked variables.

The Spearman rank-order correlation coefficient (Spearman's correlation, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. It is denoted by the symbol r_s (or the Greek letter ρ , pronounced rho). The test is used for either ordinal variables or for continuous data that has failed the assumptions necessary for conducting the Pearson's product-moment correlation. For example, you could use a Spearman's correlation to understand whether there is an association between exam performance and time spent revising; whether there is an association between depression and length of unemployment; and so forth

The Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated. However, Spearman's correlation determines the strength and direction of the **monotonic relationship** between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. Examples of monotonic and non-monotonic relationships are presented in the diagram below:



Why is a monotonic relationship important to Spearman's correlation?

Spearman's correlation measures the strength and direction of monotonic association between two variables. Monotonicity is "less restrictive" than that of a linear relationship. For example, the middle image above shows a relationship that is monotonic, but not linear.

A monotonic relationship is not strictly an assumption of Spearman's correlation. That is, you can run a Spearman's correlation on a non-monotonic relationship to determine if there is a **monotonic component** to the association. However, you would normally pick a measure of association, such as Spearman's correlation, that fits the pattern of the observed data. That is, if a scatterplot shows that the relationship between your two variables looks monotonic you would run a Spearman's correlation because this will then measure the strength and direction of this monotonic relationship. On the other hand if, for example, the relationship appears linear (assessed via scatterplot) you would run a Pearson's correlation because this will measure the strength and direction of any linear relationship. You will not always be able to visually check whether you have a monotonic relationship, so in this case, you might run a Spearman's correlation anyway.

What values can the Spearman correlation coefficient, r_s , take?

The Spearman correlation coefficient, r_s , can take values from +1 to -1. A r_s of +1 indicates a perfect association of ranks, a r_s of zero indicates no association between ranks and a r_s of -1 indicates a perfect negative association of ranks. The closer r_s is to zero, the weaker the association between the ranks.

The general form of a null hypothesis for a Spearman correlation is:

H_0 : There is no [monotonic] association between the two variables [in the population].

What is the definition of Spearman's rank-order correlation?

There are two methods to calculate Spearman's correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and n = number of cases.

The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Assumptions

Assumption #1: Your two variables should be measured on an ordinal, interval or ratio scale. Examples of ordinal variables include Likert scales (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-point scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of interval/ratio variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

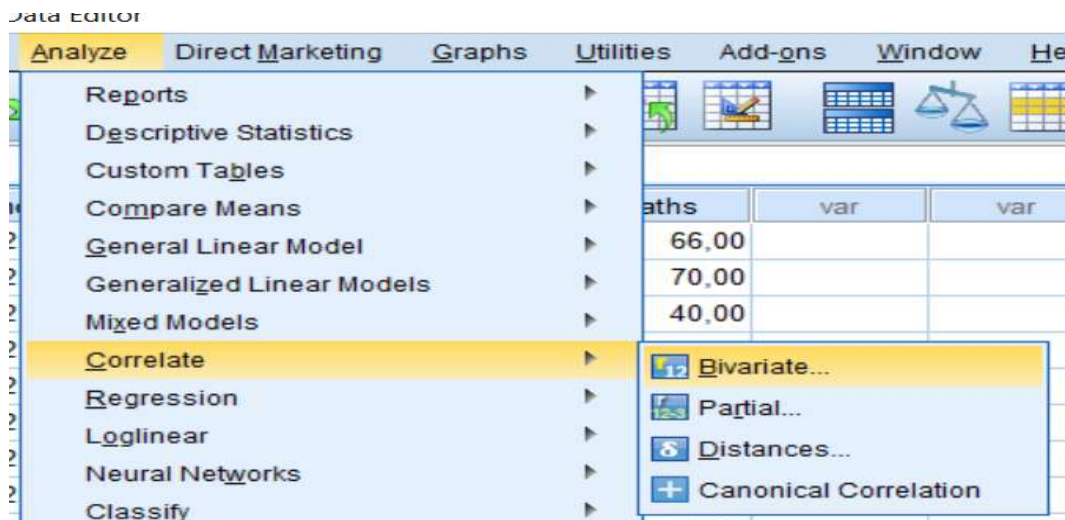
Assumption #2: Your two variables represent **paired observations**. For example, imagine that you were interested in the relationship between daily cigarette consumption and amount of exercise performed each week. A single paired observation reflects the score on each variable for a single participant (e.g., the daily cigarette consumption of "Participant 1" and the amount of exercise performed each week by "Participant 1"). With 30 participants in the study, this means that there would be 30 paired observations.

Assumption #3: There is a **monotonic relationship** between the two variables. A monotonic relationship exists when either the variables increase in value together, or as one variable value increases, the other variable value decreases. Whilst there are a number of ways to check whether a monotonic relationship exists between your two variables, it is suggested creating a scatterplot using SPSS Statistics, where you can plot one variable against the other, and then visually inspect the scatterplot to check for monotonicity.

Example

A teacher is interested in whether those who do better at English also do better in maths. To test whether this is the case, the teacher records the scores of her 10 students in their end-of-year examinations for both English and maths. Therefore, one variable records the English scores and the second variable records the maths scores for the 10 pupils.

| English (marks) | Maths(marks) | English (marks) | Maths(marks) |
|-----------------|--------------|-----------------|--------------|
| 56 | 66 | 64 | 56 |
| 75 | 70 | 58 | 59 |
| 45 | 40 | 80 | 77 |
| 71 | 60 | 76 | 67 |
| 62 | 65 | 61 | 63 |



Correlations

| | | | english | maths |
|----------------|---------|-------------------------|-------------------|-------------------|
| Spearman's rho | english | Correlation Coefficient | 1,000 | ,673 [*] |
| | | Sig. (2-tailed) | . | ,033 |
| | | N | 10 | 10 |
| | maths | Correlation Coefficient | ,673 [*] | 1,000 |
| | | Sig. (2-tailed) | ,033 | . |
| | | N | 10 | 10 |

*. Correlation is significant at the 0.05 level (2-tailed).

A Spearman's rank-order correlation was run to determine the relationship between 10 students' English and maths exam marks. There was a strong, positive correlation between English and maths marks, which was statistically significant ($r_s(8) = .669, p = .035$).

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) | d | d ² |
|----------------|--------------|----------------|--------------|---|----------------|
| 56 | 66 | 9 | 4 | 5 | 25 |
| 75 | 70 | 3 | 2 | 1 | 1 |
| 45 | 40 | 10 | 10 | 0 | 0 |
| 71 | 60 | 4 | 7 | 3 | 9 |
| 62 | 65 | 6 | 5 | 1 | 1 |
| 64 | 56 | 5 | 9 | 4 | 16 |
| 58 | 59 | 8 | 8 | 0 | 0 |
| 80 | 77 | 1 | 1 | 0 | 0 |
| 76 | 67 | 2 | 3 | 1 | 1 |
| 61 | 63 | 7 | 6 | 1 | 1 |

Where d = difference between ranks and d² = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\begin{aligned}\rho &= 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \\ \rho &= 1 - \frac{6 \times 54}{10(10^2 - 1)} \\ \rho &= 1 - \frac{324}{990} \\ \rho &= 1 - 0.33 \\ \rho &= 0.67\end{aligned}$$

Kendall's Tau-b

Kendall's tau-b (τ_b) correlation coefficient (Kendall's tau-b, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. It is considered a nonparametric alternative to the [Pearson's product-moment correlation](#) when your data has failed one or more of the assumptions of this test. It is also considered an alternative to the nonparametric [Spearman rank-order correlation coefficient](#) (especially when you have a small sample size with many tied ranks). If you consider one of your variables as an independent variable and the other as a dependent variable, you might consider running a [Somers' d](#) test instead.

For example, you could use Kendall's tau-b to understand whether there is an association between exam grade and time spent revising (i.e., where there were six possible exam grades – A, B, C, D, E and F – and revision time was split into five categories: less than 5 hours, 5-9 hours, 10-14 hours, 15-19 hours, and 20 hours or more). Alternately, you could use Kendall's tau-b to understand whether there is an association between customer satisfaction and delivery time (i.e., where delivery time had four categories – next day, 2 working days, 3-5 working days, and more than 5 working days – and customer satisfaction was measured in terms of the level of agreement customers had with the following statement, "I am satisfied with the time it took for my parcel to be delivered", where the level of agreement had five categories: strongly agree, agree, neither agree nor disagree, disagree and strongly disagree).

Assumption #1: Your two variables should be measured on an **ordinal**. Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 5-point scale explaining how much a customer liked a product, ranging from "Not very much" to "Yes, a lot"). Examples of **continuous variables** (i.e., **interval** or **ratio** variables) include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

Assumption #2: Kendall's tau-b determines whether there is a **monotonic relationship** between your two variables. As such, it is desirable if your data would appear to follow a monotonic relationship, so that formally testing for such an association makes sense, but it is not a strict assumption or one that you are often able to assess.

Note: Your variables of interest are **continuous with outliers or ordinal you may prefer**.

Spearman's Rho and Kendall's Tau are very similar tests and are used in similar scenarios. It is recommended to use Kendall's Tau first and Spearman's Rho as a backup.

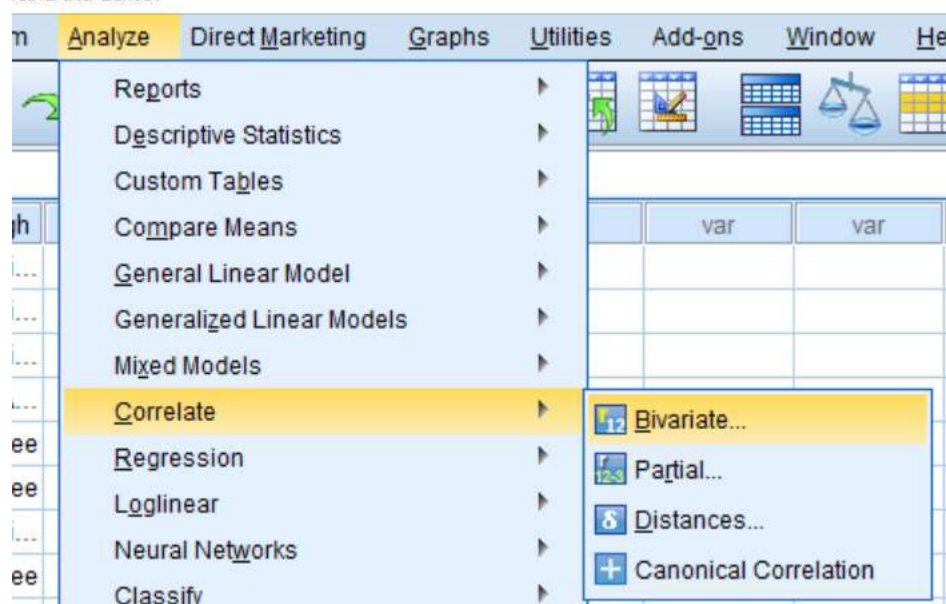
EXAMPLE:

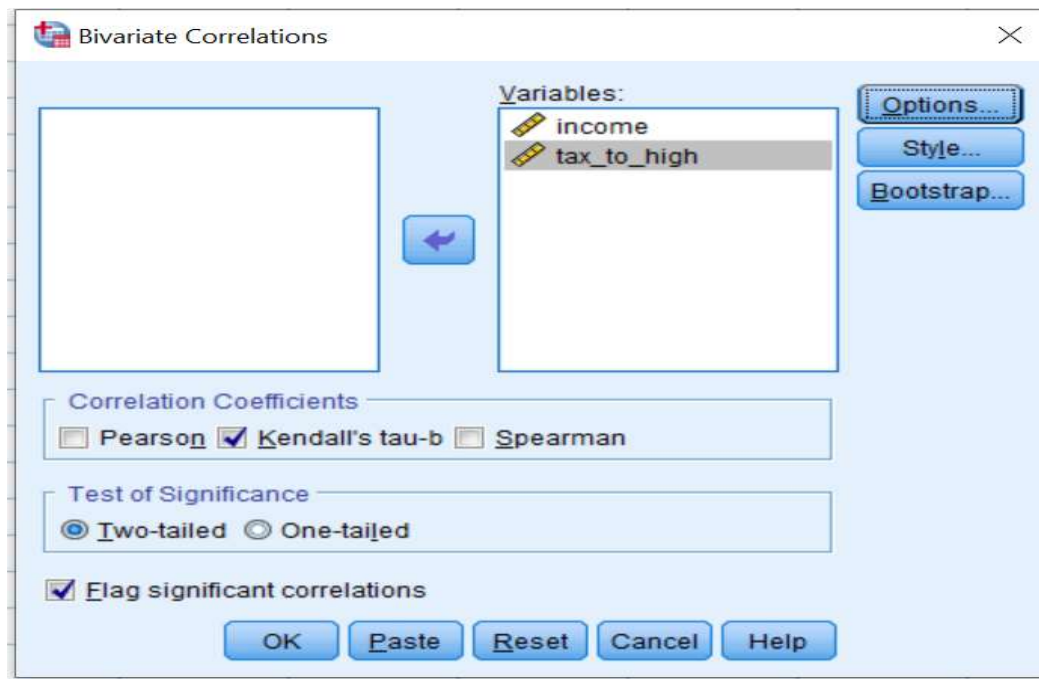
Taxes have the ability to elicit strong responses in many people, with some thinking they are too high, whilst others think they should be higher. A researcher conducted a simple study where they presented participants with the statement: "Tax is too high in this country", and asked them how much they agreed with this statement. They had four options how to respond: "Strongly Disagree", "Disagree", "Agree" or "Strongly Agree". These ordered responses were the categories of the dependent variable, `tax_too_high`. The researcher also asked participants to state whether they had a "low", "middle" or "high" income, where each of these categories

had specified income ranges (e.g., a low income was any income under £18,000 per annum). The income level of participants was recorded in the variable, income.

| Income | Tax to high | Income | Tax to high |
|---------------|-------------------|---------------|-------------------|
| Low income | Strongly Disagree | Middle income | Agree |
| Low income | Strongly Disagree | High income | Strongly Agree |
| Low income | Strongly Disagree | Low income | Strongly Disagree |
| High income | Strongly Agree | Low income | Disagree |
| Middle income | Agree | Low income | Disagree |
| Middle income | Agree | Middle income | Agree |
| Low income | Strongly Disagree | High income | Agree |
| High income | Agree | Low income | Strongly Disagree |
| Middle income | Disagree | High income | Strongly Agree |
| Middle income | Disagree | Low income | Disagree |
| High income | Agree | Middle income | Strongly Agree |
| Middle income | Strongly Disagree | Middle income | Strongly Agree |

ics Data Editor





Correlations

| | | | income | tax_to_high |
|-----------------|-------------|-------------------------|--------|-------------|
| Kendall's tau_b | income | Correlation Coefficient | 1,000 | ,701** |
| | | Sig. (2-tailed) | . | ,000 |
| | | N | 24 | 24 |
| | tax_to_high | Correlation Coefficient | ,701** | 1,000 |
| | | Sig. (2-tailed) | ,000 | . |
| | | N | 24 | 24 |

** . Correlation is significant at the 0.01 level (2-tailed).

A Kendall's tau-b correlation was run to determine the relationship between income level and views towards income taxes amongst 24 participants. There was a strong, positive correlation between income level and the view that taxes were too high, which was statistically significant ($\tau_b = .701, p = .000$).

Point-Biserial Correlation

A point-biserial correlation is used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous variable. It is a special case of the Pearson's product-moment correlation, which is applied when you have two variables, whereas in this case one of the variables is measured on a dichotomous scale.

For example, you could use a point-biserial correlation to determine whether there is an association between salaries, measured in US dollars, and gender (i.e., your continuous variable would be "salary" and your dichotomous variable would be "gender", which has two categories: "males" and "females"). Alternately, you could use a point-biserial correlation to determine whether there is an association between cholesterol concentration, measured in mmol/L, and

smoking status (i.e., your continuous variable would be "cholesterol concentration", a marker of heart disease, and your dichotomous variable would be "smoking status", which has two categories: "smoker" and "non-smoker").

A dichotomous variable is **one that takes on one of only two possible values when observed or measured.**

Assumptions

Assumption #1: One of your two variables should be measured on a continuous scale. Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

Assumption #2: Your other variable should be **dichotomous**. Examples of **dichotomous variables** include gender (two groups: male or female), employment status (two groups: employed or unemployed), smoker (two groups: yes or no), and so forth.

Assumption #3: There should be no outliers for the continuous variable for each category of the dichotomous variable. You can test for outliers using boxplots.

Assumption #4: Your continuous variable should be **approximately normally distributed** for each category of the dichotomous variable. You can test this using the Shapiro-Wilk test of normality.

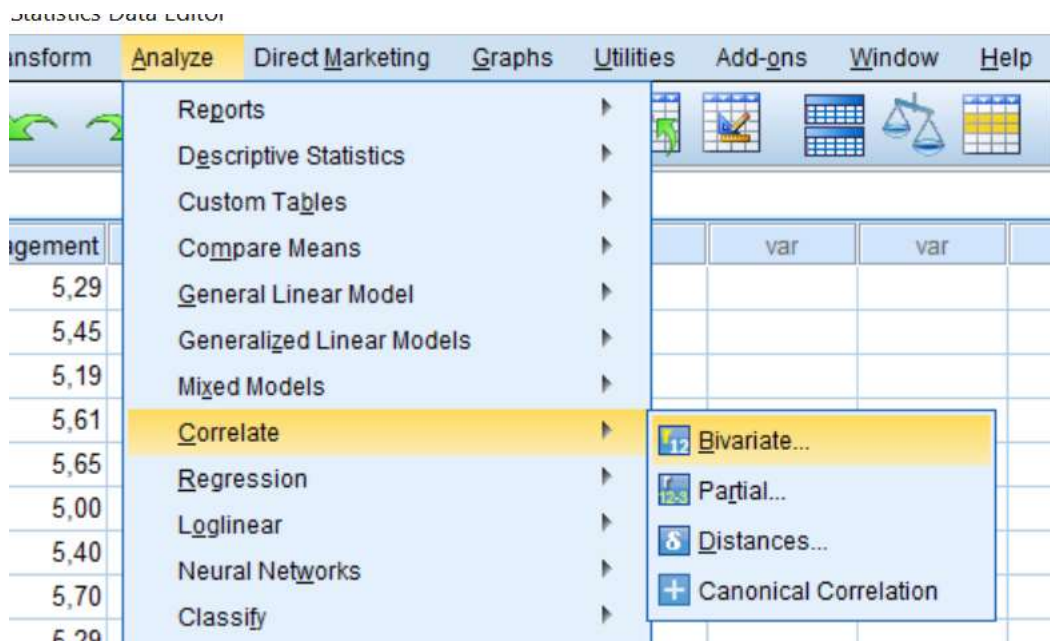
Assumption #5: Your continuous variable should have **equal variances** for each category of the dichotomous variable. You can test this using Levene's test of equality of variances.


EXAMPLE:

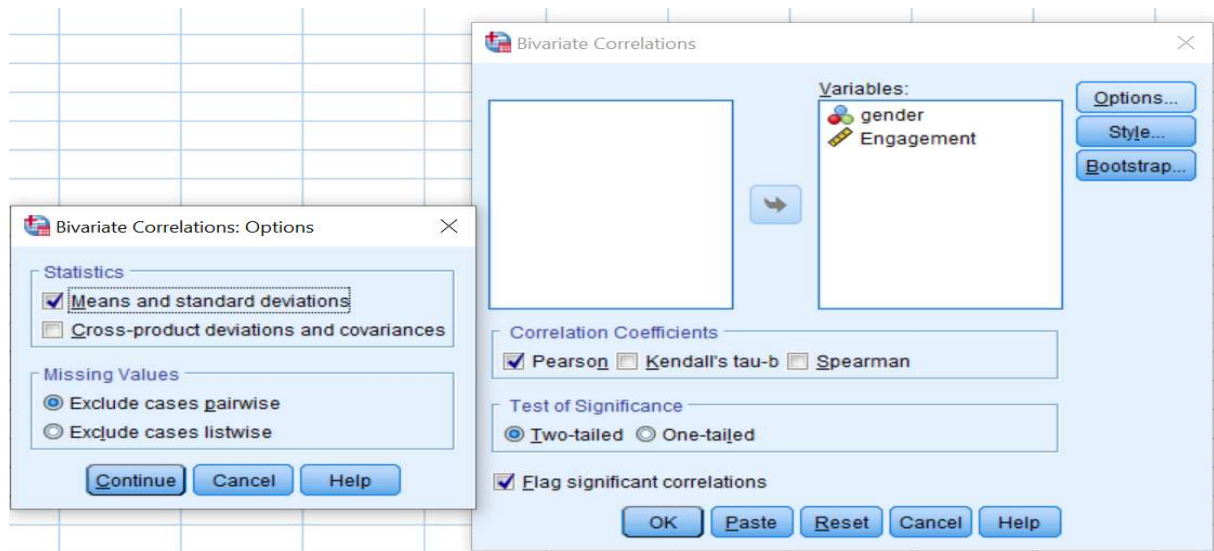
An Advertising Agency wants to determine whether there is a relationship between gender and engagement in the Internet advert. To achieve this, the Internet advert is shown to 20 men and 20 women who are then asked to complete an online survey that measures their engagement with the advertisement. The online survey results in an overall engagement score. After the data is collected, the Advertising Agency decide to use SPSS Statistics to examine the relationship between engagement and gender.

| Gender | Engagement | Gender | Engagement |
|--------|------------|--------|------------|
| Male | 5,29 | Female | 4,82 |
| Male | 5,45 | Female | 5,23 |
| Male | 5,19 | Female | 4,65 |
| Male | 5,61 | Female | 5,65 |
| Male | 5,65 | Female | 5,59 |
| Male | 5,00 | Female | 4,55 |
| Male | 5,40 | Female | 5,14 |
| Male | 5,70 | Female | 5,66 |
| Male | 5,29 | Female | 4,89 |
| Male | 5,49 | Female | 5,20 |
| Male | 5,83 | Female | 5,44 |
| Male | 6,01 | Female | 5,78 |

| | | | |
|------|------|--------|------|
| Male | 5,20 | Female | 5,12 |
| Male | 5,38 | Female | 5,23 |
| Male | 5,54 | Female | 5,46 |
| Male | 5,79 | Female | 5,57 |
| Male | 5,63 | Female | 5,51 |
| Male | 6,08 | Female | 5,55 |
| Male | 5,72 | Female | 5,71 |
| Male | 5,96 | Female | 5,80 |



Transfer the variables gender and engagement into the Variables: box by dragging-and-dropping or by clicking on the  button. You will end up with a screen similar to the one below:



Make sure that the Pearson checkbox is checked in the –Correlation Coefficients– area (although it is selected by default in SPSS Statistics)

Click on the **Options...** button. If you wish to generate some descriptives, you can do it here by clicking on the relevant checkbox in the –Statistics– area.

Descriptive Statistics

| | Mean | Std. Deviation | N |
|------------|--------|----------------|----|
| gender | 1,5000 | ,50637 | 40 |
| Engagement | 5,4513 | ,35588 | 40 |

Correlations

| | | gender | Engagement |
|------------|---------------------|--------|------------|
| gender | Pearson Correlation | 1 | -,311 |
| | Sig. (2-tailed) | | ,051 |
| | N | 40 | 40 |
| Engagement | Pearson Correlation | -,311 | 1 |
| | Sig. (2-tailed) | ,051 | |
| | N | 40 | 40 |

A point-biserial correlation was run to determine the relationship between engagement in an Internet advert and gender. There was a negative correlation between engagement and gender, which was statistically significant ($r_{pb} = -.311$, $n = 40$, $p = .051$).