# **Introduction:**

The Pearson product-moment correlation generates a coefficient called the Pearson correlation coefficient, denoted as r (i.e., the italic lowercase letter r). This coefficient measures the strength and direction of a linear relationship between two continuous variables. Its value can range from -1 for a perfect negative linear relationship to +1 for a perfect positive linear relationship. A value of 0 (zero) indicates no relationship between two variables. This statistical test is often known by its shorter title of the Pearson correlation or Pearson's correlation.

A Pearson correlation is most often used to discover the relationship between two variables. In this type of study design, you have measured two variables that are paired observations and you wish to determine the strength of the linear relationship between these two paired variables. For example, is there a linear relationship between height and basketball performance?

# **Assumptions of the Correlation:**

## 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 needs to be a linear relationship between the two variables. The best way of checking this assumption is to plot a scatterplot and visually inspect the graph.

## Assumption #3: There should be no significant outliers. Outliers are data points within your sample that do not follow a similar pattern to the other data points. Pearson's correlation coefficient, r, is sensitive to outliers, meaning that outliers can have an exaggerated influence on the value of r. This can lead to Pearson's correlation efficient not having a value that best represents the data as a whole. Therefore, it is best if there are no outliers or that they are kept to a minimum.

## Assumption #4: Data is normally distributed. If you wish to run inferential statistics (null hypothesis significance testing), you also need to satisfy the assumption of bivariate normality. You will find that this is particularly difficult to test for and so a simpler method is more commonly used, which will be demonstrated in this guide.

## Assumption #5: You will also need to show that the data has homoscedasticity. When talking about groups, we covered homogeneity or equal variances between groups. Homoscedasticity is equal variance of X for every value of Y. You can look at a scatterplot to investigate homoscedasticity.

If your data fails assumption #1, you will need to use a different statistical test. They are discussed in this order because if a violation to an assumption is not correctable, you will no longer be able to use Pearson's correlation (although you may be able to run another statistical test on your data instead).

## **Null and Alternative Hypotheses:**

The Pearson correlation coefficient, r, is a sample coefficient; that is, its value represents the strength and direction of the linear relationship that exists in the sample you studied. The null hypothesis for this test is as follows:

H0: ρ = 0; the population correlation coefficient is equal to zero. There is no relationship between Variable X and Variable Y.

And the alternative hypothesis is:

HA: ρ ≠ 0; the population correlation coefficient is not equal to zero. There is a relationship between Variable X and Variable Y.

## **Example:**

Studies show that exercising can help prevent heart disease. Within reasonable limits, the more you exercise, the less risk you have of suffering from heart disease. One way in which exercise reduces your risk is by reducing a fat in your blood called cholesterol. The more you exercise, the lower the cholesterol concentration in your blood. It has been shown that the amount of time you spend watching TV, an indicator of a sedentary lifestyle, might be a good predictor of heart disease; that is, the more TV you watch, the greater your risk of heart disease. Therefore, a researcher decided to determine if cholesterol concentration was related to time spent watching TV in otherwise healthy 45 to 65 year old men (a category of people that are at higher risk of heart disease). They believed that there would be a positive relationship; that is, men who spent more time watching TV would have a higher cholesterol concentration in their blood than those who spent less time watching TV. Daily time spent watching TV was recorded in the variable time\_tv and cholesterol concentration recorded in the variable cholesterol. Expressed in variable terms, the researcher wants to know if there is a correlation between time\_tv and cholesterol. (note: this data is fictitious. In addition, they did not decide to predict the direction of the relationship in the statistical analysis).

To get started, open the dataset for this example in JASP. Remember, you can always use the previous help guides for greater detail in case you do not remember how to do something.

File 🡪 Open 🡪 Computer 🡪 Browse 🡪 Pick the Correlation Data.

## 

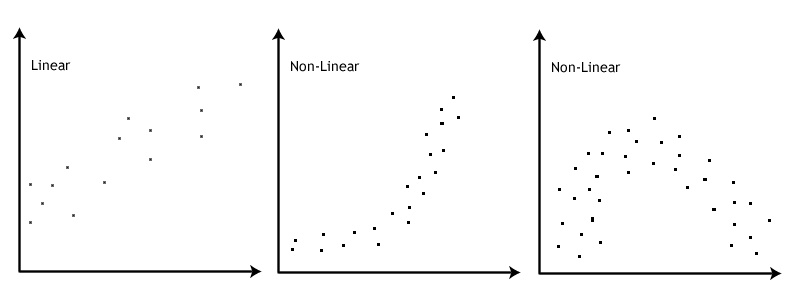
## **Check your assumptions:**

**Is the dependent variable at least scale (ratio or interval)?**

Yes, we are using ratio style data.

**Is there a linear relationship between the variables?**

Pearson's correlation is only appropriate when there is a linear relationship between your two variables; in this example, a linear relationship between "time spent watching tv" and "cholesterol concentration" (i.e., between time\_tv and cholesterol, respectively). To determine if a linear relationship exists, you need to visually inspect a scatterplot of the two variables. If the relationship approximately follows a straight line, you have a linear relationship. However, if you have something other than a straight line, for example, a curved line, you do not have a linear relationship. An example of a linear and two non-linear relationships is presented in the scatterplots below:

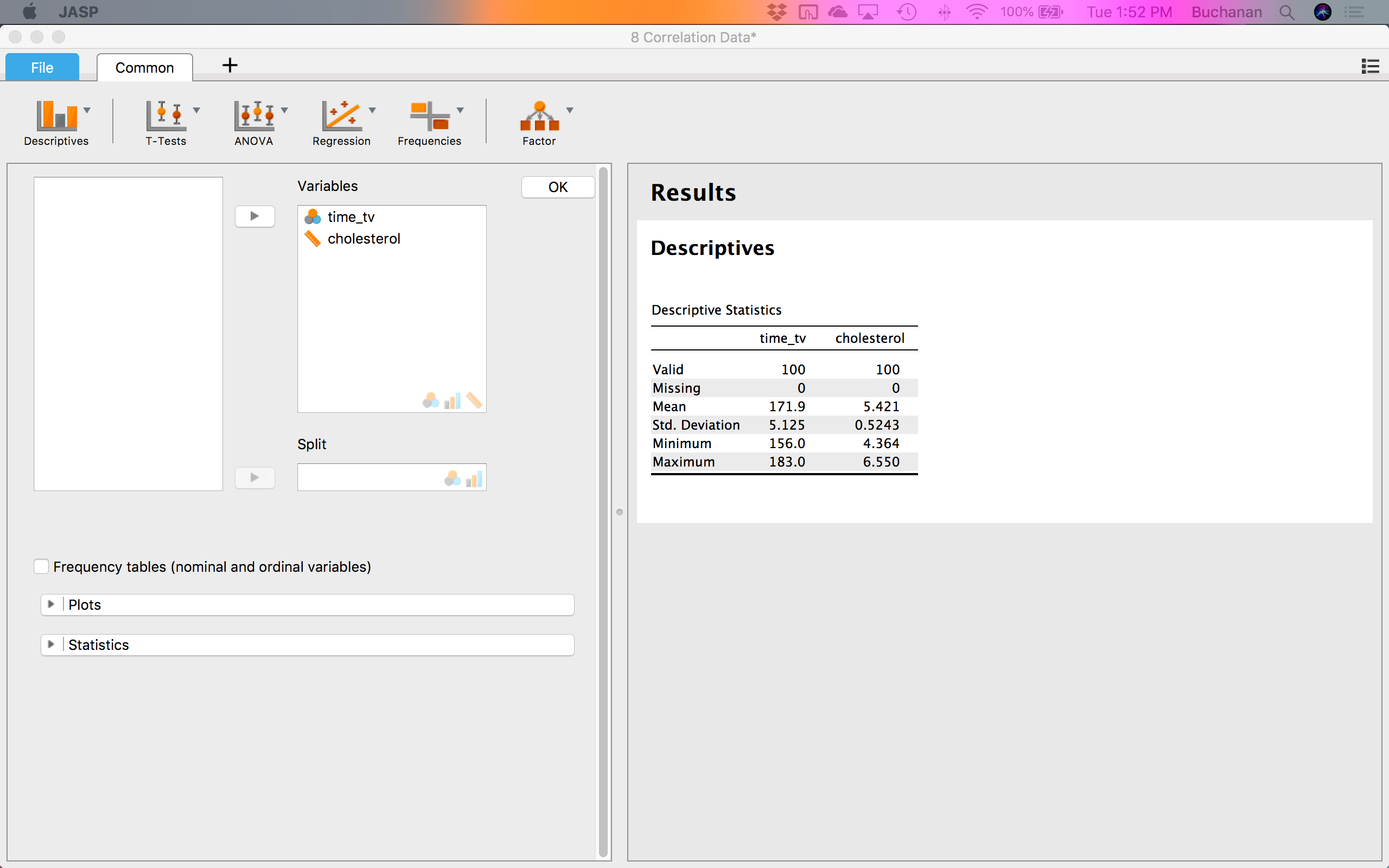


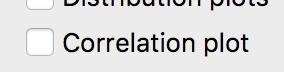
To get a scatterplot, we could use Excel as described in earlier chapters. However, there are graph options in JASP that allow you to answer this question and the outlier question below.

Click Descriptives 🡪 Descriptive Statistics.

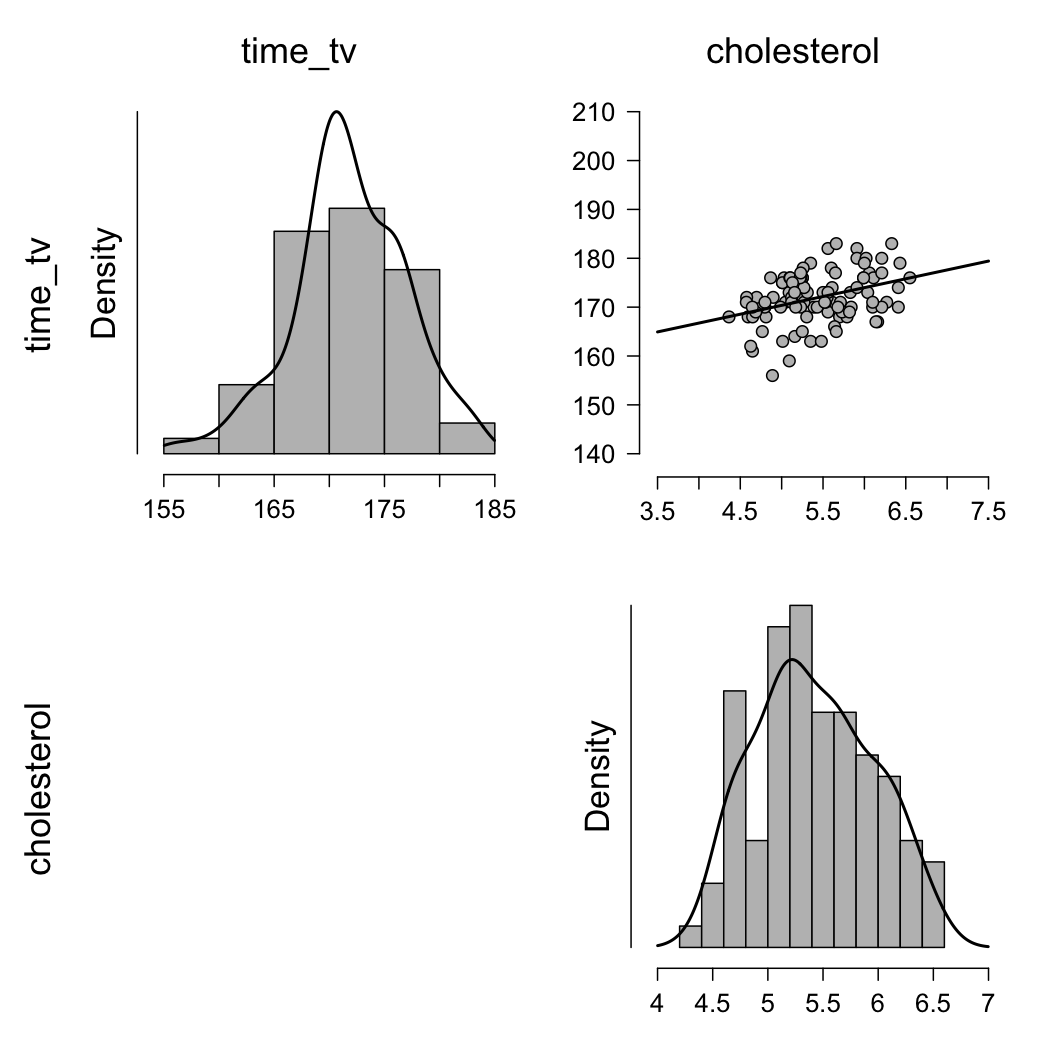


In this window, we want to click on both time\_tv and cholestorol and click the arrow  to move it over to the right hand side under Variables.



Click on the plots options:  to see more available options. Click on correlation plot . If you want to switch the X and Y axis, you can reorder the variables in the Variables window.

#### Correlation plot

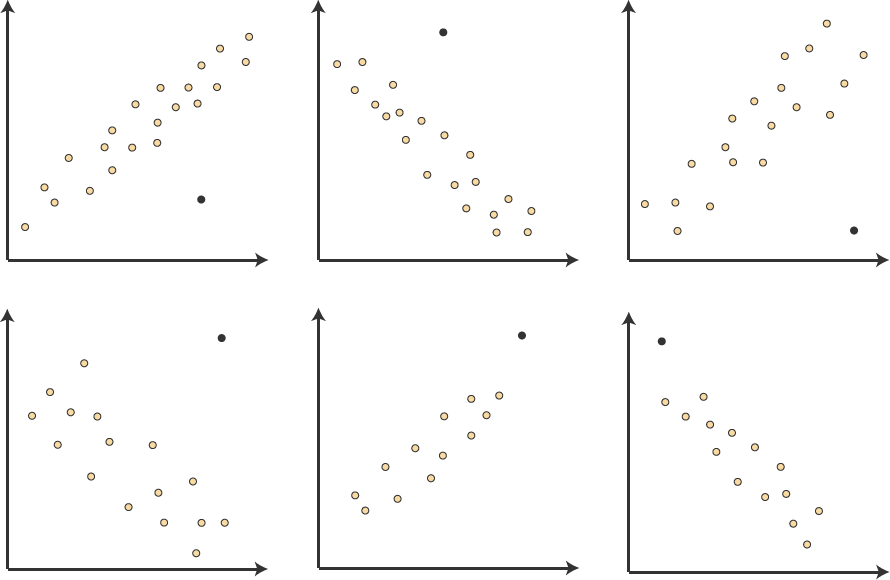


The scatterplot is presented in the top right corner. You should inspect the scatterplot above and form an opinion as to whether you believe there is enough evidence to suggest the relationship is linear. The human brain is very good at visualizing straight lines and often you can rely on your own visual inspection to determine whether the relationship is a linear one or not. For this example, you can conclude from visual inspection of the above scatterplot that there is a linear relationship between cholesterol concentration and time spent watching TV. In other situations, the relationship can sometimes be a little bit more tricky to evaluate and more care will have to be taken (especially with setting the correct scales for the x-axis and y-axis).

In this example, the linear relationship between our variables is positive; that is, as the value of time\_tv increases, so does the value of cholesterol. However, when testing your own data, you might discover a negative relationship (i.e., as the value of one variable increases, the value of the other variable decreases). You might also find that your line/relationship might be more steep or more shallow than the line/relationship in this particular example. However, for assessing linearity, all that matters is whether or not the relationship is linear (i.e., a straight line) in order to proceed.

**Are there any outliers in the sample?**

Good news! You already have the graph for outliers. You can look at the scatterplot in the top right corner for outliers. When conducting a Pearson's correlation analysis, outliers are data points that do fit the pattern of the rest of the data set. These data points can often be easily identified from the scatterplot you plotted when testing for linearity. For example, see the six scatterplots below that show six variations of outliers (identified as the black dots):



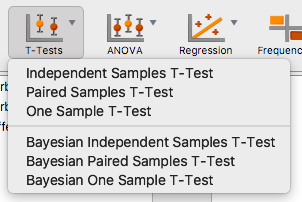
All the black dots are somewhat removed from the rest of the data set. This is a real problem for Pearson's correlation, which is particularly susceptible to outliers. The problem results in the value of Pearson's correlation coefficient being unduly altered, exerting a negative influence on the value of the correlation coefficient. As such, it is important to try to identify outliers in your data.

Further, you are given the distribution plots for the data, which have used before to determine if there are outliers. These plots are the top left and bottom right plots. From looking at these, we can see that there are not any bars that are far away from the others.

**Is the dependent variable normally distributed?**

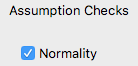
We can view the histogram created earlier to look at if the data appears normal, but we might also consider using the Shapiro-Wilk test to determine if the data is normal. To assess the statistical significance of Pearson's correlation coefficient, you need to have bivariate normality, but this assumption is difficult to assess. Therefore, in practice, a property of bivariate normality is relied upon; that is, if bivariate normality exists, both variables will be normally distributed. However, this does not work in reverse; two normally distributed variables do not mean you have bivariate normality, but it is a level of assurance that can be lived with. Therefore, you need to test both variables for normality, as instructed below – we have to use a t-test to get that output, even though we are not using a t-test for this analysis:

Click on t-tests  🡪 One Sample T-test



In this window, we want to click on time\_tv and cholesterol and click the arrow  to move it over to the right hand side under Variables.

To get the normality assumption test, click on Normality, under Assumptions:



### Assumption Checks

| **Test of Normality (Shapiro-Wilk)** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | **W** | | **p** | |
| time\_tv |  | 0.980 |  | 0.130 |  |
| cholesterol |  | 0.976 |  | 0.064 |  |
|  | | | | | |
| Note.  Significant results suggest a deviation from normality. | | | | | |

The Shapiro-Wilk test ran for each variable separately, so we can tell if the assumptions were met for each variable. We see that our data is normally distributed because *p* > .05 for both groups. Ignore the rest of the output!

If the assumption of normality has been violated, the p value will be less than .05 (i.e., the test is significant at the p < .05 level). If the assumption of normality has not been violated, the p value will be greater than .05 (i.e., p > .05). This is because the Shapiro-Wilk test is testing the null hypothesis that your data's distribution is equal to a normal distribution. Rejecting the null hypothesis means that your data's distribution is not equal to a normal distribution. In the table above, you can see that both the p values are greater than .05 (they are .130 and .064). Therefore, your variables, time\_tv and cholesterol, are normally distributed. However, be aware that larger sample sizes (e.g., above 50 cases) can lead to a statistically significant result (i.e., data are non-normal) even when they are normal. For larger sample sizes, graphical interpretation is often preferred (looking at histograms, etc.).

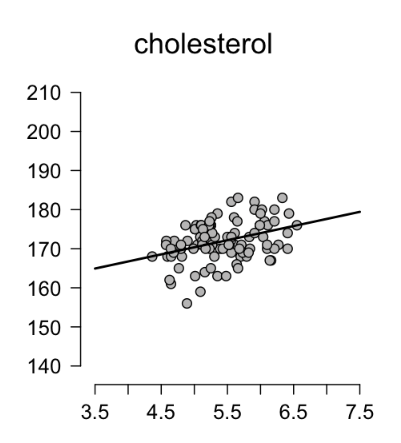
You can report these findings as follows:

Both variables were normally distributed, as assessed by Shapiro-Wilk’s test (TV time *p* = .130, Cholesterol *p* = .064).

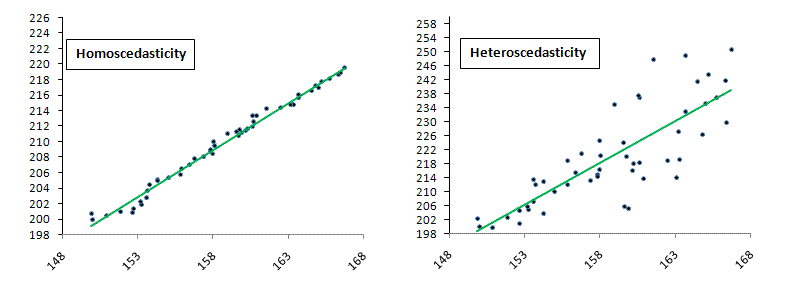
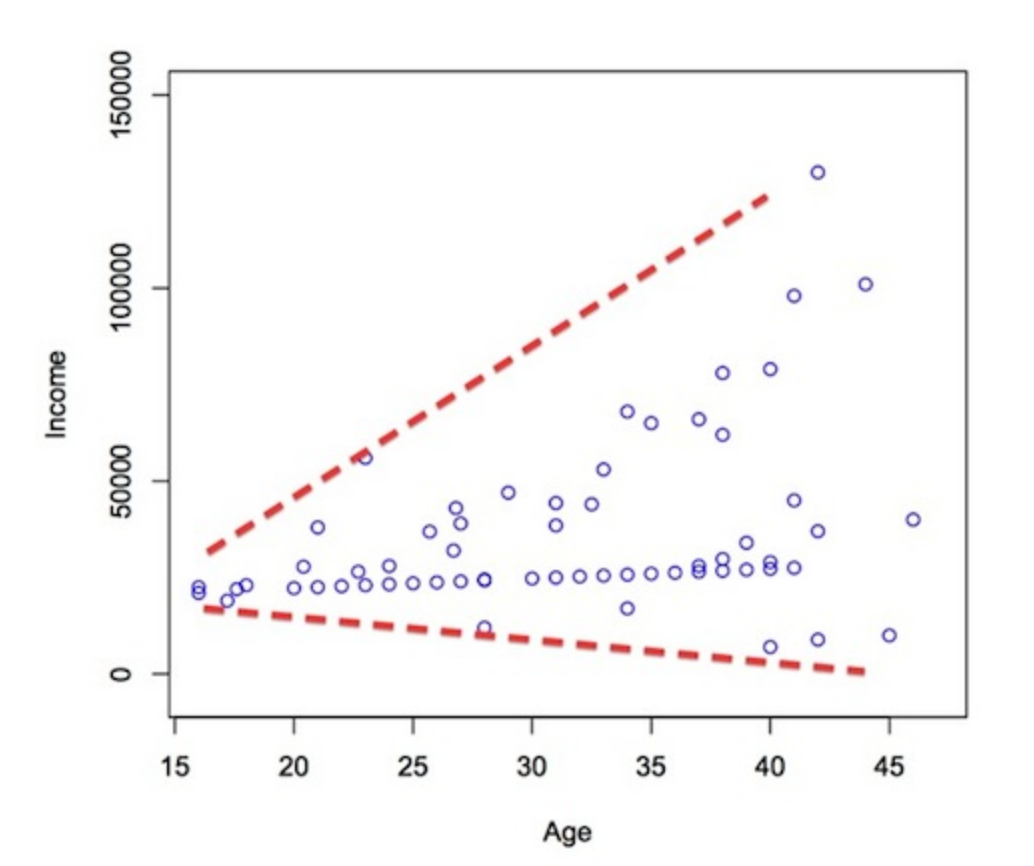
**Do you have homoscedasticity?**

Homoscedasticity occurs when the spread of X is the same all the down Y and vice versa. You do not want X and Y on a scatterplot to be shaped like a megaphone, triangle, or other odd shapes, which would indicate heteroscadasticity (that’s bad).

To investigate, we can use our scatterplot from earlier. By drawing a line around the dots, we can see that the shape is mostly square, give or take a little. This shape would indicate homoscedasticity. We will have to be careful if we do not have a lot of data because scatterplots with small amounts of dots can be hard to interpret.

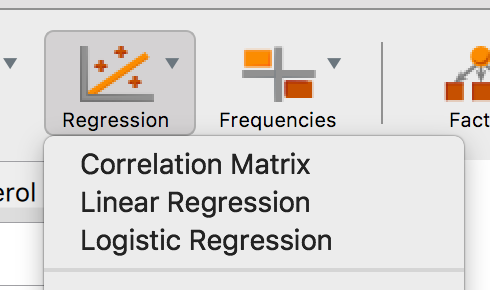


Here are some examples of bad graphs – those that have heteroscedasticity:



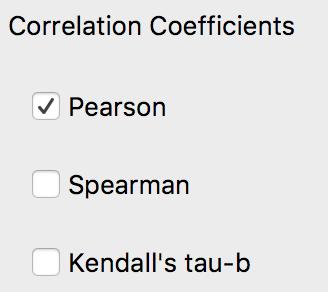
## **The correlation test:**

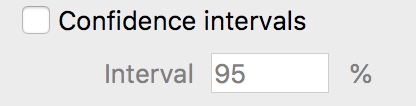
To get the correlation, you will click on Regression  🡪 Correlation Matrix:



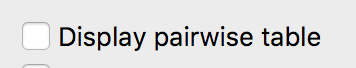
Move both variables over to the right hand window by clicking on the arrow . Note: If you are looking to calculate more than one correlation (i.e., you have more than two variables), simply transfer more variables into the Variables: box.

Under correlation coefficients, you can leave Pearson checked but if you need to run Spearman’s or Kendall’s, you can also check those.



The correlation table is automatically presented on the right for you. If you want to include the confidence interval, you can click confidence intervals: 

| **Pearson Correlations** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | **cholesterol** | | **time\_tv** | |
| cholesterol |  | Pearson's r |  | — |  |  |  |
| p-value |  | — |  |  |  |
| Upper 95% CI |  | — |  |  |  |
| Lower 95% CI |  | — |  |  |  |
| time\_tv |  | Pearson's r |  | 0.371 |  | — |  |
| p-value |  | < .001 |  | — |  |
| Upper 95% CI |  | 0.529 |  | — |  |
| Lower 95% CI |  | 0.188 |  | — |  |
|  | | | | | | | |

You will see that there are dashes --- in some of the boxes – this indicates that the correlation was not calculated because it would have been the correlation with itself (i.e., cholesterol to cholesterol). You can also view the correlations in a slightly easier table by clicking on Display pairwise table .

| **Pearson Correlations** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | |  | | **Pearson's r** | | **p** | | **Lower 95% CI** | | **Upper 95% CI** | |
| cholesterol |  | - |  | time\_tv |  | 0.371 |  | < .001 |  | 0.188 |  | 0.529 |  |
|  | | | | | | | | | | | | | |

This view shows you the correlations one at a time with the confidence interval next to it rather than beneath it.

The actual correlation is shown under Pearson’s *r*, the *p*-value shows the two-tailed significance level of the correlation coefficient, and then the lower and upper confidence interval of the *r* value.

In this example, the Pearson correlation coefficient, r, is .371. As the sign of the Pearson correlation coefficient is positive, you can conclude that there is a positive correlation between the daily time spent watching TV (tv\_time) and cholesterol concentration (cholesterol); that is, cholesterol concentration increases as time spent watching TV increases.

Important Note: Some would object to the description, "cholesterol concentration increases as time spent watching TV increases". The reason for this objection is rooted in the meaning of "increases". The use of this verb might suggest that the effect of this variable is causal and/or manipulatable such that you could increase the time spent watching TV (tv\_time) in your partipicants and this would lead to an increase in their cholesterol concentration (cholesterol). This is not to say this might not be possible. However, this knowledge is not contained in the correlation, but in theory. As such, you might prefer to state the relationship as, "higher values of cholesterol concentration are associated/related to greater time spent watching TV".

The magnitude of the Pearson correlation coefficient determines the strength of the correlation. Although there are no hard-and-fast rules for assigning strength of association to particular values, some general guidelines are provided by Cohen (1988):

|  |  |
| --- | --- |
| **Coefficient Value** | **Strength of Association** |
| 0.1 < | *r* | < .3 | small correlation |
| 0.3 < | *r* | < .5 | medium/moderate correlation |
| | *r* | > .5 | large/strong correlation |

where | *r* | means the absolute value or *r* (e.g., | *r* | > .5 means *r* > .5 and *r* < -.5). Therefore, the Pearson correlation coefficient in this example (*r* = .371) suggests a medium strength correlation. If instead, *r* = -.371, you would also have had a medium strength correlation, albeit a negative one.

The results you have reported so far have only used the Pearson correlation coefficient to describe the relationship between the two variables in your sample. If you wish to test hypotheses about the linear relationship between your variables in the population your sample is from, you need to test the statistical significance. Remember that statistical significance does not determine the strength of the relationship (r or ρ does that), but whether the correlation coefficient is statistically significantly different from zero. You could write this result as follows:

There was a moderate positive correlation between daily time spent watching TV and cholesterol concentration in males aged 45 to 65 years, r = .371, *p* < .001.

**Effect Size:**

The coefficient of determination is the proportion of variance in one variable that is "explained" by the other variable and is calculated as the square of the correlation coefficient (r2). In this example, you have a coefficient of determination, r2, equal to 0.3712 = 0.14. This can also be expressed as a percentage (i.e., 14%). Remember that this "explained" refers to being explained statistically, not causally. You could write this as:

Daily time spent watching TV statistically explained 14% of the variability in cholesterol concentration.

## **Reporting All Together:**

A Pearson's product-moment correlation was run to assess the relationship between cholesterol concentration and daily time spent watching TV in males aged 45 to 65 years. Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test (TV time *p* = .130, Cholesterol *p* = .064), and there were no outliers. There was a moderate positive correlation between daily time spent watching TV and cholesterol concentration, *r* = .371, *p* < .001, with time spent watching TV explaining 14% of the variation in cholesterol concentration**.**