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.

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.

**Variables required**

In order to run a Pearson correlation, you require the following variables:

» Two variables that are **continuous** (e.g., weight, cholesterol) and are paired (i.e., each case has two values - one for each variable).

## What problems can you solve using a Pearson 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 & order of testing**

* **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:** 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:** Data is normally distributed.

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).

## Flowchart depicting order of testing

Carrying out, interpreting and reporting Pearson's correlation (and its assumptions) is a simple, nine-step process (haha). For a Pearson's correlation to be a valid test of the strength of a linear relationship, you might have to make adjustments to your data to satisfy the assumptions underlying this test (e.g., transforming your data) or to use another test entirely (e.g., a Spearman's correlation). These nine steps are highlighted in the flowchart below:

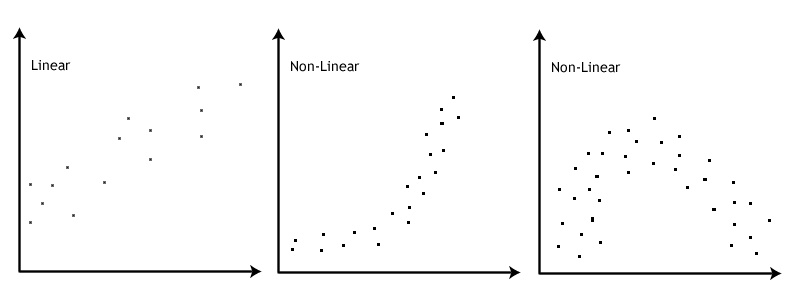
Correlation Flowchart

**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.)

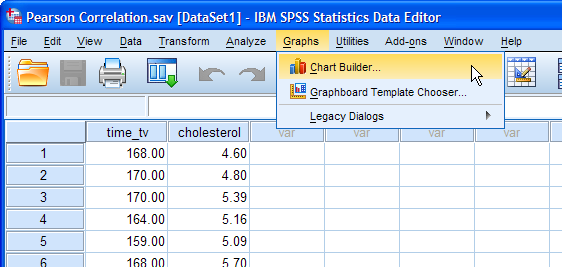
## TESTING OF ASSUMPTIONS

**Establishing if a linear relationship exists**

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:

In order to run a Pearson correlation, you are ideally looking for a linear relationship similar to the first scatterplot presented above. The following instructions will show you how to produce a scatterplot in SPSS to establish if a linear relationship exists:

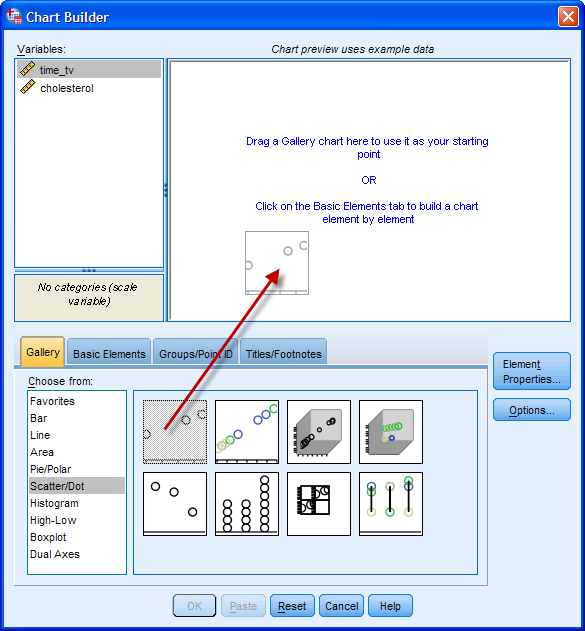
1. Click **Graphs > Chart Builder...** on the main menu, as shown below:



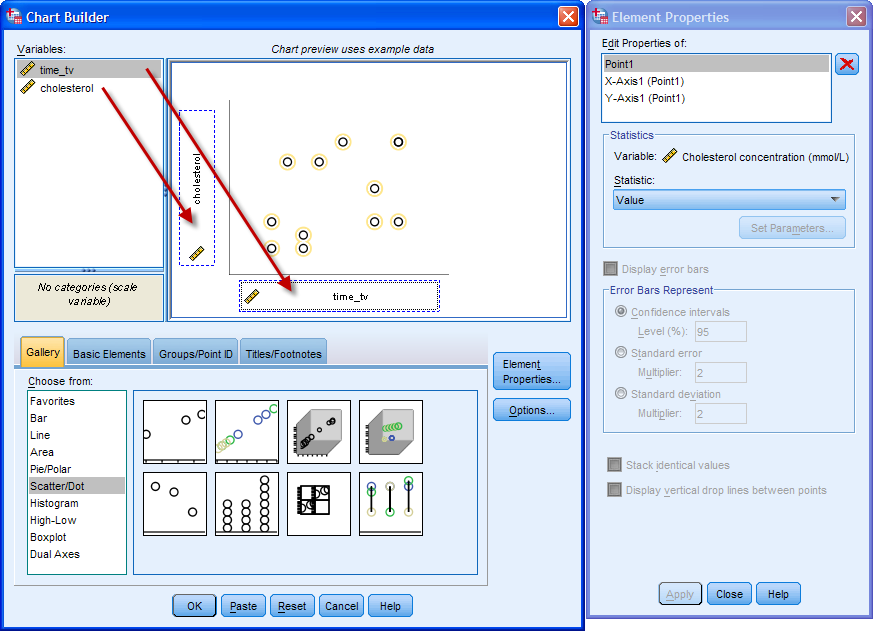
2. Select "Scatter/Dot" from the Choose from: box in the bottom-left-hand corner of the **Chart Builder** dialogue box.

3. Selecting "Scatter/Dot" will present eight different scatter/dot options in the lower-middle section of the **Chart Builder** dialogue box (as shown above and below). Drag-and-drop the top left-hand option (you will see it labeled as "Simple Scatter" if you hover your mouse over the box) into the main chart preview pane, as shown below:

4. You will be presented with the screen below, which shows a simple scatterplot in the main chart preview pane with boxes for the y-axis ("Y-Axis?") and x-axis ("X-Axis?") for you to populate with the appropriate variables.

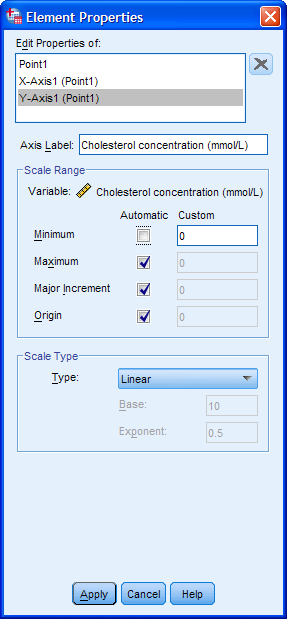
5. Drag-and-drop time\_tv from the Variables: box into the "X-axis?" box in the main chart preview screen and do the same for cholesterol but into the "Y-axis?" box. You should end up with a screen like below:

Note 1: The chart preview screen does not accurately plot the variable data that you have dragged across into the preview pane, even though it might appear that it does due to the scatterplot dots changing when you add your variables. Therefore, do not get confused and think that you have done something wrong. You will only see your true data when you actually generate the scatterplot.



Note 2: Although the Pearson's correlation does not make any distinction between dependent and independent variables, it is still customary to consider the graph's x- and y-axes in such a way. For example, this study is assuming that time spent watching TV influences cholesterol concentration, not the other way around. Therefore, time spent watching TV goes on the x-axis and cholesterol concentration on the y-axis, even though the Pearson's correlation does not make this distinction.

6. Click on "Y-Axis1 (Point1)" in the Edit Properties of: box located in the **Element Properties** dialogue box (the box on the right-hand side of the main **Chart Builder** dialogue box).

7. Uncheck the Minimum option in the -Scale Range- area so that the Custom box is highlighted and has a value of 0 (zero).

8. Click the https://statistics.laerd.com/premium/pc/img/apply-button.pngbutton to confirm these changes.

9. Click the https://statistics.laerd.com/premium/pc/img/ok-button.pngbutton in the **Chart Builder** dialogue box to generate the scatterplot.

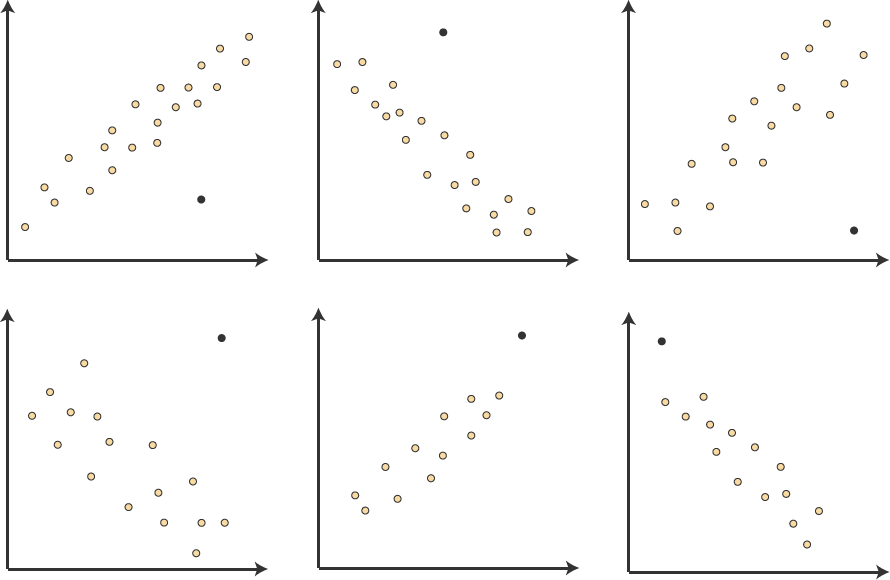
## https://statistics.laerd.com/premium/pc/img/graph-scatterplot-basic-scale.pngInterpretation of linearity

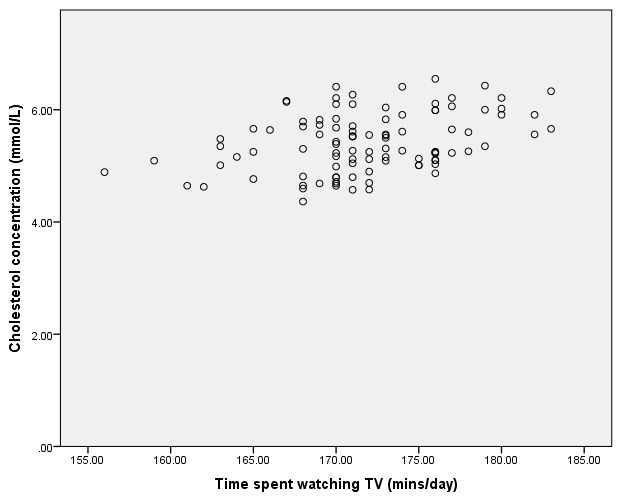
You should inspect the scatterplot above and form an opinion as to whether you believe there is enough evidence to suggest the relationship is linear. The human brain is very good at visualizing straight lines and often you can rely on your own visual inspection to determine whether the relationship 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.

## Testing 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 unduely 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. This can be accomplished by inspecting the scatterplot of your two variables. In this example, the following scatterplot was plotted:

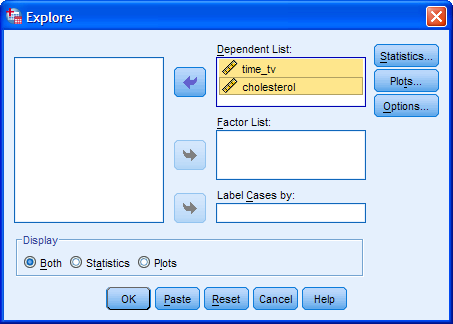
From inspection of this scatterplot it can be concluded that there are no outliers in this data set.

## Testing for normality

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:

1. Click **Analyze > Descriptive Statistics > Explore...** on the main menu.

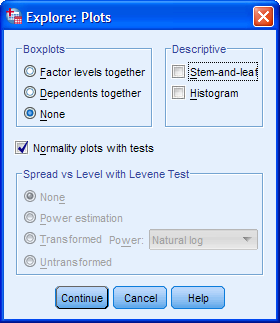
2. Transfer the variables, time\_tv and cholesterol, into the Dependent List: box by clicking on both of them whilst holding down the shift-key and then clicking the top https://statistics.laerd.com/premium/pc/img/right-arrow-button.pngbutton. You will end up with the following screen:



Note: If you have more than two variables, you need to transfer all the variables you are analyzing into the Dependent List: box.

3. Click the https://statistics.laerd.com/premium/pc/img/plots-button.pngbutton and you will be presented with the **Explore: Plots** dialogue box.

4. Select None in the -Boxplots- area and deselect Stem-and-leaf in the -Descriptive- area, but select Normality plots with tests, so that you end up with the following screen:



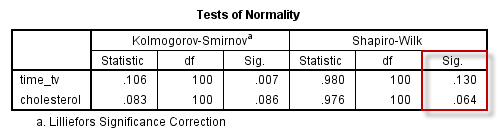
5. Click the https://statistics.laerd.com/premium/pc/img/continue-button.pngbutton and you will be returned to the **Explore** dialogue box.

Note: If you use skewness and kurtosis to determine normality, rather than the tests illustrated here, you will need to keep the default of Both in the -Display- area, as these statistics are produced by default when the "Statistics" options are also selected. **Utilizing the skewness and kurtosis values and seeing if they are <1 may be a more appropriate indicator of normality than the normality tests for real world data.**

6. Click the https://statistics.laerd.com/premium/pc/img/ok-button.pngbutton. There will be a great deal of output, including many statistics and graphs that you will not need. The next section will explain which particular parts of the output you actually need and how to interpret them.

**Determining if your data is normally distributed**

The Shapiro-Wilk test is recommended if you are not particularly familiar with statistics and/or if you have a small sample size. The results from the Shapiro-Wilk test are presented in the **Tests of Normality** table. You can see that a Shapiro-Wilk test has been run for each variable. If you look under the "**Sig.**" column located under the "**Shapiro-Wilk**" column, you will find the significance value for this test for each variable, as highlighted below:



If the assumption of normality has been violated, the "**Sig.**" 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 "**Sig.**" 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 "**Sig.**" 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.).

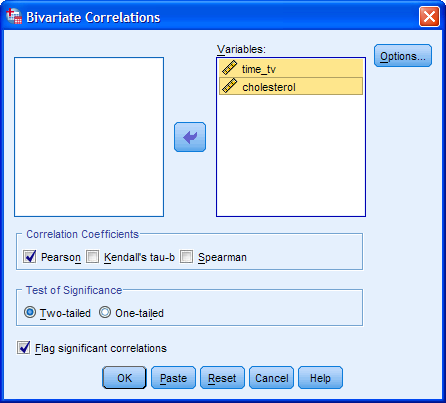
You can report these findings as follows: Both variables were normally distributed, as assessed by Shapiro-Wilk’s test (*p* > .05). OR Not all variables were normally distributed, as assessed by Shapiro-Wilk’s test (*p* < .05).

## Pearson's correlation procedure

The following procedure shows you how to run a Pearson's correlation in SPSS.

1. Click **Analyze > Correlate > Bivariate...** on the main menu.

2. Highlight both variables by clicking on both time\_tv and cholesterol whilst holding down the shift-key. Then transfer these variables into the Variables: box by clicking on the https://statistics.laerd.com/premium/pc/img/right-arrow-button.pngbutton. You will end up with the following screen:



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.

3. Make sure that the Pearson checkbox is selected in the -Correlation Coefficients- area, although this should be selected by default.

4. Click the https://statistics.laerd.com/premium/pc/img/options-button.pngbutton. You will be presented with the **Bivariate Correlations: Options** dialogue box.

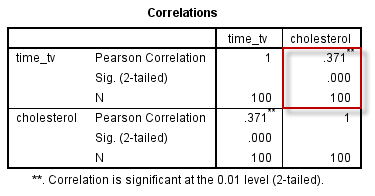
5. Leave the options in the -Statistics- area unchecked unless you have a specific reason to want the information it provides; they will not be used in this guide. Keep the Exclude cases pairwise option selected in the -Missing Values- area.

6. Click the https://statistics.laerd.com/premium/pc/img/continue-button.pngbutton. You will be returned to the **Bivariate Correlations** dialogue box.

7. Click the https://statistics.laerd.com/premium/pc/img/ok-button.pngbutton to generate the output.

## Determining the correlation coefficient

The main reason for running a Pearson correlation analysis is to obtain the value of the Pearson correlation coefficient. This value can be obtained from the **Correlations** table, which is the only table produced by SPSS. You will notice from the table above that the results are in a matrix such that each result is repeated. However, you are only concerned with the table cells that contains the Pearson correlation between the variables tv\_time and cholesterol, such as the table cell shown below:



The information in each row of the **Correlations** table (shown above) has the following meaning:

|  |  |
| --- | --- |
| **Row Title** | **Row Meaning** |
| Pearson Correlation | Pearson's correlation coefficient |
| Sig. (2-tailed) | Two-tailed significance level (*p*-value) of the correlation coefficient. |
| N | Number of paired observations (e.g., participants included in correlation).  In other words, the sample size. |

In this example, the Pearson correlation coefficient, r, is .371 (the "**Pearson Correlation**" row). 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. 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.

## Determining Statistical Significance

## 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. The statistical significance (p-value) of the correlation coefficient in this example would appear to be .000 (obtained from the "****Sig. (2-tailed)****" row). However, if you ever see SPSS Statistics print out a p-value of .000, do not interpret this as a significance level that is actually zero; it does, in fact, indicate that p < .0005. As p < .05 in this example (it is p < .0005), it can be concluded that the correlation coefficient is statistically significantly different from zero. 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 can 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*(98) = .371, *p* < .001).

## OR An increase in daily time spent watching TV was moderately correlated with an increase in cholesterol concentration in males aged 45 to 65 years, *r*(98) = .371, *p* < .001.

## Reporting

## Putting it all together: you can report your findings as follows:

## 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 (p > .05), and there were no outliers. There was a moderate positive correlation between daily time spent watching TV and cholesterol concentration, r(98) = .371, p < .001, with time spent watching TV explaining 14% of the variation in cholesterol concentration.

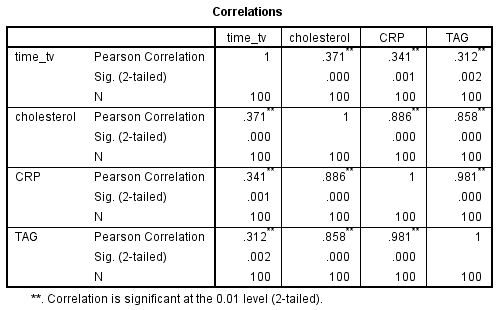
## Effect Size

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.

**APA Tables**

If you are reporting multiple correlations, you will generate a larger correlation matrix in the SPSS Statistics output that can start to get rather unwieldy, such as the one below:



You can present the correlation matrix above in a much simpler table, as shown below:

