



**JASP**

**STATISTICAL ANALYSIS IN JASP:  
A GUIDE FOR STUDENTS**



# JASP

Copyright © 2018 by Mark A Goss-Sampson.

All rights reserved. This book or any portion thereof may not be reproduced or used in any manner whatsoever without the express written permission of the author except for the purposes of research, education or private study.



## CONTENTS

PREFACE .....	1
USING THE JASP INTERFACE.....	2
DESCRIPTIVE STATISTICS .....	8
EXPLORING DATA INTEGRITY .....	15
ONE SAMPLE T-TEST .....	22
BINOMIAL TEST .....	25
MULTINOMIAL TEST.....	28
CHI-SQUARE 'GOODNESS-OF-FIT' TEST.....	30
MULTINOMIAL AND $\chi^2$ 'GOODNESS-OF-FIT' TEST.....	31
COMPARING TWO INDEPENDENT GROUPS.....	32
INDEPENDENT T-TEST .....	32
MANN-WITNEY U TEST .....	36
COMPARING TWO RELATED GROUPS.....	38
PAIRED SAMPLES T-TEST .....	38
WILCOXON'S SIGNED RANK TEST.....	40
CORRELATION ANALYSIS.....	42
REGRESSION .....	48
SIMPLE REGRESSION .....	51
MULTIPLE REGRESSION.....	54
LOGISTIC REGRESSION .....	61
COMPARING MORE THAN TWO INDEPENDENT GROUPS .....	66
ANOVA .....	66
KRUSKAL-WALLIS .....	71
COMPARING MORE THAN TWO RELATED GROUPS .....	74
RMANOVA.....	74
FRIEDMAN'S REPEATED MEASURES ANOVA .....	80
TWO-WAY INDEPENDENT ANOVA .....	82
MIXED FACTOR ANOVA USING JASP .....	87
CHI-SQUARE TEST FOR ASSOCIATION .....	95
EXPERIMENTAL DESIGN AND DATA LAYOUT IN EXCEL FOR JASP IMPORT.....	102
Independent t-test.....	102
Paired samples t-test .....	103



Correlation .....	104
Logistic Regression .....	106
One-way Independent ANOVA .....	107
One-way repeated measures ANOVA.....	108
Two-way Independent ANOVA.....	109
Two-way Mixed Factor ANOVA.....	110
Chi-squared - Contingency tables .....	111
SOME CONCEPTS IN FREQUENTIST STATISTICS .....	112
WHICH TEST SHOULD I USE? .....	116
Comparing one sample to a known or hypothesized population mean.....	116
Testing relationships between two or more variables .....	116
Predicting outcomes .....	117
Testing for differences between two independent groups.....	117
Testing for differences between two related groups .....	118
Testing for differences between three or more independent groups.....	118
Testing for differences between three or more related groups.....	119
Test for interactions between 2 or more independent variables.....	119



## PREFACE

JASP stands for **Jeffrey's Amazing Statistics Program** in recognition of the pioneer of Bayesian inference Sir Harold Jeffreys. This is a **free** multi-platform open-source statistics package, developed and continually updated (currently v 0.9.0.1 as of June 2018) by a group of researchers at the University of Amsterdam. Their aim was to develop a free, open-source programme that includes both standard and more advanced statistical techniques with a major emphasis on providing a simple intuitive user interface.

In contrast to many statistical packages, JASP provides a simple drag and drop interface, easy access menus, intuitive analysis with real-time computation and display of all results. All tables and graphs are presented in APA format and can be copied directly and/or saved independently. Tables can also be exported from JASP in LaTeX format

JASP can be downloaded free from the website <https://jasp-stats.org/> and is available for Windows, Mac OS X and Linux. You can also download a pre-installed Windows version that will run directly from a USB or external hard drive without the need to install it locally. The programme also includes a data library with an initial collection of over 50 datasets from Andy Fields book, *Discovering Statistics using IBM SPSS statistics*<sup>1</sup> and *The Introduction to the Practice of Statistics*<sup>2</sup> by Moore, McCabe and Craig.

Since May 2018 JASP can also be run directly in your browser via rollApp without having to install it on your computer (<https://www.rollapp.com/app/jasp>). However, this may not be the latest version of JASP.

Keep an eye on the JASP site since there are regular updates as well as helpful videos and blog posts!!

This document is a collection of standalone handouts covering the most common standard (frequentist) statistical analyses used by students studying Biological Sciences. Datasets used in this document are available for download from <http://bit.ly/2wlbMvf>.

Dr Mark Goss-Sampson  
Centre for Science and Medicine in Sport  
University of Greenwich  
2018

---

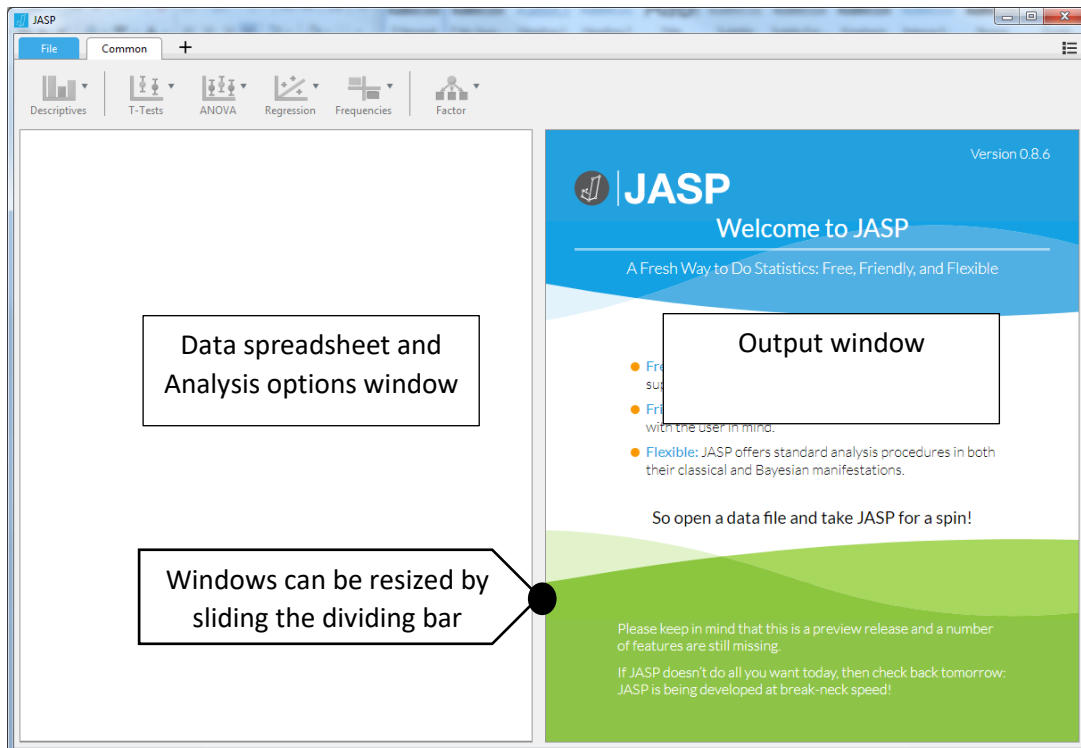
<sup>1</sup> A Field. (2017) *Discovering Statistics Using IBM SPSS Statistics* (5<sup>th</sup> Ed.) SAGE Publications.

<sup>2</sup> D Moore, G McCabe, B Craig. (2011) *Introduction to the Practice of Statistics* (7<sup>th</sup> Ed.) W H Freeman.



## USING THE JASP INTERFACE

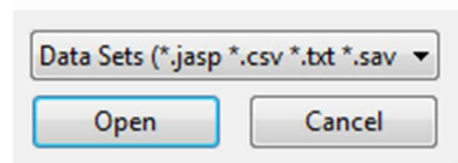
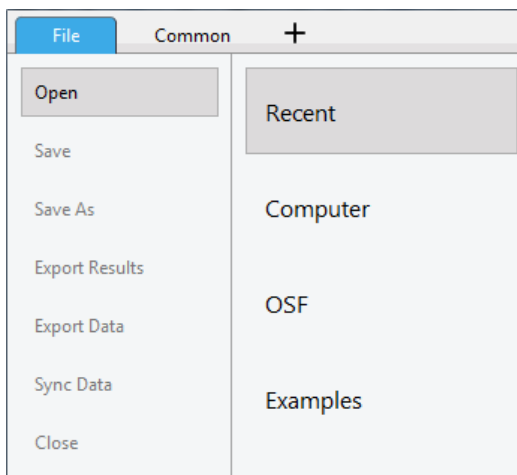
Open JASP.



JASP has its own **.jasp** format but can open a variety of different dataset formats such as:




- **.csv** (comma separated values) normally saved in Excel
- **.txt** (plain text) also can be saved in Excel
- **.sav** (IBM SPSS data file)
- **.ods** (Open Document spreadsheet)

Clicking on the File tab or “So open a data file and take JASP for a spin” in the welcome screen allows you to open recent files, browse your computer files, and access the Open Science Framework (OSF) or the wide range of examples that are packaged with JASP.





All files must have a header label in the first row. Once loaded, the dataset appears in the left window:

	 Game	 Country code	 Number of England Injuries
1	1	France	7
2	2	Tonga	4
3	3	New Zealand	2
4	4	France	5
5	5	Tonga	1
6	6	Wales	2
7	7	Wales	5
8	8	New Zealand	4
9	9	Wales	4
10	10	Tonga	3
11	11	Wales	5
12	12	Wales	3
13	13	France	6

For large datasets, there is a hand icon which allows easy scrolling through the data.

On import JASP makes a best guess at assigning data to the different variable types:

**Nominal**



**Ordinal**



**Continuous**



If JASP has incorrectly identified the data type just click on the appropriate variable data icon in the column title to change it to the correct format.

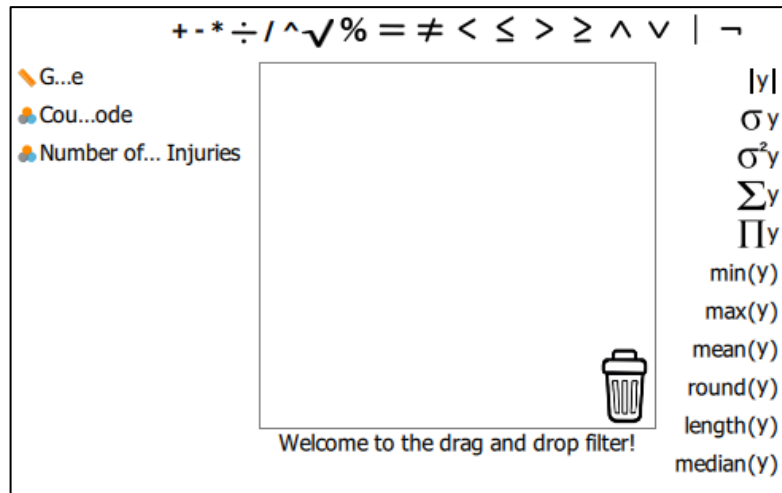
If you have coded the data you can click on the variable name to open up the following window in which you can label each code. These labels now replace the codes in the spreadsheet view. If you save this as a **.jasp** file these codes, as well as all analyses and notes, will be saved automatically. This makes the data analysis fully reproducible.

Filter	Value	Label
<input checked="" type="checkbox"/>	1	Tonga
<input checked="" type="checkbox"/>	2	New Zealand
<input checked="" type="checkbox"/>	3	France
<input checked="" type="checkbox"/>	4	Wales

In this window you can also carry out simple filtering of data, for example, if you untick the Wales label it will not be used in subsequent analyses.



Clicking this icon in the spreadsheet window opens up a much more comprehensive set of data filtering options:



Using this option will not be covered in this document. For detailed information on using more complex filters refer to the following link: <https://jasp-stats.org/2018/06/27/how-to-filter-your-data-in-jasp/>

By default, JASP plots data in the Value order (i.e. 1-4). The order can be changed by highlighting the label and moving it up or down using the using the appropriate arrows:

Filter	Value	Label
<input checked="" type="checkbox"/>	1	Tonga
<input checked="" type="checkbox"/>	2	New Zealand
<input checked="" type="checkbox"/>	3	France
<input checked="" type="checkbox"/>	4	Wales

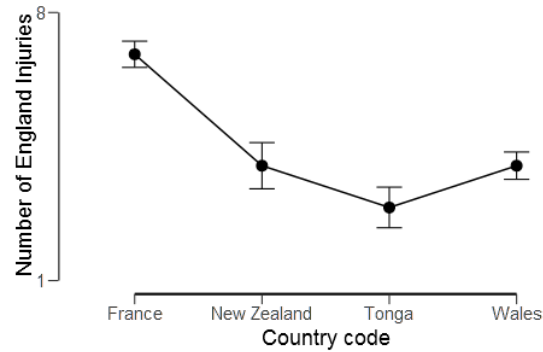
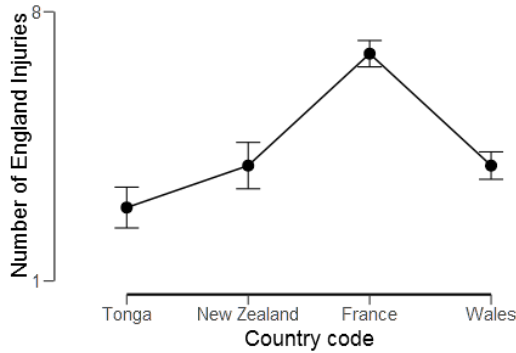
Move up

Move down

Reverse order


Close

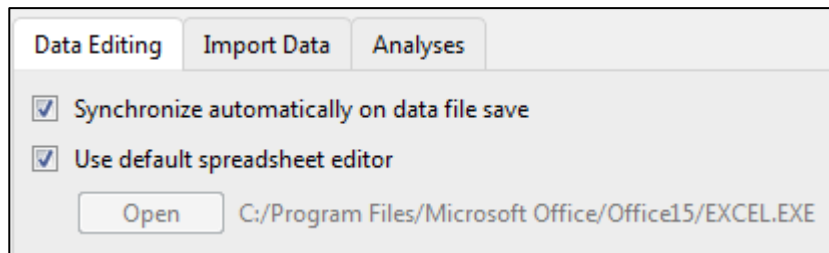




Filter	Value	Label
✓	1	Tonga
✓	2	New Zealand
✓	3	France
✓	4	Wales

Filter	Value	Label
✓	3	France
✓	2	New Zealand
✓	1	Tonga
✓	4	Wales

If you need to edit the data in the spreadsheet just double click on a cell and the data should open up in the original spreadsheet i.e. Excel. You can change the option of which spreadsheet editor that you use by clicking on  icon in the top right corner of the JASP window and select **Preferences**.



In this window you can change the spreadsheet option to SPSS, ODS etc. We will come back to preferences later.

Once you have edited your data and saved the original spreadsheet JASP will automatically update to reflect the changes that were made provided that you have not changed the file name.

## JASP ANALYSIS MENU



The **common** analysis options can be accessed from the main toolbar. Currently (v0.9.0.1) offers the following frequentist (standard statistics) and alternative Bayesian tests:



<b>Descriptives</b> <ul style="list-style-type: none"> <li>• Descriptive stats</li> <li>• Reliability analysis*</li> </ul>	<b>Regression</b> <ul style="list-style-type: none"> <li>• Correlation</li> <li>• Linear regression</li> <li>• Logistic regression</li> </ul>
<b>T-Tests</b> <ul style="list-style-type: none"> <li>• Independent</li> <li>• Paired</li> <li>• One sample</li> </ul>	<b>Frequencies</b> <ul style="list-style-type: none"> <li>• Binomial test</li> <li>• Multinomial test</li> <li>• Contingency tables</li> <li>• Log-linear regression*</li> </ul>
<b>ANOVA</b> <ul style="list-style-type: none"> <li>• Independent</li> <li>• Repeated measures</li> <li>• ANCOVA*</li> </ul>	<b>Factor</b> <ul style="list-style-type: none"> <li>• Principal Component Analysis (PCA)*</li> <li>• Exploratory Factor Analysis (EFA)*</li> </ul>

\* Not covered in this document

BY clicking on the + icon on the top menu you can also access advanced options including; Network analysis, Meta-Analysis, Structural Equation Modelling and Bayesian Summary stats.

Once you have selected your required analysis all the possible statistical options appear in the left window and output in the right window.

The screenshot shows the JASP software interface. On the left, the 'Descriptives' dialog box is open, showing the variable 'len' selected for analysis. The 'Statistics' section is expanded, showing options for Percentile Values, Central Tendency, Dispersion, and Distribution. On the right, the 'Descriptives' output window displays a table of descriptive statistics for the variable 'len' across three dose levels (500, 1000, 2000) and a corresponding boxplot.

	len		
	500	1000	2000
Valid	20	20	20
Missing	0	0	0
Mean	10.605	19.735	26.100
Std. Deviation	4.500	4.415	3.774
Minimum	4.200	13.600	18.500
Maximum	21.500	27.300	33.900

**Click in this window to toggle between analysis options and spreadsheet in the left window**

If you hover the cursor over the Results a ▼ icon appears, clicking on this provides a range of options including:



- **Remove all** analyses from the output window
- **Remove** selected analysis
- **Collapse** the output
- **Add notes** to each output
- **Copy**
- **Copy special (LaTeX code)**
- **Save image as**

The 'add notes' option allows the results output to be easily annotated and then exported to an HTML file by going to File > Export Results.

ANOVA - Number of England Injuries


Cases	Sum of Squares	df	Mean Square	F	p
Country code	97.09	3	32.364	13.23	< .001
Residual	97.82	40	2.445		

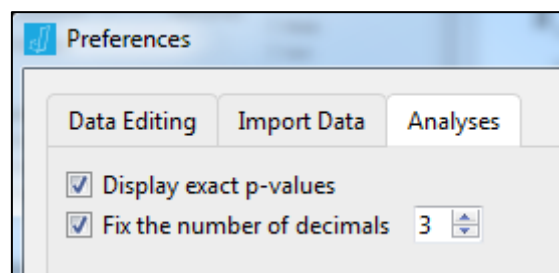
Note. Type III Sum of Squares

One way ANOVA of injuries received by England rugby players against Tonga, New Zealand, France and Wales

As previously mentioned, all tables and figures are APA standard and can just be copied into any other document.

You can change the size of all the tables and graphs using ctrl+ (increase) ctrl- (decrease) ctrl= (back to default size). Graphs can also be resized by dragging the bottom right corner of the graph.

One final tip: to make your tables less cluttered you can go to preferences  in the top right of the window and adjust the number of decimal places shown as well as displaying the exact p values i.e. from  $p < .001$  to  $p < .00084$ .



There are many further resources on using JASP on the website <https://jasp-stats.org/>



## DESCRIPTIVE STATISTICS

Presentation of all the raw data is very difficult for a reader to visualise or to draw any inference on. Descriptive statistics and related plots are a succinct way of describing and summarising data but do not test any hypotheses. There are various types of statistics that are used to describe data:

- Measures of central tendency
- Measures of dispersion
- Percentile values
- Measures of distribution
- Descriptive plots

In order to explore these measures, load **Descriptive data.csv** into JASP. Go to Descriptives > Descriptive statistics and move the Variable data to the Variables box on the right.

### CENTRAL TENDENCY.

This can be defined as the tendency for variable values to cluster around a central value. The three ways of describing this central value are mean, median or mode. If the whole population is considered we the term population mean / median/mode is used. If a sample/subset of the population is being analysed the term sample mean/ median/mode is used. The measures of central tendency move toward a constant value when the sample size is sufficient to be representative of the population.

In the Statistics options make sure that everything is unticked apart from mean, median and mode.

Central Tendency	
<input checked="" type="checkbox"/>	Mean
<input checked="" type="checkbox"/>	Median
<input checked="" type="checkbox"/>	Mode
<input type="checkbox"/>	Sum

Descriptive Statistics	
Variable	
Valid	810
Missing	0
Mean	17.71
Median	17.90
Mode	20.00

The **mean, M or  $\bar{x}$**  (17.71) is equal to the sum of all the values divided by the number of values in the dataset i.e. the average of the values. It is used for describing continuous data. It provides a simple statistical model of the centre of distribution of the values and is a theoretical estimate of the 'typical value'. However, it can be influenced heavily by 'extreme' scores.

The **median, Mdn** (17.9) is the middle value in a dataset that has been ordered from the smallest to largest value and is the normal measure used for ordinal or non-parametric continuous data. Less sensitive to outliers and skewed data

The **mode** (20.0) is the most frequent value in the dataset and is usually the highest bar in a distribution histogram



## DISPERSION

In the Statistics options make sure that everything is unticked apart from standard deviation, variance and standard error of the mean.

**Dispersion**

Std. deviation     Minimum

Variance             Maximum

Range                     S. E. mean

Descriptive Statistics	
	Variable
Valid	810
Missing	0
Std. Error of Mean	0.24
Std. Deviation	6.94
Variance	48.10

**Standard deviation, S or SD** (6.94) is used to quantify the amount of dispersion of data values around the mean. A low standard deviation indicates that the values are close to the mean, while a high standard deviation indicates that the values are dispersed over a wider range.

**Variance** ( $S^2 = 48.1$ ) is another estimate of how far the data is spread from the mean. It is also the square of the standard deviation.

**The standard error of the mean, SE** (0.24) is a measure of how far the sample mean of the data is expected to be from the true population mean. As the size of the sample data grows larger the SE decreases compared to S and the true mean of the population is known with greater specificity.

**Confidence intervals (CI)**, although not shown in the general Descriptive statistics output, these are used in many other statistical tests. When sampling from a population to get an estimate of the mean, confidence intervals are a range of values within which you are n% confident the true mean is included. A 95% CI is, therefore, a range of values that one can be 95% certain contains the true mean of the population. This is **not** the same as a range that contains 95% of **ALL** the values.

For example, in a normal distribution, 95% of the data are expected to be within  $\pm 1.96$  SD of the mean and 99% within  $\pm 2.576$  SD.

$$95\% \text{ CI} = M \pm 1.96 * \text{the standard error of the mean.}$$

Based on the data so far,  $M = 17.71$ ,  $SE = 0.24$ , this will be  $17.71 \pm (1.96 * 0.24)$  or  $17.71 \pm 0.47$ .

Therefore the 95% CI for this dataset is 17.24 - 18.18 and suggests that the true mean is likely to be within this range 95% of the time



## QUARTILES

In the Statistics options make sure that everything is unticked apart from Quartiles.

**Percentile Values**

Quartiles

Cut points for:  equal groups

Percentiles:

Descriptive Statistics	
	Variable
Valid	810
Missing	0
25th percentile	13.05
50th percentile	17.90
75th percentile	22.30

Quartiles are where datasets are split into 4 equal quarters, normally based on rank ordering of median values. For example, in this dataset

1	1	2	2	3	3	4	4	4	4	5	5	5	6	7	8	8	9	10	10	10
				25%						50%					75%					

The median value that splits data by 50% = 50th percentile = 5

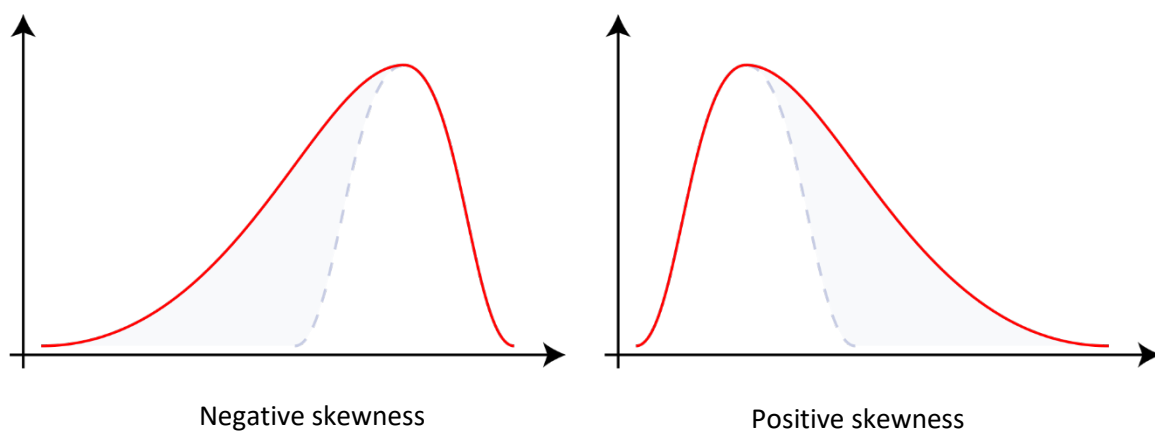
The median value of left side = 25th percentile = 3

The median value of right side = 75th percentile = 8

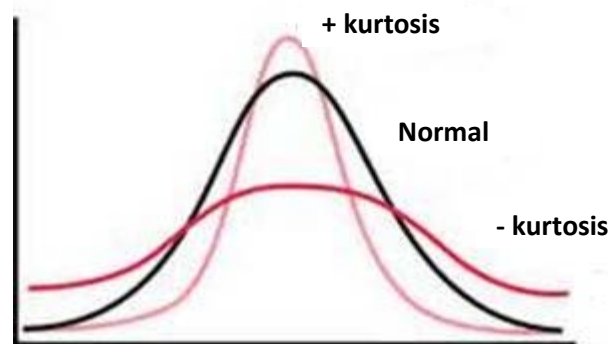
From this the Interquartile range (IQR) range can be calculated, this is the difference between the 75th and 25th percentiles i.e. 5. These values are used to construct the descriptive boxplots later.

## DISTRIBUTION

Skewness describes the shift of the distribution away from a normal distribution. Negative skewness shows that the mode moves to the right resulting in a dominant left tail. Positive skewness shows that the mode moves to the left resulting in a dominant right tail.



Kurtosis describes how heavy or light the tails are. Positive kurtosis results in an increase in the “pointiness” of the distribution with heavy (longer) tails while negative kurtosis exhibit a much more uniform or flatter distribution with light (shorter) tails.



In the Statistics options make sure that everything is unticked apart from skewness and kurtosis.

**Distribution**

Skewness

Kurtosis

**Descriptive Statistics**

	Variable
Valid	810
Missing	0
Skewness	-0.004
Std. Error of Skewness	0.086
Kurtosis	-0.410
Std. Error of Kurtosis	0.172

We can use the Descriptives output to calculate skewness and kurtosis. For a normal data distribution both values should be close to zero (see - Exploring data integrity in JASP for more details).

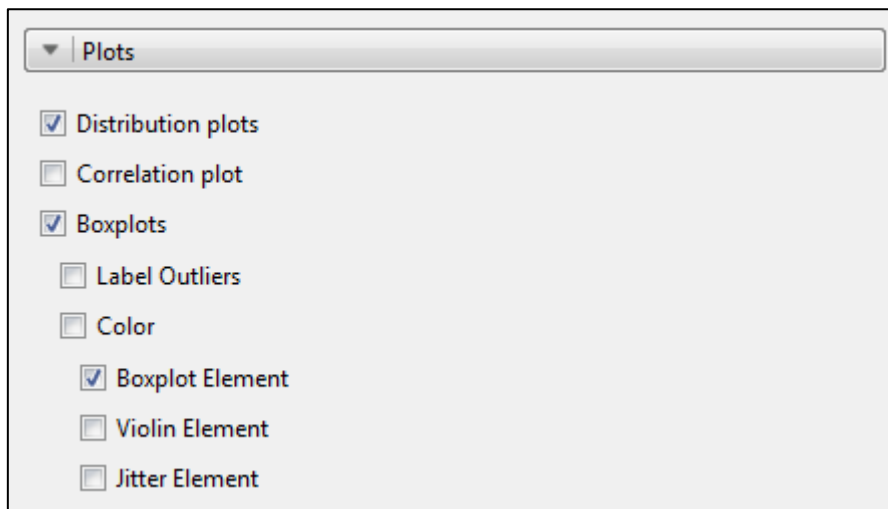
## DESCRIPTIVE PLOTS IN JASP

Currently, JASP produces three main types of descriptive plots:

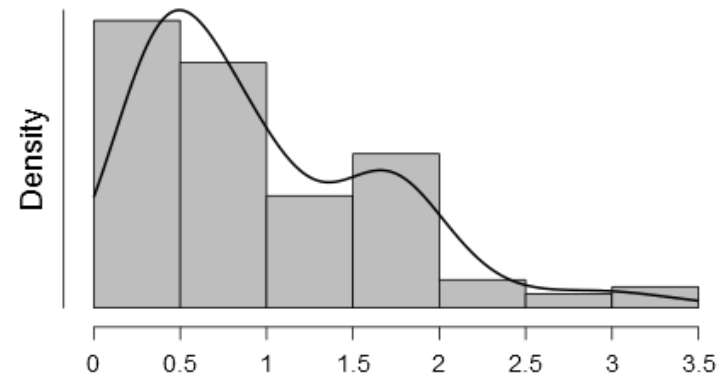
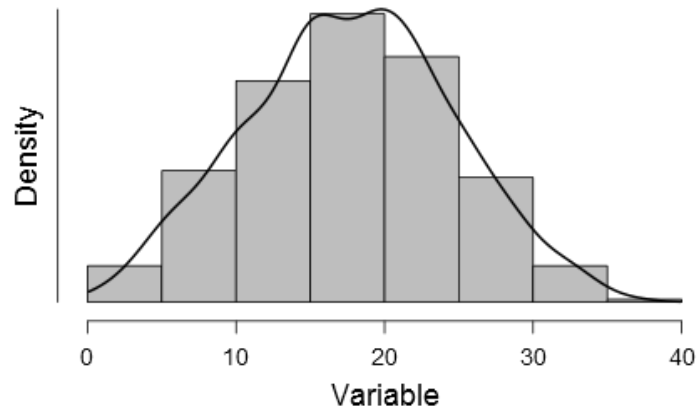
- Distribution plots
- Correlation plot
- Boxplots – with 3 options
  - Boxplot Element
  - Violin Element
  - Jitter Element



Again, using **Descriptive data.csv** with the variable data in the Variables box, go to the statistics options and under Plots tick Distribution plots and Boxplots – Boxplot Element.



The Distribution plot is based on splitting the data into frequency bins, this is then overlaid with the distribution curve. As mentioned before, the highest bar is the mode (most frequent value of the dataset). In this case, the curve looks approximately symmetrical suggesting that the data is approximately normally distributed. The second distribution plot is from another dataset which shows that the data is positively skewed.

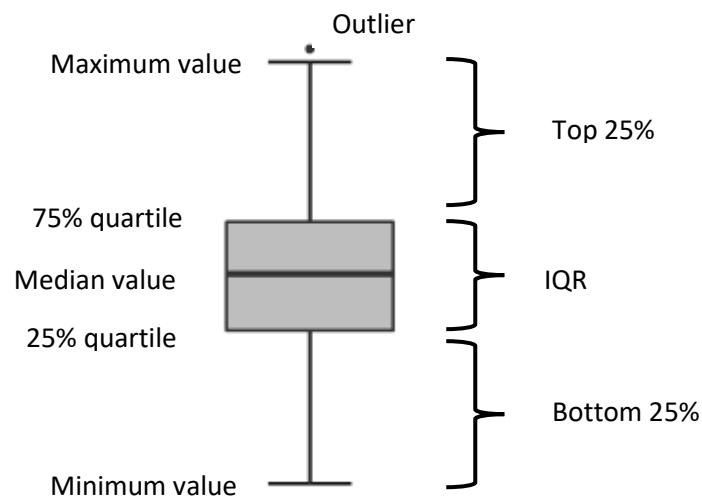




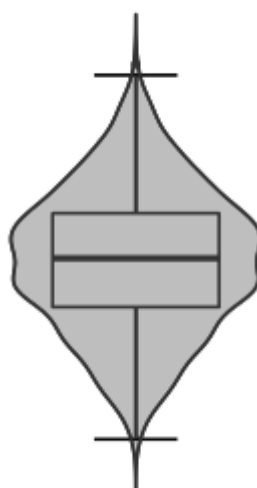
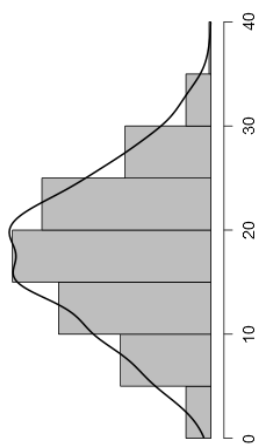


The boxplots visualise a number of statistics described above in one plot:

- Median value
- 25 and 75% quartiles
- Interquartile range (IQR) i.e. 75% - 25% quartile values
- Maximum and minimum values plotted with outliers excluded
- Outliers are shown if requested



Go back to the statistics options, in Descriptive plots tick both Boxplot and Violin Element, look at how the plot has changed. Next tick Boxplot, Violin and Jitter Elements. The Violin plot has taken the smoothed distribution curve from the Distribution plot, rotated it 90° and superimposed it on the boxplot. The jitter plot has further added all the data points.



Boxplot + Violin plot



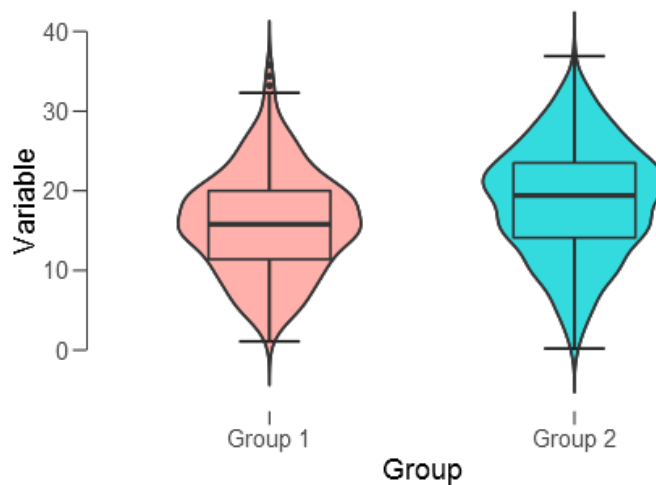
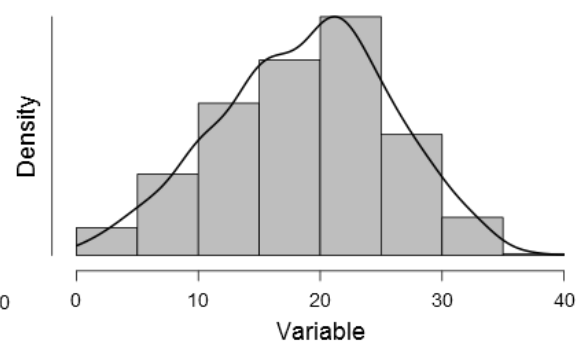
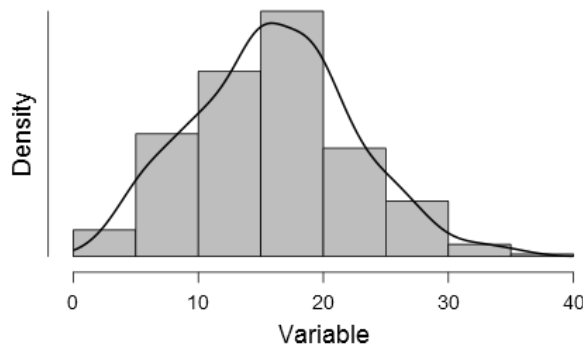
Boxplot + Violin + Jitter plot



## SPLITTING DATA FILES

If there is a grouping variable (categorical or ordinal) descriptive statistics and plots can be produced for each group. Using **Descriptive data.csv** with the variable data in the Variables box now add Group to the Split box. The output will be as follows:

	Variable	
	Group 1	Group 2
Valid	315	495
Missing	0	0
Mean	16.021	18.787
Median	15.800	19.400
Mode	20.000	20.200
Std. Deviation	6.424	7.040
Variance	41.269	49.556
Skewness	0.200	-0.176
Std. Error of Skewness	0.137	0.110
Kurtosis	-0.101	-0.397
Std. Error of Kurtosis	0.274	0.219
Minimum	1.100	0.200
Maximum	35.800	36.900





## EXPLORING DATA INTEGRITY

Sample data is used to estimate parameters of the population whereby a parameter is a measurable characteristic of a population, such as a mean, standard deviation, standard error or confidence intervals etc.

What is the difference between a statistic and a parameter? If you randomly polled a selection of students about the quality of their student bar and you find that 75% of them were happy with it. That is a sample **statistic** since only a sample of the population were asked. You calculated what the population was likely to do based on the sample. If you asked **all** the students in the university and 90% were happy you have a **parameter** since you asked the whole university population.

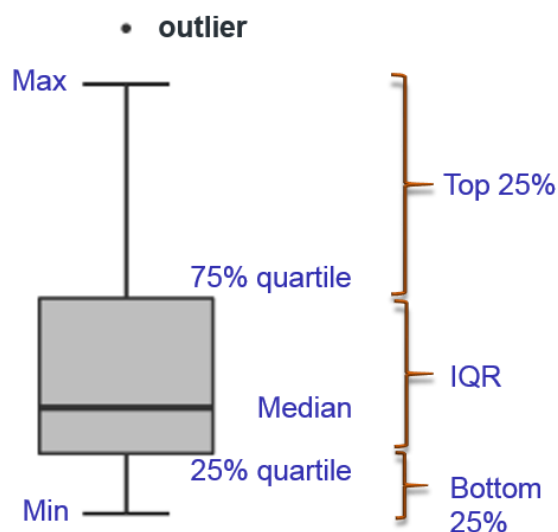
Bias can be defined as the tendency of a measurement to over- or under-estimate the value of a population parameter. There are many types of bias that can appear in research design and data collection including:

- Participant selection bias – some being more likely to be selected for study than others
- Participant exclusion bias - due to the systematic exclusion of certain individuals from the study
- Analytical bias - due to the way that the results are evaluated

However statistical bias can affect a) parameter estimates, b) standard errors and confidence intervals or c) test statistics and *p* values. So how can we check for bias?

## IS YOUR DATA CORRECT?

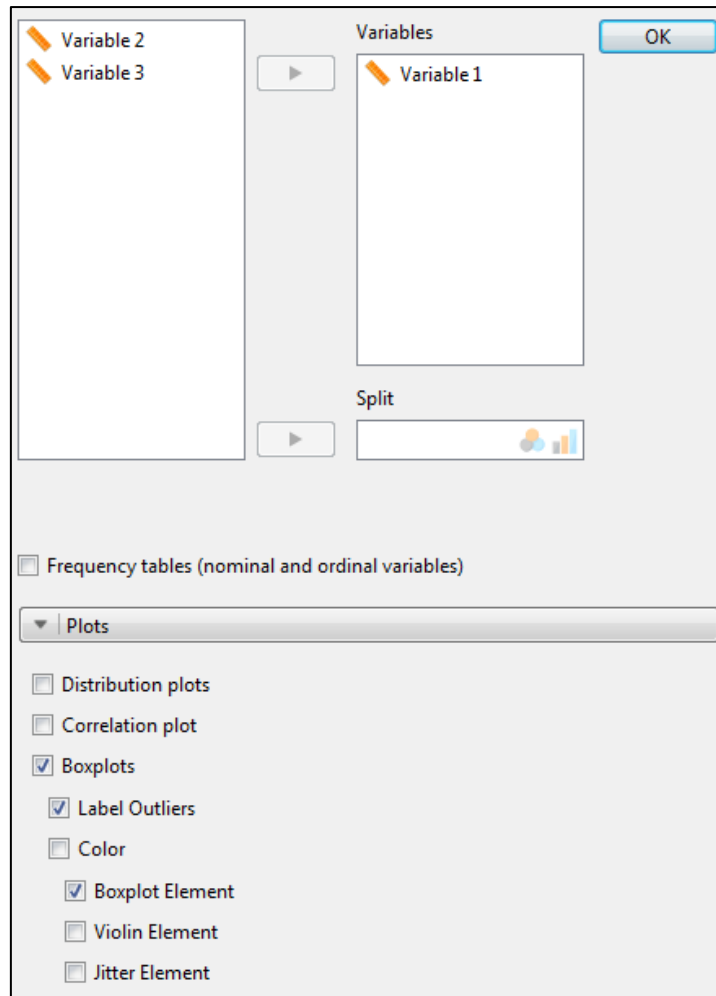
Outliers are data points that are abnormally outside all other data points. Outliers can be due to a variety of things such as errors in data input or analytical errors at the point of data collection. Boxplots are an easy way to visualise such data points where outliers are outside the upper ( $75\% + 1.5 * IQR$ ) or lower ( $25\% - 1.5 * IQR$ ) quartiles



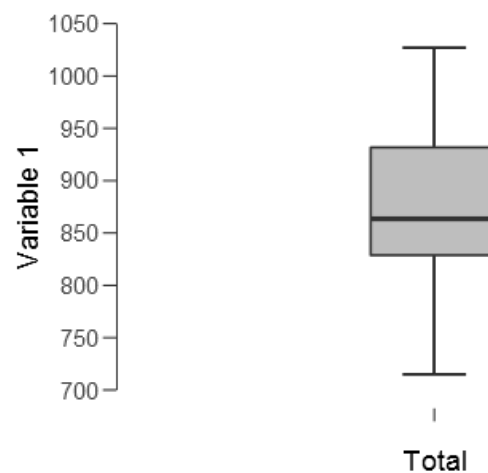
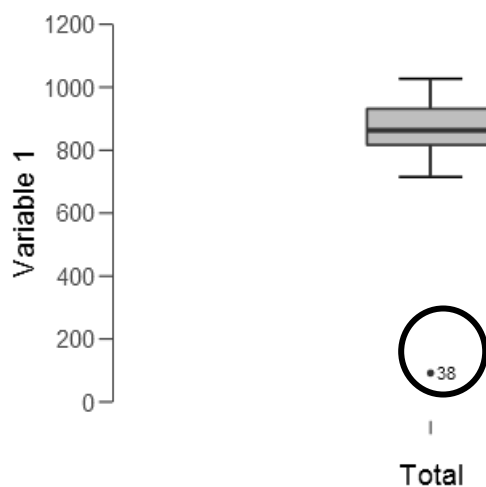
Boxplots show:

- Median value
- 25 & 75% quartiles
- IQR – Inter quartile range
- Max & min values plotted with outliers excluded
- Outliers shown if requested

Load **Exploring Data.csv** into JASP. Under Descriptives > Descriptive Statistics, add Variable 1 to the Variables box. In Plots tick the following Boxplots, Label Outliers, and BoxPlot Element.



The resulting Boxplot on the left looks very compressed and an obvious outlier is labelled as being in row 38 of the dataset. This can be traced back to a data input error in which 91.7 was input instead of 917. The graph on the right shows the BoxPlot for the 'clean' data.



How you deal with an outlier depends on the cause. Most parametric tests are highly sensitive to outliers while non-parametric tests are generally not.



*Correct it?* – Check the original data to make sure that it isn't an input error, if it is, correct it and rerun the analysis.

*Keep it?* - Even in datasets of normally distributed, data outliers may be expected for large sample sizes and should not automatically be discarded if that is the case.

*Delete it?* – This is a controversial practice in small datasets where a normal distribution cannot be assumed. Outliers resulting from an instrument reading error may be excluded but it should be verified first.

*Replace it?* – Also known as winsorizing. This technique replaces the outlier values with the relevant maximum and/or minimum values found after excluding the outlier.

Whatever method you use must be justified in your statistical methodology and subsequent analysis.

### WE MAKE MANY ASSUMPTIONS ABOUT OUR DATA.

When using parametric tests we make a series of assumptions about our data and bias will occur if these assumptions are violated, in particular:

- Normality
- Homogeneity of variance or homoscedasticity

Many statistical tests are actually an omnibus of tests of which some will check these assumptions.

### TESTING THE ASSUMPTION OF NORMALITY

Normality does not mean necessarily that the data is normally distributed per se but it is whether or not the dataset can be well modelled by a normal distribution. Normality can be explored in a variety of ways:

- Numerically
- Visually / graphically
- Statistically

Numerically we can use the Descriptives output to calculate skewness and kurtosis. For a normal data distribution, both values should be close to zero. To determine the significance of skewness or kurtosis we calculate their z-scores by dividing them by their associated standard errors:

$$\text{Skewness } z = \frac{\text{skewness}}{\text{Skewness standard error}} \quad \text{Kurtosis } z = \frac{\text{kurtosis}}{\text{kurtosis standard error}}$$

**Z score significance:**    **p<0.05 if z >1.96**    **p<0.01 if z >2.58**    **p<0.001 if z >3.29**

Using **Exploring data.csv**, go to Descriptives>Descriptive Statistics move Variable 3 to the Variables box, in the Statistics drop down menu select Mean, Std deviation, Skewness and Kurtosis as shown below with the corresponding output table.



▼ Statistics

**Percentile Values**

Quartiles

Cut points for:  equal groups

Percentiles:

**Central Tendency**

Mean

Median

Mode

Sum

**Dispersion**

Std. deviation    Minimum

Variance    Maximum

Range    S. E. mean

**Distribution**

Skewness

Kurtosis

	Variable 3
Valid	50
Missing	0
Mean	0.893
Std. Deviation	0.673
Skewness	0.839
Std. Error of Skewness	0.337
Kurtosis	-0.407
Std. Error of Kurtosis	0.662

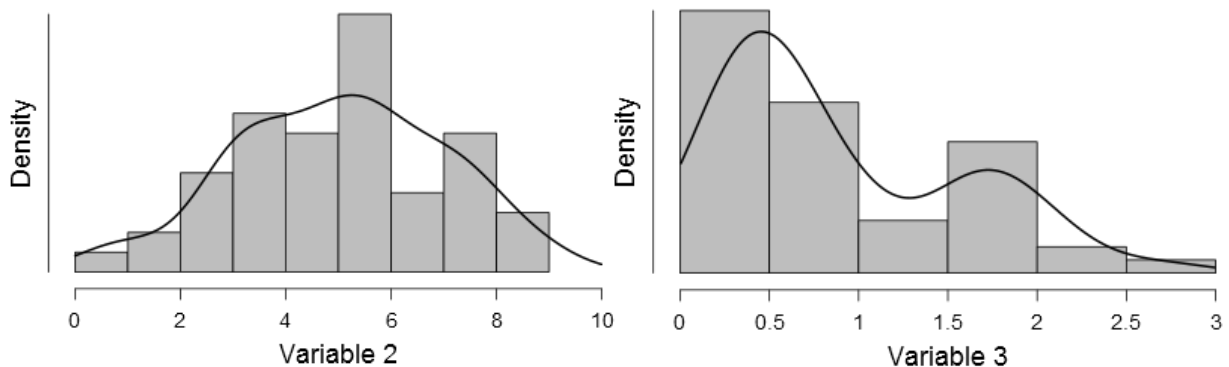
It can be seen that both skewness and kurtosis are not close to 0. The positive skewness suggests that data is distributed more on the left (see graphs later) while the negative kurtosis suggests a flat distribution. When calculating their z scores it can be seen that the data is significantly skewed  $p < 0.05$ .

$$\text{Skewness } Z = \frac{0.839}{0.337} = 2.49$$

$$\text{Kurtosis } Z = \frac{-0.407}{0.662} = -0.614$$

[As a note of caution skewness and kurtosis many appear significant in large datasets even though the distribution is normal.]

Now add Variable 2 to the Variables box and in Plots tick Distribution plot. This will show the following two graphs:

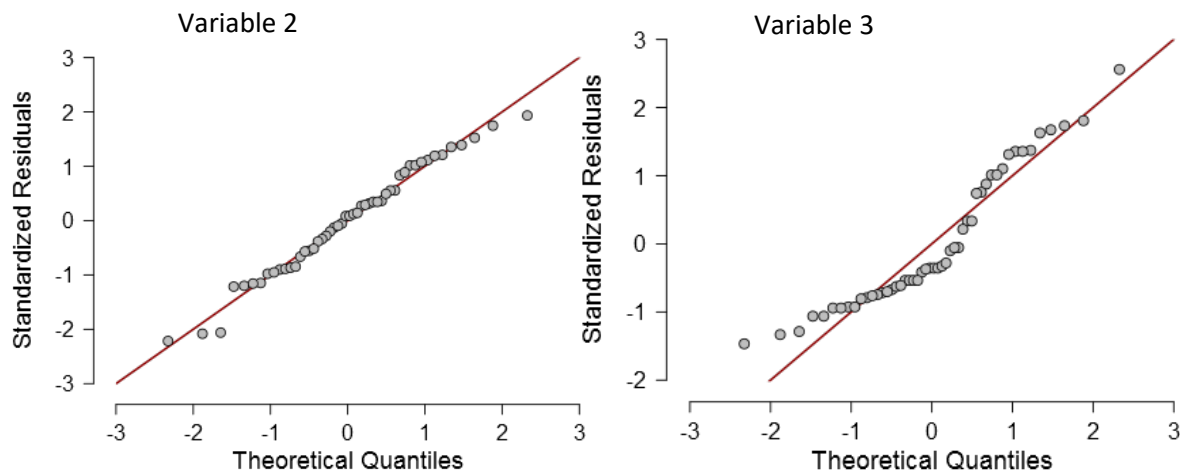


It is quite easy to visualise that Variable 2 has a symmetrical distribution. Variable 3 is skewed to the left as confirmed by the skewness Z score.

Another graphical check for normality is a Q-Q plot. This is produced as part of the Assumption Checks used in **linear regression** and **ANOVA**. Q-Q plots show the quantiles of the actual data against those expected for a normal distribution.



If data are normally distributed all the points will be close to the diagonal reference line. If the points 'sag' above or below the line there is a problem with kurtosis. If the points snake around the line then the problem is skewness. Below are Q-Q plots for Variables 2 and 3. Compare these to the previous distribution plots and the skewness/kurtosis z scores above.



The Shapiro-Wilk test is a statistical way used by JASP to check the assumption of normality. It is used in the **Independent** (distribution of the two groups) and **Paired** (distribution of differences between pairs) **t-tests**. The test results in a W value; where small values indicate your sample is not normally distributed (the null hypothesis that your population is normally distributed if your values are under a certain threshold can therefore be rejected). The table below is an example of the Shapiro-Wilk output table showing no significant deviation in normality in the 2 groups.

Test of Normality (Shapiro-Wilk)		W	p
Variable 2	Control	0.971	0.691
	Test	0.961	0.408

*Note.* Significant results suggest a deviation from normality.

The most important limitation is that the test has can be biased by sample size. The larger the sample, the more likely you'll get a statistically significant result.

### Testing the assumption of normality – A cautionary note!

For most parametric tests to be reliable, one of the assumptions is that the data is **approximately** normally distributed. A normal distribution peaks in the middle and is symmetrical about the mean. However, data does not need to be perfectly normally distributed for the tests to be reliable.

So, having gone on about testing for normality – is it necessary?

The Central Limit Theorem states that as the sample size gets larger i.e. >30 data points the distribution of the sampling means approaches a normal distribution. So the more data points you



have the more normal the distribution will look and the closer your sample mean approximates the population mean.

Large datasets may result in significant tests of normality i.e. Shapiro-Wilk or significant skewness and kurtosis z-scores when the distribution graphs look fairly normal. Conversely, small datasets will reduce the statistical power to detect non-normality.

However, data that definitely does not meet the assumption of normality is going to result in poor results for certain types of test (i.e. ones that state that the assumption must be met!). How closely does your data need to be normally distributed? This is a judgment call best made by eyeballing the data.

### WHAT DO I DO IF MY DATA IS REALLY NOT NORMALLY DISTRIBUTED?

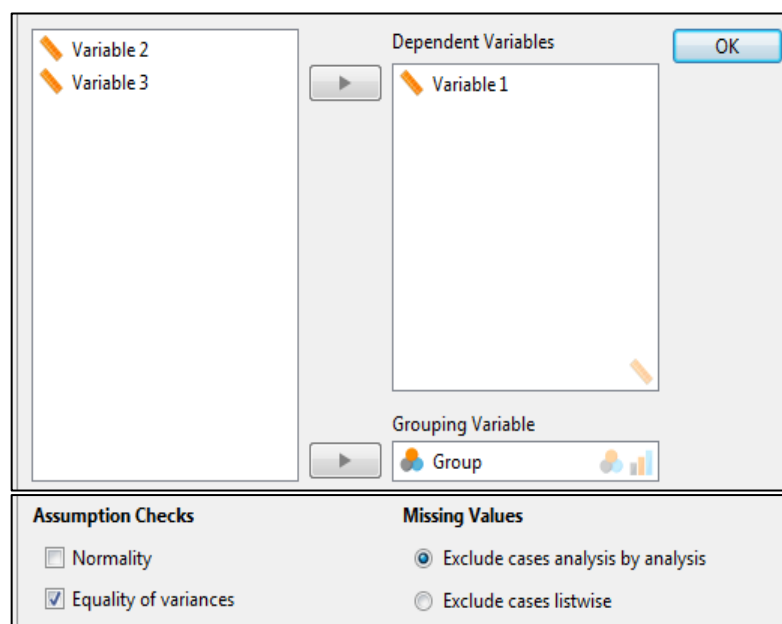
Transform the data and redo the normality checks on the transformed data. Common transformations include taking the log or square root of the data.

Use non-parametric tests since these are distribution free tests and can be used instead of their parametric equivalent.

### TESTING HOMOGENEITY OF VARIANCE

Levene's test is commonly used to test the null hypothesis that variances in different groups are equal. The result from the test (F) is reported as a p value, if not significant then you can say that the null hypothesis stands — that the variances are equal; if the p value is significant then the implication is that the variances are unequal. Levene's test is included in the **Independent t-test** and **ANOVA** in JASP as part of the Assumption Checks.

Using **Exploring data.csv**, go to T-Tests>Independent Samples t-test move Variable 1 to the Variables box and Group to the Grouping variable and tick Assumption Checks > Equality of variances.





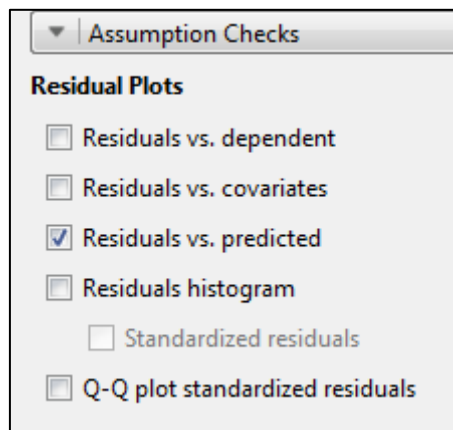


### Test of Equality of Variances (Levene's)

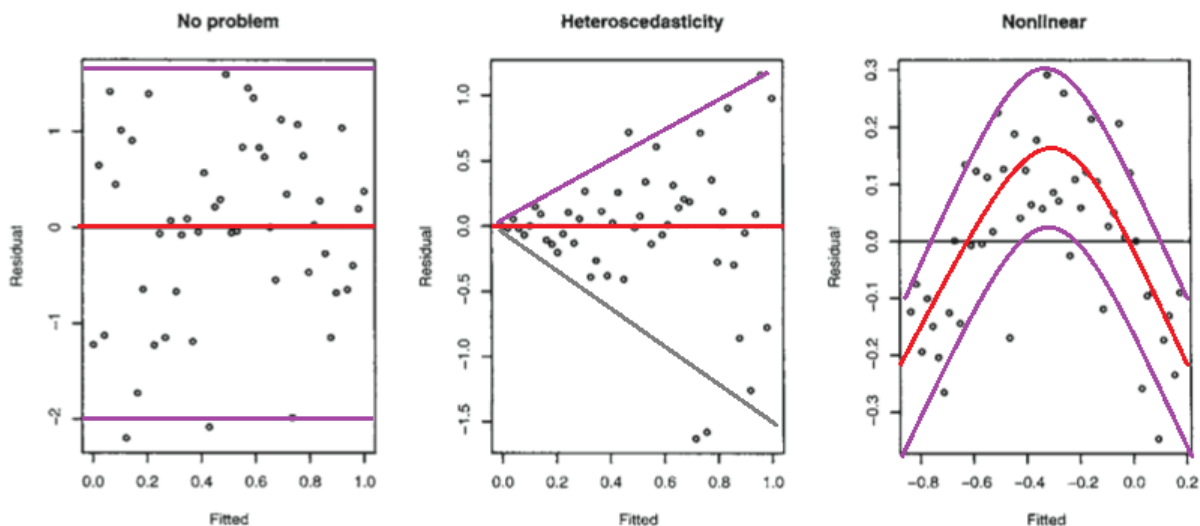
	F	df	p
Variable 1	0.218	1	0.643

In this case, there is no significant difference in variance between the two groups  $F(1) = 0.218, p = .643$ .

The assumption of homoscedasticity (equal variance) is important in **linear regression** models as is linearity. It assumes that the variance of the data around the regression line is the same for all predictor data points. Heteroscedasticity (the violation of homoscedasticity) is present when the variance differs across the values of an independent variable. This can be visually assessed in linear regression by plotting actual residuals against predicted residuals



If homoscedasticity and linearity are not violated there should be no relationship between what the model predicts and its errors as shown in the graph on the left. Any sort of funnelling (middle graph) suggests that homoscedasticity has been violated and any curve (right graph) suggests that linearity assumptions have not been met.





## ONE SAMPLE T-TEST

Research is normally carried out in sample populations, but how close does the sample reflect the whole population? The parametric one sample t-test determines whether the sample mean is statistically different from a known or hypothesized population mean.

**The null hypothesis ( $H_0$ ) tested is that the sample mean is equal to the population mean.**

## ASSUMPTIONS

Three assumptions are required for a one-sample t-test to provide a valid result:

- The test variable should be measured on a **continuous** scale.
- The test variable data should be **independent** i.e. no relationship between any of the data points.
- The data should be approximately **normally distributed**
- There should be no significant **outliers**.

## RUNNING THE ONE SAMPLE T-TEST

Open **one sample t-test.csv**, this contains two columns of data representing the height (cm) and body masses (kg) of a sample population of males used in a study. In 2017 the average adult male in the UK population was **178 cm** tall and has a body mass of **83.6 kg**.

Go to T-Tests > One Sample t-test and in the first instance add height to the analysis box on the right. Then tick the following options and add **178** as the test value:

<b>Tests</b>	<b>Additional Statistics</b>
<input checked="" type="checkbox"/> Student	<input checked="" type="checkbox"/> Location parameter
<input type="checkbox"/> Wilcoxon signed-rank	<input type="checkbox"/> Confidence interval 95 %
<input type="checkbox"/> Z test	<input checked="" type="checkbox"/> Effect size
Test value: 178	<input type="checkbox"/> Confidence interval 95 %
<b>Hypothesis</b>	<input checked="" type="checkbox"/> Descriptives
<input checked="" type="radio"/> ≠ Test value	<input type="checkbox"/> Descriptives plots
<input type="radio"/> > Test value	Confidence interval 95 %
<input type="radio"/> < Test value	<input type="checkbox"/> Vovk-Sellke maximum p-ratio
<b>Assumption Checks</b>	<b>Missing Values</b>
<input checked="" type="checkbox"/> Normality	<input checked="" type="radio"/> Exclude cases analysis by analysis
	<input type="radio"/> Exclude cases listwise



## UNDERSTANDING THE OUTPUT

The output should contain three tables.

Test of Normality (Shapiro-Wilk)

	W	p
height	0.969	0.507

Note. Significant results suggest a deviation from normality.

The assumption check of normality (Shapiro-Wilk) is not significant suggesting that the heights are normally distributed, therefore this assumption is not violated. If this showed a significant difference the analysis should be repeated using the non-parametric equivalent, **Wilcoxon’s signed rank test** tested against the population median height.

One Sample T-Test ▼

	t	df	p	Mean Difference	Cohen’s d
height	-0.382	22	0.706	-0.391	-0.080

Note. Student’s t-test.  
 Note. For the Student t-test, location parameter is given by mean difference *d*.  
 Note. For the Student t-test, effect size is given by Cohen’s *d*.  
 Note. For all tests, the alternative hypothesis specifies that the population mean is different from 178.

This table shows that there are no significant differences between the means  $p = .706$

Descriptives

	N	Mean	SD	SE
height	23.000	177.609	4.915	1.025

The descriptive data shows that the mean height of the sample population was 177.6 cm compared to the average 178 cm UK male.

Repeat the procedure by replacing height with mass and change the test value to 83.6.

Test of Normality (Shapiro-Wilk)

	W	p
mass	0.941	0.185

Note. Significant results suggest a deviation from normality.



The assumption check of normality (Shapiro-Wilk) is not significant suggesting that the masses are normally distributed.

### One Sample T-Test

	t	df	p	Mean Difference	Cohen's d
mass	-7.159	22	< .001	-10.487	-1.493

Note. Student's t-test.

Note. For the Student t-test, location parameter is given by mean difference *d*.

Note. For the Student t-test, effect size is given by Cohen's *d*.

Note. For all tests, the alternative hypothesis specifies that the population mean is different from 83.4.

This table shows that there is a significant difference between the mean sample (72.9 kg) and population body mass (83.6 kg)  $p < .001$

### Descriptives

	N	Mean	SD	SE
mass	23.000	72.913	7.025	1.465

## REPORTING THE RESULTS

A one sample t-test showed no significant difference in height compared to the population mean ( $t(22) = -0.382, p = .706$ ), however, the participants were significantly lighter than the UK male population average ( $t(22) = -7.159, p < .001$ ).



## BINOMIAL TEST

The binomial test is effectively a non-parametric version of the one-sample t-test for use with dichotomous (i.e. yes/no) categorical datasets. This tests whether or not the sample frequency is statistically different from a known or hypothesized population frequency.

**The null hypothesis ( $H_0$ ) tested is that the sample frequency is equal to the expected population frequency.**

## ASSUMPTIONS

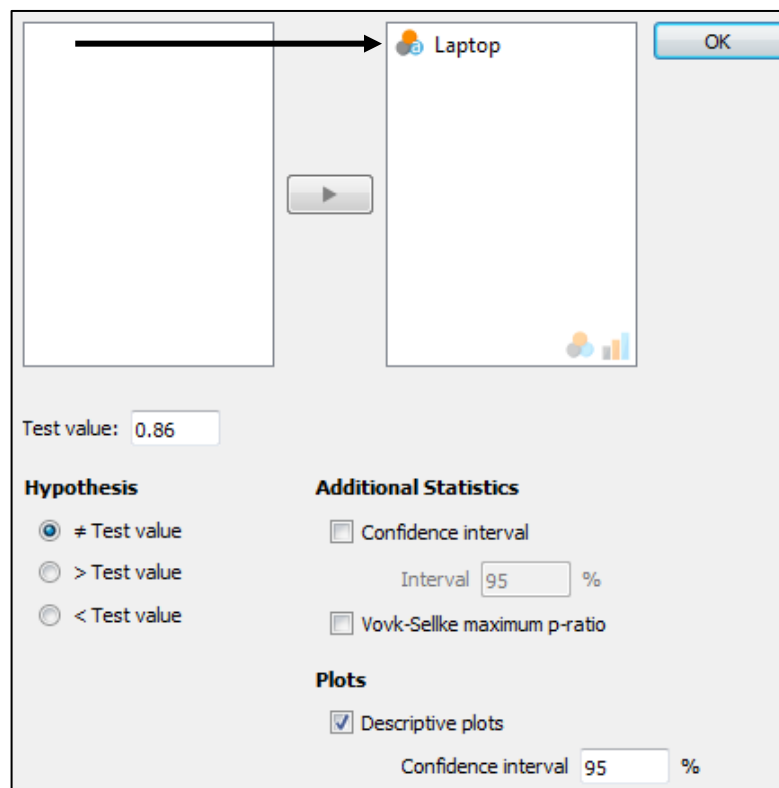
Three assumptions are required for a one-sample t-test to provide a valid result:

- The test variable should be a dichotomous scale (such as yes/no, male/female etc.).
- The sample responses should be independent
- The sample size is less, but representative of the population

## RUNNING THE BINOMIAL TEST

Open **binomial.csv**, this contains one column of data showing the number of students using either a Windows laptop or a MacBook at University. In January 2018, when comparing just the two operating systems, the UK market share of Windows was 86% and Mac IOS 14%.<sup>3</sup>

Go to Frequencies > Binomial test. Move the Laptop variable to the data window and set the Test value to 0.86 (86%). Also tick Descriptive plots.



<sup>3</sup> <https://www.statista.com/statistics/268237/global-market-share-held-by-operating-systems-since-2009/>

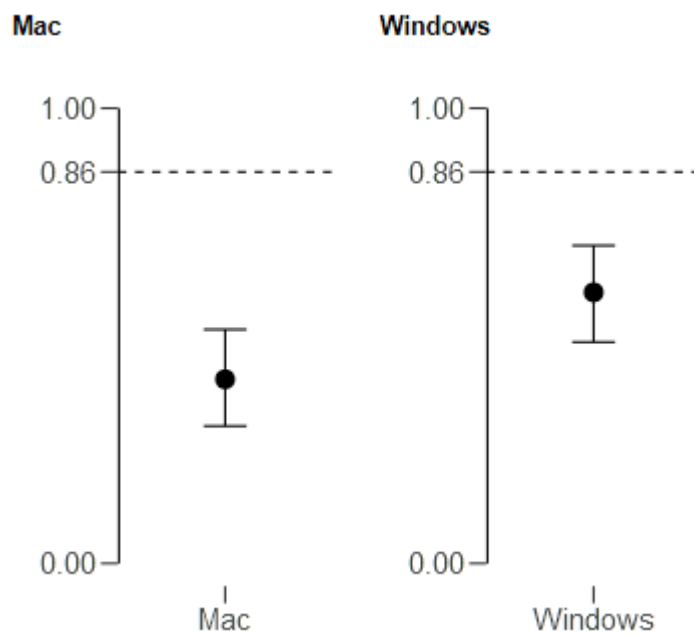


The following table and graph show that the frequencies of both laptops are significantly less than 86%. In particular, these students are using significantly fewer Windows laptops than was expected compared to the UK market share.

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.86.

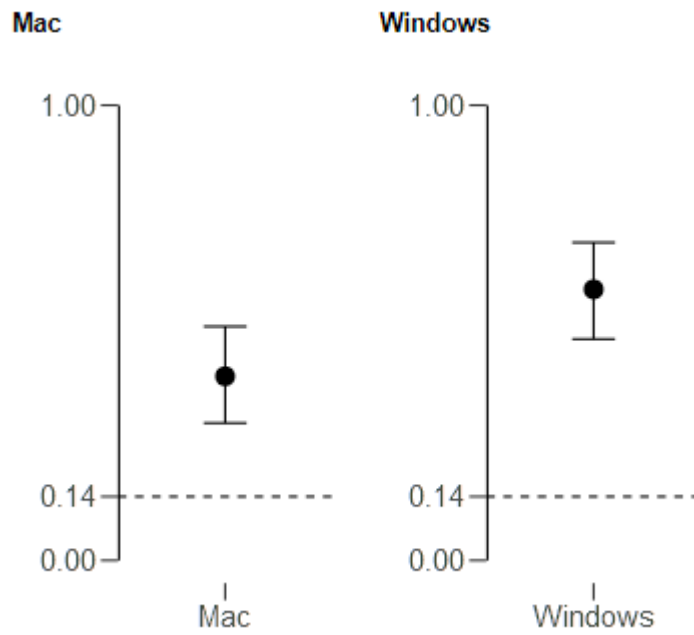


Is this the same for MacBook users? Go back to the options window and change the test value to 0.14 (14%). This time both frequencies are significantly higher than 14%. This shows that students are using significantly more MacBooks than was expected compared to the UK market share.

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.14.



### REPORTING THE RESULTS

The UK proportion of Windows and MacBook users was reported to be 86% and 14% respectively. In a cohort of University students (N=90), a Binomial test revealed that the proportion of students using Windows laptops was significantly less (59.6%,  $p < .001$ ) and those using MacBooks significantly more (40.4%,  $p < .001$ ) than expected.



## MULTINOMIAL TEST

The multinomial test is effectively an extended version of the Binomial test for use with categorical datasets containing three or more factors. This tests whether or not the sample frequency is statistically different from a hypothesized population frequency (multinomial test) or known a known frequency (Chi-square 'goodness-of-fit' test).

**The null hypothesis ( $H_0$ ) tested is that the sample frequency is equal to the expected population frequency.**

### ASSUMPTIONS

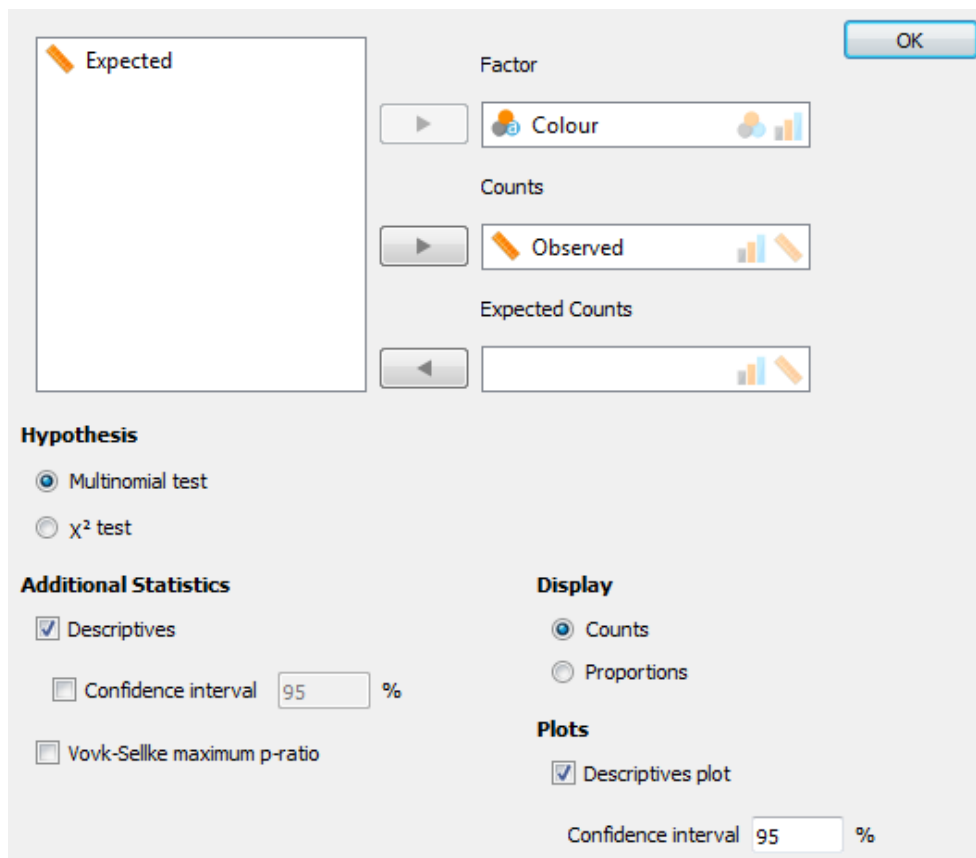
Three assumptions are required for a multinomial test to provide a valid result:

- The test variable should be a categorical scale containing 3 or more factors
- The sample responses should be independent
- The sample size is less, but representative of the population

### RUNNING THE MULTINOMIAL TEST

Open **multinomial.csv**. This contains three columns of data showing the number of different coloured M&Ms counted in five bags. Without any prior knowledge, it could be assumed that the different coloured M&Ms are equally distributed.

Go to Frequencies > Multinomial test. Move colour of the M&Ms to Factor and the observed number of M&Ms to counts. Tick Descriptives and Descriptives Plots.







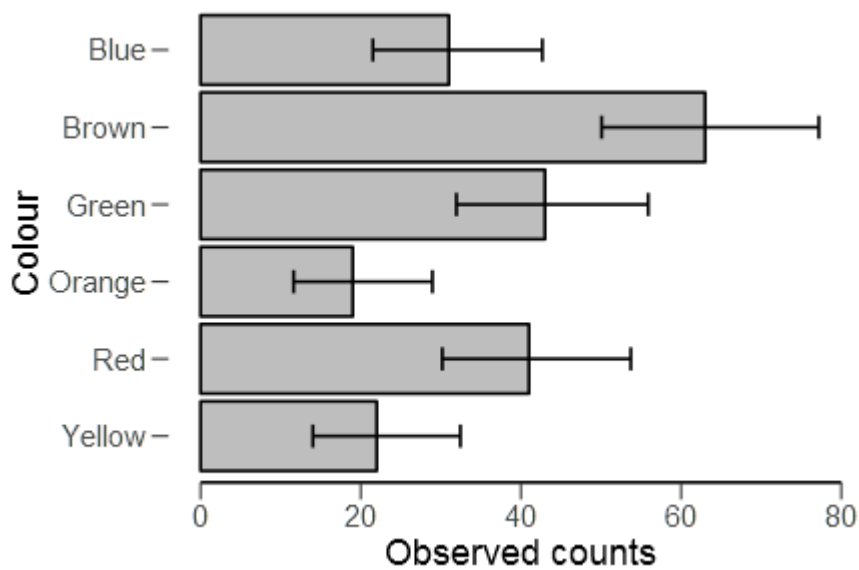
As can be seen in the Descriptive table, the test assumes an equal expectation for the proportions of coloured M&Ms (36 of each colour). The Multinomial test results show that the observed distribution is significantly different ( $p < .001$ ) to an equal distribution.

### Multinomial Test

	$\chi^2$	df	p
Multinomial	35.932	5	< .001

### Descriptives

Colour	Observed	Expected: Multinomial
Blue	31	36
Brown	63	36
Green	43	36
Orange	19	36
Red	41	36
Yellow	22	36





## CHI-SQUARE 'GOODNESS-OF-FIT' TEST.

However, further research shows that the manufacturers produce the coloured M&Ms in different ratios:

Colour	Blue	Brown	Green	Orange	Red	Yellow
Proportion	24	13	16	20	13	14

These values can now be used as the expected counts, so move the Expected variable to the Expected Counts box. This automatically runs the  $\chi^2$  'goodness-of-fit' test leaving the Hypothesis options greyed out.

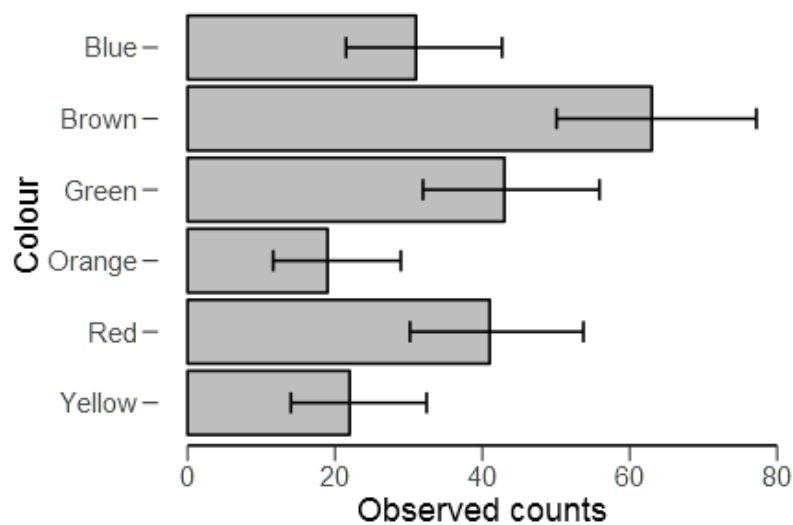
As can be seen in the Descriptives table, JASP has calculated the expected numbers of the different coloured M&Ms based on the manufacturers reported production ratio. The results of the test show that the observed proportions of the different coloured M&Ms are significantly different ( $\chi^2 = 74.5$ ,  $p < .001$ ) to those proportions stated by the manufacturer.

### Multinomial Test

	$\chi^2$	df	p
Expected	74.535	5	< .001

### Descriptives

Colour	Observed	Expected: Expected
Blue	31	52
Brown	63	28
Green	43	35
Orange	19	43
Red	41	28
Yellow	22	30





## MULTINOMIAL AND $\chi^2$ 'GOODNESS-OF-FIT' TEST.

JASP also provides another option whereby both tests can be run at the same time. Go back to the Options window and only add Colour to the Factor and Observed to the Counts boxes, remove the expected counts if the variable is still there. In Hypotheses now tick the  $\chi^2$  test. This will open up a small spreadsheet window showing the colour and  $H_0$  (a) with each cell have 1 in it. This is assuming that the proportions of each colour are equal (multinomial test).

In this window, add another column which will automatically be labelled  $H_0$  (b). The expected proportions of each colour can now be typed in.

	$H_0$ (a)	$H_0$ (b)
Brown	1	13
Green	1	16
Orange	1	20
Red	1	13
Yellow	1	14

Add column

Delete column

Reset

Now when the analysis is run, the results of the tests for the two hypotheses are shown.  $H_0$  (a) is testing the null hypothesis that the proportions of each colour are equally distributed, while  $H_0$  (b) is testing the null hypothesis that the proportions are the same as those expected. As can be seen, both hypotheses are rejected. In particular, evidence indicates that the colours of plain M&M's do not match the manufacturers published proportions.

### Multinomial Test

	$\chi^2$	df	p
$H_0$ (a)	35.932	5	< .001
$H_0$ (b)	74.535	5	< .001

### Descriptives

Colour	Observed	Expected	
		$H_0$ (a)	$H_0$ (b)
Blue	31	36	52
Brown	63	36	28
Green	43	36	35
Orange	19	36	43
Red	41	36	28
Yellow	22	36	30



## COMPARING TWO INDEPENDENT GROUPS

### INDEPENDENT T-TEST

The parametric independent t-test, also known as Student's t-test, is used to determine if there is a statistical difference between the means of two independent groups. The test requires a continuous dependent variable (i.e. body mass) and an independent variable comprising 2 groups (i.e. males and females).

This test produces a t-score which is a ration of the differences between the two groups and the differences within the two groups:

$$t = \frac{\text{mean group 1} - \text{mean group 2}}{\text{standard error of the mean differences}}$$

$$t = \frac{(X_1 - X_2)}{\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}}$$

**X** = mean

**S** = standard deviation

**n** = number of data points

A large t-score indicates that there is a greater difference between groups. The smaller the t-score, the more similarity there is between groups. A t-score of 5 means that the groups are five times as different from each other as they are within each other.

***The null hypothesis (H<sub>0</sub>) tested is that the population means from the two unrelated groups are equal***

### ASSUMPTIONS OF THE PARAMETRIC INDEPENDENT T-TEST

#### Group independence:

Both groups must be independent of each other. Each participant will only provide one data point for one group only. For example participant 1 can only be in either a male or female group – not both. Repeated measures are assessed using the **Paired t-test**.

#### Normality of the dependent variable:

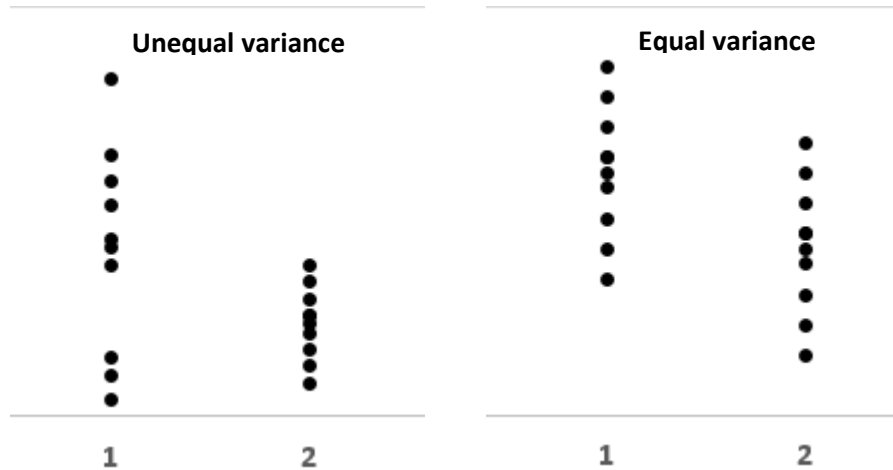
The dependent variable should also be measured on a continuous scale and be approximately normally distributed with no significant outliers. This can be checked using the Shapiro-Wilk test. The t-test is fairly robust and small deviations from normality are normally acceptable. However, this is not the case if the group sizes are very different. A rule of thumb is that the ratio between the group sizes should be <1.5 (i.e. group A = 12 participants and group B = >8 participants).

If normality is violated you can try transforming your data (for example log values, square root values) or, and if the group sizes are very different, use the **Mann-Whitney U** test which is a non-parametric equivalent that does not require the assumption of normality (see later).



## Homogeneity of variance:

The variances of the dependent variable should be equal in each group. This can be tested using Levene's Test of Equality of Variances.

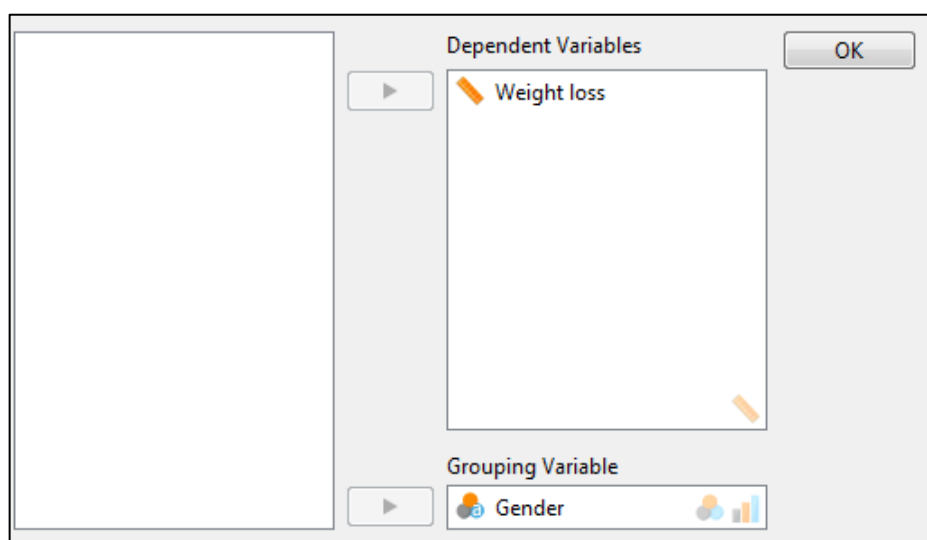


If the Levene's Test is statistically significant, indicating that the group variances are unequal we can correct for this violation by using an adjusted t-statistic based on the **Welch** method.

## RUNNING THE INDEPENDENT T-TEST

Open **Independent t-test.csv**, this contains weight loss on a self-controlled 10-week diet between men and women. Its good practice to check the Distribution and boxplots in Descriptives to visually check for distribution and outliers.

Go to T-Tests > Independent Samples t-test and put weight loss in the Dependent variable box and gender (independent variable) in the Grouping Variable box.





In the analysis window tick the following options:

<p><b>Tests</b></p> <input checked="" type="checkbox"/> Student <input checked="" type="checkbox"/> Welch <input type="checkbox"/> Mann-Whitney <p><b>Hypothesis</b></p> <input checked="" type="radio"/> Group 1 ≠ Group 2 <input type="radio"/> Group 1 > Group 2 <input type="radio"/> Group 1 < Group 2 <p><b>Assumption Checks</b></p> <input checked="" type="checkbox"/> Normality <input checked="" type="checkbox"/> Equality of variances	<p><b>Additional Statistics</b></p> <input checked="" type="checkbox"/> Location parameter <input type="checkbox"/> Confidence interval 95 % <input checked="" type="checkbox"/> Effect size <input type="checkbox"/> Confidence interval 95 % <input checked="" type="checkbox"/> Descriptives <input checked="" type="checkbox"/> Descriptives plots Confidence interval 95 % <input type="checkbox"/> Vovk-Sellke maximum p-ratio <p><b>Missing Values</b></p> <input checked="" type="radio"/> Exclude cases analysis by analysis <input type="radio"/> Exclude cases listwise
--	---

## UNDERSTANDING THE OUTPUT

The output should consist of four tables and one graph. Firstly we need to check that the parametric assumptions required are not violated.

		W	p
Weight loss	Females	0.968	0.282
	Males	0.971	0.310

Note. Significant results suggest a deviation from normality.

Shapiro-Wilk test shows that both groups have normally distributed data, there for the assumption of normality is not violated. If one or both were significant you should consider using the non-parametric equivalent **Mann-Whitney** test.

	F	df	p
Weight loss	2.278	1	0.135

Levene's test shows that there is no difference in variance, therefore, the assumption of homogeneity of variance is not violated. If Levene's test was significant **Welch's** adjusted t-statistic, degrees of freedom and p values should be reported.



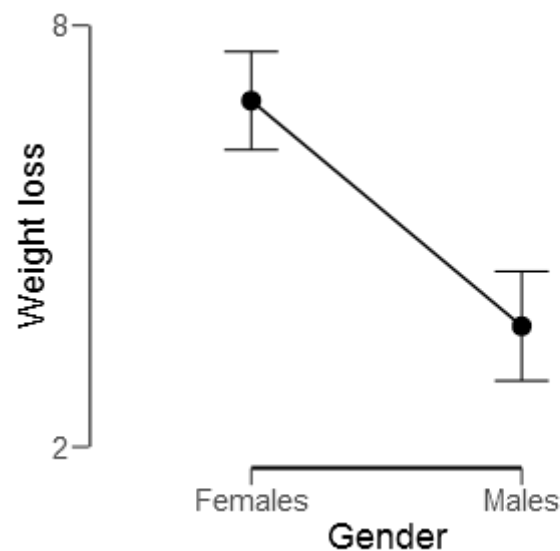
Independent Samples T-Test

	Test	Statistic	df	p	Mean Difference	SE Difference	Cohen's d
Weight loss	Student	6.160	85.000	< .001	3.209	0.521	1.322
	Welch	6.191	84.544	< .001	3.209	0.518	1.325

This table shows the two computed t-statistics (**Student and Welch**). Remember the t-statistic is derived from the mean difference divided by the standard error of the difference. Both show that there is a significant statistical difference between the two groups ( $p < .001$ ) and Cohen's d suggests that this is a large effect.

Group Descriptives

	Group	N	Mean	SD	SE
Weight loss	Females	42	6.929	2.242	0.346
	Males	45	3.720	2.588	0.386



From the descriptive data, it can be seen that females had a higher weight loss than males.

### REPORTING THE RESULTS

An independent t-test showed that females lost significantly more weight over 10 weeks dieting than males  $t(85)=6.16, p<.001$ . Cohen's d (1.322) suggests that this is a large effect.



## MANN-WITNEY U TEST

If you find that your data is not normally distributed (significant Shapiro-Wilk test result) or is ordinal by nature, the equivalent non-parametric independent test is the **Mann-Whitney U test**.

Open **Mann-Whitney pain.csv** which contains subjective pain scores (0-10) with and without ice therapy. NOTE: make sure that Treatment is categorical and pain score is ordinal. Go to T-Tests > Independent t-tests and put pain score in the Dependent variable box and use Treatment as the grouping variable.

In the analysis options only tick:

- ✓ Mann-Whitney
- ✓ Location parameter
- ✓ Effect size

There is no reason to repeat the assumption checks since Mann-Whitney does not require the assumption of normality or homogeneity of variance required by parametric tests.

## UNDERSTANDING THE OUTPUT

This time you will only get one table:

Independent Samples T-Test

	W	p	Hodges-Lehmann Estimate	Rank-Biserial Correlation
Pain score	207.000	< .001	3.000	0.840

Note. Mann-Whitney U test.

The Mann-Whitney U-statistic (JASP reports this as W since it is an adaptation of Wilcoxon’s signed rank test) is highly significant. **U=207, p<.001**.

The location parameter, the Hodges–Lehmann estimate, is the **median** difference between the two groups. The rank-biserial correlation ( $r_b$ ) can be considered as an effect size and is interpreted the same as Pearson’s  $r$ , so 0.84 is a large effect size.

For non-parametric data, you should report **median** values as your descriptive statistics and use boxplots instead of line graphs and confidence intervals, SD/SE bars. Go to Descriptive statistics, put Pain score into the variable box and Split the file by Treatment.





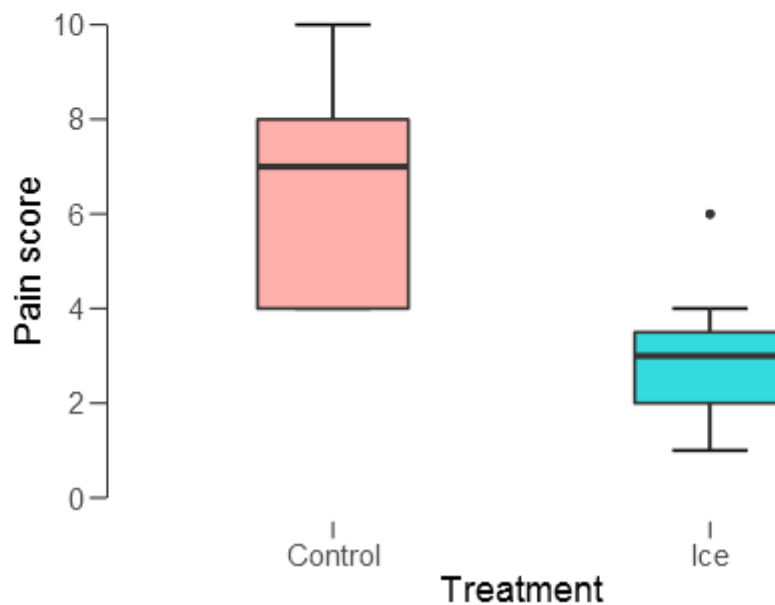
## Descriptive Statistics

	Pain score	
	Control	Ice
Valid	15	15
Missing	0	0
Median	7.000	3.000
Minimum	4.000	1.000
Maximum	10.000	6.000

## Plots

### Boxplots

Pain score



## REPORTING THE RESULTS

A Mann-Whitney test showed that Ice therapy significantly reduces pain scores (Mdn = 3) compared to the control group (Mdn = 7),  $U=207$ ,  $p<.001$ .



## COMPARING TWO RELATED GROUPS

### PAIRED SAMPLES T-TEST

As with the Independent t-test, there are both parametric and non-parametric options available in JASP. The parametric paired-samples t-test (also known as the dependent sample t-test or repeated measures t-test) compares the means between two related groups on the same continuous, dependent variable. For example, looking at weight loss pre and post 10 weeks dieting.

The paired  $t$  statistic =  $\frac{\text{mean of the differences between group pairs}}{\text{standard error of the mean differences}}$

*With the paired t-test, the null hypothesis ( $H_0$ ) is that the pairwise difference between the two groups is zero.*

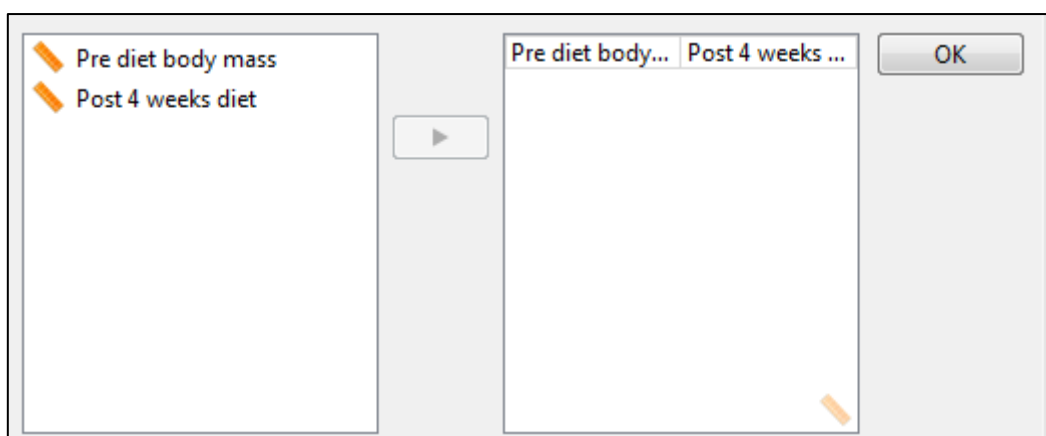
### ASSUMPTIONS OF THE PARAMETRIC PAIRED SAMPLES T-TEST

Four assumptions are required for a paired t-test to provide a valid result:

- The **dependent variable** should be measured on a continuous scale.
- The **independent variable** should consist of 2 categorical related/matched groups, i.e. each participant is matched in both groups
- The differences between the matched pairs should be approximately **normally distributed**
- There should be no significant **outliers** in the differences between the 2 groups.

### RUNNING THE PAIRED SAMPLES T-TEST

Open **Paired t-test.csv** in JASP. This contains two columns of paired data, pre-diet body mass and post 4 weeks of dieting. Go to T-Tests > Paired Samples t-test. Ctrl-click both variables and add them to the analysis box on the right.





In the analysis options tick the following:

<p><b>Tests</b></p> <p><input checked="" type="checkbox"/> Student</p> <p><input type="checkbox"/> Wilcoxon signed-rank</p> <p><b>Hypothesis</b></p> <p><input checked="" type="radio"/> Measure 1 <math>\neq</math> Measure 2</p> <p><input type="radio"/> Measure 1 &gt; Measure 2</p> <p><input type="radio"/> Measure 1 &lt; Measure 2</p> <p><b>Assumption Checks</b></p> <p><input checked="" type="checkbox"/> Normality</p>	<p><b>Additional Statistics</b></p> <p><input checked="" type="checkbox"/> Location parameter</p> <p><input type="checkbox"/> Confidence interval 95 %</p> <p><input checked="" type="checkbox"/> Effect size</p> <p><input type="checkbox"/> Confidence interval 95 %</p> <p><input checked="" type="checkbox"/> Descriptives</p> <p><input checked="" type="checkbox"/> Descriptives plots</p> <p>Confidence interval 95 %</p> <p><input type="checkbox"/> Vovk-Sellke maximum p-ratio</p> <p><b>Missing Values</b></p> <p><input checked="" type="radio"/> Exclude cases analysis by analysis</p> <p><input type="radio"/> Exclude cases listwise</p>
---	--

## UNDERSTANDING THE OUTPUT

The output should consist of three tables and one graph.

Test of Normality (Shapiro-Wilk)

	W	p
Pre diet body mass - Post 4 weeks diet	0.975	0.124

Note. Significant results suggest a deviation from normality.

The assumption check of normality (Shapiro-Wilk) is not significant suggesting that the pairwise differences are normally distributed, therefore the assumption is not violated. If this showed a significant difference the analysis should be repeated using the non-parametric equivalent, **Wilcoxon's signed rank test**.

Paired Samples T-Test

	t	df	p	Mean Difference	SE Difference	Cohen's d
Pre diet body mass - Post 4 weeks diet	13.039	77	< .001	3.782	0.290	1.476

Note. Student's t-test.

This shows that there is a significant difference in body mass between the pre and post dieting conditions, with a mean difference (location parameter) of 3.783kg. Cohen's d states that this is a large effect.



The descriptive statistics and plot show that there was a reduction in body mass following 4 weeks of dieting.

Descriptives

	N	Mean	SD	SE
Pre diet body mass	78	72.526	8.723	0.988
Post 4 weeks diet	78	68.744	9.009	1.020



## REPORTING THE RESULTS

On average participants lost 3.78 kg (SE: 0.29 kg) body mass following a 4-week diet plan. A paired samples t-test showed this decrease to be significant ( $t(77) = 13.039, p < .001$ ). Cohen's d suggests that this is a large effect

## RUNNING THE NON-PARAMETRIC PAIRED SAMPLES TEST

### WILCOXON'S SIGNED RANK TEST

If you find that your data is not normally distributed (significant Shapiro-Wilk test result) or is ordinal by nature, the equivalent non-parametric independent test is the Wilcoxon's signed rank test. Open **Wilcoxon's rank.csv**. This has two columns one with pre-anxiety and post hypnotherapy anxiety scores (from 0 - 50). In the dataset view make sure that both variables are assigned to the ordinal data type.

Go to T-Tests > Paired Samples t-test and follow the same instructions as above but now only tick the following options:

- ✓ Wilcoxon signed rank
- ✓ Location parameter
- ✓ Effect size



There will be only one table in the output:

Paired Samples T-Test

	W	p	Hodges-Lehmann Estimate	Rank-Biserial Correlation
Pre-anxiety - Post-anxiety	322.000	< .001	8.000	0.480

Note. Wilcoxon signed-rank test.

The Wilcoxon W-statistic is highly significant,  $p < 0.001$ .

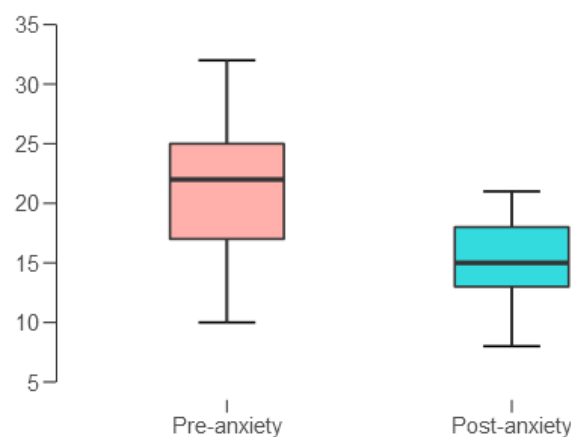
The location parameter, the Hodges–Lehmann estimate, is the median difference between the two groups. The rank-biserial correlation ( $r_B$ ) can be considered as an effect size and is interpreted the same as Pearson’s  $r$ , so 0.48 is a medium to large effect size.

Effect size	Trivial	Small	Medium	Large
Rank -biserial ( $r_B$ )	<0.1	0.1	0.3	0.5

For non-parametric data, you should report median values as your descriptive statistics and use boxplots instead of line graphs and confidence intervals, SD/SE bars.

Descriptive Statistics

	Pre-anxiety	Post-anxiety
Valid	29	29
Missing	0	0
Median	22.0	15.0
Minimum	10.0	8.0
Maximum	32.0	21.0



## REPORTING THE RESULTS

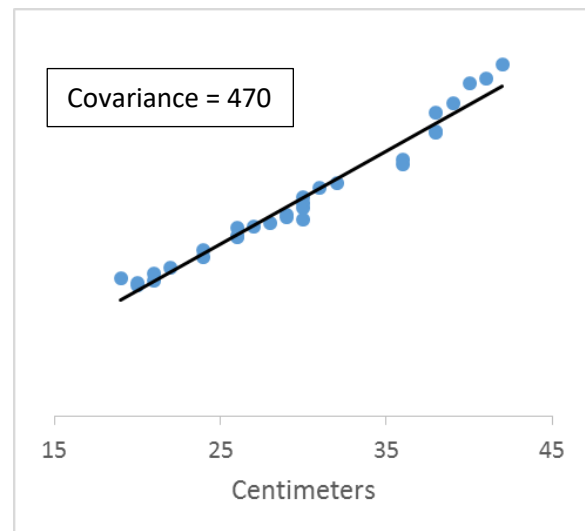
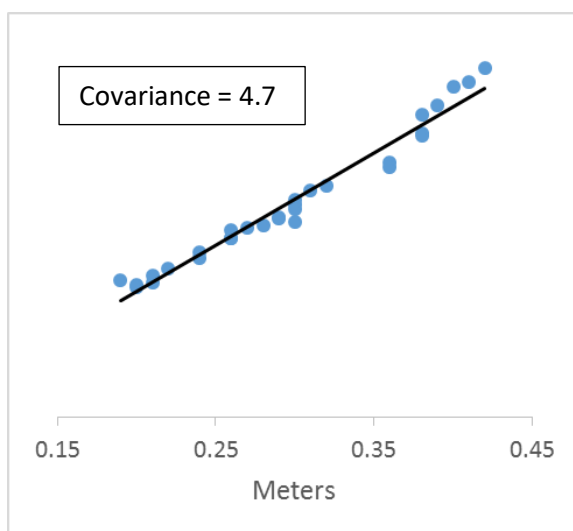
A Wilcoxon’s signed rank test showed that hypnotherapy significantly reduces anxiety scores (Mdn = 15) compared to pre-therapy (Mdn =22) scores,  $W=322$ ,  $p < .001$ .



## CORRELATION ANALYSIS

Correlation is a statistical technique that can be used to determine if, and how strongly, pairs of variables are associated. Correlation is only appropriate for quantifiable data in which numbers are meaningful, such as continuous or ordinal data. It cannot be used for purely categorical data for which we have to use contingency table analysis (see Chi-square analysis in JASP).

Essentially do different variables co-vary? i.e. are changes in one variable reflected in similar changes to another variable? If one variable deviates from its mean does the other variable deviate from its mean in either the same or opposite direction? This can be assessed by measuring covariance, however, this is not standardised. For example, we can measure the covariance of two variables which are measured in meters, however, if we convert the same values to centimetres, we get the same relationship but with a completely different covariance value.



In order to overcome this, standardised covariance is used which is known as **Pearson's correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. If r is close to 0, there is no relationship. If r is (+) then as one variable increases the other also increases. If r is (-) then as one increases, the other decreases (sometimes referred to as an "inverse" correlation).

The correlation coefficient (r) should not be confused with  $R^2$  (coefficient of determination) or R (multiple correlation coefficient as used in regression).

The main assumption in this analysis is that the data have a normal distribution and are linear. This analysis will not work well with curvilinear relationships.



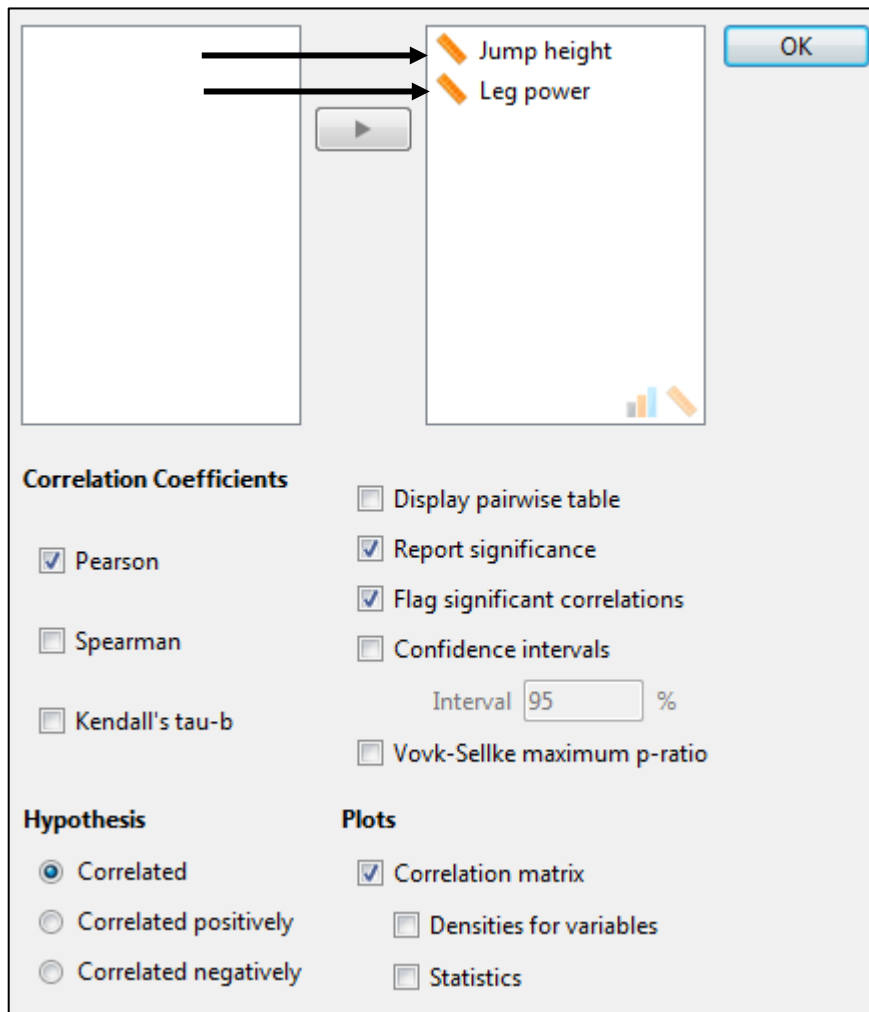
## RUNNING CORRELATION

*The analysis tests the null hypothesis ( $H_0$ ) that there is no association between the two variables*

From the example data open **Jump height correlation.csv** which contains 2 columns of data, jump height (m) and explosive leg power (W). Firstly run the Descriptive statistics and check the boxplots for any outliers.

To run the correlation analysis go to Regression > Correlation matrix. Move the 2 variables to the analysis box on the right. Tick

- ✓ Pearson,
- ✓ Report significance,
- ✓ Flag significant correlations and
- ✓ Correlation matrix under Plots.





## UNDERSTANDING THE OUTPUT

The first table shows the correlation matrix with Pearson's r value and its p value. This shows a highly significant correlation ( $p < .001$ ) with a large r value close to 1 ( $r = 0.984$ ) and that we can reject the null hypothesis.

Pearson Correlations

		Jump height	Leg power
Jump height	Pearson's r	—	
	p-value	—	
Leg power	Pearson's r	0.984***	—
	p-value	< .001	—

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

For simple correlations like this it is easier to look at the pairwise table (go back to analysis and tick the Display pairwise table option. This replaces the correlation matrix in the results which may be easier to read.

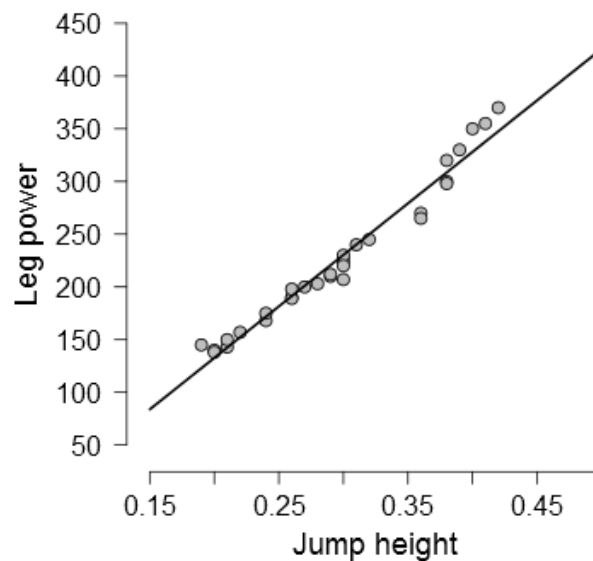
Pearson Correlations

		Pearson's r	p
Jump height	- Leg power	0.984***	< .001

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

The Pearson's r value is actually an effect size where  $< 0.1$  is trivial,  $0.1 - 0.3$  is a small effect,  $0.3 - 0.5$  a moderate effect and  $> 0.5$  a large effect.

The plot provides a simple visualisation of this strong positive correlation ( $r = 0.984$ ,  $p < .001$ )







## GOING ONE STEP FURTHER.

If you take the correlation coefficient  $r$  and square it you get the coefficient of determination ( $R^2$ ). This is a statistical measure of the proportion of variance in one variable that is explained by the other variable. Or:

$$R^2 = \text{Explained variation} / \text{Total variation}$$

$R^2$  is always between 0 and 100% where:

- 0% indicates that the model explains none of the variability of the response data around its mean and
- 100% indicates that the model explains all the variability of the response data around its mean.

In the example above  $r = 0.984$ , so  $R^2 = 0.968$ . This suggests that jump height accounts for 96.8% of the variance in explosive leg power.

## REPORTING THE RESULTS

Pearson's correlation showed a significant correlation between jump height and leg power ( $r = 0.984$ ,  $p < .001$ ) jump height accounting for 96.8% of the variance in leg power.

## RUNNING NON-PARAMETRIC CORRELATION – Spearman's and Kendall's tau

If your data is ordinal or is continuous data that has violated the assumptions required for parametric testing (normality and/or variance) you need to use the non-parametric alternatives to Pearson's correlation coefficient.

The alternatives are Spearman's ( $\rho$ ) or Kendall's ( $\tau$ ) correlation coefficients. Both are based on ranking data and are affected by outliers or normality/variance violations.

Spearman's  $\rho$  is usually used for ordinal scale data and Kendall's  $\tau$  is used in small samples or when many values with the same score (ties). In most cases, Kendall's  $\tau$  and Spearman's rank correlation coefficients are very similar and thus invariably lead to the same inferences.

The effect sizes are the same as Pearson's  $r$ . The main difference is that  $\rho^2$  can be used as an approximate non-parametric coefficient of determination but the same is not true for Kendall's  $\tau$ .

From the example data open **Non-parametric correlation.csv** which contains 2 columns of data, a creativity score and position in the 'World's biggest liar' competition (thanks to Andy Field).

Run the analysis as before but now using Spearman and Kendall's tau-b coefficients instead of Pearson's.



**Correlation Coefficients**

Pearson

Spearman

Kendall's tau-b

Display pairwise table

Report significance

Flag significant correlations

Confidence intervals

Interval  %

Vovk-Sellke maximum p-ratio

**Hypothesis**

Correlated

Correlated positively

Correlated negatively

**Plots**

Correlation matrix

Densities for variables

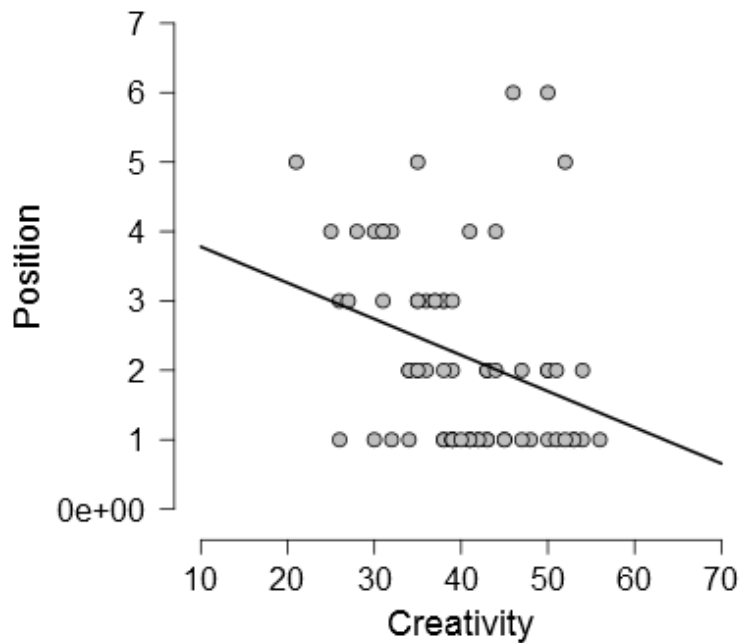
Statistics

Correlation Table

		Spearman		Kendall	
		rho	p	tau B	p
Creativity	- Position	-0.373**	0.002	-0.300**	0.001

\* p < .05, \*\* p < .01, \*\*\* p < .001

As can be seen there is a significant correlation between creativity scores and final position in the 'World's biggest liar' competition, the higher the score the better the final competition position. However, the effect size is only moderate.





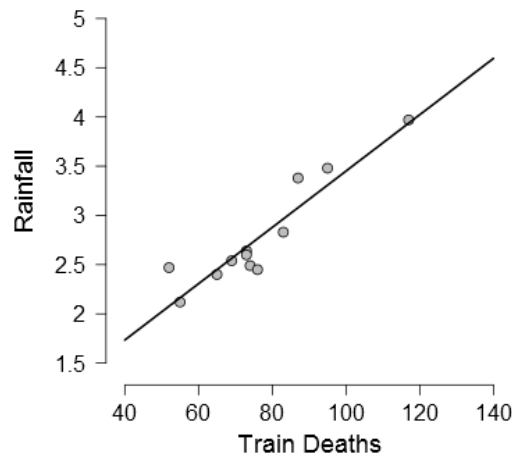
### NOTE OF CAUTION.

Correlation really only give information on the strength of association. It gives no information on the direction i.e. which variable causes the other to change. So it cannot be used to state the one thing causes the other. Often a significant correlation means absolutely nothing and is purely by chance especially if you correlate thousands of variables. This can be seen in the following strange correlations:

#### **Pedestrians killed in a collision with a railway train correlates with rainfall in Missouri:**

Pearson Correlations

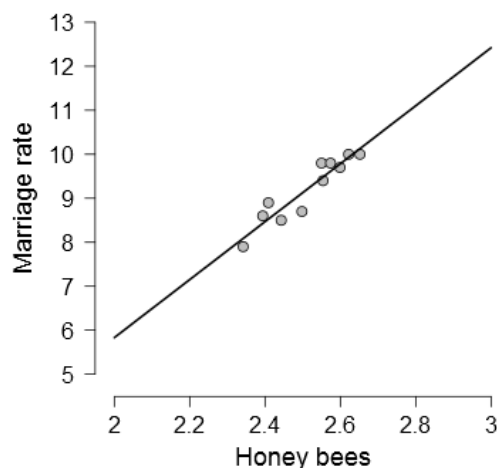
		Pearson's r	p
Train Deaths	- Rainfall	0.928	< .001



#### **Number of honey-producing bee colonies (1000's) correlates strongly with the marriage rate in South Carolina (per 1000 marriages)**

Pearson Correlations

		Pearson's r	p
Honey bees	- Marriage rate	0.938	< .001





## REGRESSION

Whereas correlation tests for associations between variables, regression is the next step commonly used for predictive analysis, i.e. to predict a dependent outcome variable from one (simple regression) or more (multiple regression) independent predictor variables.

Regression results in a hypothetical model of the relationship between the outcome and predictor variable(s). The model used is a linear one defined by the formula;

$$y = c + b*x + \epsilon$$

- $y$  = estimated dependent outcome variable score,
- $c$  = constant,
- $b$  = regression coefficient and
- $x$  = score on the independent predictor variable
- $\epsilon$  = random error component (based on residuals)

**Linear regression provides both the constant and regression coefficient(s).**

Linear regression makes the following assumptions:

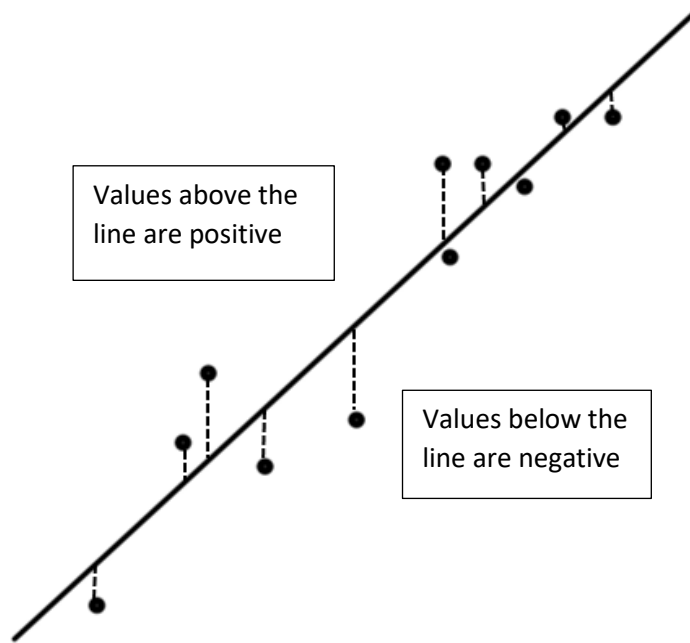
1. **Linear relationship:** important to check for outliers since linear regression is sensitive to their effects.
2. **Independence** of variables
3. **Multivariate normality:** requires all variables to be normally distributed
4. **Homoscedasticity:** homogeneity of variance of the residuals
5. **Minimal multicollinearity /autocorrelation:** when the independent variables/residuals are too highly correlated with each other.

With regard to sample sizes, there are many different 'rules of thumb' in the literature ranging from 10-15 data points per predictor in the model i.e. 4 predictor variables will each require between 40 and 60 data points each to  $50 + (8 * \text{number of predictors})$  for each variable. So for 4 variables that would require 82 data point for each variable. Effectively the bigger your sample size the better your model.

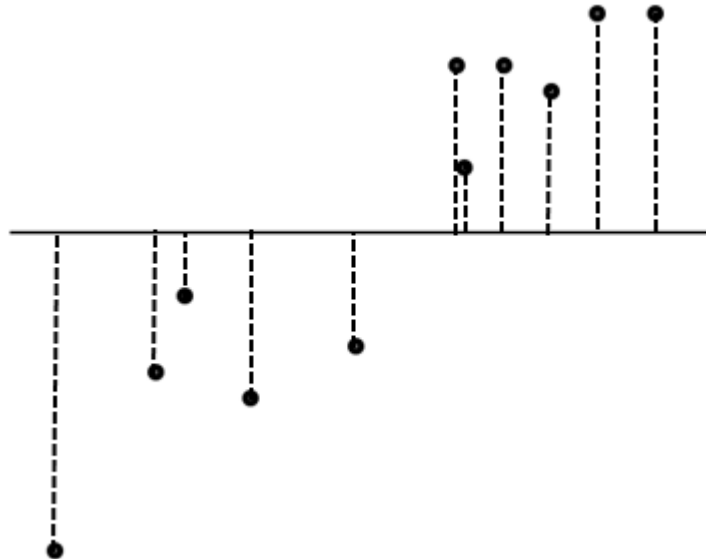
## SUMS OF SQUARES (Boring, but the basis of evaluating the regression model.)

Most regression analysis will produce the best model available, but how good is it actually and how much error is in the model?

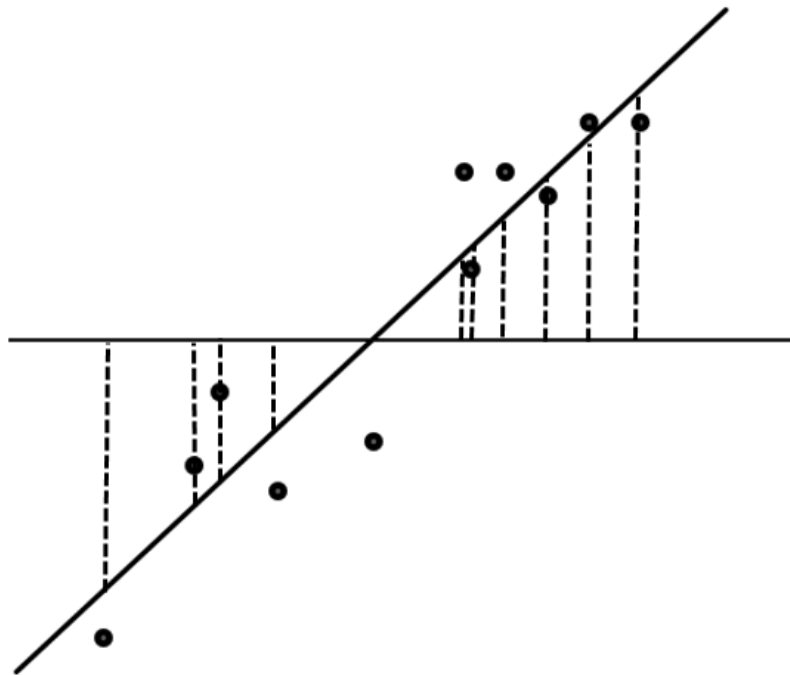
This can be determined by looking at 'the goodness of fit' using the sums of squares. This is a measure of how close the actual data points are close to the modelled regression line.



The vertical difference between the data points and the predicted regression line are known as the **residuals**. These values are squared to remove the negative numbers and then summed to give  $SS_R$ . This is effectively the error of the model or the '**goodness of fit**', obviously the smaller the value the less error in the model.



The vertical difference between the data points and the mean of the outcome variable can be calculated. These values are squared to remove the negative numbers and then summed to give the **total** sum of the squares  $SS_T$ . This shows how good the mean value is as a model of the outcome scores.



The vertical difference between the mean of the outcome variable and the predicted regression line are now determined. Again these values are squared to remove the negative numbers and then summed to give the **model** sum of squares (**SS<sub>M</sub>**). This indicates how better the model is compared to just using the mean of the outcome variable.

So, the larger the **SS<sub>M</sub>** the better the model is at predicting the outcome compared to the mean value alone. If this is accompanied by a small **SS<sub>R</sub>** the model also has a small error.

**R<sup>2</sup>** is similar to the coefficient of determination in correlation in that it shows how much of the variation in the outcome variable can be predicted by the predictor variable(s).

$$R^2 = \frac{SS_M}{SS_R}$$

In regression, the model is assessed by the F statistic based on the improvement in prediction of the model **SS<sub>M</sub>** and the residual error **SS<sub>R</sub>**. The larger the F value the better the model.

$$F = \frac{\text{Mean } SS_M}{\text{Mean } SS_R}$$



## SIMPLE REGRESSION

*Regression tests the null hypothesis ( $H_0$ ) that there will be no significant prediction of the dependent (outcome) variable by the predictor variable(s).*

Open **Rugby kick regression.csv**. This dataset contains rugby kick data including distance kicked, right/left leg strength and flexibility and bilateral leg strength.

Firstly go to Descriptives > Descriptive statistics and check the boxplots for any outliers. In this case, there should be none, though it is good practice to check.

For this simple regression go to Regression > Linear regression and put distance into the Dependent Variable (outcome) and R\_Strength into the Covariates (Predictor) box. Tick the following options in the Statistics options:

Statistics

**Regression Coefficients**

- Estimates
- Confidence intervals
- Interval  %
- Covariance matrix
- Model fit
- R squared change
- Descriptives
- Part and partial correlations
- Collinearity diagnostics

**Residuals**

- Durbin-Watson
- Casewise diagnostics
- Outliers outside  standard deviations
- All cases

## UNDERSTANDING THE OUTPUT

You will now get the following outputs:

### Model Summary

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	RMSE	Durbin-Watson
1	0.784	0.614	0.579	55.285	1.524

Here it can be seen that the correlation (R) between the two variables is high (0.784). The R<sup>2</sup> value of 0.614 tells us that right leg strength accounts for 61.4% of the variance in kick distance. Durbin-Watson checks for correlations between residuals, which can invalidate the test. This should be above 1 and below 3 and ideally around 2.



### ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	53589.863	1	53589.863	17.533	0.002
	Residual	33621.061	11	3056.460		
	Total	87210.923	12			

The ANOVA table shows all the sums of squares mentioned earlier. With regression being the model and Residual being the error. The F-statistic is significant  $p=0.002$ . This tells us that the model is a significantly better predictor of kicking distance than the mean distance.

Report as **F (1, 11) = 17.53,  $p<.001$ .**

### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	57.105	103.588		0.551	0.592
	R_Strength	6.425	1.534	0.784	4.187	0.002

This table gives the coefficients (unstandardized) that can be put into the linear equation.

$$y = c + b*x$$

y = estimated dependent outcome variable score,

c = constant (**intercept**)

b = regression coefficient (**R\_strength**)

x = score on the independent predictor variable

For example for a leg strength of 60 kg the distance kicked can be predicted by the following:

$$\text{Distance} = 57.105 + (6.452 * 60) = 454.6 \text{ m}$$

### FURTHER ASSUMPTION CHECKS

In Assumption checks, tick the following two options:

▼ Assumption Checks

**Residual Plots**

Residuals vs. dependent

Residuals vs. covariates

Residuals vs. predicted

Residuals histogram

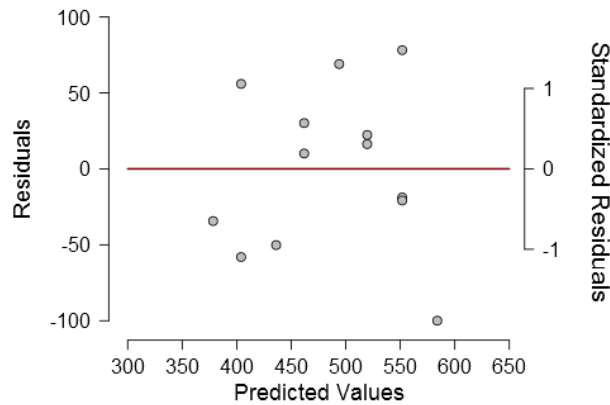
Standardized residuals

Q-Q plot standardized residuals

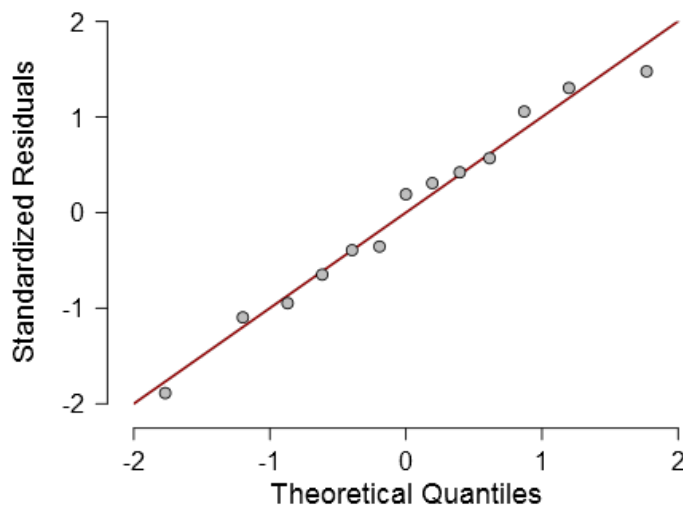




This will result in two graphs:



This graph shows a balanced random distribution of the residuals around the baseline suggesting that the assumption of homoscedasticity has not been violated. (See Exploring data integrity in JASP for further details.)



The Q-Q plot shows that the standardized residuals fit nicely along the diagonal suggesting that both assumptions of normality and linearity have also not been violated.

## REPORTING THE RESULTS

Linear regression shows that right leg strength can significantly predict kicking distance  $F(1, 11) = 17.53, p < .001$  using the following regression equation:

$$\text{Distance} = 57.105 + (6.452 * \text{Right leg strength})$$



## MULTIPLE REGRESSION

The model used is still a linear one defined by the formula;

$$y = c + b*x + \epsilon$$

- $y$  = estimated dependent outcome variable score,
- $c$  = constant,
- $b$  = regression coefficient and
- $x$  = score on the independent predictor variable
- $\epsilon$  = random error component (based on residuals)

However, we now have more than 1 regression coefficient and predictor score i.e.

$$y = c + b_1*x_1 + b_2*x_2 + b_3*x_3 \dots\dots\dots b_n*x_n$$

### Data entry methods.

If predictors are uncorrelated their order of entry has little effect on the model. In most cases, predictor variables are correlated to some extent and thus, the order in which the predictors are entered can make a difference. The different methods are subject to much debate in the area.

**Forced entry (Enter):** This is the **default method** in which all the predictors are forced into the model in the order they appear in the Covariates box. This is considered to be the best method.

**Blockwise entry (Hierarchical entry):** The researcher, normally based on prior knowledge and previous studies, decides the order in which the known predictors are entered first depending on their importance in predicting the outcome.. Additional predictors are added in further steps.

**Stepwise (Backward entry):** All predictors are initially entered in the model and then the contribution of each is calculated. Predictors with less than a given level of contribution ( $p < 0.1$ ) are removed. This process repeats until all the predictors are statistically significant.

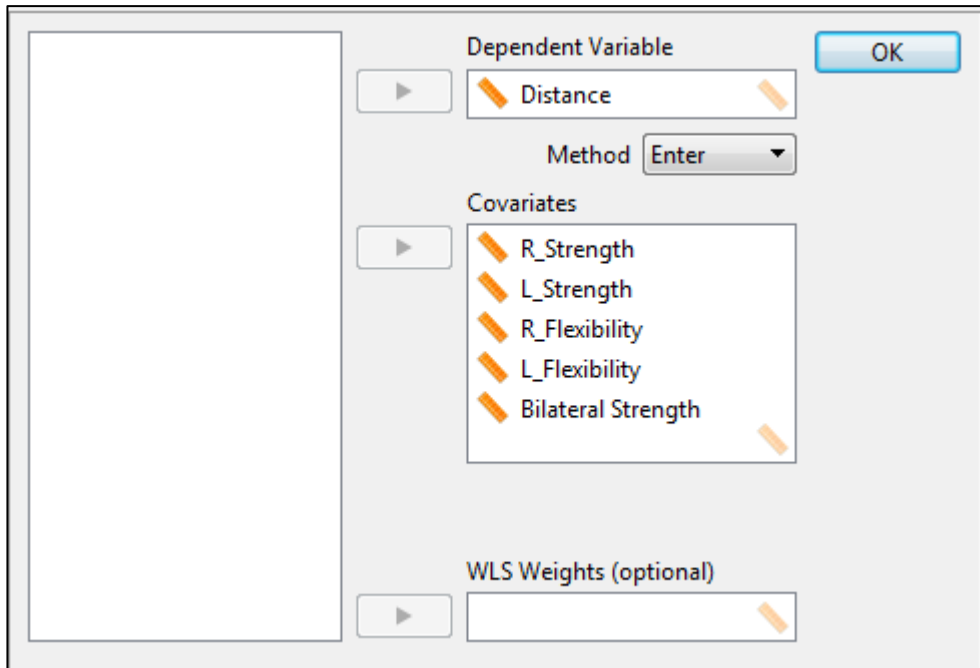
**Stepwise (Forward entry):** The predictor with the highest simple correlation with the outcome variable is entered first. Subsequent predictors selected on the basis of the size of their semi-partial correlation with the outcome variable. This is repeated until all predictors that contribute significant unique variance to the model have been included in the model.

**Stepwise entry:** Same as the Forward method, except that every time a predictor is added to the model, a removal test is made of the least useful predictor. The model is constantly reassessed to see whether any redundant predictors can be removed.

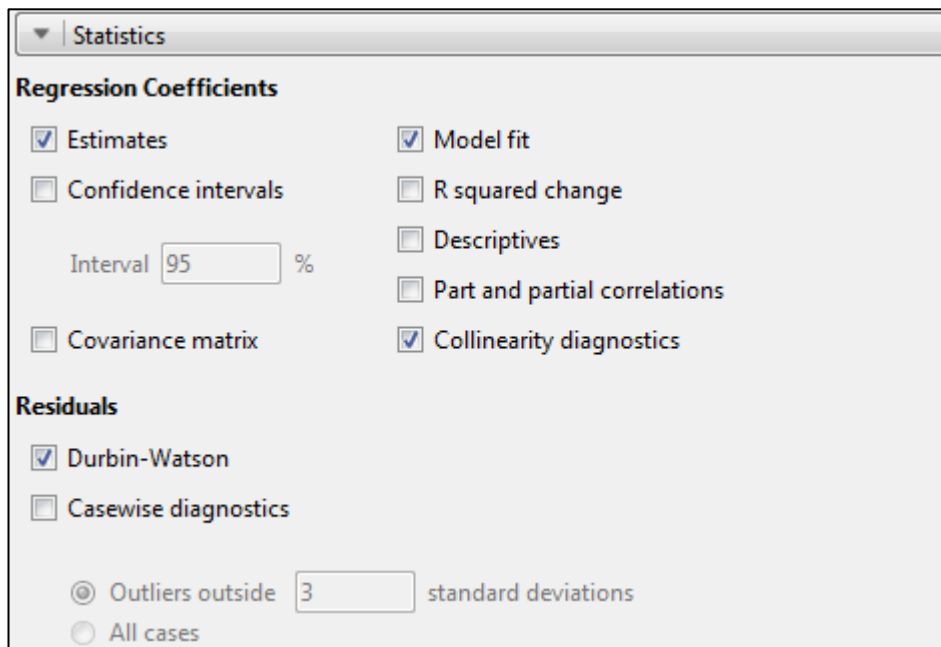
There are many reported disadvantages of using stepwise data entry methods, however, **Backward** entry methods can be useful for exploring previously unused predictors or for fine-tuning the model to select the best predictors from the available options.

## RUNNING MULTIPLE REGRESSION

Open **Rugby kick regression.csv** that we used for simple regression. Go to Regression > Linear regression and put distance into the Dependent Variable (outcome) and now add all the other variables into the Covariates (Predictor) box.



In the Variable section leave the Method as **Enter**. **Tick** the following options in the Statistics options, Estimates, Model fit, Collinearity diagnostics and Durbin-Watson.





## UNDERSTANDING THE OUTPUT

You will now get the following outputs:

Model Summary

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	RMSE	Durbin-Watson
1	0.902	0.814	0.681	48.132	1.328

The adjusted R<sup>2</sup> (used for multiple predictors) shows that they can predict 68.1% of the outcome variance. Durbin-Watson checks for correlations between residuals is between 1 and 3 as required.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	70994.078	5	14198.816	6.129	0.017
	Residual	16216.845	7	2316.692		
	Total	87210.923	12			

The ANOVA table shows the F-statistic to be significant p=0.017 suggesting that the model is a significantly better predictor of kicking distance than the mean distance.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
1	(Intercept)	-92.367	218.389		-0.423	0.685		
	R_Strength	1.747	3.321	0.213	0.526	0.615	0.162	6.180
	L_Strength	0.703	3.590	0.086	0.196	0.850	0.138	7.231
	R_Flexibility	4.078	4.759	0.373	0.857	0.420	0.140	7.125
	L_Flexibility	-1.339	2.447	-0.135	-0.547	0.601	0.438	2.281
	Bilateral Strength	1.665	0.946	0.423	1.759	0.122	0.458	2.181

This table shows one model and the constant (intercept) and regression coefficients (unstandardized) for all the predictors forced into the model. Even though the ANOVA shows the model to be significant none of the predictor regression coefficients are significant!

The collinearity statistics, Tolerance and VIF (Variance Inflation Factor) check the assumption of multicollinearity. As a rule of thumb if VIF >10 and tolerance <0.1 the assumptions have been greatly violated. If the **average** VIF >1 and tolerance <0.2 the model may be biased. In this case, the average VIF is quite large (around 5).



As a comparison re-run the analyses but now choose Backward as the method of data entry.

The outputs are as follows:

Model Summary

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	RMSE	Durbin-Watson
1	0.902	0.814	0.681	48.132	
2	0.902	0.813	0.720	45.146	
3	0.897	0.805	0.740	43.505	
4	0.884	0.782	0.738	43.618	1.676

JASP has now calculated 4 potential regression models. It can be seen that each consecutive model increases the adjusted R<sup>2</sup>, with model 4 accounting for 73.5% of the outcome variance. The Durbin-Watson score is also higher than with the forced entry method.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	70994.078	5	14198.816	6.129	0.017
	Residual	16216.845	7	2316.692		
	Total	87210.923	12			
2	Regression	70905.329	4	17726.332	8.697	0.005
	Residual	16305.594	8	2038.199		
	Total	87210.923	12			
3	Regression	70176.855	3	23392.285	12.359	0.002
	Residual	17034.068	9	1892.674		
	Total	87210.923	12			
4	Regression	68185.712	2	34092.856	17.920	< .001
	Residual	19025.211	10	1902.521		
	Total	87210.923	12			

The ANOVA table indicates that each successive model is better as shown by the increasing F-value and improving p value.



### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics		
							Tolerance	VIF	
1	(Intercept)	-92.367	218.389		-0.423	0.685			
	R_Strength	1.747	3.321	0.213	0.526	0.615	0.162	6.180	
	L_Strength	0.703	3.590	0.086	0.196	0.850	0.138	7.231	
	R_Flexibility	4.078	4.759	0.373	0.857	0.420	0.140	7.125	
	L_Flexibility	-1.339	2.447	-0.135	-0.547	0.601	0.438	2.281	
	Bilateral Strength	1.665	0.946	0.423	1.759	0.122	0.458	2.181	
2	(Intercept)	-110.347	185.840		-0.594	0.569			
	R_Strength	2.218	2.148	0.271	1.033	0.332	0.340	2.938	
	R_Flexibility	4.501	3.978	0.411	1.131	0.291	0.177	5.658	
	L_Flexibility	-1.370	2.291	-0.138	-0.598	0.566	0.440	2.272	
	Bilateral Strength	1.605	0.840	0.408	1.910	0.092	0.512	1.954	
	3	(Intercept)	-116.892	178.772		-0.654	0.530		
R_Strength		2.710	1.911	0.331	1.418	0.190	0.399	2.505	
R_Flexibility		2.886	2.814	0.264	1.026	0.332	0.328	3.048	
Bilateral Strength		1.642	0.807	0.418	2.033	0.073	0.515	1.944	
4		(Intercept)	46.251	81.820		0.565	0.584		
		R_Strength	3.914	1.512	0.478	2.588	0.027	0.641	1.561
	Bilateral Strength	2.009	0.725	0.511	2.770	0.020	0.641	1.561	

Model 1 is the same as the forced entry method first used. The table shows that as the least significantly contributing predictors are sequentially removed, we end up with a model with two significant predictor regression coefficients, right leg strength and bilateral leg strength. Both tolerance and VIF are acceptable.

We now can report the Backward predictor entry results in a highly significant model  $F(2, 10) = 17.92$ ,  $p < .001$  and a regression equation of

$$\text{Distance} = 57.105 + (3.914 * R\_Strength) + (2.009 * Bilateral Strength)$$

### TESTING FURTHER ASSUMPTIONS.

As for the simple linear regression example, tick the following options.

▼ Assumption Checks

**Residual Plots**

Residuals vs. dependent

Residuals vs. covariates

Residuals vs. predicted

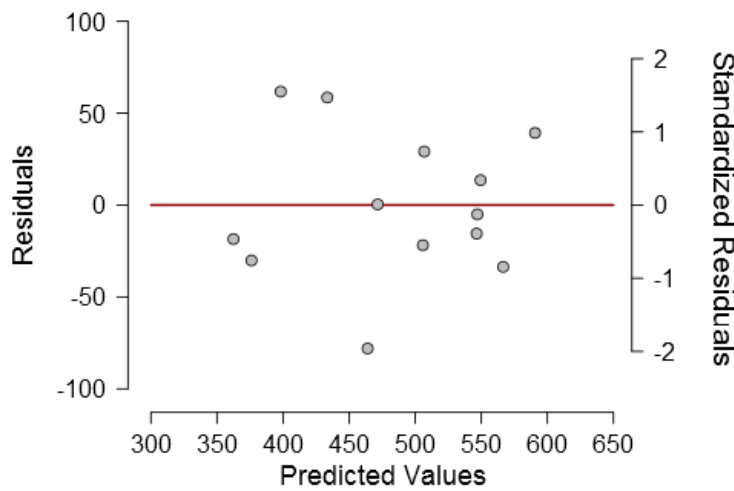
Residuals histogram

Standardized residuals

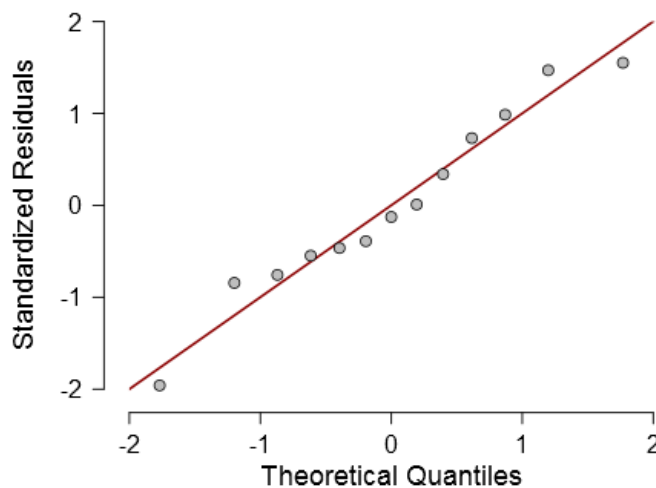
Q-Q plot standardized residuals



### Residuals vs. Predicted



### Q-Q Plot Standardized Residuals ▼



The balanced distribution of the residuals around the baseline suggests that the assumption of homoscedasticity has not been violated.

The Q-Q plot shows that the standardized residuals fit along the diagonal suggesting that both assumptions of normality and linearity have also not been violated.

### REPORTING THE RESULTS

Multiple linear regression using backward data entry shows that right leg and bilateral strength can significantly predict kicking distance  $F(2,10) = 17.92, p < .001$  using a regression equation of

$$\text{Distance} = 57.105 + (3.914 * R\_Strength) + (2.009 * \text{Bilateral Strength})$$



## IN SUMMARY

$R^2$  provides information on how much variance is explained by the model using the predictors provided.

F-statistic provides information as to how good the model is.

The unstandardized (b)-value provides a constant which reflects the strength of the relationship between the predictor(s) and the outcome variable.

Violation of assumptions can be checked using Durbin-Watson value, tolerance/VIF values, Residual vs predicted and Q-Q plots.



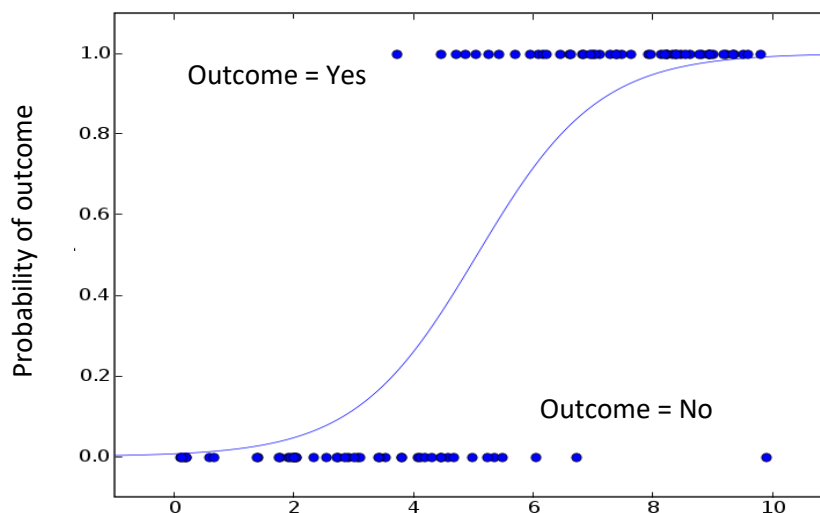


## LOGISTIC REGRESSION

In simple and multiple linear regression outcome and predictor variable(s) were continuous data. What if the outcome was a binary/categorical measure? Can, for example, a yes or no outcome be predicted by other categorical or continuous variables? The answer is yes if binary logistic regression is used. This method is used to predict the probability of the binary yes or no outcome.

The null hypothesis tested is that there is no relationship between the outcome and predictor variable(s).

As can be seen in the graph below, a linear regression line between the yes and no responses would be meaningless as a prediction model. Instead, a sigmoidal logistic regression curve is fitted with a minimum of 0 and a maximum of 1. It can be seen that some predictor values overlap between yes and no. For example, a prediction value of 5 would give an equal 50% probability of being a yes or no outcome. Thresholds are therefore calculated to determine if a predictor data value will be classified as a yes or no outcome.



## ASSUMPTIONS FOR BINARY LOGISTIC REGRESSION

- The dependent variable must be binary i.e. yes or no, male or female, good or bad.
- One or more independent (predictor variables) which can be continuous or categorical variables.
- A linear relationship between any continuous independent variables and the logit transformation (natural log of the odds that the outcome equals one of the categories) of the dependent variable.

## LOGISTIC REGRESSION METRICS

**AIC** (Akaike Information Criteria) and **BIC** (Bayesian Information Criteria) are measures of fit for the model, the best model will have the lowest AIC and BIC values.



Three **pseudo R<sup>2</sup>** values are calculated in JASP, McFadden, Nagelkerke and Tjur. These are analogous to R<sup>2</sup> in linear regression and all give different values. What constitutes a good R<sup>2</sup> value varies, however, they are useful when comparing different models for the same data. The model with the largest R<sup>2</sup> statistic is considered to be the best.

The **confusion matrix** is a table showing actual vs predicted outcomes and can be used to determine the accuracy of the model. From this sensitivity and specificity can be derived.

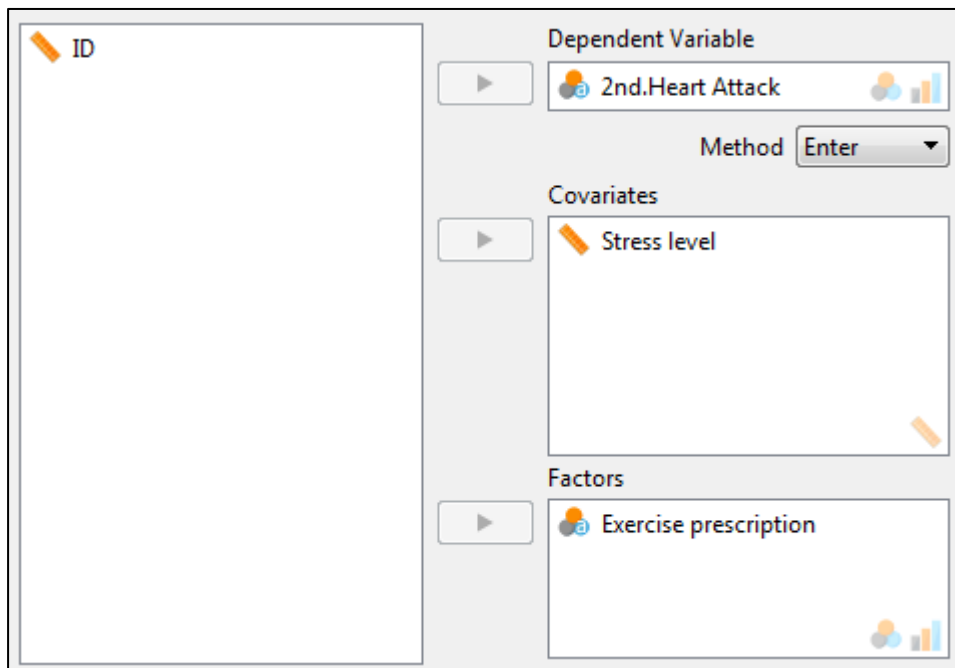
**Sensitivity** is the percentage of cases that had the observed outcome was correctly predicted by the model (i.e., true positives).

**Specificity** is the percentage of observations that were also correctly predicted as not having the observed outcome (i.e., true negatives).

## RUNNING LOGISTIC REGRESSION

Open **Heart attack.csv** in JASP. This contains 4 columns of data, Patient ID, did they have a second heart attack (yes/no), whether they were prescribed exercise (yes/no) and their stress levels (high value = high stress).

Put the outcome variable (2<sup>nd</sup> heart attack) into the Dependent variable, add the stress levels to Covariates and Exercise prescription to Factors. Leave the data entry method as Enter.





In the Statistics options tick Estimates, Odds ratios, Confusion matrix, Sensitivity and Specificity.

▼ Statistics

**Descriptives**

Factor descriptives

**Regression Coefficients**

Estimates

Standardized coefficients

Odds ratios

Confidence intervals

Interval  %

Odds ratio scale

Robust standard errors

Vovk-Sellke maximum p-ratio

**Performance Diagnostics**

Confusion matrix

Proportions

**Performance metrics**

AUC

Sensitivity / Recall

Specificity

Precision

F-measure

Brier score

H-measure

## UNDERSTANDING THE OUTPUT

The initial output should comprise of 4 tables.

The model summary shows that H1 (with the lowest AIC and BIC scores) suggests a significant

Model summary

Model	Deviance	AIC	BIC	df	X <sup>2</sup>	p	McFadden R <sup>2</sup>	Nagelkerke R <sup>2</sup>	Tjur R <sup>2</sup>
H <sub>0</sub>	55.452	57.452	59.141	39					
H <sub>1</sub>	34.195	40.195	45.261	37	21.257	< .001	0.383	0.550	0.126

relationship ( $X^2(37) = 17.82, p < .001$ ) between the outcome (2<sup>nd</sup> heart attack) and the predictor variables (exercise prescription and stress levels).

McFadden's R<sup>2</sup> = 0.383. It is suggested that a range from 0.2 to 0.4 indicates a good model fit.

Coefficients

	Estimate	Standard Error	Odds Ratio	z	p
(Intercept)	-4.368	2.550	0.013	-1.713	0.087
Stress level	0.089	0.041	1.093	2.159	0.031
Exercise prescription (Yes)	-2.043	0.890	0.130	-2.295	0.022

Note. 2nd.Heart Attack level 'Yes' coded as class 1.

Both stress level and exercise prescription are significant predictor variables (p=.031 and .022 respectively). The most important values in the coefficients table are the odds ratios. For the continuous predictor, an odds ratio of greater than 1 suggests a positive relationship while < 1 implies



a negative relationship. This suggests that high stress levels are significantly related to an increased probability of having a second heart attack. Having an exercise intervention is related to a significantly reduced probability of a second heart attack. The odds ratio of 0.13 can be interpreted as only having a 13% probability of a 2<sup>nd</sup> heart attack if undergoing an exercise intervention.

Confusion matrix			Performance metrics	
Observed	Predicted		Value	
	No	Yes	Sensitivity	Specificity
No	15.000	5.000	0.750	
Yes	5.000	15.000	0.750	

The confusion matrix shows that the 15 true negative and positive cases were predicted by the model while the error, false negatives and positives, were found in 5 cases. This is confirmed in the Performance metrics where both sensitivity (% of cases that had the outcome correctly predicted) and specificity (% of cases correctly predicted as not having the outcome (i.e., true negatives) are both 75%.

### PLOTS

These findings can be easily visualised through the inferential plots.

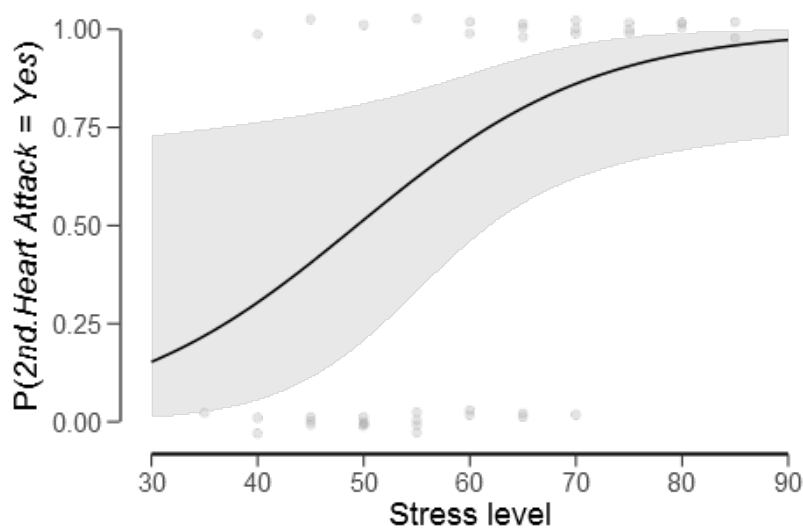
Plots

**Inferential plots**

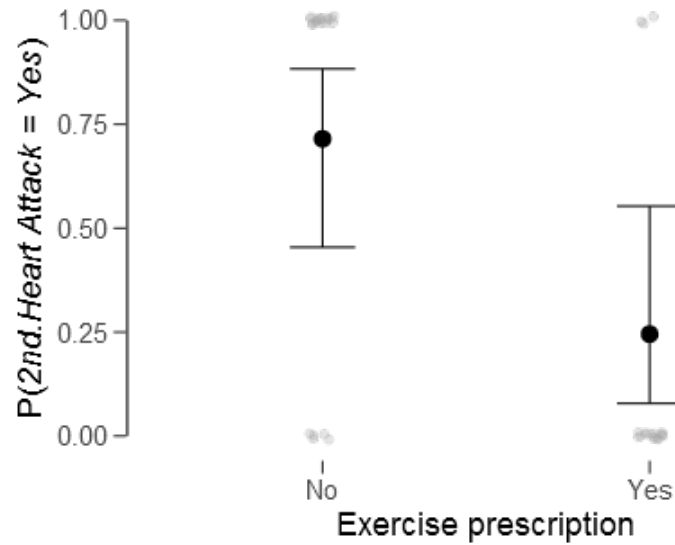
- Display conditional estimates plots
- Confidence Interval  %
- Show data points

**Residual plots**

- Predicted - residuals plot
- Predictor - residuals plots
- Squared Pearson residuals plot



As stress levels increase the probability of having a second heart attack increases.



No exercise intervention increases the probability of a 2<sup>nd</sup> heart attack while it is reduced when it had been put in place.

### REPORTING THE RESULTS

A logistic regression was performed to ascertain the effects of stress and exercise intervention on the likelihood that participants have a 2<sup>nd</sup> heart attack. The logistic regression model was statistically significant,  $\chi^2(37) = 21.257, p < .001$ . The model correctly classified 75.0% of cases. Increasing stress was associated with an increased likelihood of a 2<sup>nd</sup> heart attack, but decreasing stress was associated with a reduction in the likelihood. The presence of an exercise intervention programme reduced the probability of a 2<sup>nd</sup> heart attack to 13%.



## COMPARING MORE THAN TWO INDEPENDENT GROUPS

### ANOVA

Whereas t-tests compare the means of two groups/conditions, one-way analysis of variance (**ANOVA**) compares the means of 3 or more groups/conditions. There are both independent and repeated measures ANOVAs available in JASP. ANOVA has been described as an ‘omnibus test’ which results in an F-statistic that compares whether the datasets overall explained variance is significantly greater than the unexplained variance. **The null hypothesis tested is that there is no significant difference between the means of all the groups.** If the null hypothesis is rejected, ANOVA just states that there is a significant difference between the groups but not where those differences occur. In order to determine where the group differences are, post hoc (From the Latin *post hoc*, "after this") tests are subsequently used.

Why not just multiple pairwise comparisons? If there are 4 groups (A, B, C, D) for example and the differences were compared using multiple t-tests:

- A vs. B            P<0.05            95% no type I error
- A vs. C            P<0.05            95% no type I error
- A vs. D            P<0.05            95% no type I error
- B vs. C            P<0.05            95% no type I error
- B vs. D            P<0.05            95% no type I error
- C vs. D            P<0.05            95% no type I error

Assuming that each test was independent, the overall probability would be:

$$0.95 * 0.95 * 0.95 * 0.95 * 0.95 * 0.95 = 0.735$$

This is known as familywise error or, cumulative Type I error, and in this case results in only a 73.5% probability of no Type I error whereby the null hypothesis could be rejected when it is in fact true. This is overcome by using post hoc tests that make multiple pairwise comparisons with stricter acceptance criteria to prevent familywise error.

### ASSUMPTIONS

The independent ANOVA makes the same assumptions as most other parametric tests.

- The independent variable must be categorical and the dependent variable must be continuous.
- The groups should be independent of each other.
- The dependent variable should be approximately normally distributed.
- There should be no significant outliers.
- There should be homogeneity of variance between the groups otherwise the p value for the F-statistic may not be reliable.

The first 2 assumptions are usually controlled through the use of appropriate research method design.

If the last three assumptions are violated then the non-parametric equivalent, Kruskal-Wallis should be considered instead.



## POST HOC TESTING

JASP provides 4 alternatives for use with the independent group ANOVA tests:

**Bonferroni** – can be very conservative but gives guaranteed control over Type I error at the risk of reducing statistical power.

**Holm** – the Holm-Bonferroni test which is a sequential Bonferroni method that is less conservative than the original Bonferroni test.

**Tukey** – one of the most commonly used tests and provides controlled Type I error for groups with the same sample size and equal group variance.

**Scheffe** – controls for the overall confidence level when the group sample sizes are different.

## EFFECT SIZE

JASP provides 3 alternative effect size calculations for use with the independent group ANOVA tests:

**Eta squared ( $\eta^2$ )** - accurate for the sample variance explained but overestimates the population variance. This can make it difficult to compare the effect of a single variable in different studies.

**Partial Eta squared ( $\eta_p^2$ )** – this solves the problem relating to population variance overestimation allowing for comparison of the effect of the same variable in different studies.

**Omega squared ( $\omega^2$ )** – Normally, statistical bias gets very small as sample size increases, but for small samples ( $n < 30$ )  $\omega^2$  provides an unbiased effect size measure.

Test	Measure	Trivial	Small	Medium	Large
ANOVA	Eta	<0.1	0.1	0.25	0.37
	Partial Eta	<0.01	0.01	0.06	0.14
	Omega squared	<0.01	0.01	0.06	0.14

## RUNNING THE INDEPENDENT ANOVA

Load **Independent ANOVA diet.csv**. This contains A column containing the 3 diets used (A, B and C) and another column containing the absolute amount of weight loss after 8 weeks on one of 3 different diets For good practice check the descriptive statistics and the boxplots for any extreme outliers.

Go to ANOVA > ANOVA, put weight loss into the Dependent Variable and the Diet groupings into the Fixed Factors box. In the first instance tick both Assumption Checks and in Additional Options tick Descriptive statistics and  $\omega^2$  as the effect size;



**Dependent Variable**

▶

OK

**Fixed Factors**

▶

**WLS Weights**

▶

---

▶ Model

▼ Assumption Checks

Homogeneity tests

Q-Q plot of residuals

---

▼ Additional Options

**Display**

Descriptive statistics

Estimates of effect size

$\eta^2$   partial  $\eta^2$    $\omega^2$

Vovk-Sellke maximum p-ratio

This should result in 3 tables and one Q-Q plot.

## UNDERSTANDING THE OUTPUT

ANOVA - Weight loss kg ▼

Cases	Sum of Squares	df	Mean Square	F	p	$\omega^2$
Diet	92.369	2.000	46.184	10.826	< .001	0.214
Residual	294.371	69.000	4.266			

Note. Type III Sum of Squares

The main ANOVA table shows that the F-statistic is significant ( $p < .001$ ) and that there is a large effect size. Therefore, there is a significant difference between the means of the 3 diet groups.





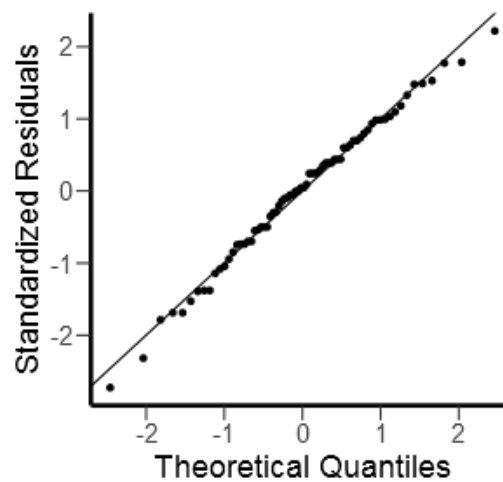
## TESTING ASSUMPTIONS

Before accepting this any violations in the assumptions required for an ANOVA should be checked.

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.298	2.000	69.000	0.280

Levene's test shows that homogeneity of variance is not significant.



The Q-Q plot shows that the data appear to be normally distributed and linear.

Descriptives - Weight loss kg

Diet	Mean	SD	N
Diet A	3.008	1.668	24.000
Diet B	3.413	2.361	24.000
Diet C	5.588	2.108	24.000

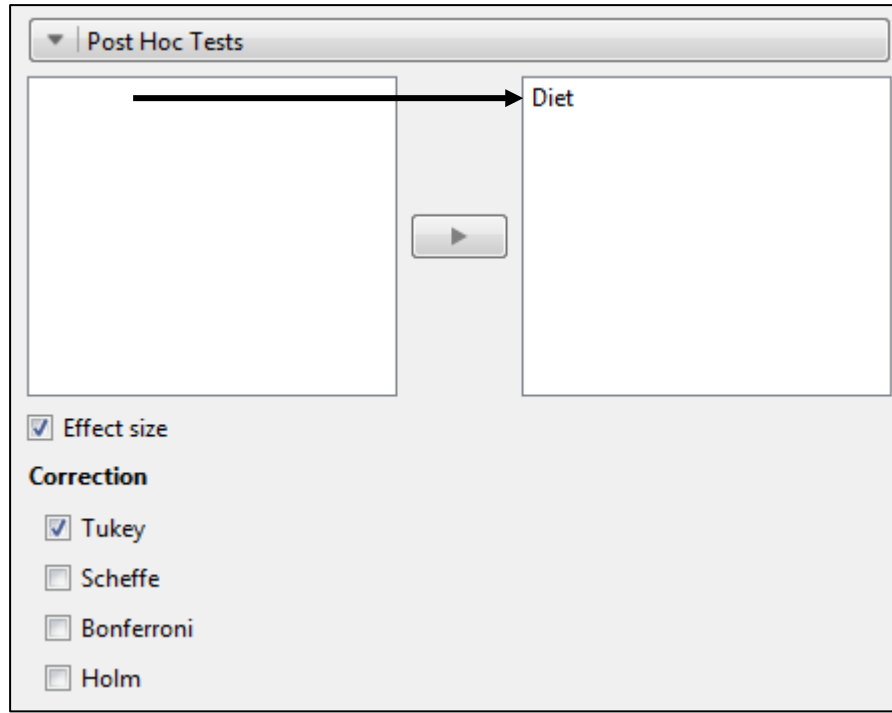
The descriptive statistics suggest that Diet 3 results in the highest weight loss after 8 weeks.

**If the ANOVA reports no significant difference you can go no further in the analysis.**



## POST HOC TESTING

If the ANOVA is significant post hoc testing can now be carried out. In Post Hoc Tests add Diet to the analysis box on the right, tick Effect size and, in this case, use Tukey for the post hoc correction.



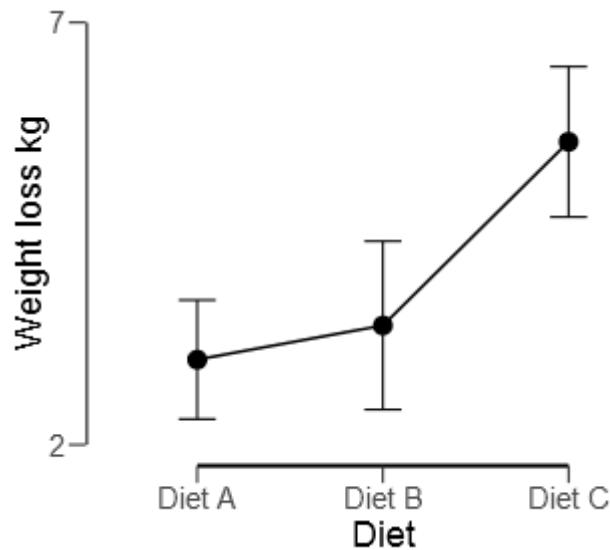
Also in Descriptive Plots add the Factor – Diet to the horizontal axis and tick display error bars.

Post Hoc Comparisons - Diet

		Mean Difference	SE	t	Cohen's d	P <sub>Tukey</sub>
Diet A	Diet B	-0.404	0.596	-0.678	-0.198	0.777
	Diet C	-2.579	0.596	-4.326	-1.357	< .001
Diet B	Diet C	-2.175	0.596	-3.648	-0.972	0.001

Note. Cohen's d does not correct for multiple comparisons.

Post hoc testing shows that there is no significant difference between weight loss on diets A and B. However, it is significantly higher in diet C compared to diet A ( $p < .001$ ) and diet B ( $p = .001$ ). Cohen's d shows that these differences have a large effect size.



## REPORTING THE RESULTS

Independent one way ANOVA showed a significant effect of the type of diet on weight loss after 10 weeks ( $F(2, 69) = 46.184, p < .001, \omega^2 = 0.214$ ).

Post hoc testing using Tukey's correction revealed that diet C resulted in significantly greater weight loss than diet A ( $p < .001$ ) or diet B ( $p = .001$ ). There were no significant differences in weight loss between diets A and B ( $p = .777$ ).

## KRUSKAL-WALLIS – NON-PARAMETRIC ANOVA

If your data fails parametric assumption tests or is nominal in nature, the Kruskal-Wallis H test is a non-parametric equivalent to the independent samples ANOVA. It can be used for comparing two or more independent samples of equal or different sample sizes. Like the Mann-Whitney and Wilcoxon's tests, it is a rank based test.

As with the ANOVA, Kruskal-Wallis H test (also known as the "one-way ANOVA on ranks") is an omnibus test which does not specify which specific groups of the independent variable are statistically significantly different from each other. To do this, JASP provides the option for running Dunn's post hoc test. This multiple comparisons test can be very conservative in particular for large numbers of comparisons.

Load **Kruskal-Wallis ANOVA.csv** dataset into JASP. This dataset contains subjective pain scores for participants undergoing no treatment (control), cryotherapy or combined cryotherapy-compression for delayed onset muscle soreness after exercise.



## RUNNING THE KRUSKAL-WALLIS TEST

Go to ANOVA > ANOVA. In the analysis window add Pain score to the dependent variable and treatment to the fixed factors. Check that the pain score is set to ordinal. This will automatically run the normal independent ANOVA. Under Assumption Checks tick both Homogeneity tests and Q-Q plots.

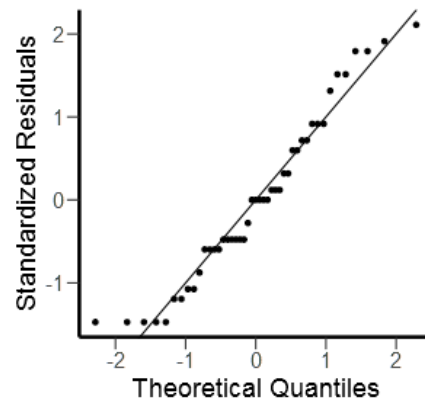
ANOVA - Pain Score

Cases	Sum of Squares	df	Mean Square	F	p
Treatment	98.844	2.000	49.422	16.457	< .001
Residual	126.133	42.000	3.003		

Note. Type III Sum of Squares

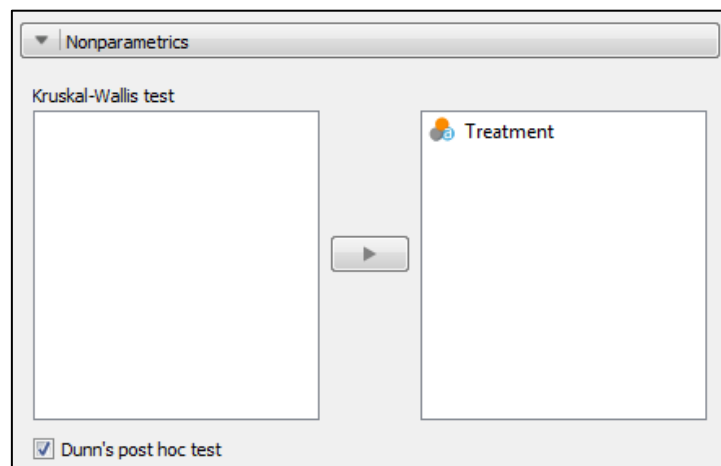
Test for Equality of Variances (Levene's)

F	df1	df2	p
3.832	2.000	42.000	0.030



Although the ANOVA indicates a significant result, the data has not met the assumptions of homogeneity of variance as seen by the significant Levene's test and only shows linearity in the middle of the Q-Q plot and curves off at the extremities indicating more extreme values. Added to the fact that the dependent variable is based on subjective pain scores suggest the use of a non-parametric alternative.

Return to the statistics options and open up the Nonparametrics option at the bottom. For the Kruskal-Wallis test Move the Treatment variable to the box on the right and tick Dunn's post hoc test.





## UNDERSTANDING THE OUTPUT

Two tables are shown in the output. The Kruskal-Wallis test shows that there is a significant difference between the three treatment modalities.

Kruskal-Wallis Test

Factor	Statistic	df	p
Treatment	19.693	2	< .001

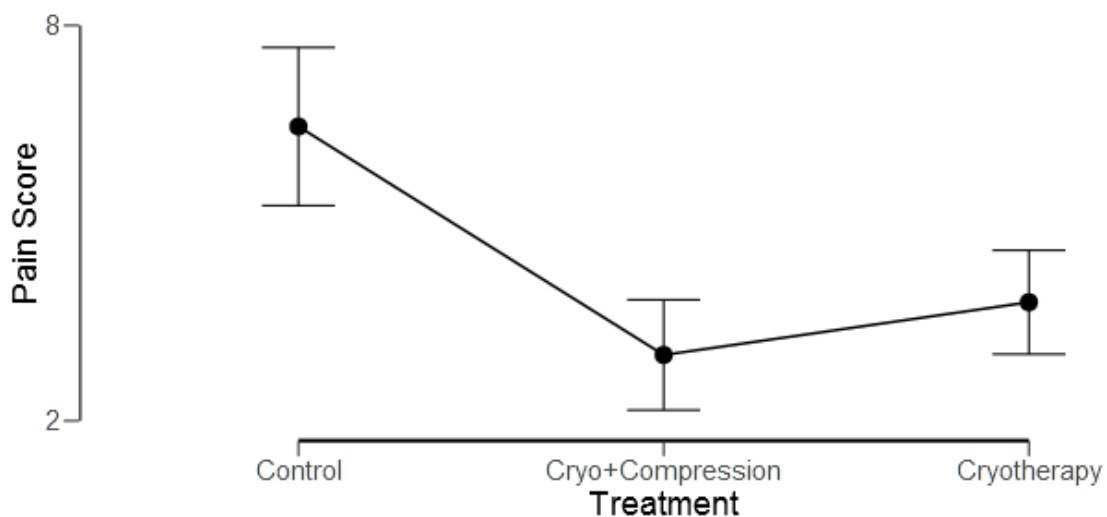
Dunn's Post Hoc Comparisons - Treatment

		z	W <sub>i</sub>	W <sub>j</sub>	p	P <sub>bonf</sub>	P <sub>holm</sub>
Control	Cryo+Compression	4.317	34.600	14.200	< .001	< .001	< .001
	Cryotherapy	3.048	34.600	20.200	0.001	0.003	0.002
Cryo+Compression	Cryotherapy	-1.270	14.200	20.200	0.102	0.306	0.102

The Dunn's post hoc test provides its own p value as well as those for Bonferroni and Holm's Bonferroni correction. As can be seen, both treatment conditions are significantly different from the controls but not from each other.

## REPORTING THE RESULTS

Pain scores were significantly affected by treatment modality  $H(2) = 19.693, p < .001$ . Pairwise comparisons showed that both cryotherapy and cryotherapy with compression significantly reduces pain scores ( $p = .001$  and  $p < .001$  respectively) compared to the control group. There were no significant differences between cryotherapy and cryotherapy with compression ( $p = .102$ ).





## COMPARING MORE THAN TWO RELATED GROUPS

### RMANOVA

The one-way repeated measures ANOVA (**RMANOVA**) is used to assess if there is a difference in means between 3 or more groups (where the participants are the same in each group) that have been tested multiple times or under different conditions. Such a research design, for example, could be that the same participants were tested for an outcome measure at 1, 2 and 3 weeks or that the outcome was tested under conditions 1, 2 and 3.

**The null hypothesis tested is that there is no significant difference between the means of the differences between all the groups.**

The independent variable should be categorical and the dependent variable needs to be a continuous measure. In this analysis the independent categories are termed **levels** i.e. these are the related groups. So in the case where an outcome was measured at weeks 1, 2 and 3, the 3 levels would be week 1, week 2 and week 3.

The **F-statistic** is calculated by dividing the mean squares for the variable (variance explained by the model) by its error mean squares (unexplained variance). The larger the F-statistic, the more likely it is that the independent variable will have had a significant effect on the dependent variable.

### ASSUMPTIONS

The RMANOVA makes the same assumptions as most other parametric tests.

- The dependent variable should be approximately normally distributed.
- There should be no significant outliers.
- No significant outliers
- Sphericity, which relates to the equality of the variances of the differences between levels of the repeated measures factor.

If the assumptions are violated then the non-parametric equivalent, **Friedman's test** should be considered instead and is described later in this section.

### SPHERICITY

If a study has 3 levels (A, B and C) sphericity assumes the following:

$$\text{Variance (A-B)} \approx \text{Variance (A-C)} \approx \text{Variance (B-C)}$$

RMANOVA checks the assumption of sphericity using Mauchly's (pronounced Mockley's) test of sphericity. This tests **the null hypothesis that the variances of the differences are equal**. In many cases, repeated measures violate the assumption of sphericity which can lead to Type I error. If this is the case corrections to the F-statistic can be applied.

JASP offers two methods of correcting the F-statistic, the **Greenhouse-Geisser** and **the Huynh-Feldt** epsilon ( $\epsilon$ ) corrections. A general rule of thumb is that if the  $\epsilon$  values are  $<0.75$  then use the Greenhouse-Geisser correction and if they are  $>0.75$  then use the Huynh-Feldt correction.



## POST HOC TESTING

Post hoc testing is limited in RMANOVA, JASP provides two alternatives:

**Bonferroni** – can be very conservative but gives guaranteed control over Type I error at the risk of reducing statistical power.

**Holm** – the Holm-Bonferroni test which is a sequential Bonferroni method that is less conservative than the original Bonferroni test.

If you ask for either Tukey or Scheffe post hoc corrections JASP will return a NaN (not a number) error.

## EFFECT SIZE

JASP provides the same alternative effect size calculations that are used with the independent group ANOVA tests:

**Eta squared ( $\eta^2$ )** - accurate for the sample variance explained but overestimates the population variance. This can make it difficult to compare the effect of a single variable in different studies.

**Partial Eta squared ( $\eta_p^2$ )** – this solves the problem relating to population variance overestimation allowing for comparison of the effect of the same variable in different studies. This appears to be the most commonly reported effect size in repeated measures ANOVA

**Omega squared ( $\omega^2$ )** – Normally, statistical bias gets very small as sample size increases, but for small samples ( $n < 30$ )  $\omega^2$  provides an unbiased effect size measure.

Levels of effect size:

Test	Measure	Trivial	Small	Medium	Large
ANOVA	Eta	<0.1	0.1	0.25	0.37
	Partial Eta	<0.01	0.01	0.06	0.14
	Omega squared	<0.01	0.01	0.06	0.14

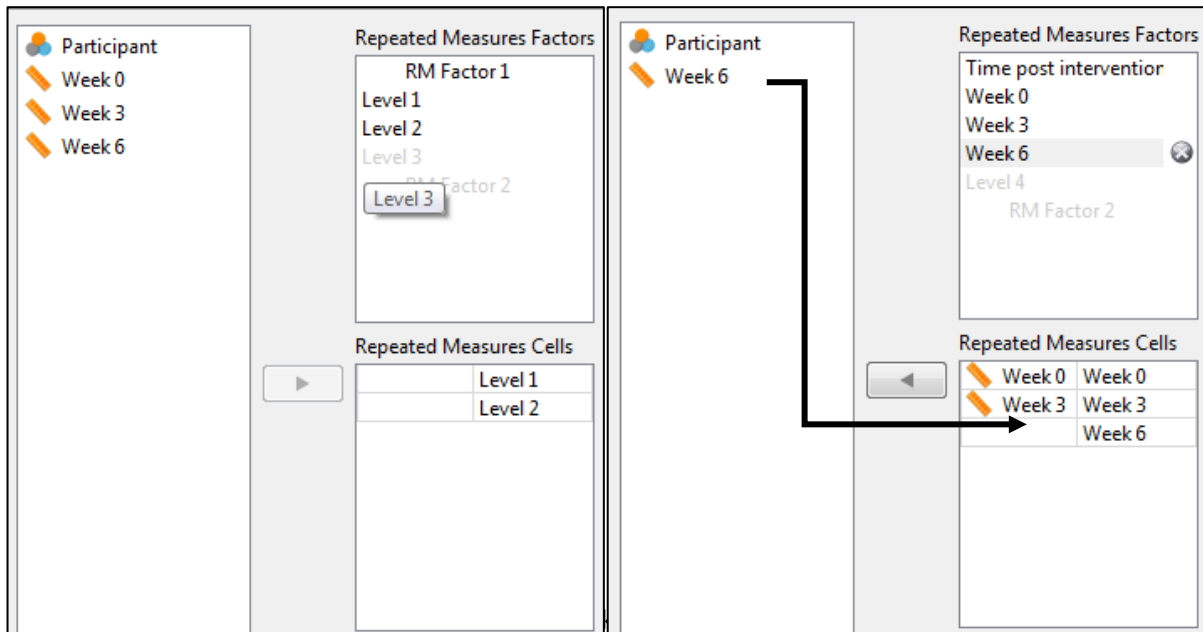
## RUNNING THE REPEATED MEASURES ANOVA

Load **Repeated ANOVA cholesterol.csv**. This contains one column with the participant IDs and 3 columns one for each repeated measurement of blood cholesterol following an intervention. For good practice check the descriptive statistics and the boxplots for any extreme outliers.

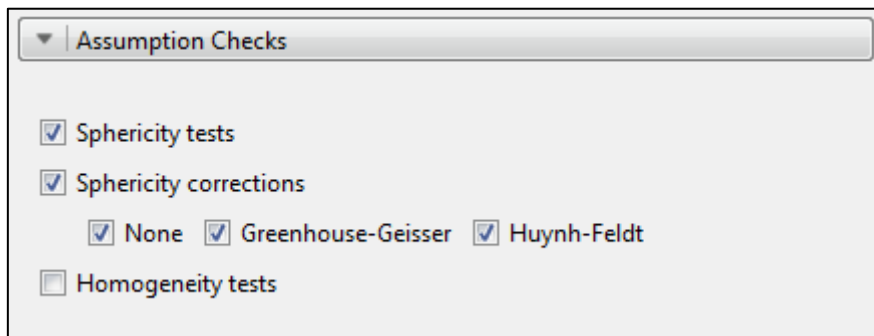
Go to ANOVA > Repeated measures ANOVA. As stated above, the independent variable (repeated measures factor) has levels, in this case, there are 3 levels. Rename RM Factor 1 to Time post intervention and then rename 3 levels to Week 0, week 3 and week 6 accordingly.



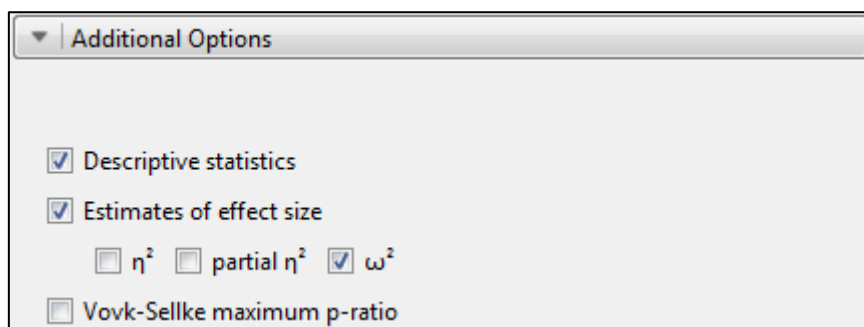
Once these have been done they will appear in the Repeated Measures Cells. Now add the appropriate data to the appropriate level.



Under Assumption Checks tick Sphericity tests and all Sphericity correction options.



Under Additional Options tick Descriptive Statistics, Estimates of effect size and  $\omega^2$ .







The output should consist of 4 tables. The third table, between subject effects, can be ignored for this analysis.

## UNDERSTANDING THE OUTPUT

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	$\omega^2$
Time post intervention	None	4.320*	2.000*	2.160*	212.321*	< .001*	0.058
	Greenhouse-Geisser	4.320*	1.235*	3.497*	212.321*	< .001*	0.058
	Huynh-Feldt	4.320*	1.284*	3.365*	212.321*	< .001*	0.058
Residual	None	0.346	34.000	0.010			
	Greenhouse-Geisser	0.346	21.001	0.016			
	Huynh-Feldt	0.346	21.822	0.016			

Note. Type III Sum of Squares

\* Mauchly's test of sphericity indicates that the assumption of sphericity is violated ( $p < .05$ ).

The within subjects effects table reports a large F-statistic which is highly significant ( $p < .001$ ) and has a small to medium effect size (0.058). This table shows the statistics for sphericity assumed (none) and the two correction methods. The main differences are in the degrees of freedom (df) and the mean squares value. Under the table, it is noted that the assumption of sphericity has been violated.

The following table gives the results of Mauchly's test of sphericity. It can be seen that there is a significant difference ( $p < .001$ ) in the variances of the differences between the groups. Greenhouse-Geisser and the Huynh-Feldt epsilon ( $\epsilon$ ) values are below 0.75. Therefore the ANOVA result should be reported based on the Greenhouse-Geisser correction:

Test of Sphericity

	Mauchly's W	p	Greenhouse-Geisser $\epsilon$	Huynh-Feldt $\epsilon$
Time post intervention	0.381	< .001	0.618	0.642

To provide a cleaner table, go back to Assumption Checks and only tick Greenhouse-Geisser for sphericity correction.

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	$\omega^2$
Time post intervention	Greenhouse-Geisser	4.320*	1.235*	3.497*	212.321*	< .001*	0.058
Residual	Greenhouse-Geisser	0.346	21.001	0.016			

Note. Type III Sum of Squares

\* Mauchly's test of sphericity indicates that the assumption of sphericity is violated ( $p < .05$ ).

There is a significant difference between the means of the differences between all the groups **F (1.235, 21.0) = 212.3,  $p < .001$ ,  $\omega^2 = 0.058$ .**



### Descriptives

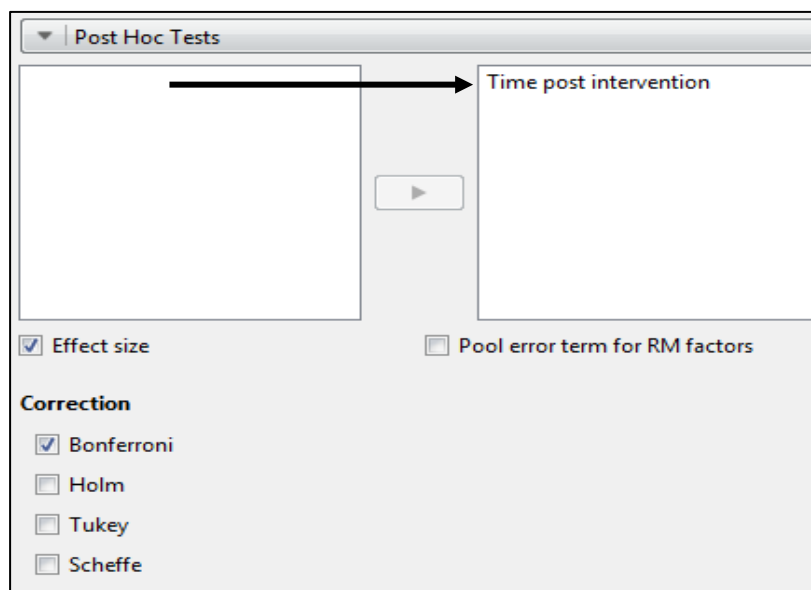
Time post intervention	Mean	SD	N
Week 0	6.408	1.191	18.000
Week 3	5.842	1.123	18.000
Week 6	5.779	1.102	18.000

The descriptive data suggest that blood cholesterol levels were higher at week 0 compared to weeks 3 and 6.

**However, if the ANOVA reports no significant difference you can go no further in the analysis.**

### POST HOC TESTING

If the ANOVA is significant, post hoc testing can now be carried out. In Post Hoc Tests add Time post-intervention to the analysis box on the right, tick Effect size and, in this case, use Bonferroni for the post hoc correction.



Also in Descriptive Plots add the Factor – Time post-intervention to the horizontal axis and tick display error bars.



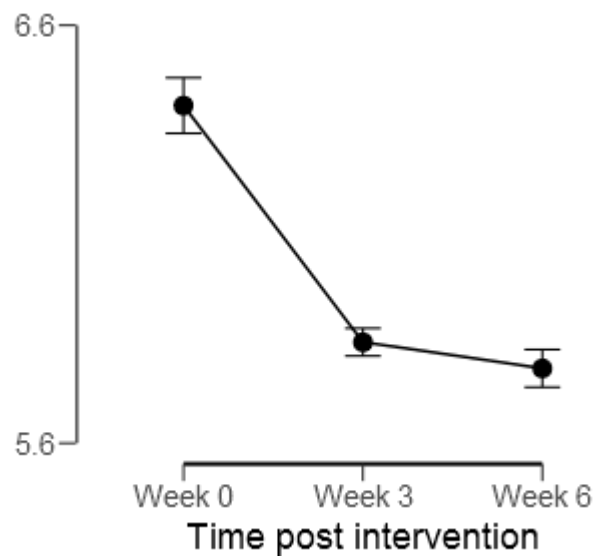
### Post Hoc Comparisons - Time post intervention

		Mean Difference	SE	t	Cohen's d	$P_{\text{bonf}}$
Week 0	Week 3	0.566	0.037	15.439	3.639	< .001
	Week 6	0.629	0.042	14.946	3.523	< .001
Week 3	Week 6	0.063	0.017	3.781	0.891	0.004

Note. Cohen's d does not correct for multiple comparisons.

Post hoc testing shows that there are significant differences in blood cholesterol levels between all of the time point combinations and are associated with large effect sizes.

### REPORTING THE RESULTS



Since Mauchly's test of sphericity was significant, the Greenhouse-Geisser correction was used. This showed that cholesterol levels differed significantly between  $F(1.235, 21.0) = 212.3, p < .001, \omega^2 = 0.058$ .

Post hoc testing using the Bonferroni correction revealed that cholesterol levels decreased significantly as time increased, weeks 0 – 3 (mean difference = 0.566 units,  $p < .001$ ) and weeks 3 – 6 (mean difference = 0.063 units,  $p = .004$ ).



## FRIEDMAN'S REPEATED MEASURES ANOVA

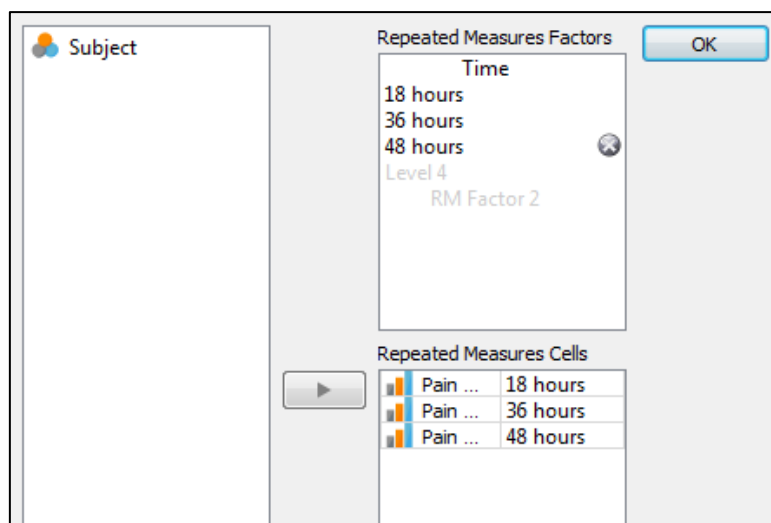
If parametric assumptions are violated or the data is ordinal in nature you should consider using the non-parametric alternative, Friedman's test. Similar to the Kruskal-Wallis test, the Friedman's test is used for one-way repeated measures analysis of variance by ranks and doesn't assume the data comes from a particular distribution. This test is another omnibus test which does not specify which specific groups of the independent variable are statistically significantly different from each other. To do this, JASP provides the option for running Conover's post hoc test if the Friedman's test is significant.

Load **Friedman RMANOVA.csv** into JASP. This has 3 columns of subjective pain ratings measured at 18, 36 and 48 hours post-exercise. Check that the pain scores are set to ordinal data.

## RUNNING THE FRIEDMAN'S TEST

Go to ANOVA > Repeated measures ANOVA. The independent variable (repeated measures factor) has 3 levels. Rename RM Factor 1 to Time and then rename 3 levels to 18 hours, 36 hours and w48 hours accordingly.

Once these have been done they will appear in the Repeated Measures Cells. Now add the appropriate dataset to the appropriate level.



This will automatically produce the standard repeated measures within subjects ANOVA table. To run the Friedman's test, expand the Nonparametrics tab, move Time to the RM factor box and tick Conover's post hoc tests.



## UNDERSTANDING THE OUTPUT

Two tables should be produced.

Friedman Test

Factor	Chi-Squared	df	p	Kendall's W
Time	26.772	2	< .001	0.764

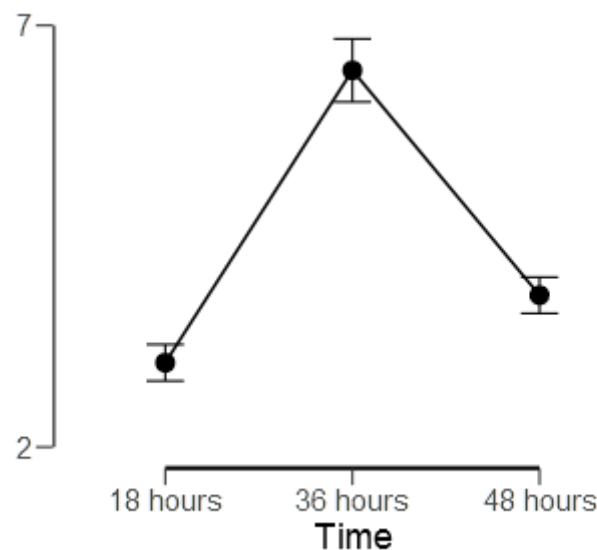
Conover's Post Hoc Comparisons - Time

	T-Stat	df	W <sub>i</sub>	W <sub>j</sub>	p	p <sub>bonf</sub>	p <sub>holm</sub>
18 hours 36 hours	15.171	28	17.000	44.500	< .001	< .001	< .001
48 hours	6.344	28	17.000	28.500	< .001	< .001	< .001
36 hours 48 hours	8.827	28	44.500	28.500	< .001	< .001	< .001

Friedman's test shows that time has a significant effect on pain perception. Connor's post hoc pairwise comparisons show that all pain perception is significantly different between each time point.

## REPORTING THE RESULTS

Time has a significant effect on subjective pain scores  $\chi^2(2) = 26.77, p < .001$ . Pairwise comparisons showed that pain perception is significantly different between each time point (all  $p < 0.001$ ).





## TWO-WAY INDEPENDENT ANOVA

One-way ANOVA tests situations when only one independent variable is manipulated, two-way ANOVA is used when more than 1 independent variable has been manipulated. In this case, independent variables are known as factors.

FACTOR 1	FACTOR 2	
CONDITION 1	Group 1	Dependent variable
	Group 2	Dependent variable
CONDITION 2	Group 1	Dependent variable
	Group 2	Dependent variable
CONDITION 3	Group 1	Dependent variable
	Group 2	Dependent variable

The factors are split into levels, therefore, in this case, Factor 1 has 3 levels and Factor 2 has 2 levels.

A “main effect” is the effect of one of the independent variables on the dependent variable, ignoring the effects of any other independent variables. There are 2 main effects tested both of which are “**between-subjects**”: in this case comparing differences between factor 1 (i.e. condition) and differences between factor 2 (i.e. groups). An **interaction** is where one factor influences the other factor.

The two-way independent ANOVA is another omnibus test that is used to test 2 null hypotheses:

- 1. There is no significant between-subject effect i.e. no significant difference between the means of the groups in either of the factors.**
- 2. There is no significant interaction effect i.e. no significant group differences across conditions.**

## ASSUMPTIONS

Like all other parametric tests, mixed factor ANOVA makes a series of assumptions which should either be addressed in the research design or can be tested for.

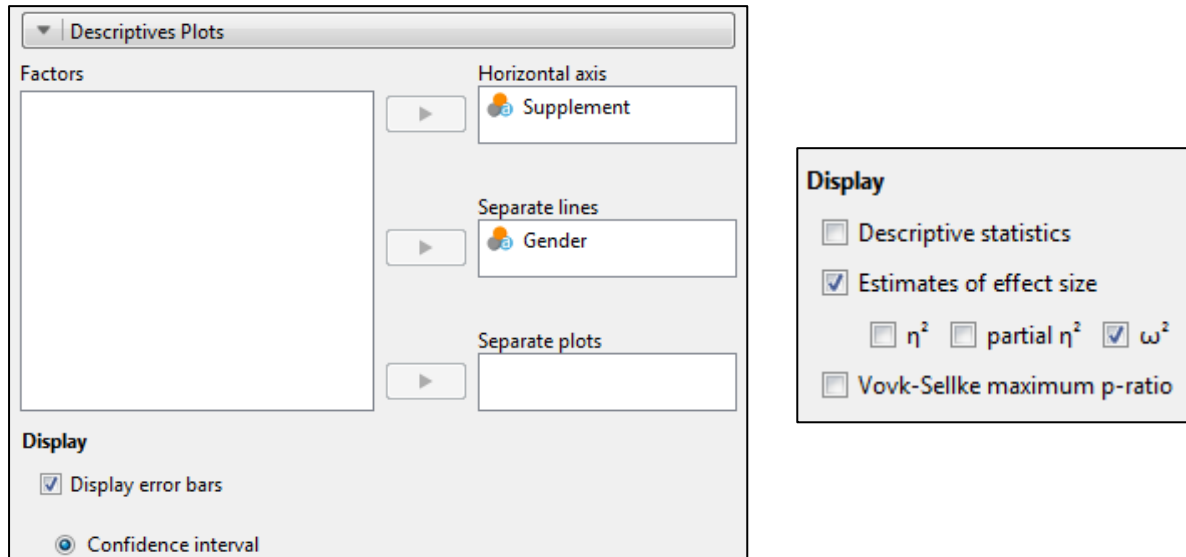
- The independent variables (factors) should have at least two categorical independent groups (levels).
- The dependent variable should be continuous and approximately normally distributed for all combinations of factors.
- There should be homogeneity of variance for each of the combination of factors.
- There should be no significant outliers.

## RUNNING TWO-WAY INDEPENDENT ANOVA

Open **2-way independent ANOVA.csv** in JASP. This comprises on 3 columns of data, Factor 1 – gender with 2 levels (male and female), Factor 2 - supplement with 3 levels (control, carbohydrate CHO and protein) and the dependent variable (explosive jump power. In Descriptive statistics check the data for significant outliers. Go to ANOVA >ANOVA, add Jump power to the Dependent variable, Gender and Supplement to the Fixed factors.



In Descriptive plots add supplement to the horizontal axis and Gender to separate lines. In Additional Options, tick Descriptive statistics and Estimates of effect size ( $\omega^2$ ).



## UNDERSTANDING THE OUTPUT

The output should comprise 2 tables and one plot.

ANOVA - Jump power

Cases	Sum of Squares	df	Mean Square	F	p	$\omega^2$
Gender	119108.037	1.000	119108.037	9.589	0.003	0.058
Supplement	896116.137	2.000	448058.068	36.071	< .001	0.477
Gender * Supplement	275806.438	2.000	137903.219	11.102	< .001	0.138
Residual	521712.054	42.000	12421.716			

Note. Type III Sum of Squares

The ANOVA table shows that there are significant main effects for both Gender and Supplement ( $p=0.003$  and  $p<.001$  respectively) with medium and large effect sizes respectively. This suggests that there is a significant difference in jump power between genders, irrespective of Supplement, and significant differences between supplements, irrespective of Gender.

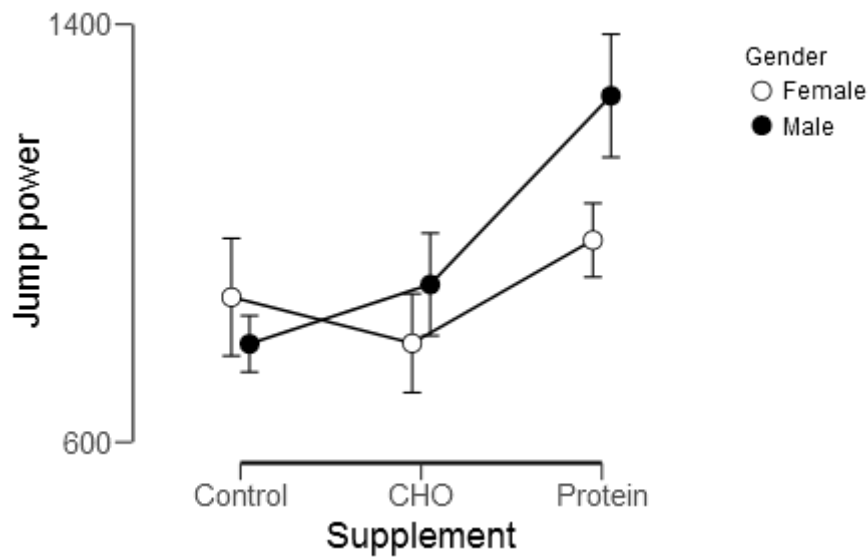
There is also a significant interaction between Gender and Supplement ( $p<.001$ ) which also has a medium to large effect size (0.138). This suggests that the differences in jump power between genders is affected somehow by the type of supplement used.

The Descriptive statistics and plot suggest that the main differences are between genders when using a protein supplement.



### Descriptives - Jump power ▼

Gender	Supplement	Mean	SD	N
Female	Control	877.500	134.563	8.000
	CHO	789.286	102.283	7.000
	Protein	986.667	91.924	9.000
Male	Control	788.125	64.417	8.000
	CHO	901.875	117.502	8.000
	Protein	1263.125	140.863	8.000



### TESTING ASSUMPTIONS

In Assumption Checks, tick Homogeneity tests and Q-Q plot of residuals.

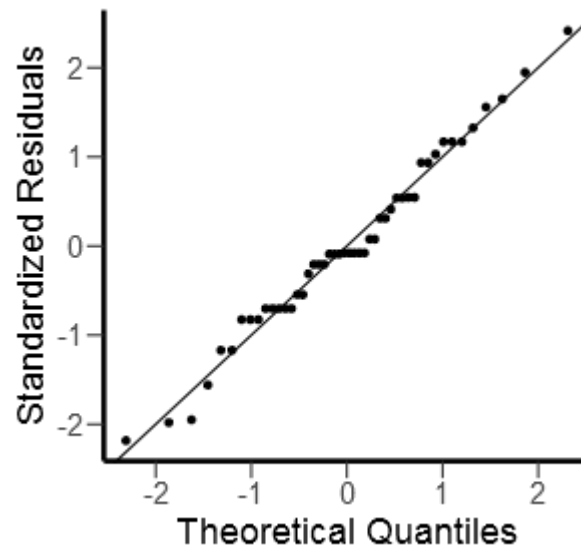
#### Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.100	5.000	42.000	0.375

Levene's test shows no significant difference in variance within the dependent variable groups, thus homogeneity of variance has not been violated.





The Q-Q plot shows that the data appear to be normally distributed and linear. We can now accept the ANOVA result since none of these assumptions have been violated.

**However, if the ANOVA reports no significant difference you can go no further with the analysis.**

### POST HOC TESTING

If the ANOVA is significant post hoc testing can now be carried out. In Post Hoc Tests add Supplement to the analysis box on the right, tick Effect size and, in this case, use Tukey for the post hoc correction.

Post hoc testing is not done for Gender since there are only 2 levels.

Post Hoc Comparisons - Supplement

		Mean Difference	SE	t	Cohen's d	P <sub>Tukey</sub>
Control	CHO	-12.768	40.102	-0.318	-0.109	0.946
	Protein	-292.083	38.853	-7.518	-1.919	< .001
CHO	Protein	-279.315	39.561	-7.060	-1.782	< .001

Note. Cohen's d does not correct for multiple comparisons.

Post hoc testing shows no significant difference between the control and CHO, supplement group, irrespective of Gender, but significant differences between Control and Protein ( $p < .001$ ) and between CHO and Protein ( $p < .001$ ).

Now go to the analysis options and Simple Main Effects. Here add Gender to the Simple effect factor and Supplement to the Moderator Factor 1. Simple main effects are effectively pairwise comparisons.



Simple Main Effects - Gender

Level of Supplement	Sum of Squares	df	Mean Square	F	p
Control	31951.563	1	31951.563	2.572	0.116
CHO	47325.030	1	47325.030	3.810	0.058
Protein	323700.184	1	323700.184	26.059	< .001

This table shows that there are no gender differences in jump power between the control or CHO groups ( $p=.116$  and  $p=0.058$  respectively). However, there is a significant difference ( $p<.001$ ) in jump power between genders in the protein supplement group.

## REPORTING THE RESULTS

A two-way ANOVA was used to examine the effect of gender and supplement type on explosive jump power. There were significant main effects for both gender ( $F(1, 42) = 9.59, p=.003, \omega^2 = 0.058$ ) and Supplement ( $F(2, 42) = 30.07, p<.001, \omega^2 = 0.477$ ). There was a statistically significant interaction between the effects of gender and supplement on explosive jump power ( $F(2, 42) = 11.1, p<.001, \omega^2 = 0.138$ ).

Tukey's post hoc correction showed that explosive leg power was significantly higher in the protein group compared to the control or CHO groups ( $t=-1.919, p<.001$  and  $t=-1.782, p<.001$  respectively).

Simple main effects showed that jump power was significantly higher in males on a protein supplement compared to females ( $F(1) = 28.06, p<.001$ ).



## MIXED FACTOR ANOVA USING JASP

Mixed factor ANOVA (another two-way ANOVA) is a combination of both independent and repeated measures ANOVA involving more than 1 independent variable (known as factors).

*Independent variable*    *Independent variable (Factor 1) = time or condition*

*(Factor 2)*

*Group 1*

*Group 2*

Time/condition 1	Time/condition 2	Time/condition 3
Dependent variable	Dependent variable	Dependent variable
Dependent variable	Dependent variable	Dependent variable

The factors are split into levels, therefore, in this case, Factor 1 has 3 levels and Factor 2 has 2 levels. This results in 6 possible combinations.

A “main effect” is the effect of one of the independent variables on the dependent variable, ignoring the effects of any other independent variables. There are 2 main effects tested: in this case comparing data across factor 1 (i.e. time) is known as the “**within-subjects**” factor while comparing differences between factor 2 (i.e. groups) is known as the “**between-subjects**” factor. An **interaction** is where one factor influences the other factor.

The main effect of time or condition tests the following i.e. irrespective of which group the data is in:

*Independent variable*    *Independent variable (Factor 1) = time or condition*

*(Factor 2)*

*Group 1*

*Group 2*

Time/condition 1	Time/condition 2	Time/condition 3
All data	All data	All data
└─── * ───┘		└─── * ───┘
└────────── * ─────────┘		

The main effect of group tests the following i.e. irrespective of which condition the data is in:

*Independent variable*    *Independent variable (Factor 1) = time or condition*

*(Factor 2)*

*Group 1*

*Group 2*

Time/condition 1	Time/condition 2	Time/condition 3
All data		└─── * ───┘
All data		└─── * ───┘

Simple main effects are effectively pairwise comparisons:

*Independent variable*    *Independent variable (Factor 1) = time or condition*

*(Factor 2)*

*Group 1*

*Group 2*

Time/condition 1	Time/condition 2	Time/condition 3
Data	└─── * ───┘	Data
Data	└─── * ───┘	Data
Data	└─── * ───┘	Data



A mixed factor ANOVA is another omnibus test that is used to test 3 null hypotheses:

3. *There is no significant within-subject effect i.e. no significant difference between the means of the differences between all the conditions/times.*
4. *There is no significant between-subject effect i.e. no significant difference between the means of the groups.*
5. *There is no significant interaction effect i.e. no significant group differences across conditions/time*

## ASSUMPTIONS

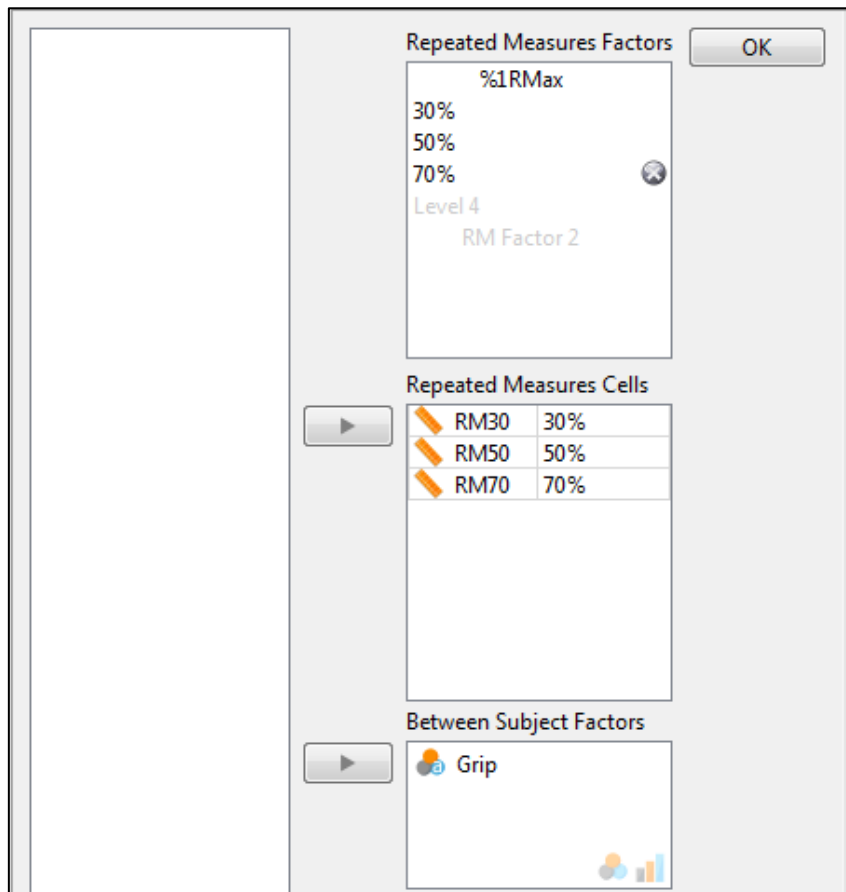
Like all other parametric tests, mixed factor ANOVA makes a series of assumptions which should either be addressed in the research design or can be tested for.

- The “**within-subjects**” factor should contain at least two related (repeated measures) categorical groups (levels)
- The “**between-subjects**” factor should have at least two categorical independent groups (levels).
- The dependent variable should be continuous and approximately normally distributed for all combinations of factors.
- There should be homogeneity of variance for each of the groups and, if more than 2 levels) sphericity between the related groups.
- There should be no significant outliers.

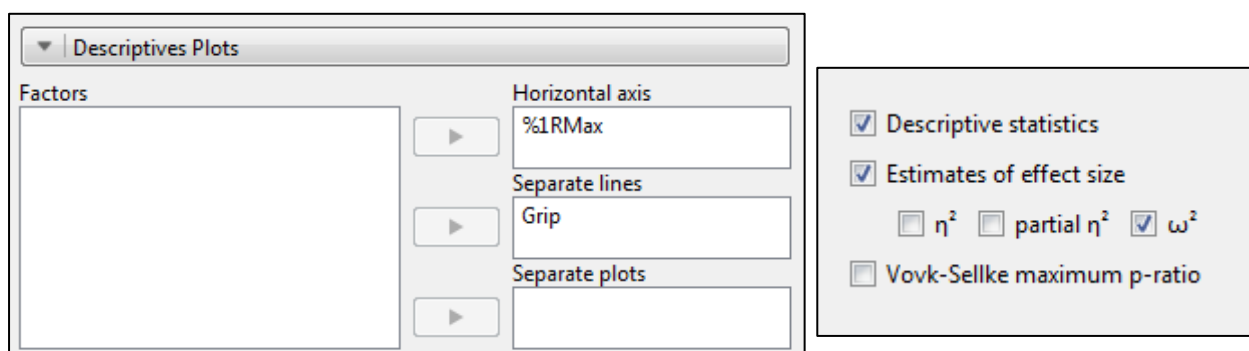
## RUNNING THE MIXED FACTOR ANOVA

Open **2-way Mixed ANOVA.csv** in JASP. This contains 4 columns of data relating to the type of weightlifting grip and speed of the lift at 3 different loads (%1RM). Column 1 contains the grip type, columns 2-4 contain the 3 repeated measures (30, 50 and 70%). Check for significant outliers using boxplots then go to ANOVA > Repeated measures ANOVA.

Define the Repeated Measures Factor, %1RM, and add 3 levels (30, 50 and 70%). Add the appropriate variable to the Repeated measures Cells and add Grip to the Between-Subjects Factors:



In Descriptive plots, move %1RM to the horizontal axis and Grip to separate lines. In Additional Options, tick Descriptive statistics and Estimates of effect size ( $\omega^2$ ).





## UNDERSTANDING THE OUTPUT

### Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	$\omega^2$
%1RM	5.605 <sup>a</sup>	2 <sup>a</sup>	2.803 <sup>a</sup>	115.450 <sup>a</sup>	< .001 <sup>a</sup>	0.744
%1RM * Grip	0.583 <sup>a</sup>	2 <sup>a</sup>	0.291 <sup>a</sup>	12.003 <sup>a</sup>	< .001 <sup>a</sup>	0.218
Residual	0.874	36	0.024			

Note. Type III Sum of Squares

<sup>a</sup> Mauchly's test of sphericity indicates that the assumption of sphericity is violated ( $p < .05$ ).

The output should initially comprise of 3 tables and 1 graph.

For the main effect with respect to %1RM, the within-subjects effects table reports a large F-statistic which is highly significant ( $p < .001$ ) and has a large effect size (0.744). Therefore, irrespective of grip type, there is a significant difference between the three %1RM loads.

However, JASP has reported under the table that the assumption of sphericity has been violated. This will be addressed in the next section.

### Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	$\omega^2$
Grip	1.095	1	1.095	20.925	< .001	0.499
Residual	0.942	18	0.052			

Note. Type III Sum of Squares

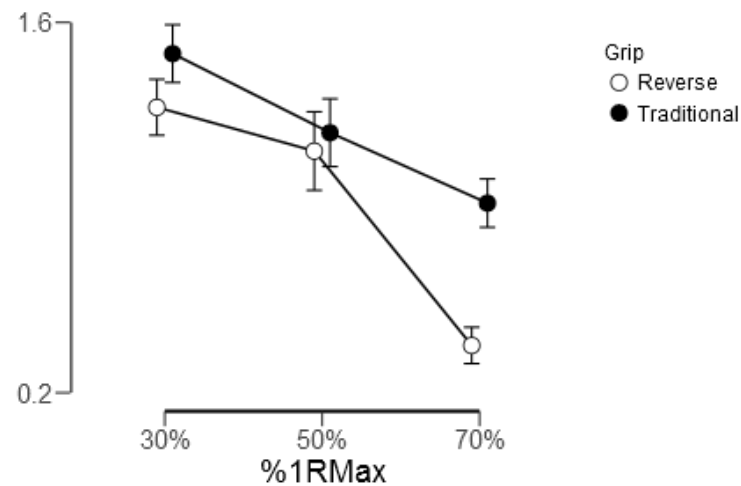
Finally, there is a significant interaction between %1RM and grip ( $p < .001$ ) which also has a large effect size (0.499). This suggests that the differences between the %1RM loads are affected somehow by the type of grip used.

For the main effect with respect to grip, the between-subjects table shows a significant difference between grips ( $p < .001$ ), irrespective of %1RM.

From the descriptive data and the plot, it appears that there is a larger difference between the two grips at the high 70% RM load.

### Descriptives

%1RMax	Grip	Mean	SD	N
30%	Reverse	1.279	0.178	10.000
	Traditional	1.482	0.217	10.000
50%	Reverse	1.114	0.198	10.000
	Traditional	1.183	0.256	10.000
70%	Reverse	0.379	0.105	10.000
	Traditional	0.917	0.086	10.000



### TESTING ASSUMPTIONS

In Assumptions Checks, tick Sphericity tests, Sphericity corrections and Homogeneity tests.

#### Test of Sphericity

	Mauchly's W	p	Greenhouse-Geisser $\epsilon$	Huynh-Feldt $\epsilon$
%1RMax	0.649	0.025	0.740	0.791

Mauchly's test of sphericity is significant so that assumption has been violated, therefore, the Greenhouse-Geisser correction should be used since epsilon is  $< 0.75$ . Go back to Assumption Checks and in Sphericity corrections leave Greenhouse-Geisser only ticked. This will result in an updated Within-Subjects Effects table:

#### Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	$\omega^2$
%1RM	Greenhouse-Geisser	5.605 <sup>a</sup>	1.480 <sup>a</sup>	3.787 <sup>a</sup>	115.450 <sup>a</sup>	$< .001^a$	0.744
%1RM * Grip	Greenhouse-Geisser	0.583 <sup>a</sup>	1.480 <sup>a</sup>	0.394 <sup>a</sup>	12.003 <sup>a</sup>	$< .001^a$	0.218
Residual	Greenhouse-Geisser	0.874	26.639	0.033			

Note. Type III Sum of Squares

<sup>a</sup> Mauchly's test of sphericity indicates that the assumption of sphericity is violated ( $p < .05$ ).

#### Test for Equality of Variances (Levene's)

	F	df1	df2	p
RM30	0.523	1.000	18.000	0.479
RM50	0.346	1.000	18.000	0.564
RM70	0.183	1.000	18.000	0.674

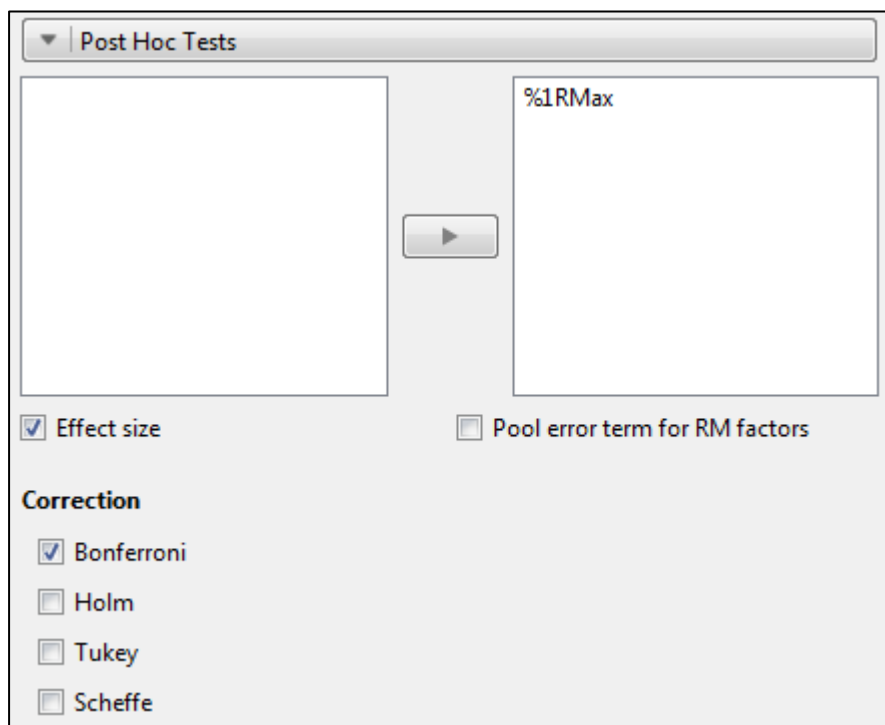
Levene's test shows that there is no difference in variance in the dependent variable between the two grip types.



**However, if the ANOVA reports no significant difference you can go no further in the analysis.**

## POST HOC TESTING

If the ANOVA is significant post hoc testing can now be carried out. In Post Hoc Tests add %1RM to the analysis box on the right, tick Effect size and, in this case, use Bonferroni for the post hoc correction. Only Bonferroni or Holm's correction are available for repeated measures.



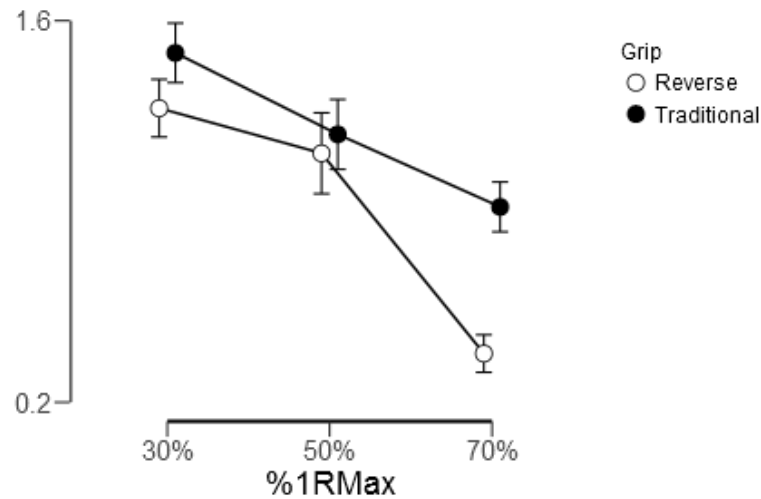
Post Hoc Comparisons - %1RMax

		Mean Difference	SE	t	Cohen's d	P <sub>bonf</sub>
30%	50%	0.232	0.060	3.856	0.862	0.003
	70%	0.733	0.050	14.583	3.261	< .001
50%	70%	0.500	0.073	6.839	1.529	< .001

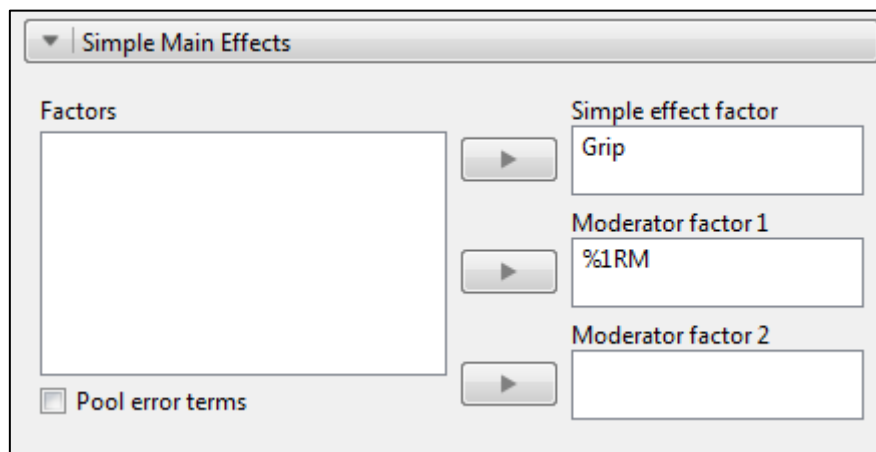
Note. Cohen's d does not correct for multiple comparisons.

The post hoc tests show that irrespective of grip type each load is significantly different from each of the other loads, and as seen from the plot, lift velocity significantly decreases as load increases.





Finally, In Simple main effects add Grip to the Simple effect factor and %1RM to Moderator factor 1



Simple Main Effects - Grip

Level of %1RM	Sum of Squares	df	Mean Square	F	p
30%	0.206	1	0.206	5.229	0.035
50%	0.024	1	0.024	0.461	0.506
70%	1.447	1	1.447	157.212	< .001

These results show that there is a significant difference in lift speed between the two grips at 30% 1RM and also at the higher 70% 1RM loads ( $p=0.035$  and  $p<0.001$  respectively).



## REPORTING THE RESULTS

Using the Greenhouse-Geisser correction, there was a significant main effect of load ( $F(1.48, 26.64) = 115.45, p < .001$ ). Bonferroni corrected post hoc testing showed that there was a significant sequential decline in lift speed from 30-50% 1RM ( $p = .035$ ) and 50-70% 1RM ( $p < .001$ ).

There was a significant main effect for grip type ( $F(1, 18) = 20.925, p < .001$ ) showing an overall higher lift speed using the traditional rather than the reverse grip.

Using the Greenhouse-Geisser correction, there was a significant %1RM x Grip interaction ( $F(1.48, 26.64) = 12.00, p < .001$ ) showing that the type of grip affected lift velocity over the %1RM loads.



## CHI-SQUARE TEST FOR ASSOCIATION

The chi-square ( $\chi^2$ ) test for independence (also known as Pearson's  $\chi^2$  test or the  $\chi^2$  test of association) can be used to determine if a relationship exists between two or more categorical variables. The test produces a contingency table, or cross-tabulation, which displays the cross-grouping of the categorical variables.

The  $\chi^2$  test checks the null hypothesis that there is no association between two categorical variables. It compares the observed frequencies of the data with frequencies which would be expected if there was no association between the two variables.

The analysis requires two assumptions to be met:

1. The two variables must be categorical data (nominal or ordinal)
2. Each variable should comprise two or more independent categorical groups

Most statistical tests fit a model to the observed data with a null hypothesis that there is no difference between the observed and modelled (expected) data. The error or deviation of the model is calculated as:

$$\text{Deviation} = \sum (\text{observed} - \text{model})^2$$

Most parametric models are based around population means and standard deviations. The  $\chi^2$  model, however, is based on expected frequencies.

How are the expected frequencies calculated? For example, we categorised 100 people into male, female, short and tall. If there was an equal distribution between the 4 categories expected frequency =  $100/4$  or 25% but the actual observed data does not have an equal frequency distribution.

<b>Equal Distribution</b>	Male	Female	Row Total
Tall	25	25	50
Short	25	25	50
Column Total	50	50	

<b>Observed Distribution</b>	Male	Female	Row Total
Tall	57	24	<b>81</b>
Short	14	5	<b>19</b>
Column Total	<b>71</b>	<b>29</b>	

The model based on expected values can be calculated by:

**Model (expected) = (row total x column total)/100**

- Model – tall male =  $(81 \times 71) / 100 = 57.5$
- Model – tall female =  $(81 \times 29) / 100 = 23.5$
- Model – small male =  $(19 \times 71) / 100 = 13.5$
- Model – small female =  $(19 \times 29) / 100 = 5.5$

These values can then be added to the contingency table:



	Male (M)	Female (F)	Row Total
Tall (T)	57	24	81
<b>Expected</b>	<b>57.5</b>	<b>23.5</b>	
Short (S)	14	5	19
<b>Expected</b>	<b>13.5</b>	<b>5.5</b>	
Column Total	71	29	

$$\chi^2 \text{ statistic is derived from } \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

### Validity

$\chi^2$  tests are only valid when you have a reasonable sample size, that is, less than 20% of cells have an expected count of less than 5 and none have an expected count of less than 1.

### RUNNING THE ANALYSIS

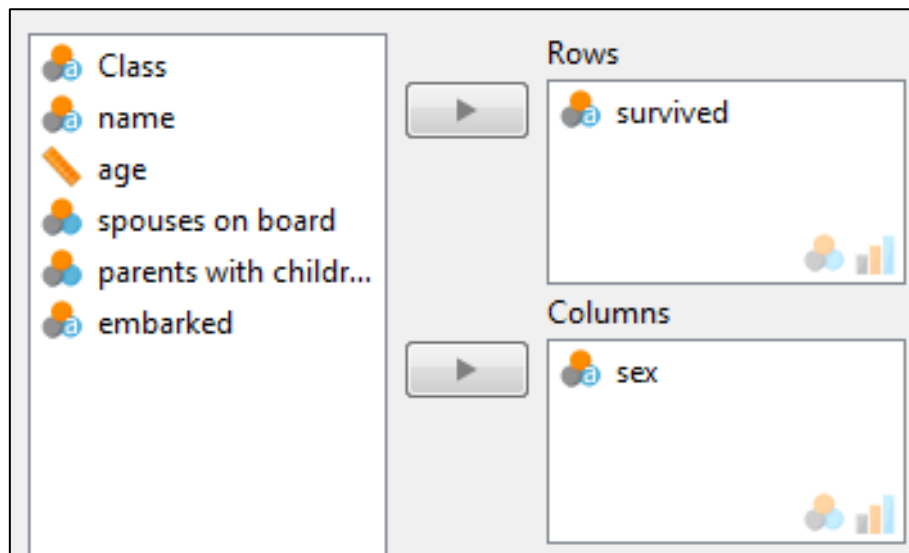
The dataset **Titanic survival** is a classic dataset used for machine learning and contains data on 1309 passengers and crew who were on board the Titanic when it sank in 1912. We can use this to look at associations between survival and other factors. The dependent variable is 'Survival' and possible independent values are all the other variables.

Class	survived	name	sex	age
Third	No	Abbing, Mr. Anthony	male	42
Third	No	Abbott, Master. Eugene Joseph	male	13
Third	No	Abbott, Mr. Rossmore Edward	male	16
Third	Yes	Abbott, Mrs. Stanton (Rosa Hunt)	female	35
Third	Yes	Abelseth, Miss. Karen Marie	female	16
Third	Yes	Abelseth, Mr. Olaus Jorgensen	male	25
Second	No	Abelson, Mr. Samuel	male	30
Second	Yes	Abelson, Mrs. Samuel (Hannah Wizosky)	female	28
Third	Yes	Abrahamsson, Mr. Abraham August Johannes	male	20
Third	Yes	Abraham, Mrs. Joseph (Sophie Halaut Easu)	female	18
Third	No	Adahl, Mr. Mauritz Nils Martin	male	30
Third	No	Adams, Mr. John	male	26

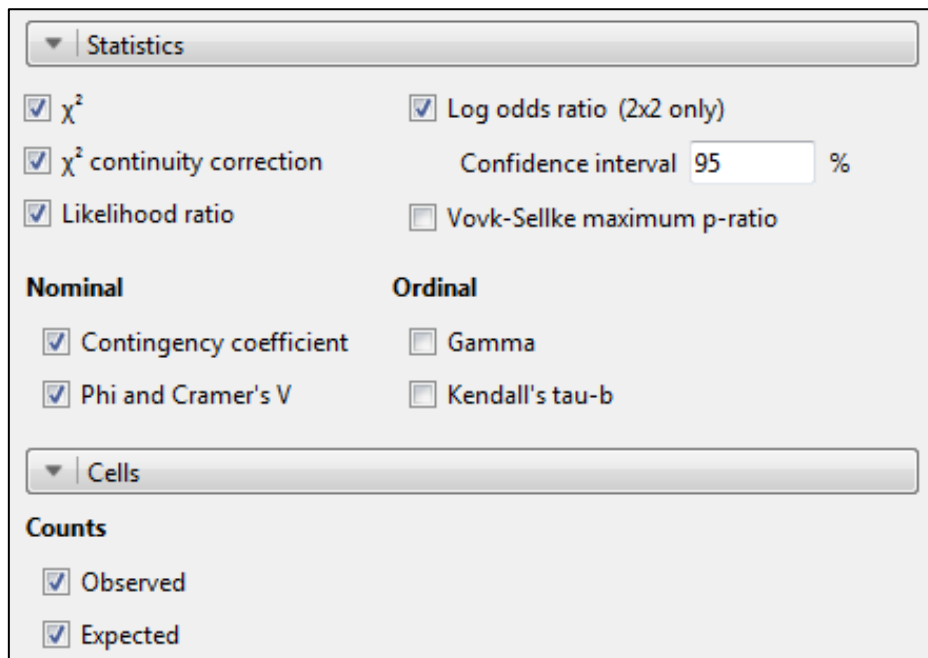


By convention, the independent variable is usually placed in the contingency table columns and the dependent variable is placed in the rows.

Open **Titanic survival.csv** in JASP, add survived to rows as the dependent variable and sex into columns as the independent variable.



Then tick all the following options:





## UNDERSTANDING THE OUTPUT

First look at the Contingency table output.

Contingency Tables

survived		sex		Total
		female	male	
No	Count	127.0	682.0	809.0
	Expected count	288.0	521.0	809.0
	% within row	15.7 %	84.3 %	100.0 %
	% within column	27.3 %	80.9 %	61.8 %
	% of Total	9.7 %	52.1 %	61.8 %
Yes	Count	339.0	161.0	500.0
	Expected count	178.0	322.0	500.0
	% within row	67.8 %	32.2 %	100.0 %
	% within column	72.7 %	19.1 %	38.2 %
	% of Total	25.9 %	12.3 %	38.2 %
Total	Count	466.0	843.0	1309.0
	Expected count	466.0	843.0	1309.0
	% within row	35.6 %	64.4 %	100.0 %
	% within column	100.0 %	100.0 %	100.0 %
	% of Total	35.6 %	64.4 %	100.0 %

**Remember that  $\chi^2$  tests are only valid when you have a reasonable sample size, i.e. less than 20% of cells have an expected count of less than 5 and none have an expected count of less than 1.**

From this table, looking at % within rows, it can be seen that more males died on the Titanic compared to females and more females survived compared to males. But is there a significant association between gender and survival?

The statistical results are shown below:

Chi-Squared Tests

	Value	df	p
$X^2$	365.9	1	< .001
$X^2$ continuity correction	363.6	1	< .001
Likelihood ratio	372.9	1	< .001
N	1309		

$\chi^2$  statistic ( $\chi^2(1) = 365.9, p < .001$ ) suggest that there is a significant association between gender and survival.

$\chi^2$  continuity correction can be used to prevent overestimation of statistical significance for small datasets. This is mainly used when at least one cell of the table has an expected count smaller than 5.



As a note of caution, this correction may overcorrect and result in an overly conservative result that fails to reject the null hypothesis when it should (a type II error).

The likelihood ratio is an alternative to the Pearson chi-square. It is based on maximum-likelihood theory. For large samples, it is identical to Pearson  $\chi^2$ . It is recommended in particular for small samples sizes i.e. <30.

Nominal measures, Phi (2 x 2 contingency tables only) and Cramer's V (most popular) are both tests of the strength of association (i.e. effect sizes). Both values are in the range of 0 (no association) to 1 (complete association). It can be seen that the strength of association between the variables is of a large effect size.

Nominal	
	Value
Contingency coefficient	0.5
Phi-coefficient	0.5
Cramer's V	0.5

The Contingency coefficient is an adjusted Phi value and is only suggested for large contingency tables such as 5 x 5 tables or larger.

Effect size <sup>4</sup>	df	Small	Moderate	Large
Phi and Cramer's V (2x2 only)	1	0.1	0.3	0.5
Cramer's V	2	0.07	0.21	0.35
Cramer's V	3	0.06	0.17	0.29
Cramer's V	4	0.05	0.15	0.25
Cramer's V	5	0.04	0.13	0.22

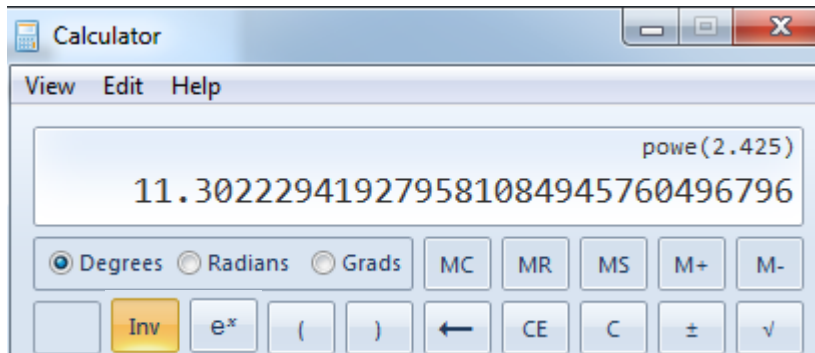
JASP also provides the Odds ratio (OR) which is used to compare the relative odds of the occurrence of the outcome of interest (survival), given exposure to the variable of interest (in this case gender).

Log Odds Ratio ▼			
	Log Odds Ratio	95% Confidence Intervals	
		Lower	Upper
Odds ratio	-2.425	-2.692	-2.159
Fisher's exact test	-2.423	-2.701	-2.150

<sup>4</sup> Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor. Dent. Endod. 2017; 42(2):152-155.



For some reason, JASP calculates OR as a natural log. To convert this from a log value calculate the natural antilog value (using Microsoft calculator, input number then click on **Inv** followed by **e<sup>x</sup>**), in this case, it is 11.3. This suggests that male passengers had 11.3 times more chance of dying than females.



How is this calculated? Use the counts from the contingency table in the following:

Odds[males] = Died/Survived = 682/162 = 4.209  
 Odds[females] = Died/Survived = 127/339 = 0.374

OR = Odds[males] / Odds [females] = 11.3

### GOING ONE STEP FURTHER.

We can also further decompose the contingency table as a form of post hoc testing by converting the counts and expected counts in each cell to a standardised residual. This can tell us if the observed counts and expected counts are significantly different in each cell.

The standardized residual for a cell in a table is a version of the standard z-score, calculated as

$$z = \frac{\text{observed} - \text{expected}}{\sqrt{\text{expected}}}$$

In the special case where df = 1, the calculation of the standardized residual incorporates a correction factor:

$$z = \frac{|\text{observed} - \text{expected}| - 0.5}{\sqrt{\text{expected}}}$$

The resulting value of z is then given a positive sign if observed > expected and a negative sign if observed < expected. Z-score significances are shown below.

z-score	P value
<-1.96 or > 1.96	<0.05
<-2.58 or > 2.58	<0.01
<-3.29 or > 3.29	<0.001





### Contingency Tables

survived		sex		Total
		female	male	
No	Count	127.0	682.0	809.0
	Expected count	288.0	521.0	809.0
	% within row	15.7 %	84.3 %	100.0 %
	% within column	27.3 %	80.9 %	61.8 %
	% of Total	9.7 %	52.1 %	61.8 %
Yes	Count	339.0	161.0	500.0
	Expected count	178.0	322.0	500.0
	% within row	67.8 %	32.2 %	100.0 %
	% within column	72.7 %	19.1 %	38.2 %
	% of Total	25.9 %	12.3 %	38.2 %
Total	Count	466.0	843.0	1309.0
	Expected count	466.0	843.0	1309.0
	% within row	35.6 %	64.4 %	100.0 %
	% within column	100.0 %	100.0 %	100.0 %
	% of Total	35.6 %	64.4 %	100.0 %

Female No z = - 9.5	Male No z = 7.0
Female Yes z = 12.0	Male Yes z = -8.9

When the z-scores are calculated for each cell in the contingency table we can see that significantly fewer women died than expected and significantly more males died than expected  $p < .001$ .



## EXPERIMENTAL DESIGN AND DATA LAYOUT IN EXCEL FOR JASP IMPORT.

Independent t-test

Design example:

Independent variable	Group 1	Group 2
Dependent variable	Data	Data

Independent variable

Dependent variable

Categorical

Continuous

	A	B
1	Group	Data
2	1	0
3	1	0
4	1	3.8
5	1	6
6	1	0.7
7	1	2.9
8	1	2.8
9	1	2
10	1	2
11	1	8.5
12	1	1.9
13	1	3.1
14	1	1.5
15	1	3
16	1	3.6
17	1	0.9
18	1	-2.1
19	2	2
20	2	1.7
21	2	4.3
22	2	7
23	2	0.6
24	2	2.7
25	2	3.6

More dependent variables can be added if required



## Paired samples t-test

Design example:

Independent variable	Pre-test	Post-test
Participant	Dependent variable	
1	Data	Data
2	Data	Data
3	Data	Data
..n	Data	Data

Pre-test

Post-test


	A	B
1	Pre-test	Post-test
2	60	60
3	103	103
4	58	54
5	60	54
6	64	63
7	64	61
8	65	62
9	66	64
10	67	65
11	69	61
12	70	68
13	70	67
14	72	71
15	72	69
16	72	68
17	82	81
18	58	60
19	58	56
20	59	57
21	61	57
22	62	55
23	63	62
24	63	60
25	63	59



## Correlation

Design example:

### Simple correlation



Participant	Variable 1	Variable 2	Variable 3	Variable 4	Variable ..n
1	Data	Data	Data	Data	Data
2	Data	Data	Data	Data	Data
3	Data	Data	Data	Data	Data
...n	Data	Data	Data	Data	Data



### Multiple correlation

	A	B	C	D	E	F
1	Participant	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
2	1	533	77	77	106	106
3	2	472	63	59	92	93
4	3	484	82	77	93	78
5	4	536	72	72	103	93
6	5	630	77	68	104	93
7	6	563	68	68	101	87
8	7	531	77	82	108	106
9	8	344	50	50	86	92
10	9	346	54	50	90	86
11	10	386	59	54	85	80
12	11	460	54	63	89	83
13	12	492	63	59	92	94



## Regression

Design example:

### Simple Regression

Participant	Outcome	Predictor 1	Predictor 2	Predictor 3	Predictor ..n
1	Data	Data	Data	Data	Data
2	Data	Data	Data	Data	Data
3	Data	Data	Data	Data	Data
...n	Data	Data	Data	Data	Data

### Multiple regression

	A	B	C	D	E	F
1	Participant	Outcome	Predictor 1	Predictor 2	Predictor 3	Predictor 4
2	1	533	77	77	106	106
3	2	472	63	59	92	93
4	3	484	82	77	93	78
5	4	536	72	72	103	93
6	5	630	77	68	104	93
7	6	563	68	68	101	87
8	7	531	77	82	108	106
9	8	344	50	50	86	92
10	9	346	54	50	90	86
11	10	386	59	54	85	80
12	11	460	54	63	89	83
13	12	492	63	59	92	94



## Logistic Regression

Design example:

	Dependent Variable (categorical)	Factor (categorical)	Covariate (continuous)
Participant	Outcome	Predictor 1	Predictor 2
1	Data	Data	Data
2	Data	Data	Data
3	Data	Data	Data
...n	Data	Data	Data

	A	B	C	D
1	ID	Outcome	Factor	Covariate
2	1	Yes	Yes	70
3	2	Yes	No	80
4	3	Yes	Yes	50
5	4	Yes	No	60
6	5	Yes	No	40
7	6	Yes	No	65
8	7	Yes	No	75
9	8	Yes	No	80
10	9	Yes	No	70
11	10	Yes	No	60
12	11	No	Yes	65
13	12	No	Yes	50
14	13	No	Yes	45
15	14	No	Yes	35
16	15	No	Yes	40
17	16	No	Yes	50
18	17	No	No	55
19	17	Yes	No	65
20	18	No	Yes	45

**More factors and covariates can be added if required**



## One-way Independent ANOVA

Design example:

<b>Independent variable</b>	<b>Group 1</b>	<b>Group 2</b>	<b>Group 3</b>	<b>Group...n</b>
<b>Dependent variable</b>	Data	Data	Data	Data

**Independent variable**

**(Categorical)**

**Dependent variable**

**(Continuous)**

	A	B
1	Group	Dependent variable
2	Group 1	3.8
3	Group 1	6
4	Group 1	0.7
5	Group 1	2.9
6	Group 1	2.8
7	Group 1	2
8	Group 1	2
9	Group 1	3.5
10	Group 2	1.9
11	Group 2	3.1
12	Group 2	1.5
13	Group 2	3
14	Group 2	3.6
15	Group 2	0.9
16	Group 2	-0.6
17	Group 3	1.1
18	Group 3	4.5
19	Group 3	6.1
20	Group 3	5
21	Group 3	2.4
22	Group 3	3.9
23	Group 3	3.5
24	Group 3	5.1
25	Group 3	3.5

**More dependent variables can be added if required**



## One-way repeated measures ANOVA

Design example:

Participant	Independent variable (Factor)			
	Level 1	Level 2	Level 3	Level..n
1	Data	Data	Data	Data
2	Data	Data	Data	Data
3	Data	Data	Data	Data
4	Data	Data	Data	Data
..n	Data	Data	Data	Data

Factor (time)

	A	B	C	D
1	Participant	Week 0	Week 3	Week 6
2	1	6.42	5.83	5.75
3	2	6.76	6.2	6.13
4	3	6.56	5.83	5.71
5	4	4.8	4.27	4.15
6	5	8.43	7.71	7.67
7	6	7.49	7.12	7.05
8	7	8.05	7.25	7.1
9	8	5.05	4.63	4.67
10	9	5.77	5.31	5.33
11	10	3.91	3.7	3.66
12	11	6.77	6.15	5.96
13	12	6.44	5.59	5.64
14	13	6.17	5.56	5.51
15	14	7.67	7.11	6.96
16	15	7.34	6.84	6.82
17	16	6.85	6.4	6.29
18	17	5.13	4.52	4.45
19	18	5.73	5.13	5.17

More levels can be added if required





## Two-way Independent ANOVA

Design example:

Factor 1	Supplement 1			Supplement 2		
Factor 2	Dose 1	Dose 2	Dose 3	Dose 1	Dose 2	Dose 3
Dependent variable	Data	Data	Data	Data	Data	Data

Factor 1    Factor 2    Dependent variable

	A	B	C
1	supp	dose	len
2	OJ	1000	19.7
3	OJ	1000	23.3
4	OJ	1000	23.6
5	OJ	1000	26.4
6	OJ	1000	20
7	OJ	1000	25.2
8	OJ	1000	25.8
9	OJ	1000	21.2
10	OJ	1000	14.5
11	OJ	1000	27.3
12	OJ	2000	25.5
13	OJ	2000	26.4
14	OJ	2000	22.4
15	OJ	2000	24.5
16	OJ	2000	24.8
17	OJ	2000	30.9
18	OJ	2000	26.4
19	OJ	2000	27.3
20	OJ	2000	29.4
21	OJ	2000	23
22	VC	1000	16.5
23	VC	1000	16.5
24	VC	1000	15.2
25	VC	1000	17.3

More factors and dependent variables can be added if required



## Two-way Mixed Factor ANOVA

Design example:

Factor 1 (Between subjects)	Group 1			Group 2		
Factor 2 levels (Repeated measures)	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
1	Data	Data	Data	Data	Data	Data
2	Data	Data	Data	Data	Data	Data
3	Data	Data	Data	Data	Data	Data
..n	Data	Data	Data	Data	Data	Data

Factor 1

Factor 2 levels

(Categorical)

(Continuous)

	A	B	C	D
1	Group	Level 1	Level 2	Level 3
2	Group 1	1.31	0.9	0.9
3	Group 1	1.29	0.89	0.72
4	Group 1	1.8	0.9	0.96
5	Group 1	1.4	1.26	0.97
6	Group 1	1.49	1.18	0.88
7	Group 1	1.35	1.15	0.92
8	Group 1	1.45	1.19	1
9	Group 1	1.21	1.2	0.85
10	Group 1	1.79	1.48	0.99
11	Group 1	1.73	1.68	0.98
12	Group 2	1.55	0.9	0.55
13	Group 2	1.27	0.95	0.41
14	Group 2	1.53	0.87	0.42
15	Group 2	1.26	1.15	0.44
16	Group 2	1.14	1.12	0.38
17	Group 2	1.11	1.08	0.34
18	Group 2	1.1	1.0758	0.18
19	Group 2	1.08	1.18	0.24
20	Group 2	1.3	1.26	0.39
21	Group 2	1.45	1.55	0.44



## Chi-squared - Contingency tables

Design example:

Participant	Response 1	Response 2	Response 3	Response...n
1	Data	Data	Data	Data
2	Data	Data	Data	Data
3	Data	Data	Data	Data
..n	Data	Data	Data	Data

All data should be categorical

	A	B	C	D	E
1	Respondant	Response 1	Response 2	Response 3	Response 4
2	1	Female	clay	Morning	yes
3	2	Male	astro	Morning	No
4	3	Female	grass	Evening	No
5	4	Male	clay	Afternoon	No
6	5	Male	clay	Morning	No
7	6	Male	grass	Evening	No
8	7	Female	grass	Evening	yes
9	8	Male	clay	Morning	yes
10	9	Female	grass	Morning	No
11	10	Male	clay	Afternoon	No
12	11	Female	clay	Afternoon	No
13	12	Male	astro	Afternoon	No
14	13	Male	astro	Afternoon	No
15	14	Male	astro	Afternoon	yes
16	15	Female	clay	Morning	No
17	16	Male	astro	Afternoon	yes
18	17	Female	astro	Afternoon	yes
19	18	Male	grass	Morning	No
20	19	Male	clay	Afternoon	No



## SOME CONCEPTS IN FREQUENTIST STATISTICS

The frequentist approach is the most commonly taught and used statistical methodology. It describes sample data based on the frequency or proportion of the data from repeated studies through which the probability of events is defined.

Frequentist statistics uses rigid frameworks including hypothesis testing, p values and confidence intervals etc.

### Hypothesis testing

A hypothesis can be defined as ***“a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation”***.

There are two simple types of hypotheses, a null hypothesis ( $H_0$ ) and an alternative or experimental hypothesis ( $H_1$ ). The **null hypothesis** is the default position for most statistical analyses in which it is stated that there is no relationship or difference between groups. The **alternative hypothesis** states that there is a relationship or difference between groups for can a direction of difference/relationship. For example, if a study was carried out to look at the effects of a supplement on sprint time in one group of participants compared to the placebo group:

$H_0$  = there is **no** difference in sprint times between the two groups

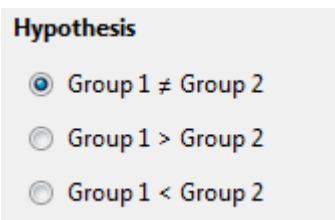
$H_1$  = there is a difference in sprint times between the two groups

$H_2$  = group 1 is greater than group 2

$H_3$  = group 1 is less than group 2

Hypothesis testing refers to the strictly predefined procedures used to accept or reject the hypotheses and the probability that this could be purely by chance. The confidence at which a null hypothesis is accepted or rejected is called the level of significance. The level of significance is denoted by  $\alpha$ , usually 0.05 (5%). This is the level of probability of accepting an effect as true (95%) and that there is only 5% of the result being purely by chance.

Different types of hypothesis can easily be selected in JASP, however, the null hypothesis is always the default.





## Type I and II errors

The probability of rejecting the null hypothesis, when it is, in fact, true, is called Type I error whereas the probability of accepting the null hypothesis when it is not true is called Type II error.

		The truth	
		Not guilty ( $H_0$ )	Guilty ( $H_1$ )
The verdict	Guilty ( $H_1$ )	<b>Type I error</b> An innocent person goes to prison	Correct decision
	Not guilty ( $H_0$ )	Correct decision	<b>Type II error</b> A guilty person goes free

Type I error is deemed the worst error to make in statistical analyses.

Statistical power is defined as the probability that the test will reject the null hypothesis when the alternative hypothesis is true. For a set level of significance, if the sample size increases, the probability of Type II error decreases, which therefore increases the statistical power.

## Testing the hypothesis

The essence of hypothesis testing is to first define **the null (or alternative) hypothesis**, set the criterion level  $\alpha$ , usually 0.05 (5%), collect and analyse sample data. Use a **test statistic** to determine how far (or the number of standard deviations) the sample mean is from the population mean stated in the null hypothesis. The test statistic is then compared to a critical value. This is a cut-off value defining the boundary where less than 5% of the sample means can be obtained if the null hypothesis is true.

If the probability of obtaining a difference between the means by chance is less than 5% when the null hypothesis has been proposed, the null hypothesis is rejected and the alternative hypothesis can be accepted.

The **p value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. If the p value is less than 5% ( $p < .05$ ) the null hypothesis is rejected. When the p value is greater than 5% ( $p > .05$ ), we accept the null hypothesis.

## Effect size

An effect size is a standard measure that can be calculated from any number of statistical analyses. If the null hypothesis is rejected the result is significant. This significance only evaluates the probability of obtaining the sample outcome by chance but does not indicate how big a difference (practical significance), nor can it be used to compare across different studies.

The effect size indicates the magnitude of the difference between the groups. So for example, if there was a significant decrease in 100m sprint times in a supplement compared to a



placebo group, the effect size would indicate how much more effective the intervention was. Some common effect sizes are shown below.

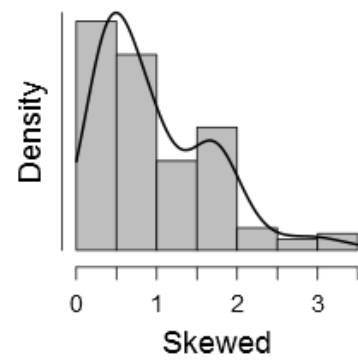
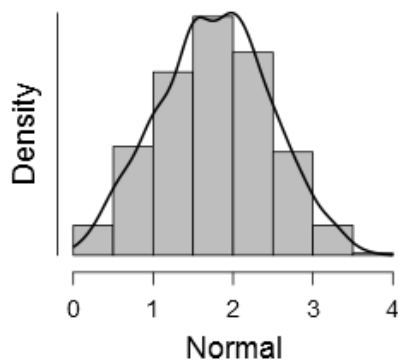
Test	Measure	Trivial	Small	Medium	Large
<b>Between means</b>	Cohen's d	<0.2	0.2	0.5	0.8
<b>Correlation</b>	Correlation coefficient (r)	<0.1	0.1	0.3	0.5
	Rank -biserial ( $r_B$ )	<0.1	0.1	0.3	0.5
	Spearman's rho	<0.1	0.1	0.3	0.5
<b>Multiple Regression</b>	Multiple correlation coefficient (R)	<0.10	0.1	0.3	0.5
<b>ANOVA</b>	Eta	<0.1	0.1	0.25	0.37
	Partial Eta	<0.01	0.01	0.06	0.14
	Omega squared	<0.01	0.01	0.06	0.14
<b>Chi-squared</b>	Phi (2x2 tables only)	<0.1	0.1	0.3	0.5
	Cramer's V	<0.1	0.1	0.3	0.5
	Odds ratio (2x2 tables only)	<1.5	1.5	3.5	9.0

In small datasets, there may be a moderate to large effect size but no significant differences. This could suggest that the analysis lacked statistical power and that increasing the number of data points may show a significant outcome. Conversely, when using large datasets, significant testing can be misleading since small or trivial effects may produce statistically significant results.

### PARAMETRIC vs NON-PARAMETRIC TESTING

Most research collects information from a sample of the population of interest, it is normally impossible to collect data from the whole population. We do, however, want to see how well the collected data reflects the population in terms of the population mean, standard deviations, proportions etc. based on parametric distribution functions. These measures are the population **parameters**. Parameter estimates of these in the sample population are statistics. Parametric statistics require assumptions to be made of the data including the normality of distribution and homogeneity of variance.

In some cases these assumptions may be violated in that the data may be noticeably skewed:





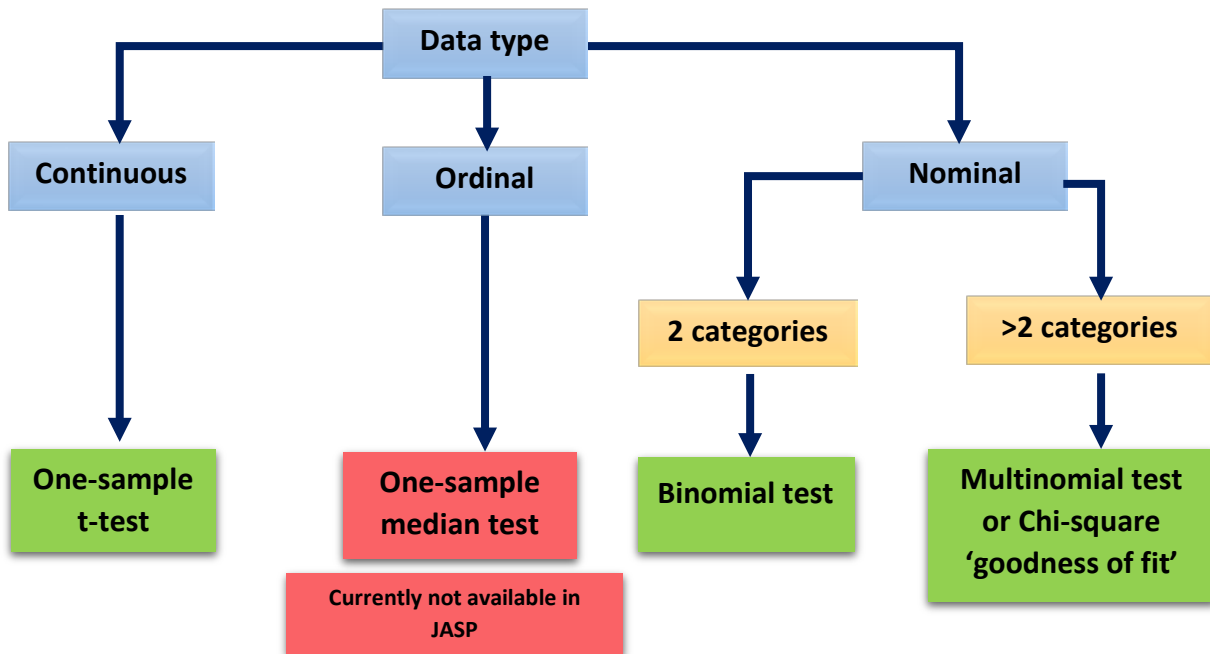
Sometimes transforming the data can rectify this but not always. It is also common to collect ordinal data (i.e. Likert scale ratings) for which terms such as mean and standard deviation are meaningless. As such there are no parameters associated with ordinal (**non-parametric**) data. The non-parametric counterparts include median values and quartiles.

In both of the cases described non-parametric statistical tests are available. There are equivalents of most common classical parametric tests. These tests that don't assume normally distributed data or population parameters and are based on sorting the data into ranks from lowest to highest values. All subsequent calculations are done with these ranks rather than with the actual data values.

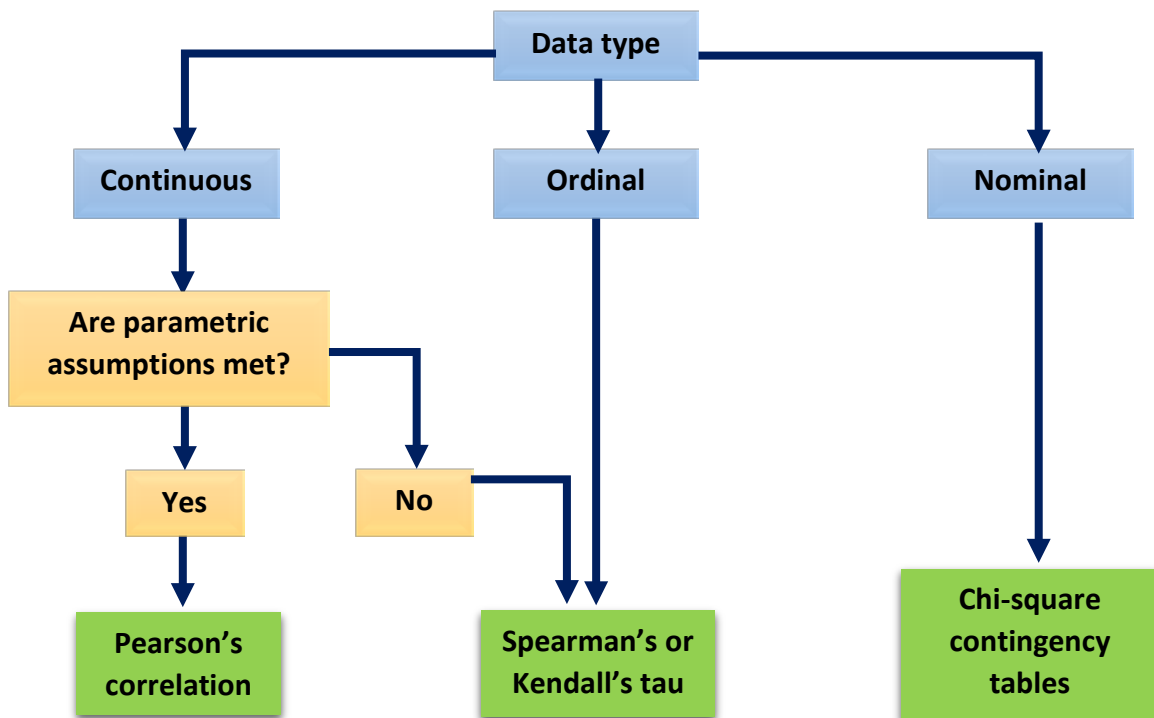


## WHICH TEST SHOULD I USE?

Comparing one sample to a known or hypothesized population mean.



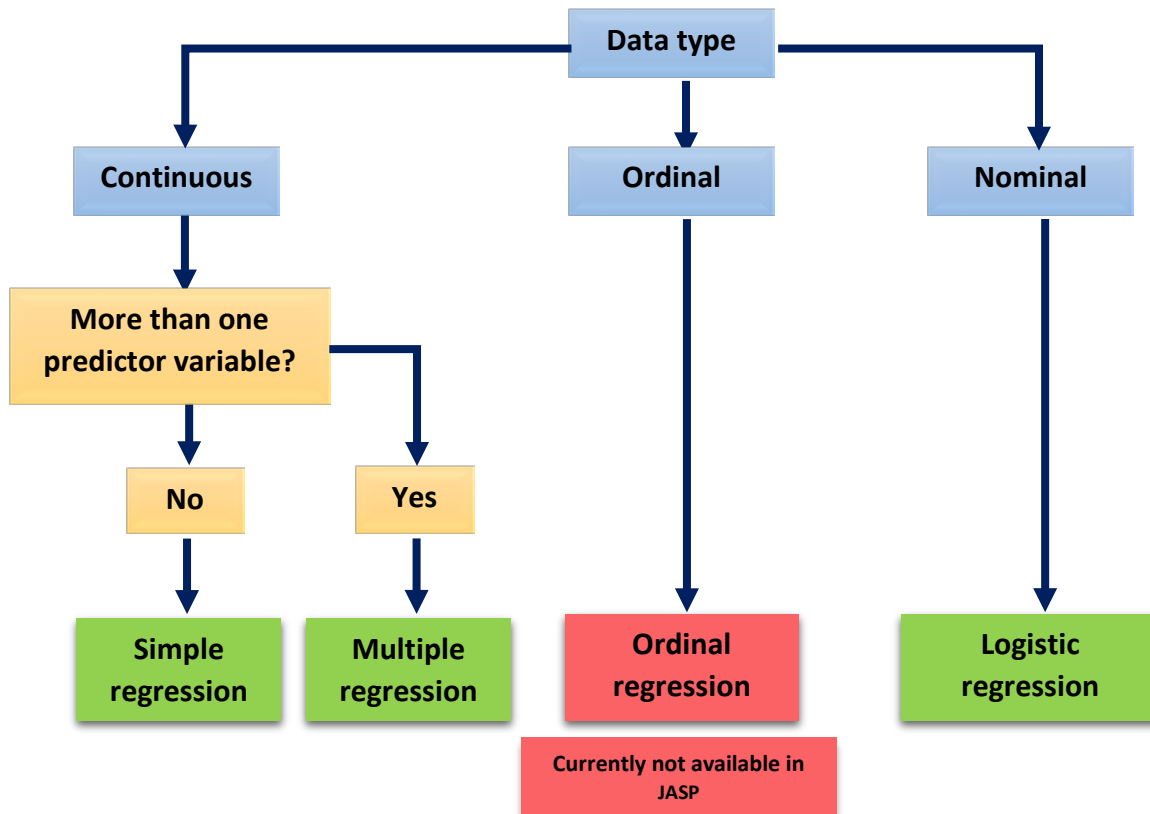
## Testing relationships between two or more variables



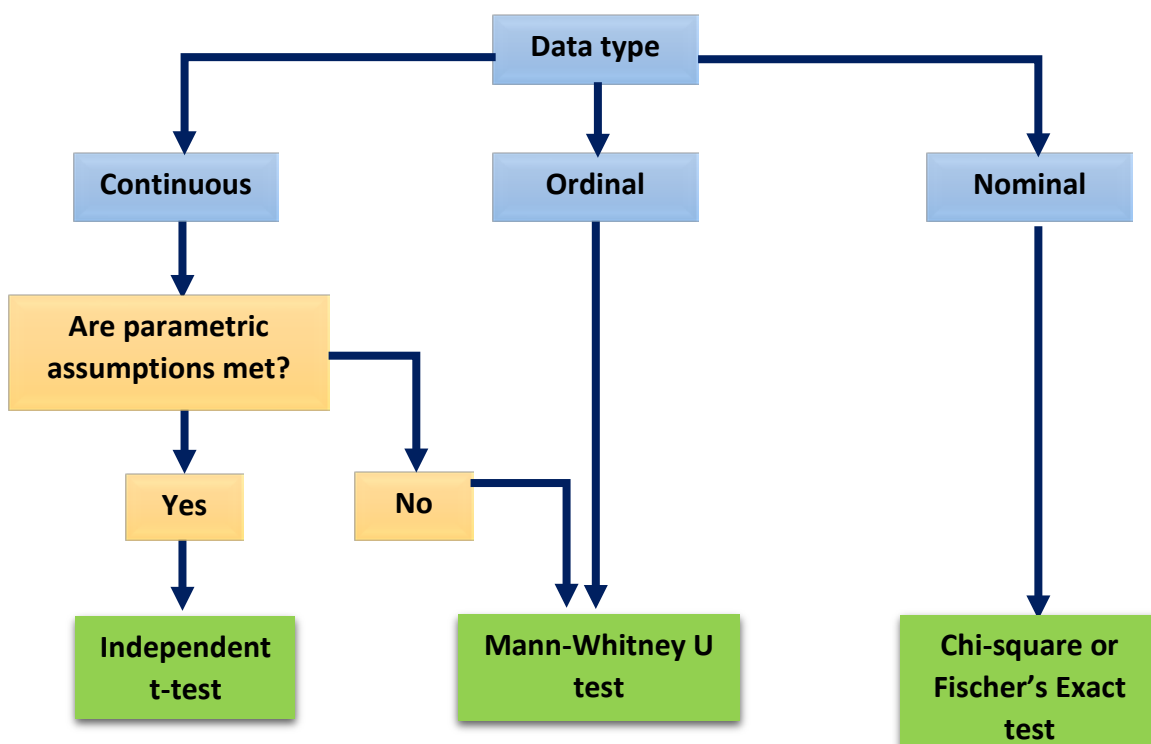




## Predicting outcomes

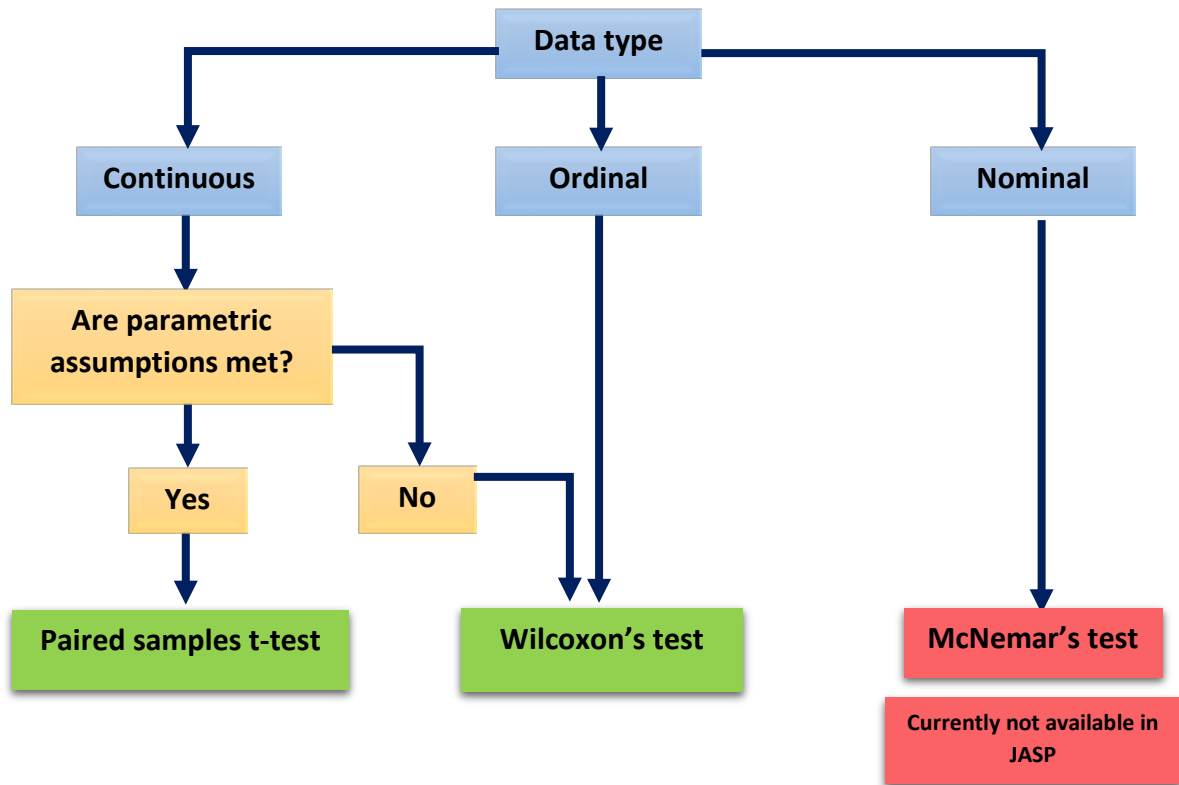


## Testing for differences between two independent groups

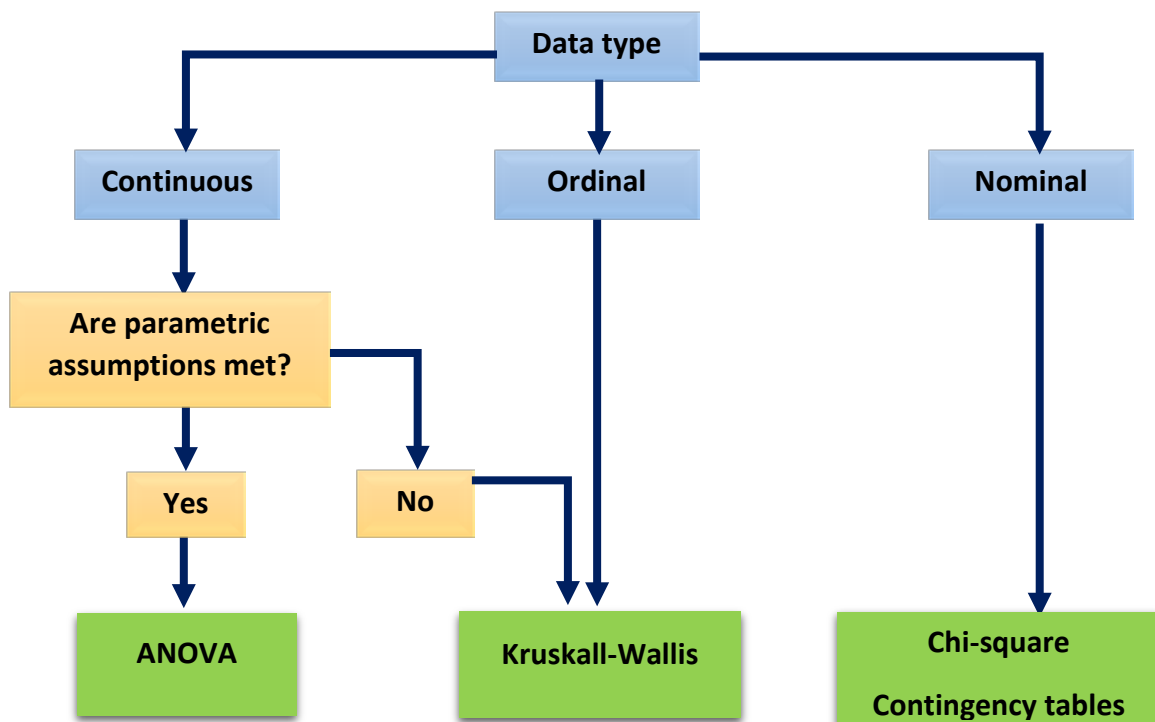




## Testing for differences between two related groups

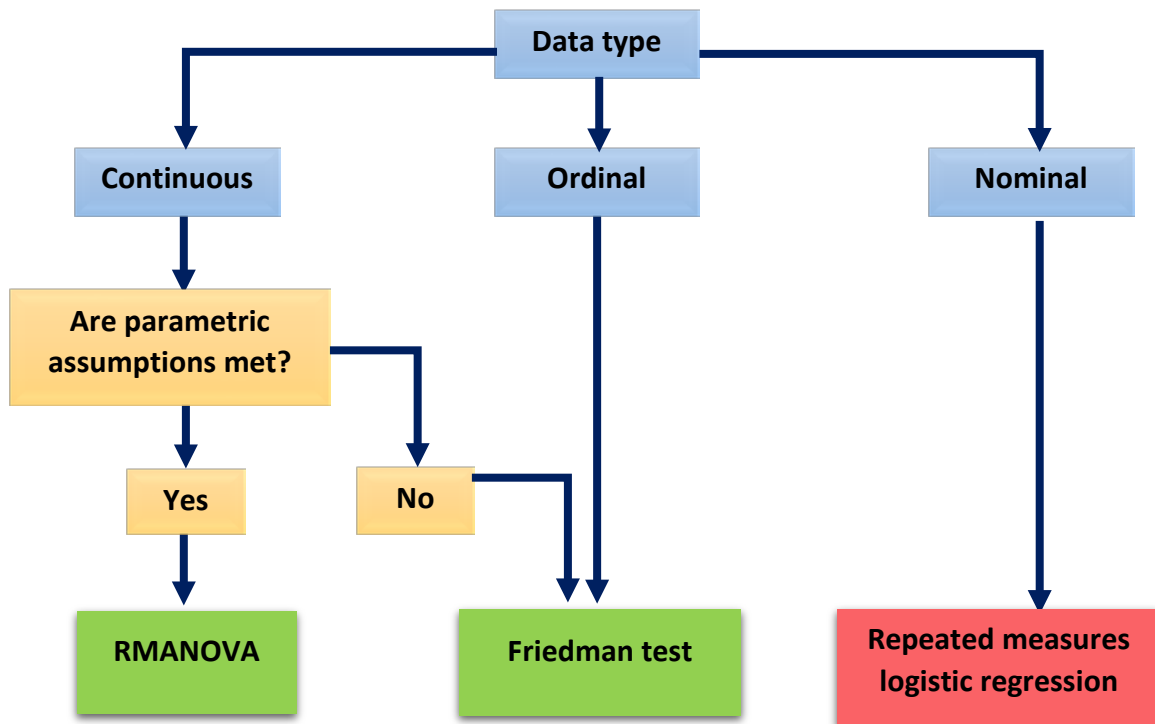


## Testing for differences between three or more independent groups





## Testing for differences between three or more related groups



## Test for interactions between 2 or more independent variables

