JASP

STATISTICAL ANALYSIS IN JASP

A GUIDE FOR STUDENTS

JASP v 0.16 (2022)

Dr Mark A Goss-Sampson

5th Edition JASP v0.16.1 2022

CONTENTS

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 by a group of researchers at the University of Amsterdam. They aimed 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 WIX installer for Windows enables you to choose a path for the installation of JASP – however, this may be blocked in some institutions by local Administrative rights.

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[1] and The Introduction to the Practice of Statistics[2] by Moore, McCabe and Craig.

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

This book is a collection of standalone handouts covering the most common standard (frequentist) statistical analyses used by students studying Human Sciences. Datasets used in this document are available for download from   https://osf.io/bx6uv/

I would also like to thank Per Palmgren from the Karolinska Institutet in Sweden for his helpful comments, suggestions and proofreading of this guide.

Dr Mark Goss-Sampson
Centre for Exercise Activity and Rehabilitation
University of Greenwich
2022

---

[1] A Field. (2017) Discovering Statistics Using IBM SPSS Statistics (5th Ed.) SAGE Publications.
[2] D Moore, G McCabe, B Craig. (2011) Introduction to the Practice of Statistics (7th Ed.) W H Freeman.

JASP 0.16.1 - Dr Mark Goss-Sampson

USING THE JASP ENVIRONMENT

Open JASP.
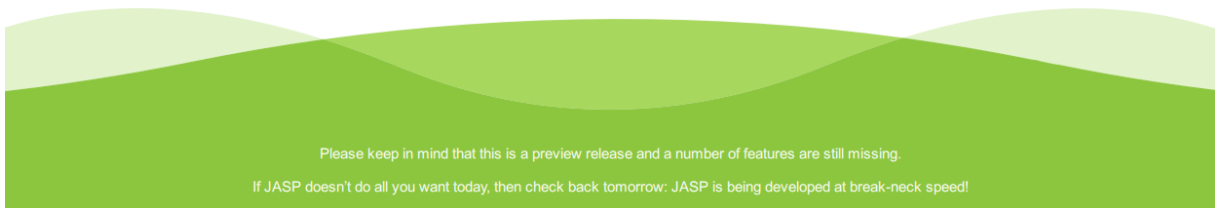
The main menu can be accessed by clicking on the top-left icon.
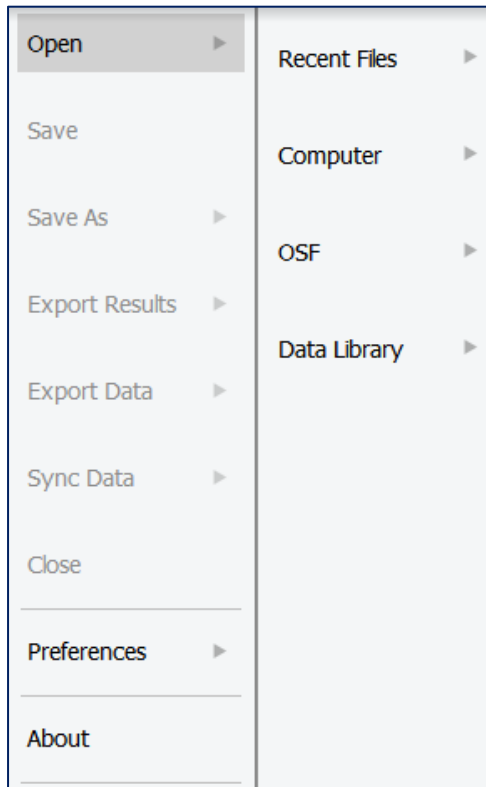
Open:

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

- **.csv** (comma separated values) can be saved in Excel
- **.txt** (plain text) also can be saved in Excel
- **.tsv** (tab-separated values) also can be saved in Excel
- **.sav** (IBM SPSS data file)
- **.ods** (Open Document spreadsheet)
- **.dta** (Stata data file)
- **.por** (SPSS ASCII file)
- **.Sas7bdat** /cat (SAS data files)
- **.xpt** (SAS transport file)

You can open recent files, browse your computer files, access the Open Science Framework (OSF) or open the wide range of examples that are packaged with the Data Library in JASP.

JASP 0.16.1 - Dr Mark Goss-Sampson
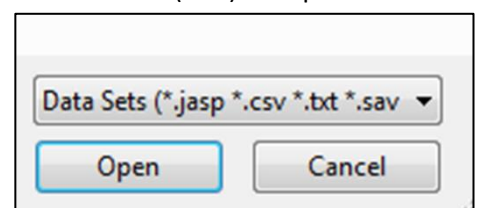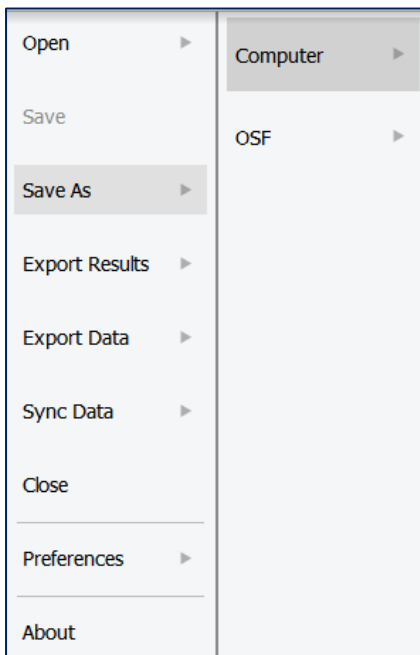
Save/Save as:

Using these options the data file, any annotations and the analysis can be saved in the .jasp format

Export:

Results can be exported to either an HTML file or as a PDF

Data can be exported to either a .csv, .tsv or .txt file

Sync data:

Used to synchronize with any updates in the current data file (also can use Ctrl-Y)

Close:

As it states - it closes the current file but not JASP

Preferences:

There are four sections that users can use to tweak JASP to suit their needs

JASP 0.16.1 - Dr Mark Goss-Sampson
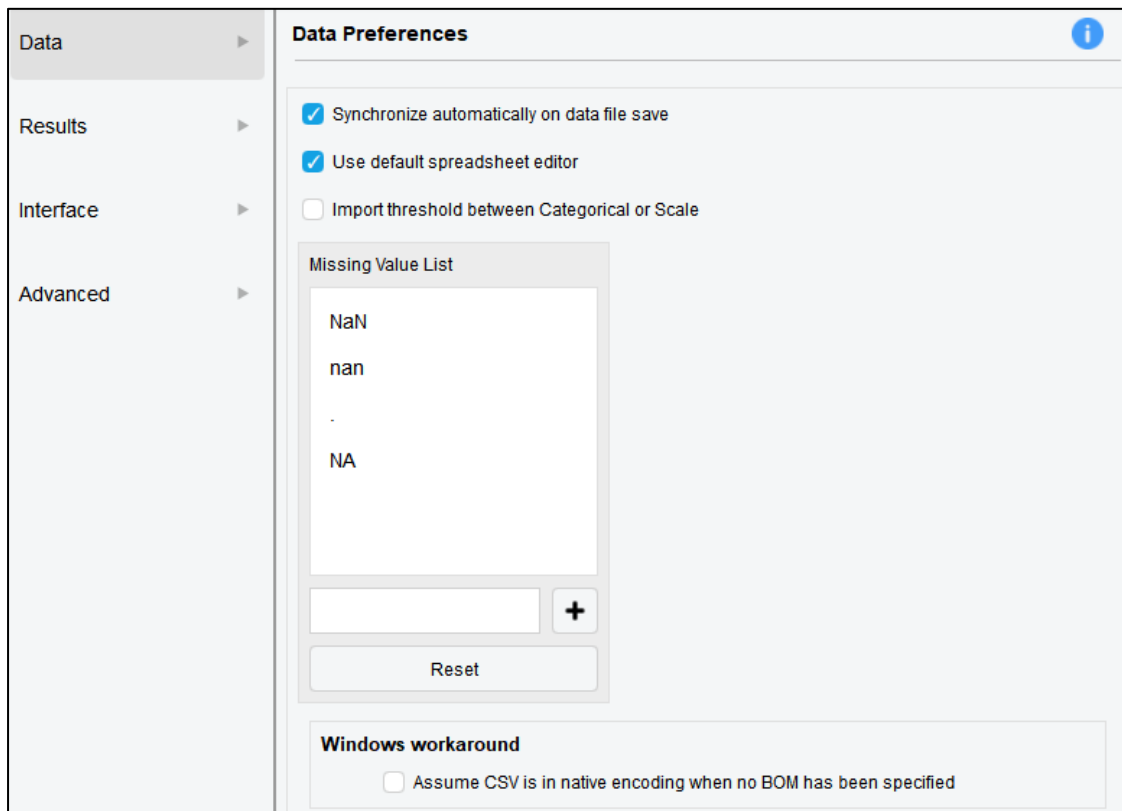
In the **Data Preferences** section users can:

- Synchronize/update the data automatically when the data file is saved (default)
- Set the default spreadsheet editor (i.e. Excel, SPSS etc)
- Change the threshold so that JASP more readily distinguishes between nominal and scale data
- Add a custom missing value code
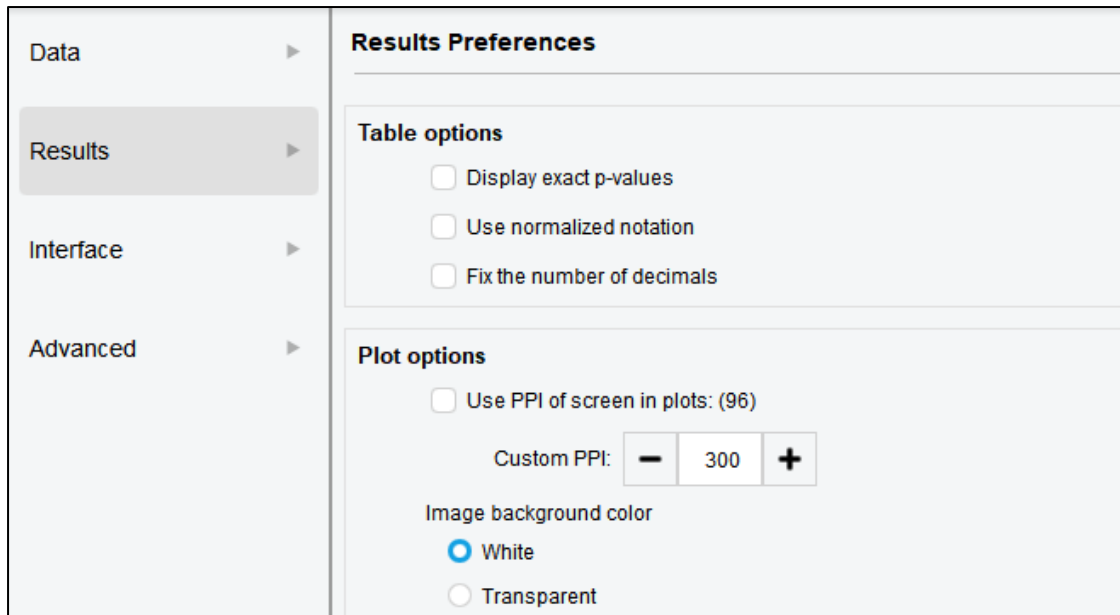
In the **Results Preferences** section users can:

- Set JASP to return exact p values i.e. P=0.00087 rather than P<.001
- Fix the number of decimals for data in tables – makes tables easier to read/publish
- Change the pixel resolution of the graph plots
- Select when copying graphs whether they have a white or transparent background.

In the **Interface Preferences** section, users can now define a user font and pick between two different themes: a light theme (default) and a dark theme.

The preferred language currently supports 10 languages English, Spanish, German, Dutch, French, Indonesian, Japanese, Portuguese, Chinese and Galician.

In this section, there is also the ability to change the system size (zoom) for accessibility and scroll speeds.

JASP 0.16.1 - Dr Mark Goss-Sampson

Comparison of the dark and light themes in JASP

In the *Advanced Preferences* section, most users will probably never have to change any of the default settings.

JASP 0.16.1 - Dr Mark Goss-Sampson

JASP has a streamlined interface to switch between the spreadsheet, analysis and results views.
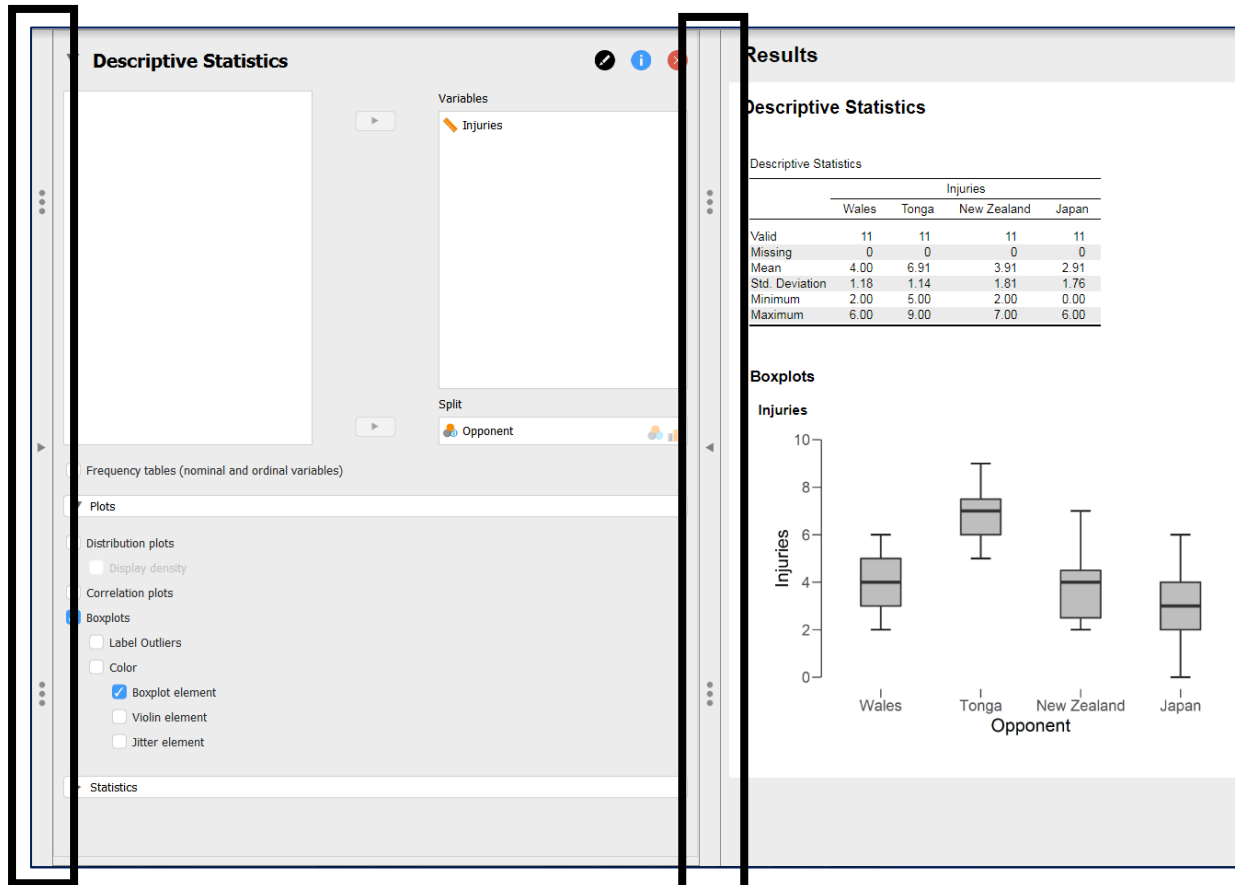
The vertical bars highlighted above allows for the windows to be dragged right or left by clicking and dragging the three vertical dots

The individual windows can also be completely collapsed using the right or left arrow icons

If you click the Results ▼ icon a range of options is provided including:

- **Edit title**
- **Copy**
- **Export results**
- **Add notes**
- **Remove all**
- **Refresh all**

JASP 0.16.1 - Dr Mark Goss-Sampson

The 'add notes' option allows the results output to be easily annotated and then exported to an HTML or PDF 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

B *I* U ⌗   Normal ⬍ ☰ ☰   A 🗚   x₂ x²   " ⬌ ⬌   Normal ⬍ 𝐼ₓ

The Add notes menu provides many options to change text font, colour size etc.

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.

As previously mentioned, all tables and figures are APA standard and can just be copied into any other document. Since all images can be copied/saved with either a white or transparent background. This can be selected in Preferences > Advanced as described earlier.

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

JASP 0.16.1 - Dr Mark Goss-Sampson

DATA HANDLING IN JASP

For this section open England injuries.csv

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

	Opponent	Injuries	+
1	Japan	4	
2	Japan	1	
3	Japan	3	
4	Japan	6	
5	Japan	2	
6	Japan	3	
7	Japan	4	
8	Japan	0	
9	Japan	5	
10	Japan	2	
11	Japan	2	
12	New Zealand	2	
13	New Zealand	4	

For large datasets, there is a hand icon that 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.

	Opponent	Injuries	
			Scale
2	Japan	1	Ordinal
3	Japan	3	Nominal

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.

JASP 0.16.1 - Dr Mark Goss-Sampson

Filter	Value	Label
✓	1	Tonga
✓	2	New Zealand
✓	3	France
✓	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:

Welcome to the drag and drop filter!

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/

JASP 0.16.1 - Dr Mark Goss-Sampson

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 appropriate arrows:

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

Move up

Move down

Reverse order

Close

Injuries
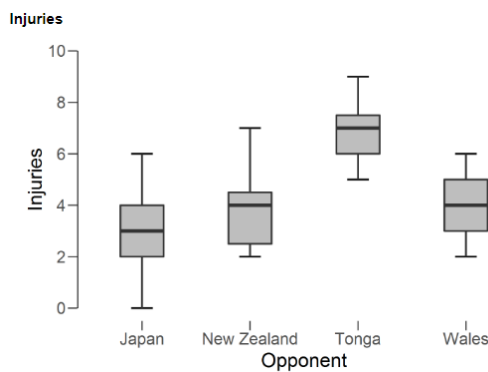
Filter	Value	Label
✓	Japan	Japan
✓	New Zealand	New Zealand
✓	Tonga	Tonga
✓	Wales	Wales

Injuries
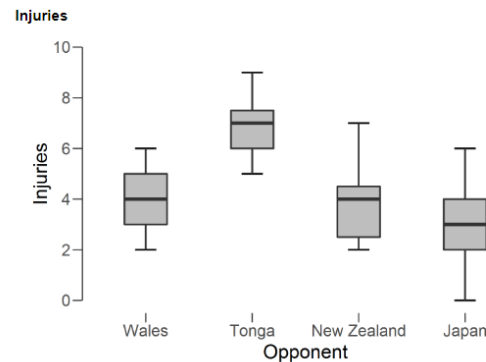
Filter	Value	Label
✓	Wales	Wales
✓	Tonga	Tonga
✓	New Zealand	New Zealand
✓	Japan	Japan

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. 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
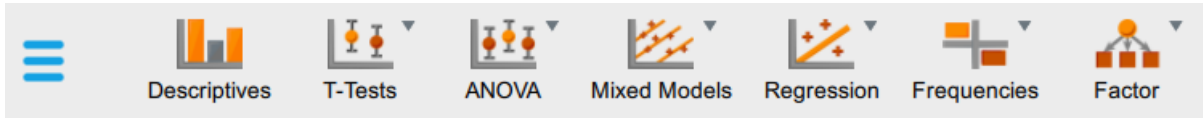
The **main** analysis options can be accessed from the main toolbar. Currently, JASP offers the following frequentist (parametric and non-parametric standard statistics) and alternative Bayesian tests:

Descriptives • Descriptive stats	**Regression** • Correlation • Linear regression • Logistic regression
T-Tests • Independent • Paired • One sample	**Frequencies** • Binomial test • Multinomial test • Contingency tables • Log-linear regression*
ANOVA • Independent • Repeated measures • ANCOVA • MANOVA *	**Factor** • Principal Component Analysis (PCA)* • Exploratory Factor Analysis (EFA)* • Confirmatory Factor Analysis (CFA)*
Mixed Models* • Linear Mixed Models Generalised linear mixed models	

* Not covered in this guide

BY clicking on the **+** icon on the top-right menu bar you can also access advanced options that allow the addition of optional modules. Once ticked they will be added to the main analysis ribbon. These include;

Audit	Network analysis
BAIN	Prophet
Circular Statistics	Reliability analysis
Distributions	SEM
Cochrane Meta-Analyses	Summary statistics
Equivalence tests	Visual modelling
JAGS	Learning Bayes
Machine learning	R (beta)
Meta-analysis (included in this guide)	

See the JASP website for more information on these advanced modules

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

JASP provides the ability to rename and 'stack' the results output thereby organising multiple analyses.

▶ **Descriptive Statistics**

▶ **ANOVA**

▼ **Correlation**
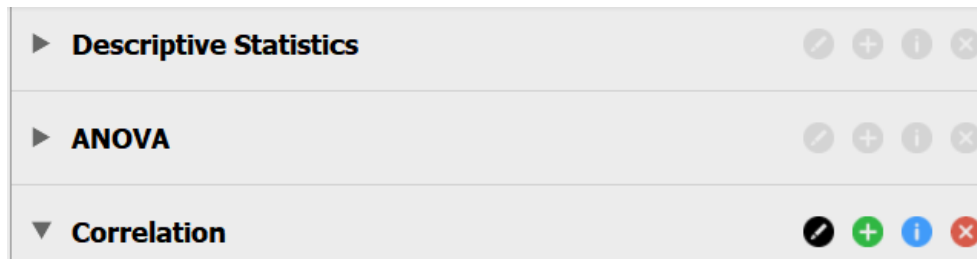
The individual analyses can be renamed using the pen icon or deleted using the red cross.

▶ **Descriptive Statistics England injuries**

▶ **ANOVA 5 countries played**

▼ **Correlation -Spearmans injury correlation**

Clicking on the analysis in this list will then take you to the appropriate part of the results output window. They can also be rearranged by dragging and dropping each of the analyses.

The green **+** icon produces a copy of the chosen analysis

▶ **Descriptive Statistics England injuries**

▶ **ANOVA 5 countries played**

▶ **Correlation -Spearmans injury correlation**

▼ **Copy of Correlation -Spearmans injury correlation**

The blue information icon provides detailed information on each of the statistical procedures used and includes a search option.

JASP 0.16.1 - Dr Mark Goss-Sampson

Correlation

The Correlation analysis allows estimation of the population correlation, as well as testing the null hypothesis that the population correlation between pairs of variables equals 0. All possible pairs of the specified variables are analyzed.

Assumptions (Pearson's rho)

- Continuous variables.
- The data are a random sample from the population.
- The pairs of variables follow a bivariate normal distribution in the population.
- The relationship between the pairs of variables is linear.

Assumptions (Spearman's rho & Kendall's tau)

- Ordinal or continuous variables.
- The data are a random sample from the population

Search for:

DESCRIPTIVE STATISTICS

Presentation of all the raw data is very difficult for a reader to visualise or 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

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.
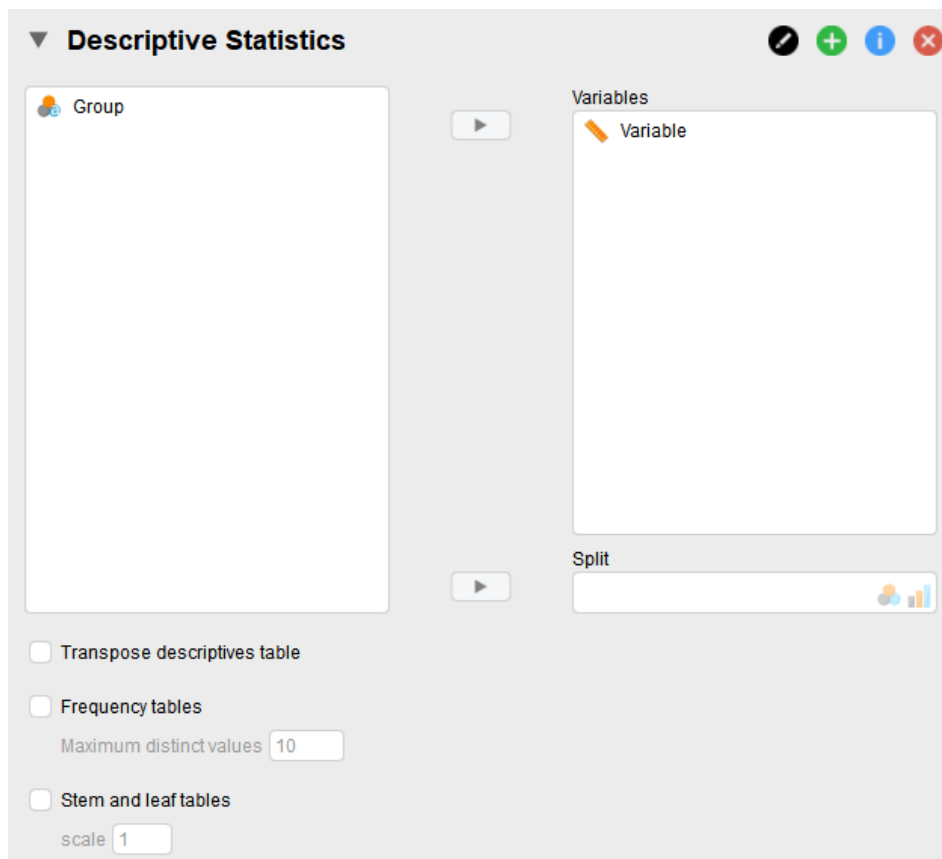
You also have options to change and add tables in this section:

- Split analyses by a categorical variable (i.e., group)
- Transpose the main descriptive table (switch columns and rows)
- Add frequency tables - important for categorical data
- Add stem and Leaf tables: (shows all numeric observations from small to large. The observations are split into a "stem", the first digit(s), and a "leaf", the subsequent digit).

JASP 0.16.1 - Dr Mark Goss-Sampson

The Statistics menu can now be opened to see the various options available.

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 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.

The **mean**, **M or x̄** (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.

JASP 0.16.1 - Dr Mark Goss-Sampson

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

Descriptive Statistics

	Variable
Std. Error of Mean	0.244
Std. Deviation	6.935
Coefficient of variation	0.392
MAD	4.700
MAD robust	6.968
IQR	9.175
Variance	48.100

The standard error of the mean, SE (0.244) 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.

Standard deviation, S or SD (6.935) 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.

The coefficient of variation (0.392) provides the relative dispersion of the data, in contrast to the standar5)d deviation, which gives the absolute dispersion.

MAD, (4.7) median absolute deviation, a robust measure of the spread of data. It is relatively unaffected by data that is not normally distributed. Reporting median +/- MAD for data that is not normally distributed is equivalent to mean +/- SD for normally distributed data.

MAD Robust: (6.968) median absolute deviation of the data points, adjusted by a factor for asymptotically normal consistency.

IQR (9.175) Interquartile Range is similar to the MAD but is less robust (see Boxplots).

Variance (48.1) is another estimate of how far the data is spread from the mean. It is also the square of the standard deviation.

Confidence intervals (CI), although not shown in the general Descriptive statistics output, they 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 ± 1.96 SD of the mean and 99% within ± 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 ± (1.96 * 0.24) or 17.71 ± 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.

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. The IQR can also be shown by ticking this option in the Dispersion menu.

JASP 0.16.1 - Dr Mark Goss-Sampson

DISTRIBUTION

Descriptive Statistics	
	Variable
Skewness	−0.004
Std. Error of Skewness	0.086
Kurtosis	−0.410
Std. Error of Kurtosis	0.172
Shapiro-Wilk	0.996
P-value of Shapiro-Wilk	0.032
Sum	14346.200

Distribution
- ✓ Skewness
- ✓ Kurtosis
- ✓ Shapiro-Wilk test
- ✓ Sum

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.

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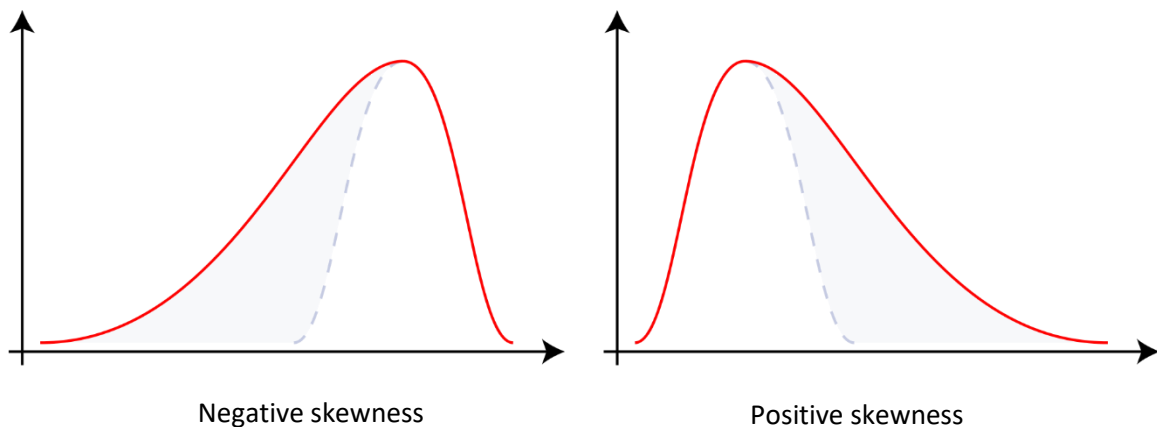
Kurtosis describes how heavy or light the tails are. Positive kurtosis results in an increase in the "pointiness" of the distribution with heavy (longer) tails while negative kurtosis exhibit a much more uniform or flatter distribution with light (shorter) tails.

+ kurtosis
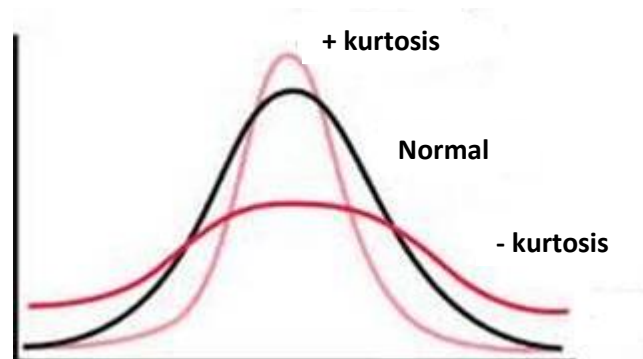
Normal

- kurtosis

In the Statistics options make sure that everything is unticked apart from skewness, kurtosis and Shapiro-Wilk test.

JASP 0.16.1 - Dr Mark Goss-Sampson

We can use the Descriptives output to calculate skewness and kurtosis. For a normal data distribution, both values should be close to zero. The Shapiro-Wilk test is used to assess if the data is significantly different from a normal distribution. (see - Exploring data integrity in JASP for more details).

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.

Descriptive Statistics

	Variable	
	Group 1	Group 2
Mean	16.021	18.787
Std. Error of Mean	0.362	0.316
Std. Deviation	6.424	7.040
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
Shapiro-Wilk	0.993	0.993
P-value of Shapiro-Wilk	0.119	0.029

JASP 0.16.1 - Dr Mark Goss-Sampson

DESCRIPTIVE DATA VISUALISATION

JASP produces a comprehensive range of descriptive and analysis specific plots. The analysis specific plots will be explained in their relevant chapters

BASIC PLOTS

Firstly, to look at examples of the basic plots, open **Descriptive data.csv** with the variable data in the Variables box, go to Plots and tick Distribution plots, Display density, Interval plots, Q-Q plots, and dot plots.

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.

JASP 0.16.1 - Dr Mark Goss-Sampson
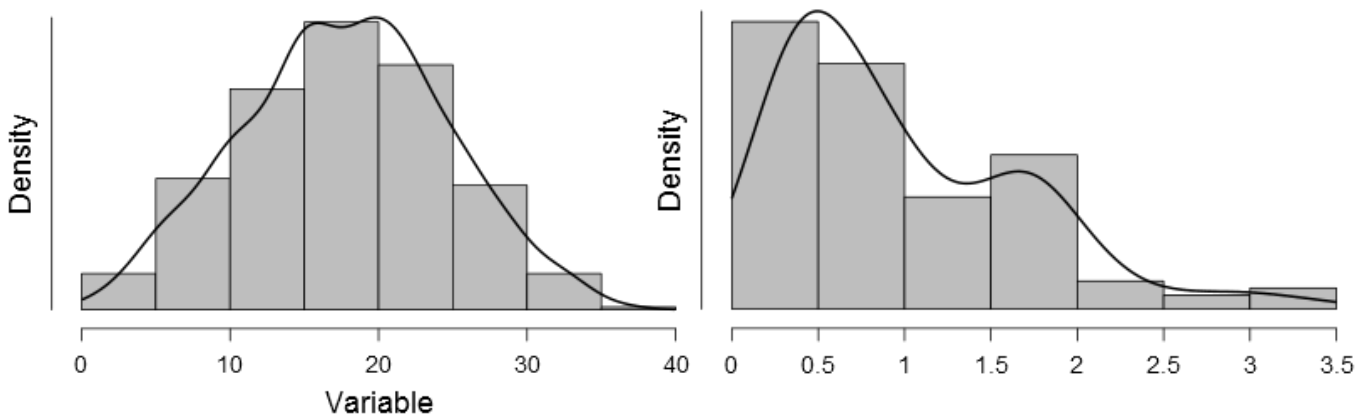
The dot plot displays the distribution where each dot represents a value. If a value occurs more than once, the dots are placed one above the other so that the height of the column of dots represents the frequency for that value.

The interval plot shows a 95% confidence interval for the mean of each variable.
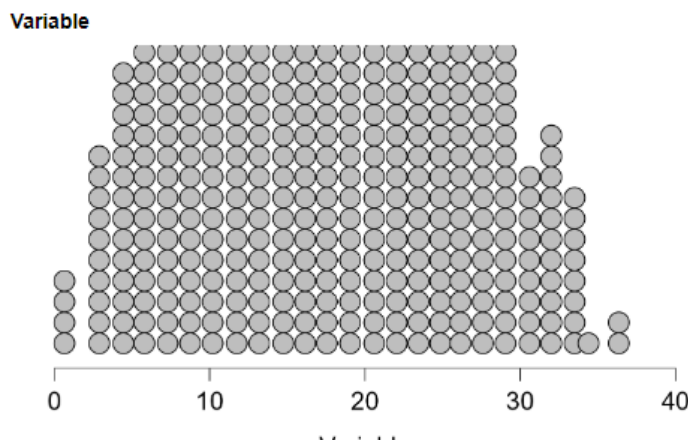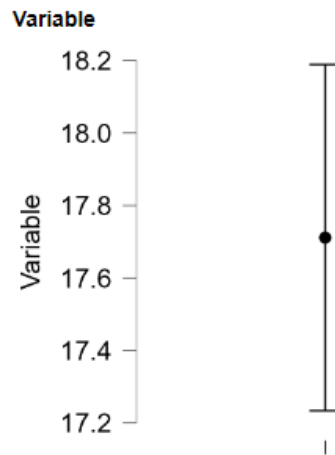
The Q-Q plot (quantile-quantile plot) can be used to visually assess if a set of data comes from a normal distribution. Q-Q plots take the sample data, sort it in ascending order, and then plot them against quantiles (percentiles) calculated from a theoretical distribution. If the data is normally distributed, the points will fall on or close to the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.

JASP 0.16.1 - Dr Mark Goss-Sampson
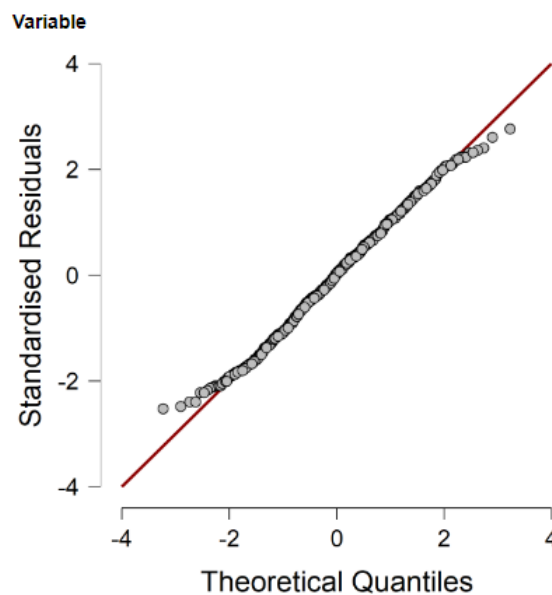
Depending on the data sets, basic correlation graphs and pie charts for non-scale data can also be produced.

Surface
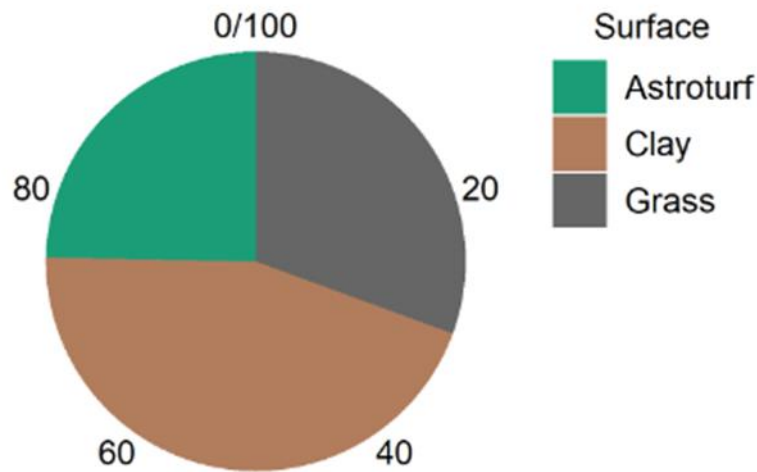
CUSTOMISABLE PLOTS

There are a variety of options depending on your datasets.

JASP 0.16.1 - Dr Mark Goss-Sampson

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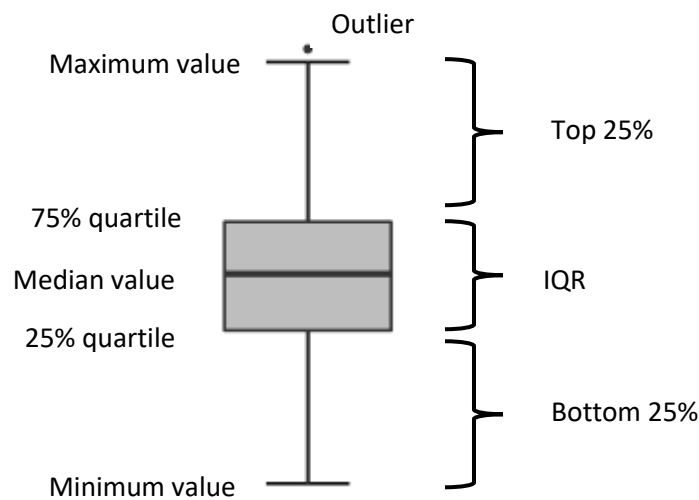
Go back to the statistics options, in Descriptive plots tick both Boxplot and Violin Element, look at how the plot has changed. Next tick Boxplot, Violin and Jitter Elements. The Violin plot has taken the smoothed distribution curve from the Distribution plot, rotated it 90° and superimposed it on the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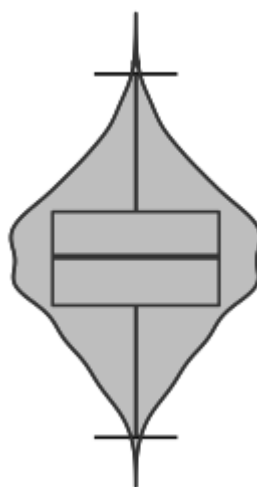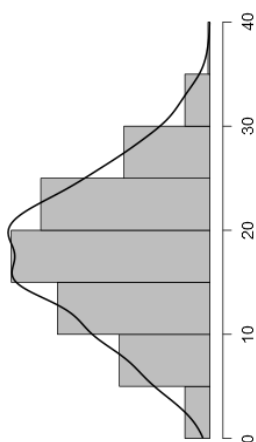
Boxplot + Violin plot   Boxplot + Violin + Jitter plot

JASP 0.16.1 - Dr Mark Goss-Sampson

If your data is split by group, for example, the boxplots for each group will be shown on the same graph, the colours of each will be different if the Colour palette is ticked. 5 colour palettes are available.
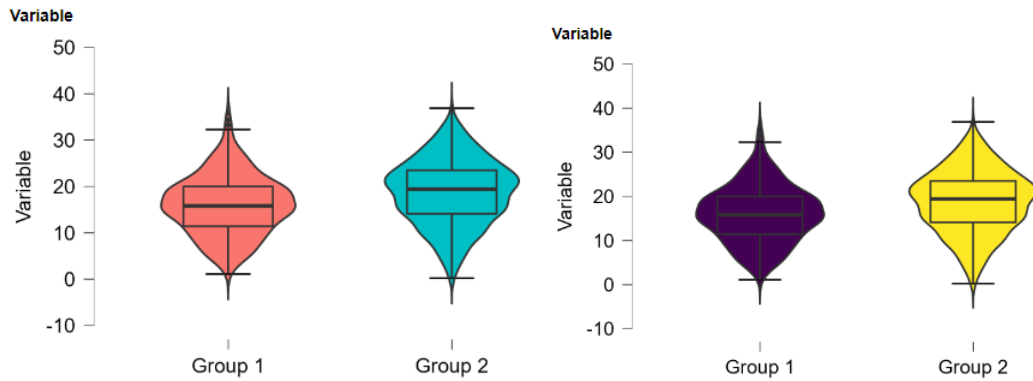
Ggplot2 palette   **Viridis pallette**

Scatter Plots

JASP can produce scatterplots of various types and can include smooth or linear regression lines. There are also options to add distributions to these either in the form of density plots or histograms.

JASP 0.16.1 - Dr Mark Goss-Sampson
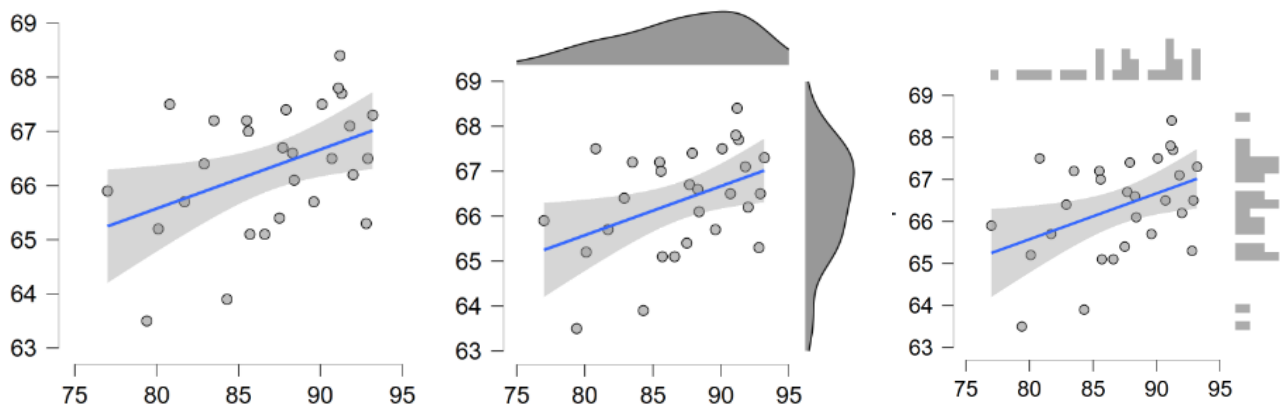
Tile Heatmap

These plots provide an alternative way of visualising data. For example, using the titanic survival dataset to look at the relationship between the class of passage and survival.

JASP 0.16.1 - Dr Mark Goss-Sampson
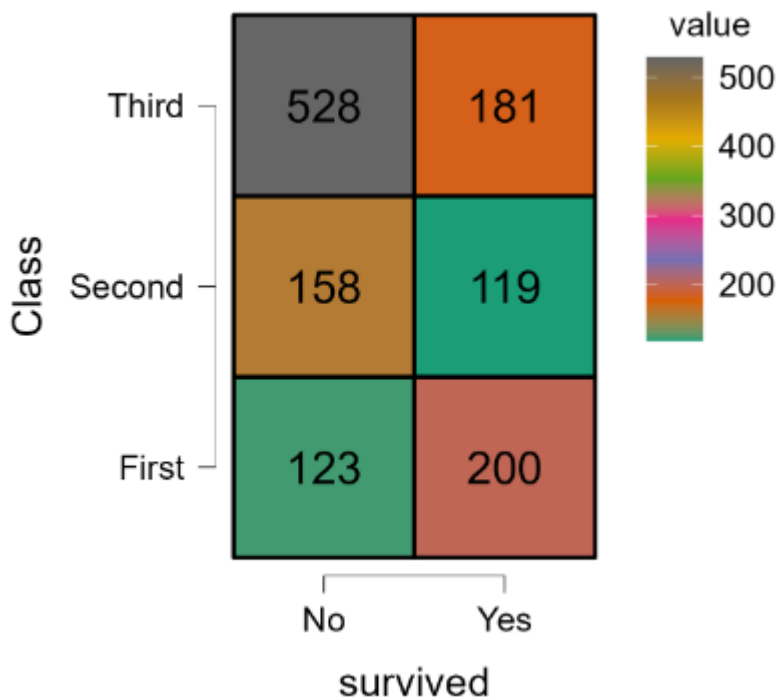
EDITING PLOTS

Clicking on the drop-down menu provided access to a range of options including Edit Image. Selecting this option provides some customisation for each graph.
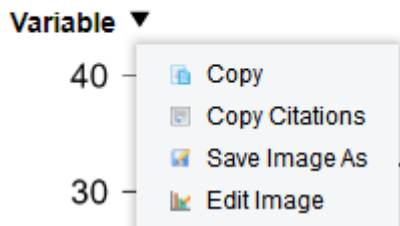
This will open the plot in a new window which allows some modifications of each axis in terms of axis title and range.
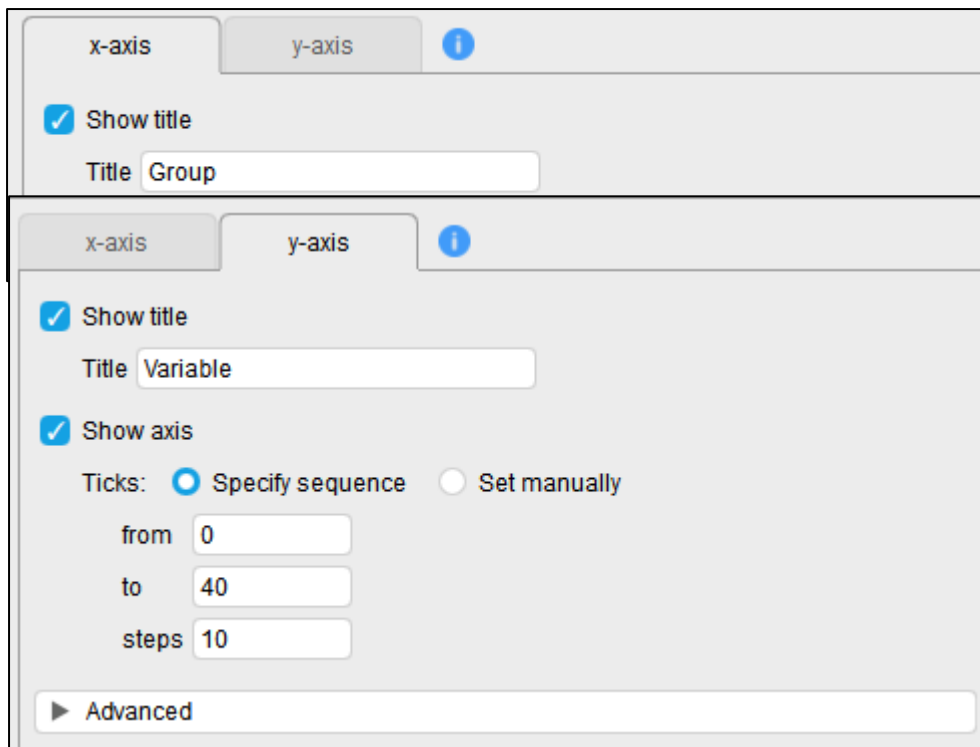
Any changes are then updated in the results window. The new plot can be saved as an image or can be reset to default values.

Do not forget that group labels can be changed in the spreadsheet editor.

JASP 0.16.1 - Dr Mark Goss-Sampson
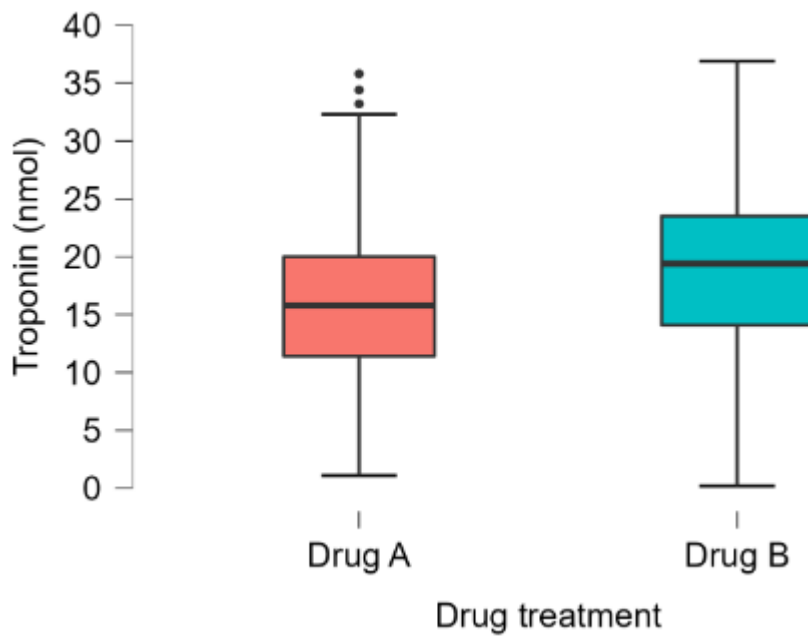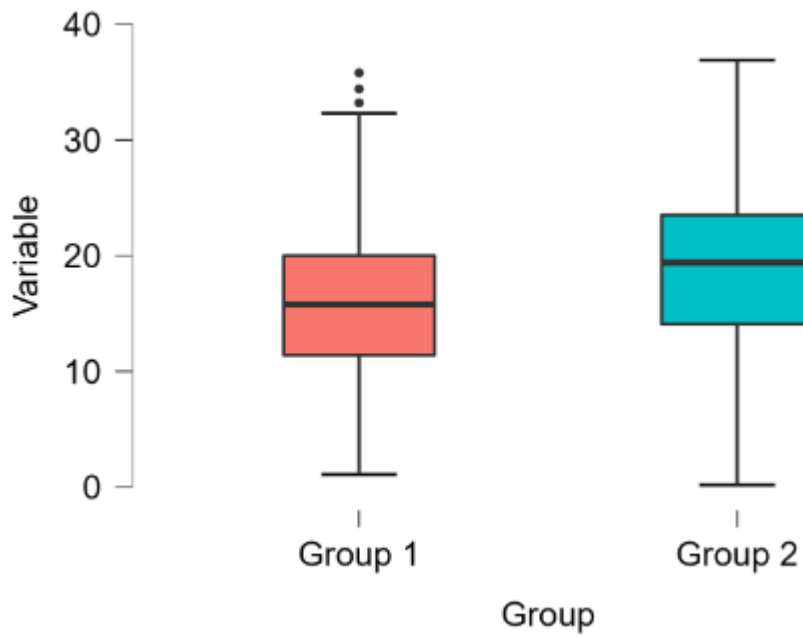
EXPLORING DATA INTEGRITY

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 underestimate 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
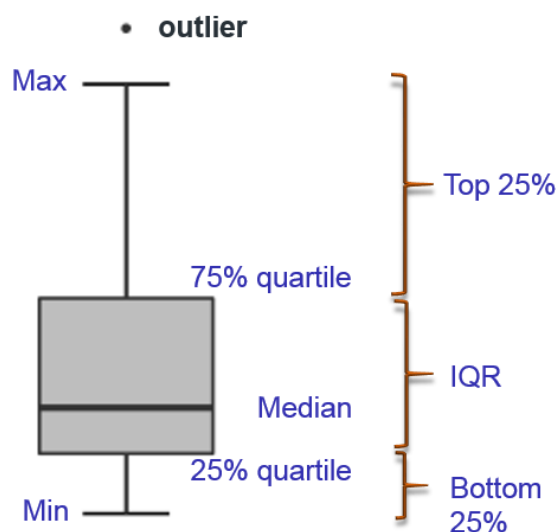
Boxplots show:

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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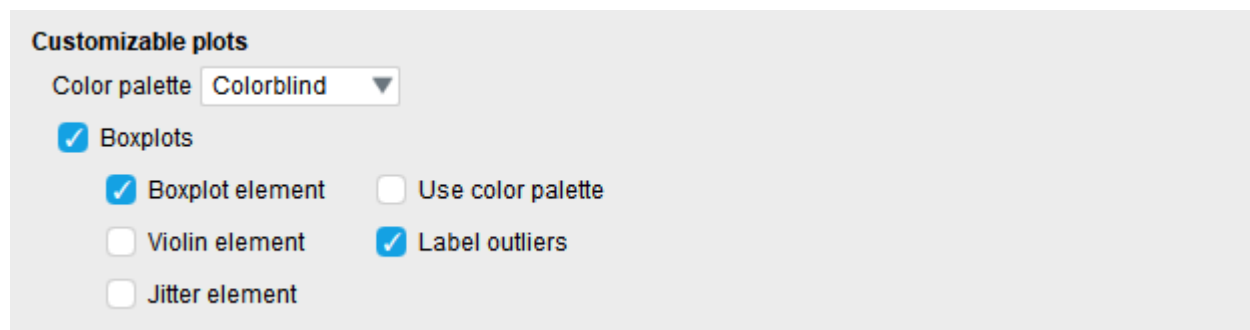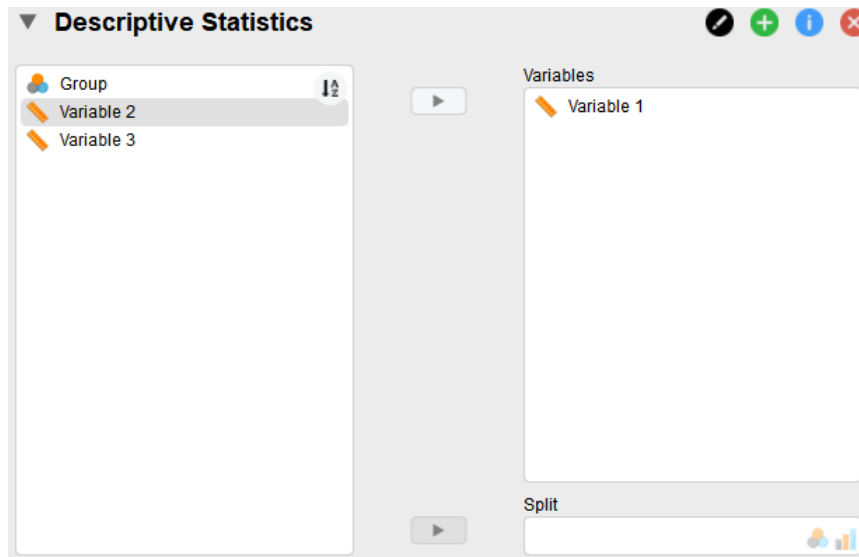
The resulting Boxplot on the left looks very compressed and an obvious outlier is labelled as being in row 38 of the dataset. This can be traced back to a data input error in which 91.7 was input instead of 917. The graph on the right shows the BoxPlot for the 'clean' data.

JASP 0.16.1 - Dr Mark Goss-Sampson
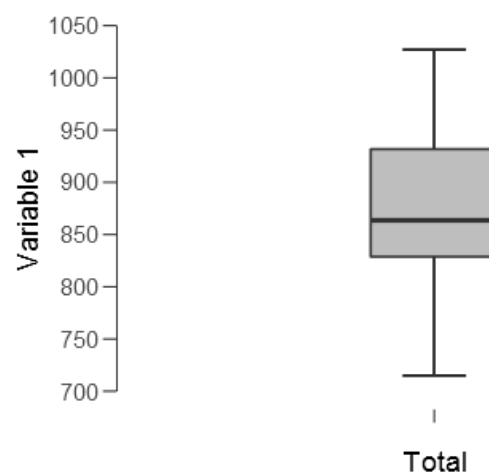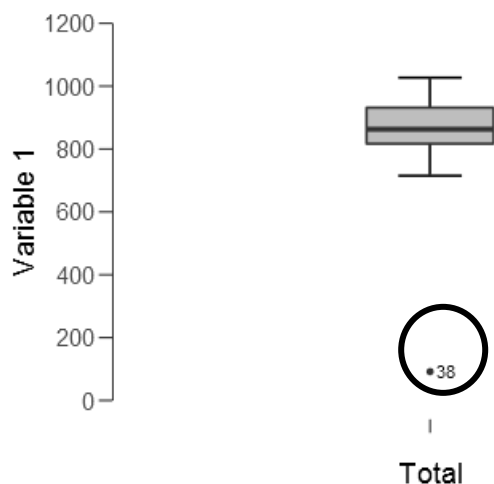
How you deal with an outlier depends on the cause. Most parametric tests are highly sensitive to outliers while non-parametric tests are generally not.

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 they 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 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:

$$\text{Skewness}_z = \frac{skewness}{Skewness\ standard\ error} \qquad \text{Kurtosis}_z = \frac{kurtosis}{kurtosis\ standard\ error}$$

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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.

Descriptive Statistics

	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

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 $p<0.05$.

Skewness Z = $\frac{0.839}{0.337}$ = 2.49   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:
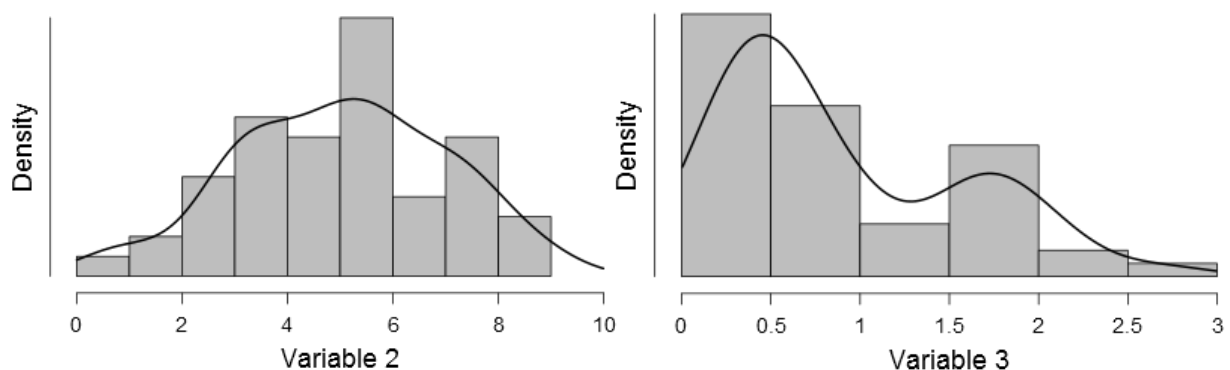
It is quite easy to visualise that Variable 2 has a symmetrical distribution. Variable 3 is skewed to the left as confirmed by the skewness Z score.

JASP 0.16.1 - Dr Mark Goss-Sampson

Another graphical check for normality is a Q-Q plot. Q-Q plots are available in Descriptives and are also 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.
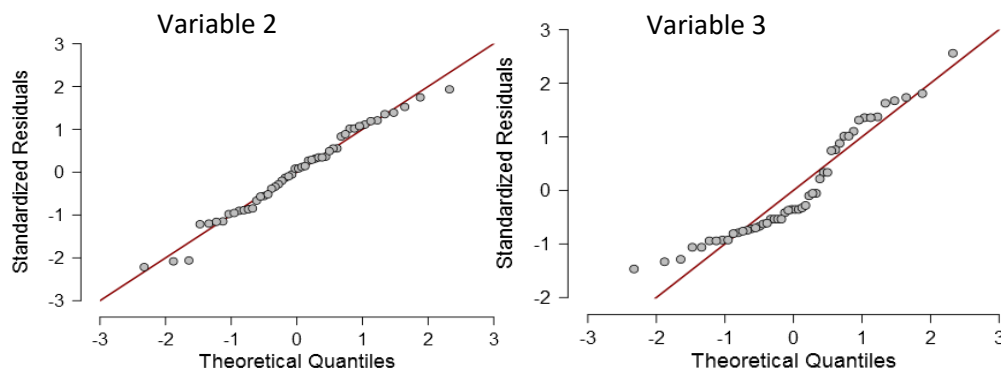
The following Q-Q plot scenarios are possible:

JASP 0.16.1 - Dr Mark Goss-Sampson
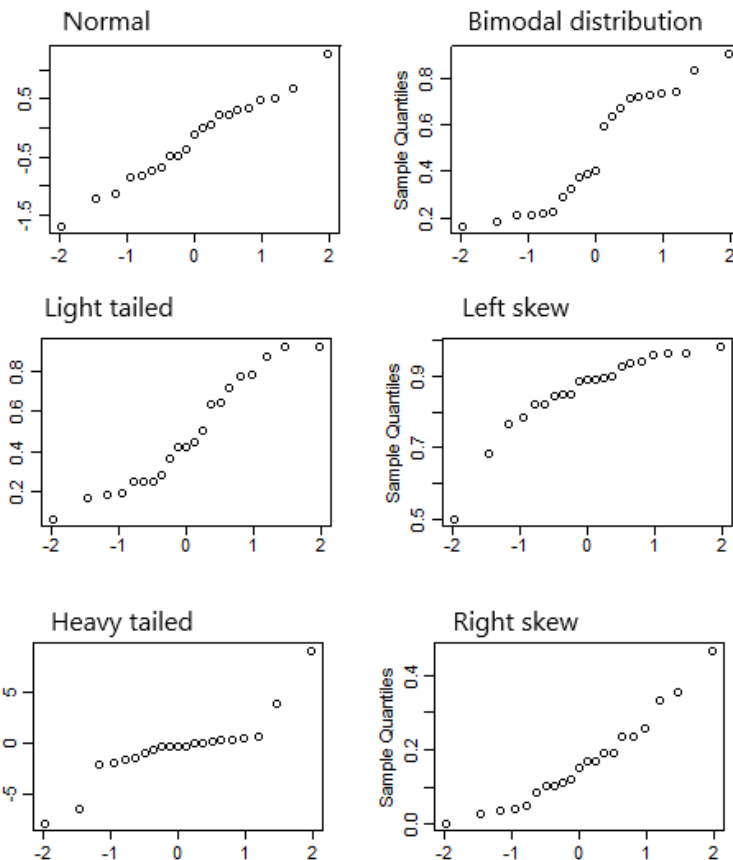
Shapiro-Wilk test is a statistical way used by JASP to check the assumption of normality. It is also used in the **Independent** (distribution of the two groups) and **Paired** (distribution of differences between pairs) **t-tests**. The test results in a W value; where small values indicate your sample 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).

Descriptive Statistics

	Variable 2	Variable 3
Valid	50	50
Missing	0	0
Shapiro-Wilk	0.982	0.885
P-value of Shapiro-Wilk	0.650	< .001

In Descriptives, the Shapiro-Wilk test can be selected in the Distribution tests. The Shapiro-Wilk output table shows no significant deviation in normality for Variable 2 but a significant deviation ($p<.001$) for Variable 3.

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 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.

JASP 0.16.1 - Dr Mark Goss-Sampson

WHAT DO I DO IF MY DATA IS 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. Please refer to the next chapter.

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.

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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.

JASP 0.16.1 - Dr Mark Goss-Sampson
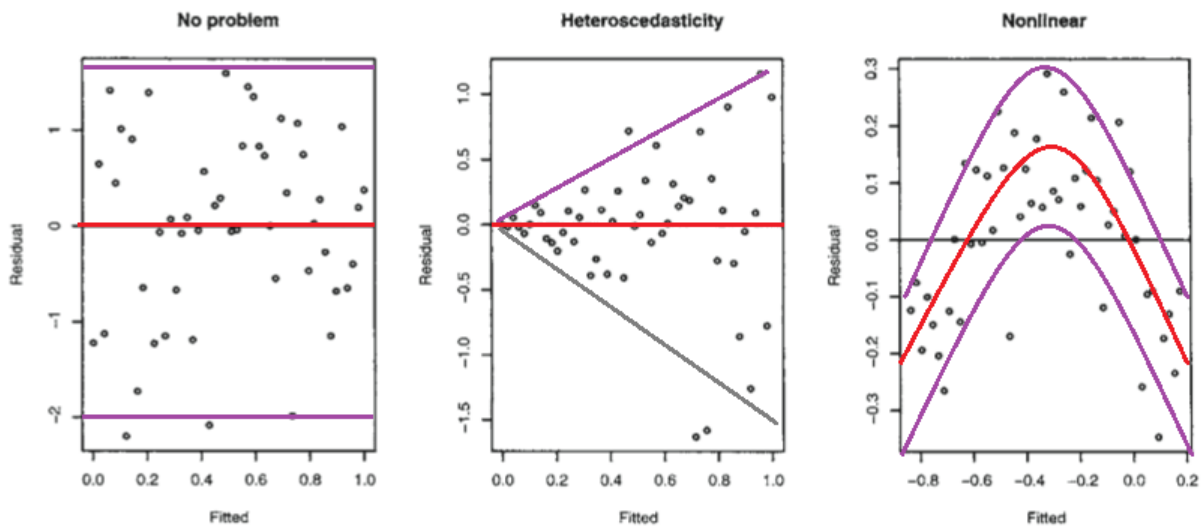
DATA TRANSFORMATION

JASP provides the ability to compute new variables or transform data. In some cases, it may be useful to compute the differences between repeated measures or, to make a dataset more normally distributed, you can apply a log transform for example.

When a dataset is opened there will be a plus sign (✚) at the end of the columns.

▼	Group	Variable 1	Variable 2	Variable 3	✚
1	1	912	2.78	0.29	
2	1	826	4.89	0.55	
3	1	1004	6.79	0.47	
4	1	982	6.24	1.58	
5	1	920	8.59	0.76	
6	1	814	5.86	0.76	

Clicking on the ✚ opens up a small dialogue window where you can;

- Enter the name of a new variable or the transformed variable
- Select whether you enter the R code directly or use the commands built into JASP
- Select what data type is required

Create Computed Column

Name: [...]

[R] [👆]

[✎ Scale] [Ordinal] [Nominal] [Text]

[ⓘ] [Create Column] [✖]

Once you have named the new variable and chosen the other options – click create.

If you choose the manual option rather than the R code, this opens all the built-in create and transform options. Although not obvious, you can scroll the left and right-hand options to see more variables or more operators respectively.
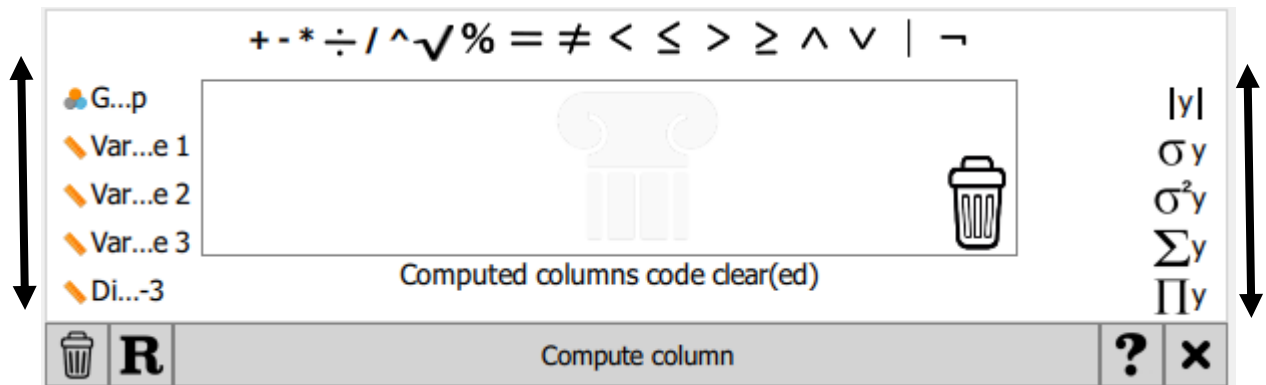
For example, we want to create a column of data showing the difference between variable 2 and variable 3. Once you have entered the column name in the Create Computed Column dialogue window, its name will appear in the spreadsheet window. The mathematical operation now needs to be defined. In this case drag variable 2 into the equation box, drag the 'minus' sign down and then drag in variable 3.

JASP 0.16.1 - Dr Mark Goss-Sampson

If you have made a mistake, i.e. used the wrong variable or operator, remove it by dragging the item into the dustbin in the bottom right corner.

When you are happy with the equation/operation, click compute column and the data will be entered.

If you decide that you do not want to keep the derived data, you can remove the column by clicking the other dustbin icon next to the **R**.

Another example is to do a log transformation of the data. In the following case variable 1 has been transformed by scrolling the operators on the left and selecting the log10(y) option. Replace the "y" with the variable that you want to transform and then click Compute column. When finished, click the **X** to close the dialogue.

JASP 0.16.1 - Dr Mark Goss-Sampson

The Export function will also export any new data variables that have been created.

The two graphs below show the untransformed and the log10 transformed data. The skewed data has been transformed into a profile with a more normal distribution

JASP 0.16.1 - Dr Mark Goss-Sampson
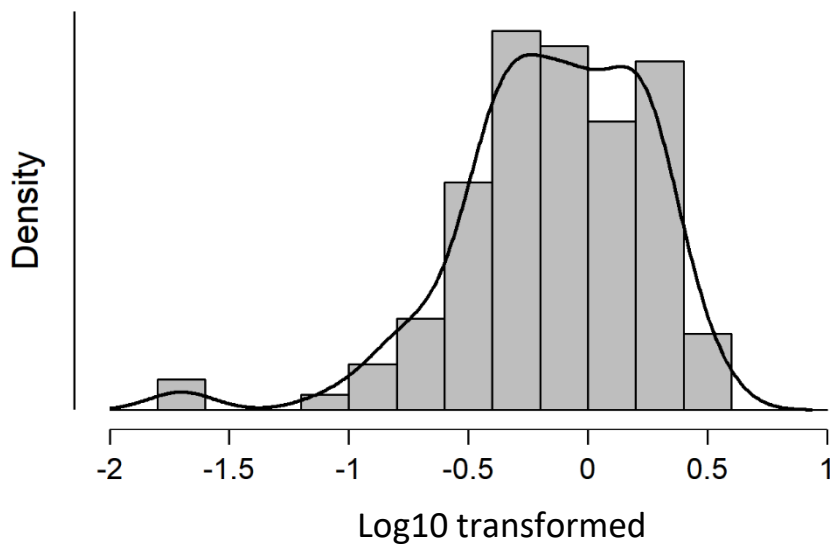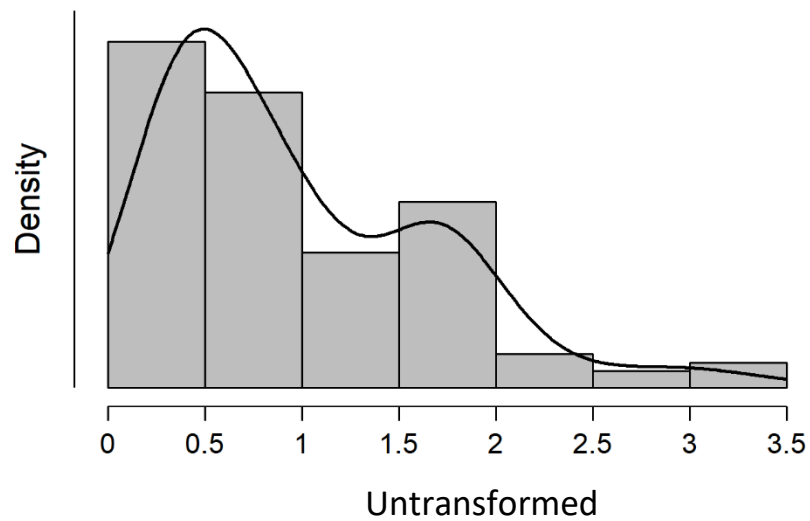
EFFECT SIZE

When performing a hypothesis test on data we determine the relevant statistic (r, t, F etc) and p-value to decide whether to accept or reject the null hypothesis. A small p-value, <0.05 in most analyses provides evidence against the null hypothesis whereas a large p-value >0.05 only means that there is not enough evidence to reject the null hypothesis. A lower p-value is sometimes incorrectly interpreted as meaning there is a stronger relationship of difference between variables. So what is needed is not just null hypothesis testing but also a method of determining precisely how large the effects seen in the data are.
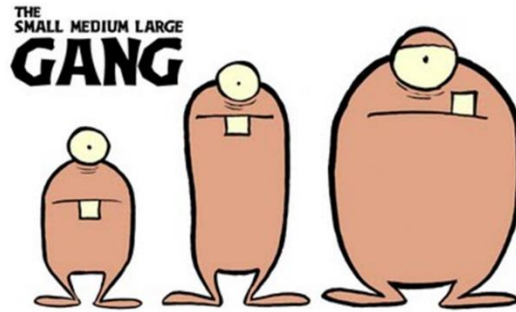
An effect size is a statistical measure used to determine the strength of the relationship or difference between variables. Unlike a p-value, effect sizes can be used to quantitatively compare the results of different studies.

For example, comparing heights between 11 and 12-year-old children may show that the 12-year-olds are significantly taller but it is difficult to visually see a difference i.e. small effect size. However, a significant difference in height between 11 and 16-year-old children is obvious to see (large effect size).

The effect size is usually measured in three ways:

- the standardized mean difference
- correlation coefficient
- odds ratio

When looking at differences between groups most techniques are primarily based on the differences between the means divided by the average standard deviations. The values derived can then be used to describe the magnitude of the differences. The effect sizes calculated in JASP for t-tests and ANOVA are shown below:

Test	Measure	Trivial	Small	Medium	Large
Between means **Parametric**	Cohen's d	<0.2	0.2	0.5	0.8
	Hedge's g	<0.2	0.2	0.5	0.8
Non-parametric	Rank -biserial (r_B)	<0.1	0.1	0.3	0.5
ANOVA	Eta squared (η^2)	<0.1	0.1	0.25	0.37
	Partial Eta squared (η^2_p)	<0.01	0.01	0.06	0.14
	Omega squared (ω^2)	<0.01	0.01	0.06	0.14

JASP 0.16.1 - Dr Mark Goss-Sampson

When analysing bivariate or multivariate relationships the effect sizes are the correlation coefficients:

Test	Measure	Trivial	Small	Medium	Large
Correlation	Correlation coefficient (r) Spearman's rho Kendall's tau	<0.1 <0.1 <0.1	0.1 0.1 0.1	0.3 0.3 0.3	0.5 0.5 0.5
Multiple Regression	Multiple correlation coefficient (R)	<0.1	0.1	0.3	0.5

When analysing categorical relationships via contingency tables i.e. chi-square test Phi is only used for 2x2 tables while Cramer's V and be used for any table size.

Contingency tables	df	Small	Medium	Large
Phi (2x2 only) Cramer's V	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

For a 2 × 2 contingency table, we can also define the odds ratio measure of effect size.

JASP 0.16.1 - Dr Mark Goss-Sampson

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$_o$) 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 **83.4 kg**.
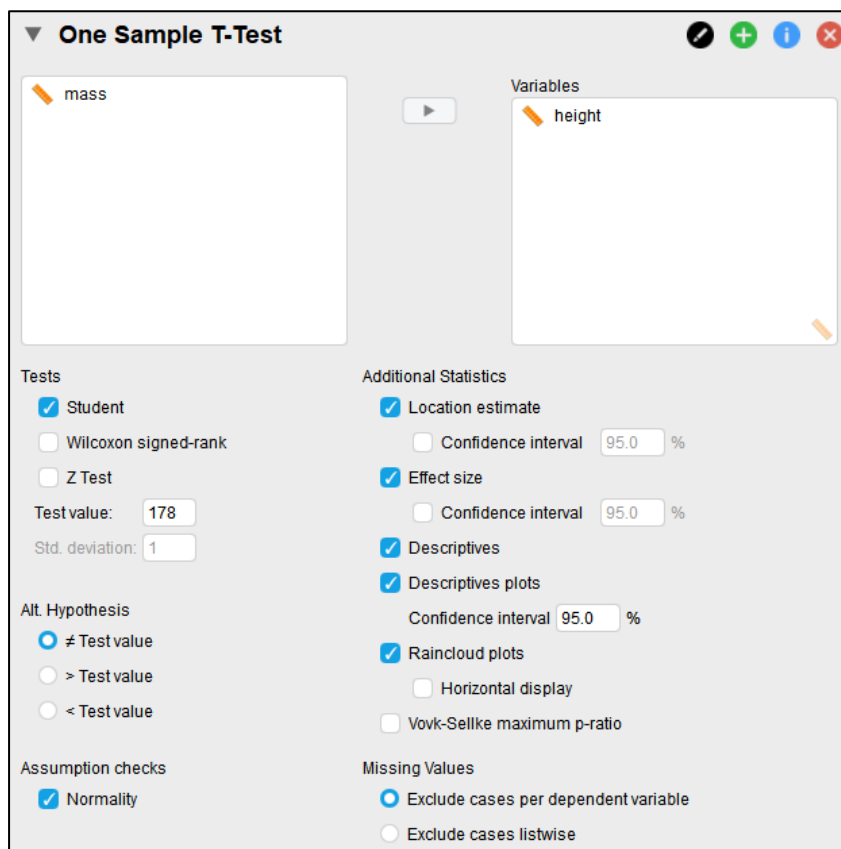
Go to T-Tests > One-Sample t-test and in the first instance add height to the analysis box on the right. Then tick the following options above and add **178** as the test value:

JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

The output should contain three tables and two graphs.

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.

JASP 0.16.1 - Dr Mark Goss-Sampson

The two plots show essentially the same data but in different ways. The standard Descriptive plot is an Interval plot showing the sample mean (black bullet), the 95% confidence interval (whiskers), relative to the test value (dashed line).
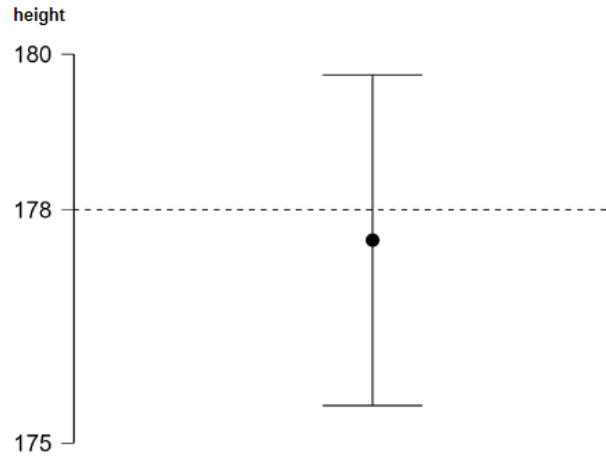
The Raincloud Plot shows the data as individual data points, boxplot, and the distribution plot. This can be shown as either a vertical or horizontal display.

JASP 0.16.1 - Dr Mark Goss-Sampson
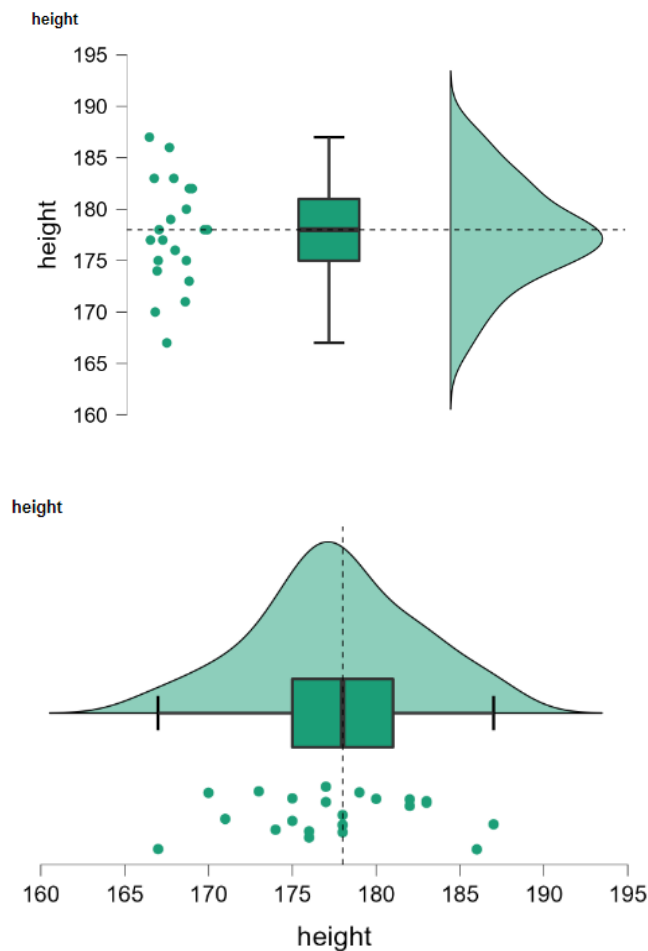
Repeat the procedure by replacing height with mass and changing the test value to 83.4.

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 (83.4 kg) $p <.001$

Descriptives

	N	Mean	SD	SE
mass	23.000	72.913	7.025	1.465
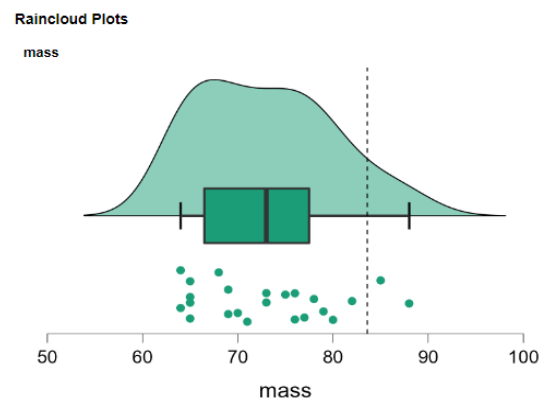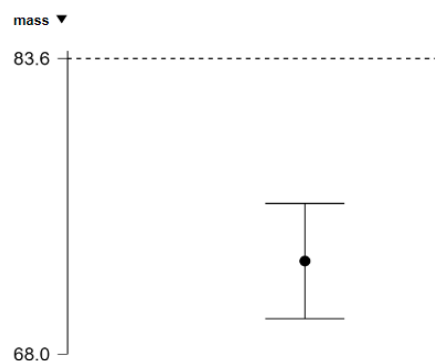
REPORTING THE RESULTS

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, $p<.001$).

JASP 0.16.1 - Dr Mark Goss-Sampson

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 if the sample frequency is statistically different from a known or hypothesized population frequency.

The null hypothesis (H$_o$) tested is that the sample data frequency is equal to the expected population frequency.

ASSUMPTIONS

Three assumptions are required for a binomial 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 2021, when comparing just the two operating systems, the relative UK market share of Windows was 72% and Mac IOS 28%.[3]

Go to Frequencies >Binomial test. Move the Laptop variable to the data window and set the Test value to 0.72 (72%). Also, tick Descriptive plots.

---

[3]https://www.statista.com/statistics/487002/market-share-of-desktop-operating-systems-uk/#:~:text=Apple%20and%20Microsoft%20own%20a,26.48%20percent%20of%20the%20market.

JASP 0.16.1 - Dr Mark Goss-Sampson
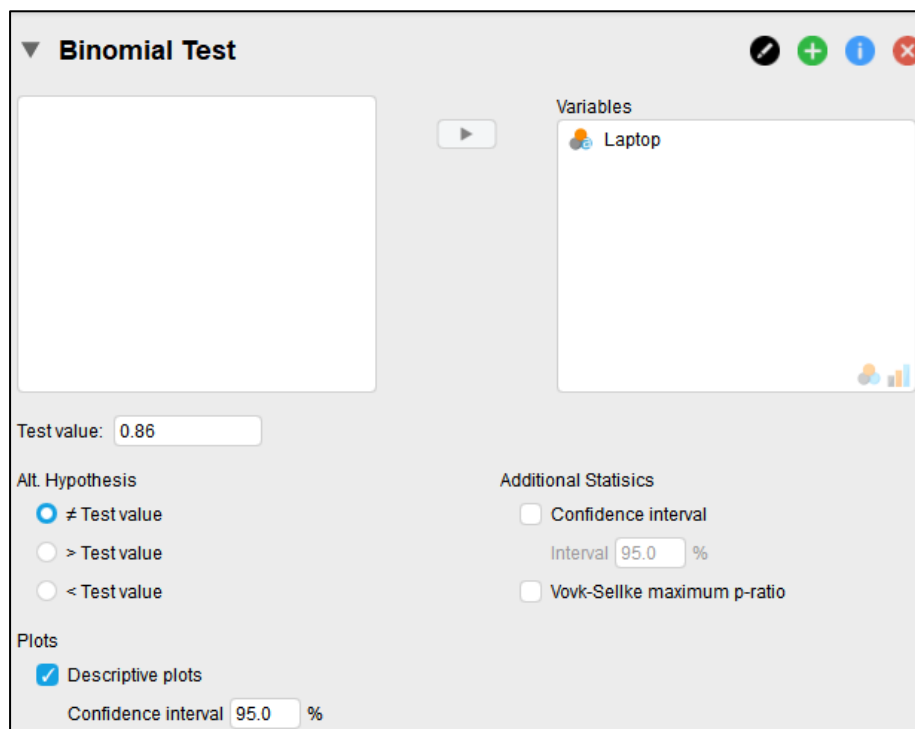
The following table and graph show that the frequencies of both laptops are significantly less than 72%. In particular, students are using significantly fewer Windows laptops than was expected compared to the UK market share.

Binomial Test

Variable	Level	Counts	Total	Proportion	p
Laptop	Mac	36	90	0.400	< .001
	Windows	54	90	0.600	0.014

Note. Proportions tested against value: 0.72.

Laptop
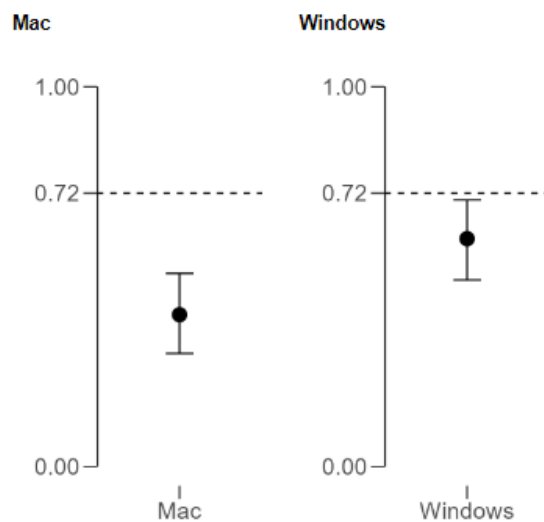
Is this the same for MacBook users? Go back to the Options window and change the test value to 0.28 (14%). This time both frequencies are significantly higher than 28%. This shows that students are using significantly more MacBooks than was expected compared to the UK market share.

Binomial Test

Variable	Level	Counts	Total	Proportion	p
Laptop	Mac	36	90	0.400	0.014
	Windows	54	90	0.600	< .001

Note. Proportions tested against value: 0.28.

JASP 0.16.1 - Dr Mark Goss-Sampson
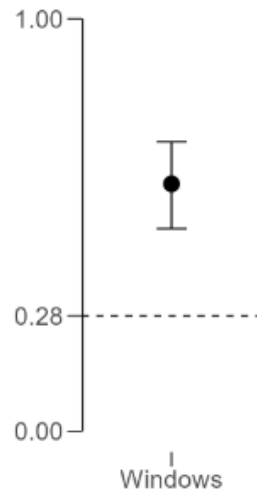
Laptop

REPORTING THE RESULTS

The UK relative proportion of Windows and MacBook users was reported to be 72% and 28% respectively. In a cohort of University students (N=90), a Binomial test revealed that the proportion of students using Windows laptops was significantly less (60%, *p*=.014) and those using MacBooks significantly more (40%, *p*<.001) than expected.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 a known (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 the 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) from an equal distribution.

Multinomial Test

	χ^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
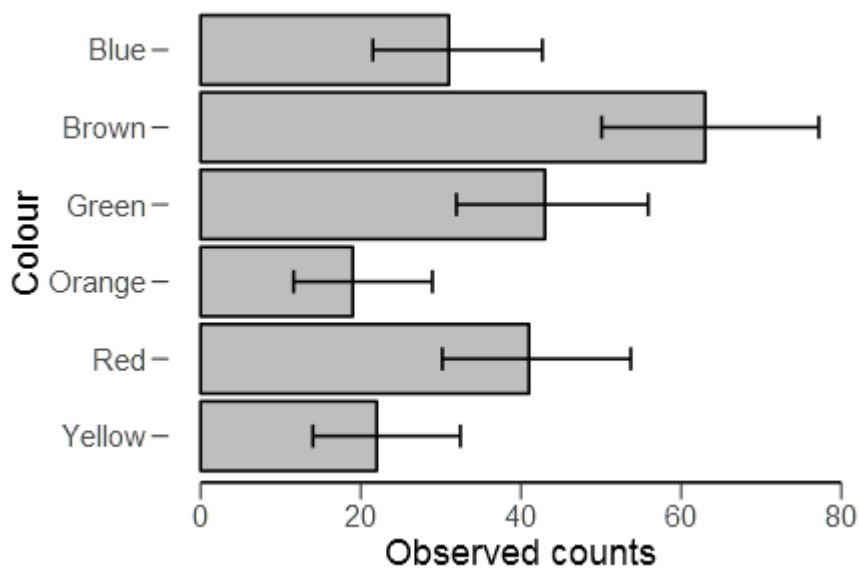
REPORTING THE RESULTS

A multinomial test showed that there was a significant difference between the observed and expected M&M colour distribution: *χ2* (5) = 35.9, *p*<.001.

JASP 0.16.1 - Dr Mark Goss-Sampson

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

However, further research shows that the manufacturer produces 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 ($\chi 2 = 74.5$, $p<.001$) from those proportions stated by the manufacturer.

Multinomial Test

	χ^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

JASP 0.16.1 - Dr Mark Goss-Sampson
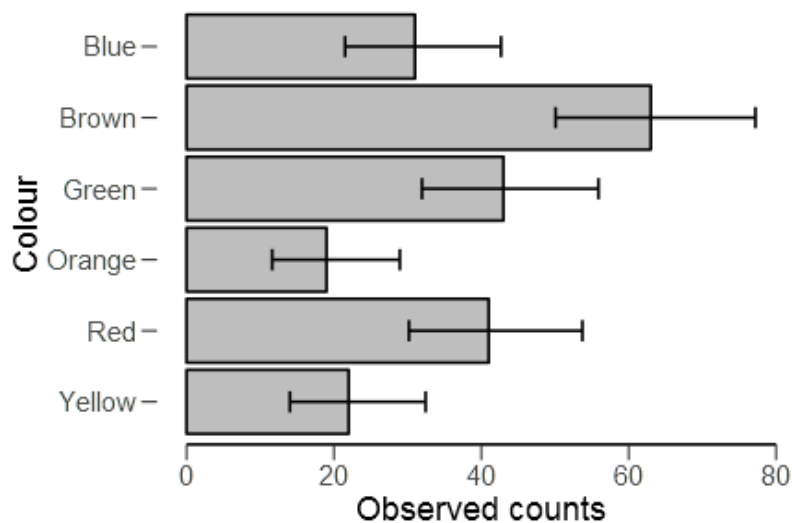
MULTINOMIAL AND X^2 'GOODNESS-OF-FIT' TEST.

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_o (a) with each cell having 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_o (b). The expected proportions of each colour can now be typed in.

Test Values
- Equal proportions (multinomial test)
- Expected proportions (χ^2 test)

	H_o (a)	H_o (b)		
Blue	1	24		Add Column
Brown	1	13		Delete Column
Green	1	16		Reset
Orange	1	20		
Red	1	13		
Yellow	1	14		

Now when the analysis is run, the results of the tests for the two hypotheses are shown. H_o (a) is testing the null hypothesis that the proportions of each colour are equally distributed, while H_o (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 M&M's do not match the manufacturer's published proportions.

Multinomial Test

	χ^2	df	p
H_o (a)	35.932	5	< .001
H_o (b)	74.535	5	< .001

Descriptives

		Expected	
Colour	Observed	H_o (a)	H_o (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

JASP 0.16.1 - Dr Mark Goss-Sampson

Reporting the results

A $\chi 2$ 'goodness of fit' test was carried out to assess the distributions of M&M's in retail packets. This showed that firstly, the proportions of each colour are not equally distributed, $\chi 2$ (5) = 35.9, *p*<.001. Secondly, the distribution of colours was significantly different to the manufacturer's stated distribution, $\chi 2$ (5) = 74.5, *p*<.001.

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 ratio of the differences between the two groups and the differences within the two groups:

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

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

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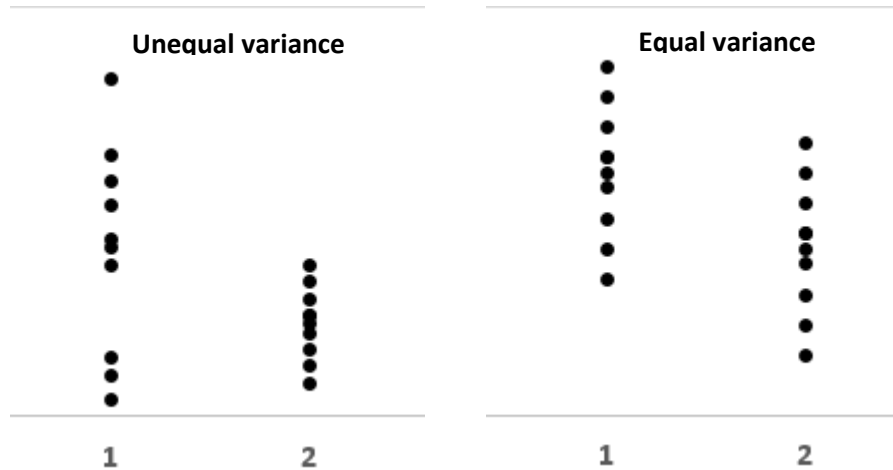
If Levene's test is statistically significant, indicating that the group variances are unequal we can correct for this violation by using an adjusted t-statistic based on the **Welch** method.

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.

JASP 0.16.1 - Dr Mark Goss-Sampson
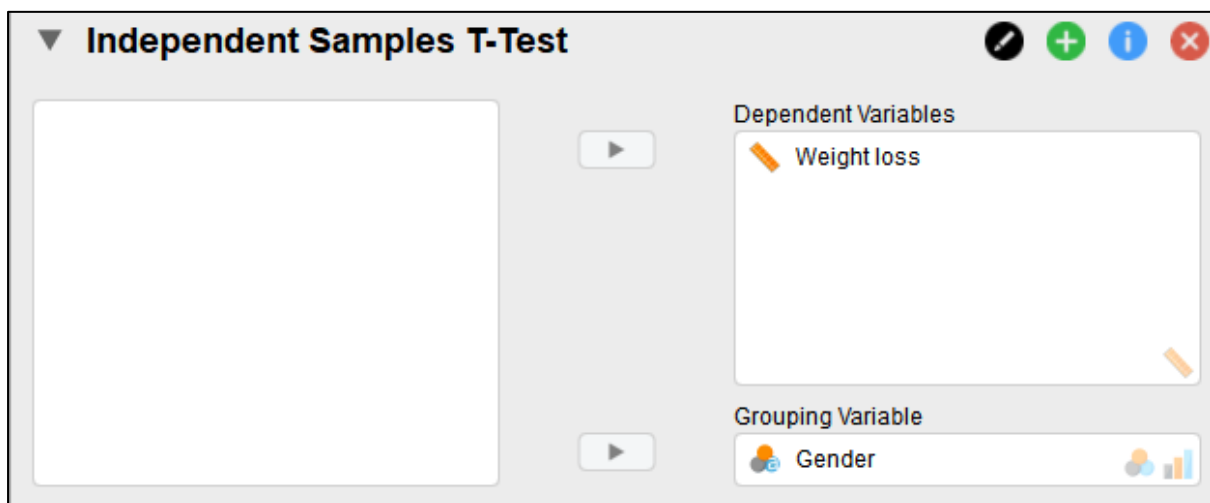
In the analysis window tick the following options:

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.

Test of Normality (Shapiro-Wilk)

		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, therefore, the assumption of normality is not violated. If one or both were significant you should consider using the non-parametric equivalent **Mann-Whitney** test.

JASP 0.16.1 - Dr Mark Goss-Sampson

Test of Equality of Variances (Levene's)

	F	df	p
Weight loss	2.278	1	0.135

Levene's test shows that there is no difference in the 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
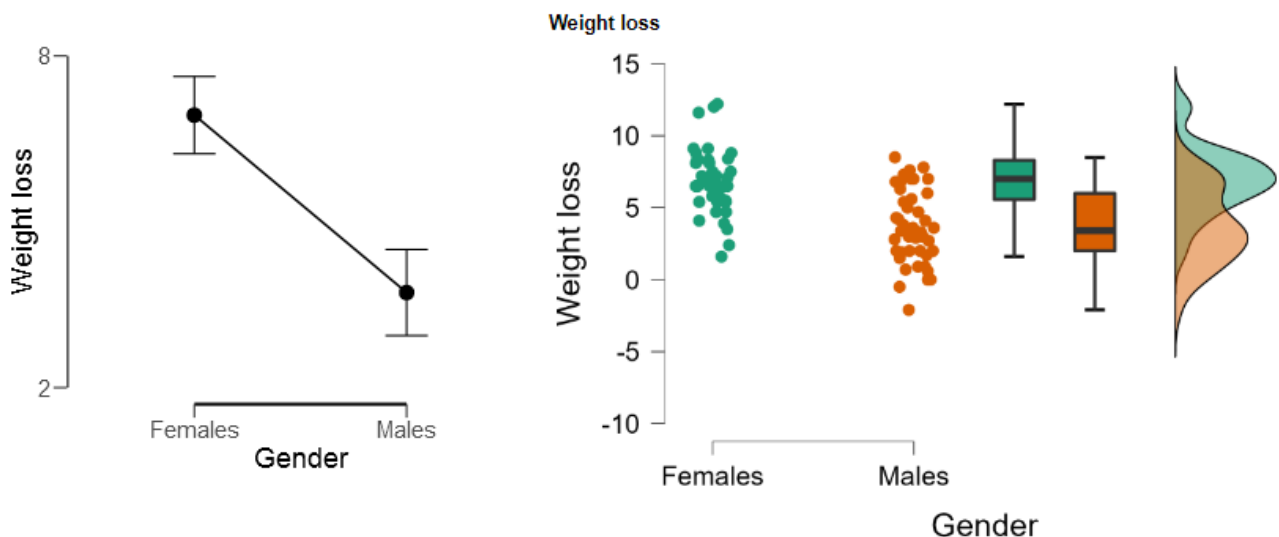
From the descriptive data and plots, it can be seen that females had a higher weight loss than males.

JASP 0.16.1 - Dr Mark Goss-Sampson

REPORTING THE RESULTS

A Levene test found that the assumption of homogeneity of variance was met: $F(1) = 2.28$, $p=.135$. A Shapiro-Wilk test shows that both groups have normally distributed data: Females ($W=0.97$, $p=.282$), Males ($W=0.97$, $p=0.31$). therefore an independent samples, Student t-test based on equal variances was carried out. This showed that females lost significantly more weight ($M = 6.93$, $SD = 2.24$) over 10 weeks of dieting than males ($M =3.72$, $SD = 2.56$): $t(85)=6.16$, $p<.001$. Cohen's d (1.322) suggests that this is a large effect.

NB if the Levene test is significant the Welch t-statistic and corrected df should be reported.

MANN-WITNEY U TEST

If you find that your data is not normally distributed (such as a 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 the 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. **W=207, p<.001**.

The location parameter, the Hodges–Lehmann estimate, is a measure of 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 the Pain score into the variable box and Split the file by Treatment.

You can alternatively show the Raincloud plots for more information on the differences and distributions.

Descriptive Statistics

	Pain score	
	Control	Ice
Median	7.000	3.000
Std. Deviation	2.167	1.512
MAD	2.000	1.000
IQR	4.000	1.500
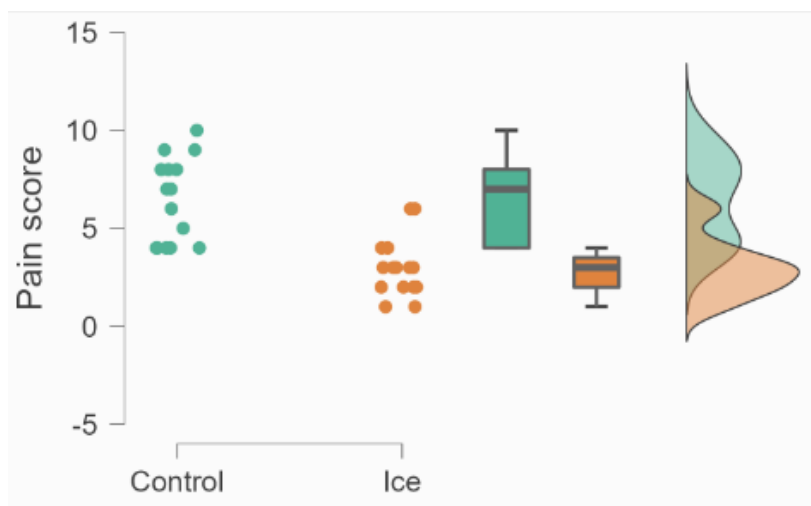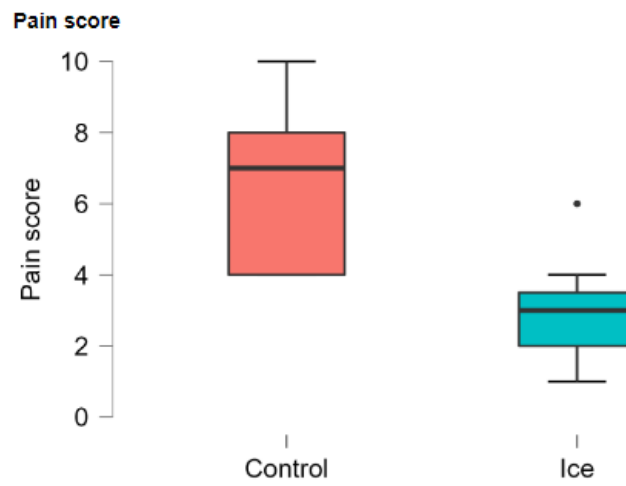
REPORTING THE RESULTS

A Mann-Whitney test showed that Ice therapy significantly reduces pain scores (*Mdn* = 3) compared to the control group (*Mdn* = 7): $W=207$, $p<.001$. The r_B= 0.84 suggests that this is a large effect size.

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 of dieting.

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

With the paired t-test, the null hypothesis (H_o) 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.
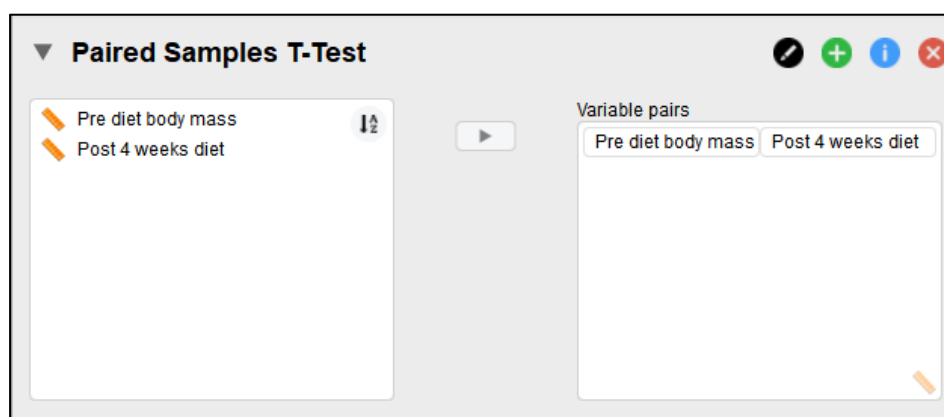
In the analysis options tick the following:

UNDERSTANDING THE OUTPUT

The output should consist of three tables and three graphs.

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.

JASP 0.16.1 - Dr Mark Goss-Sampson

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

The two Raincloud plots show either the data for both groups or the distribution of the differences between the two groups.

JASP 0.16.1 - Dr Mark Goss-Sampson
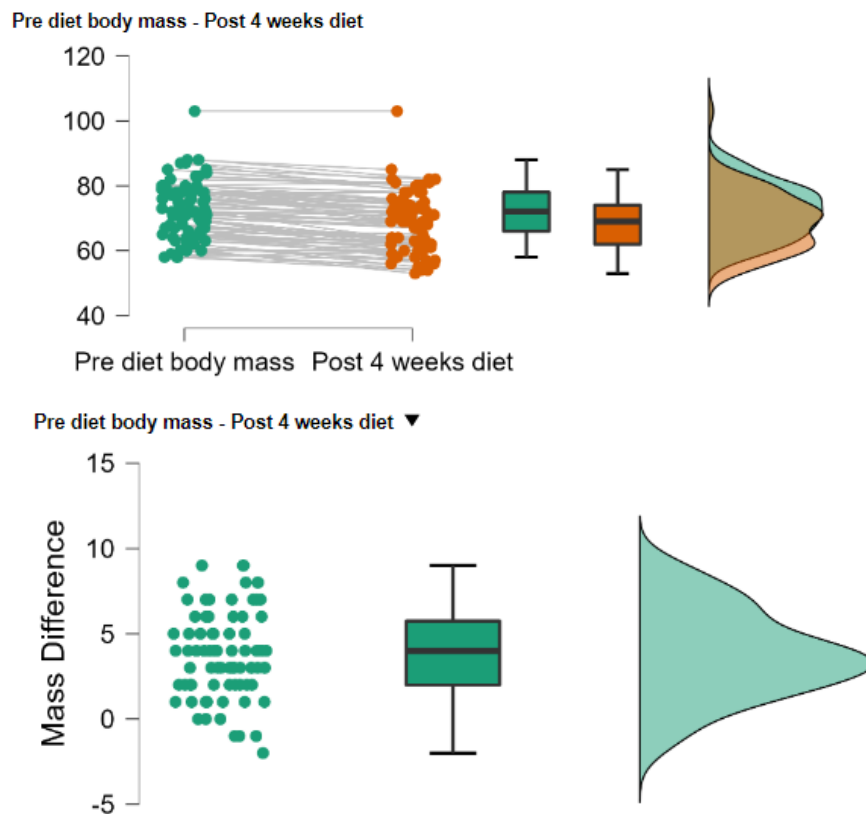
REPORTING THE RESULTS

The Shapiro-Wilk test was not significant suggesting that the pairwise differences were normally distributed, therefore the assumption is not violated: W=0.98, p = .124.

Post-diet body mass was lower (M = 68.7, SD =9.0 kg) than at the start of the diet (M =725., SD = 8.7kg). This mean difference (M = 3.78, SE = 0.29 kg) was shown to be significant: t (77) =13.04, p<.001. Cohen's d (1.48) 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 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:
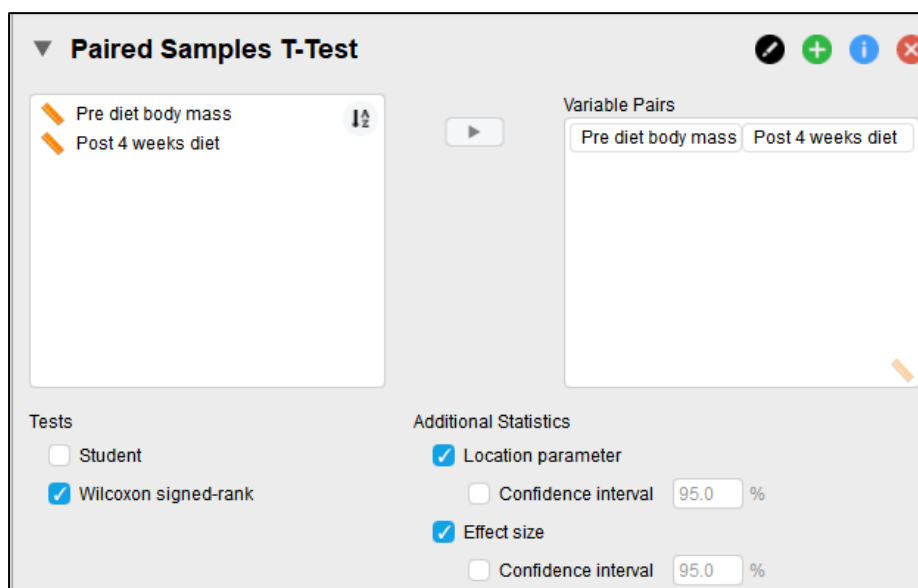
There will be only one table in the output:

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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

Descriptive Statistics

	Pre-anxiety	Post-anxiety
Valid	29	29
Missing	0	0
Median	22.000	15.000
Std. Deviation	5.766	3.307
MAD	7.413	2.965
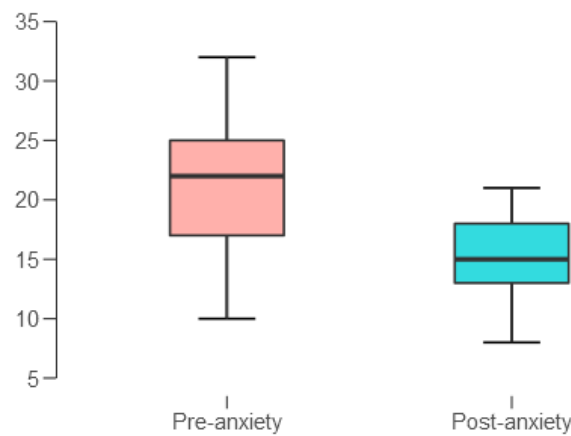
Pre diet body mass - Post 4 weeks diet
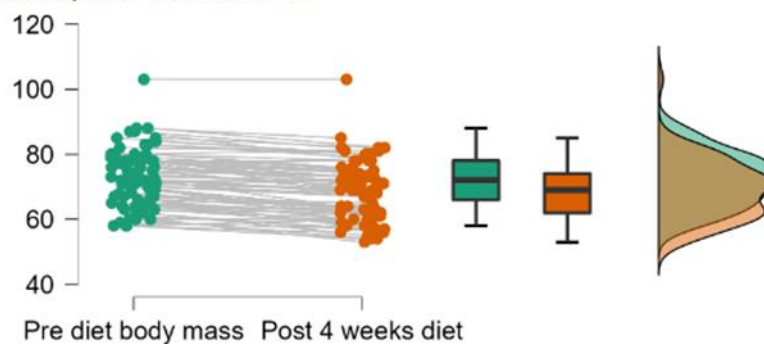
REPORTING THE RESULTS

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. The rank-biserial correlation (r_B) = 0.48 suggests that this is medium to large effect size.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 that 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.
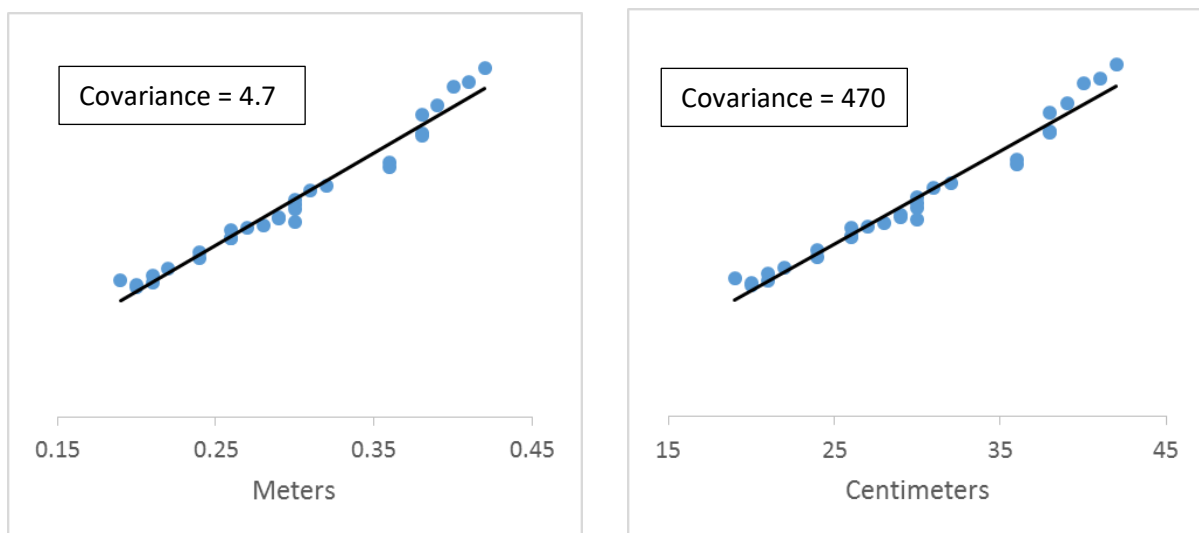
To overcome this, standardised covariance is used which is known as **Pearson's correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. If r is close to 0, there is no relationship. If r is (+) then as one variable 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 the regression analysis).

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.

JASP 0.16.1 - Dr Mark Goss-Sampson

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. Move the 2 variables to the analysis box on the right. Tick

  ✓ Pearson,
  ✓ Report significance,
  ✓ Flag significant correlations

Under Plots

  ✓ Scatter plots
  ✓ Heatmap

JASP 0.16.1 - Dr Mark Goss-Sampson

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 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 an effect size where <0.1 is trivial, 0.1 -0.3 is a small effect, 0.3 – 0.5 is a moderate effect and >0.5 is a large effect.

The plot provides a simple visualisation of this strong positive correlation ($r = 0.984$, $p<.001$) which is also highlighted by the heatmap (more relevant when looking at multiple correlations).

JASP 0.16.1 - Dr Mark Goss-Sampson
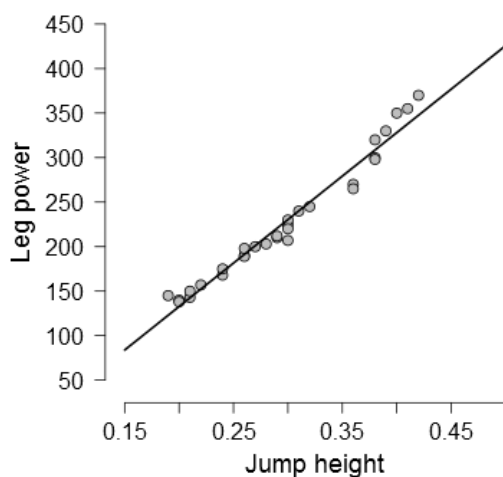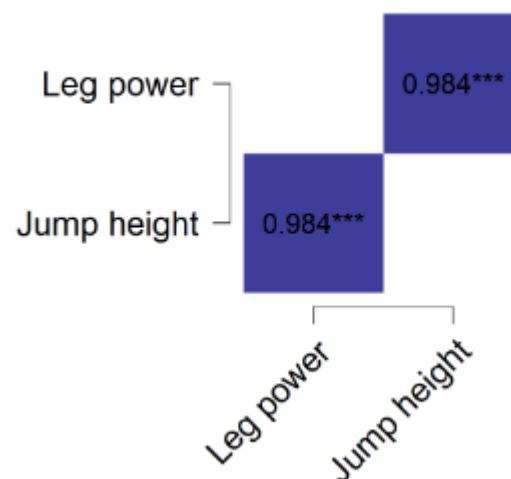
GOING ONE STEP FURTHER.

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) with jump height accounting for 96.8% of the variance in leg power.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 not 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.

Sample Correlation Coefficient
- ☐ Pearson's r
- ☑ Spearman's rho
- ☑ Kendall's tau-b

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
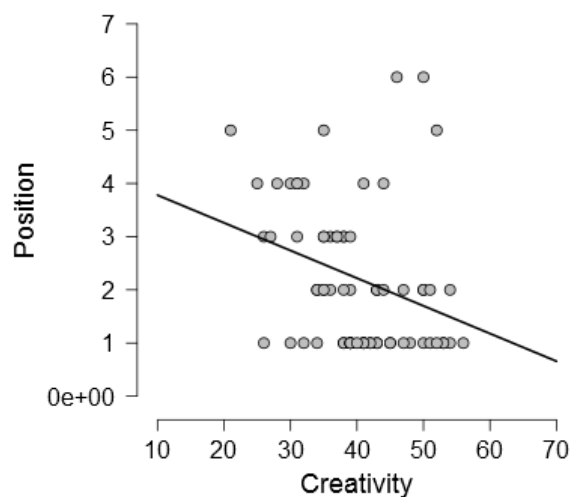
REPORTING THE RESULTS

The relationship between creativity scores and final position in the 'World's biggest liar' competition was assessed. A Spearman's correlation test showed a significant relationship: *rho* = -0.373, *p*=.002.

NOTE OF CAUTION.

The correlation only gives 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 that 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
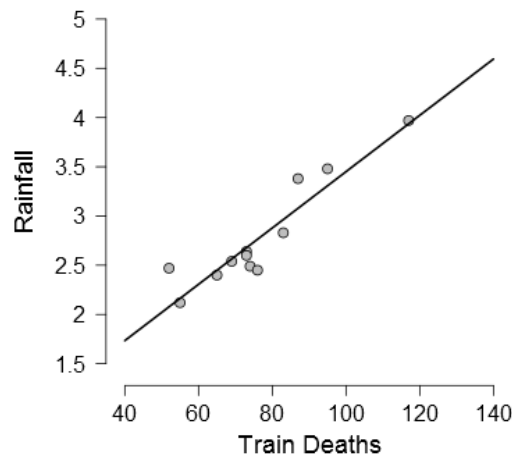
The number of honey-producing bee colonies (1000's) correlates strongly with the marriage rate in South Carolina (per 1000 marriages)

Pearson Correlations

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

JASP 0.16.1 - Dr Mark Goss-Sampson
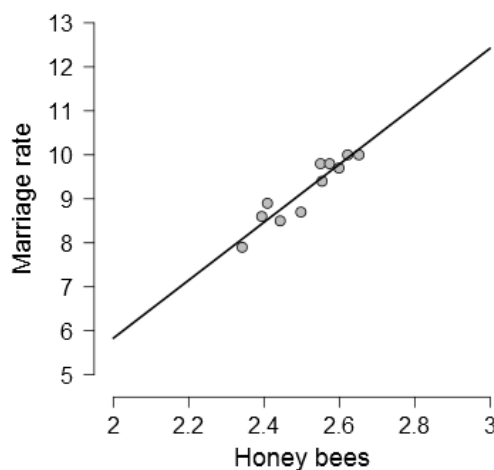
REGRESSION

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;

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

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

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

Linear regression makes the following assumptions:

- **Data** (1 dependent variable of continuous level, independent variables that are measured either at the continuous or nominal level, sample size etc)

- **Outliers,** there should be no significant outliers, high leverage points or highly influential points

- **Linear relationship**, there needs to be a linear relationship between (a) the dependent variable and each of your independent variables, and (b) the dependent variable and the independent variables collectively

- **Independence,** there should be independence of observations (i.e., independence of residuals)

- **Normality,** a) check that the residuals (errors) are approximately normally distributed; and b) check that the dependent variable is approximatively normally distributed.

- **Equal error variances**, data needs to show homoscedasticity of residuals (equal error variances)

- **Multicollinearity,** data must show minimal multicollinearity. This occurs when you have two or more independent variables that are highly correlated with each other.

Concerning 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 * number of predictors) for each variable. So for 4 variables that would require 82 data points 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 analyses 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.
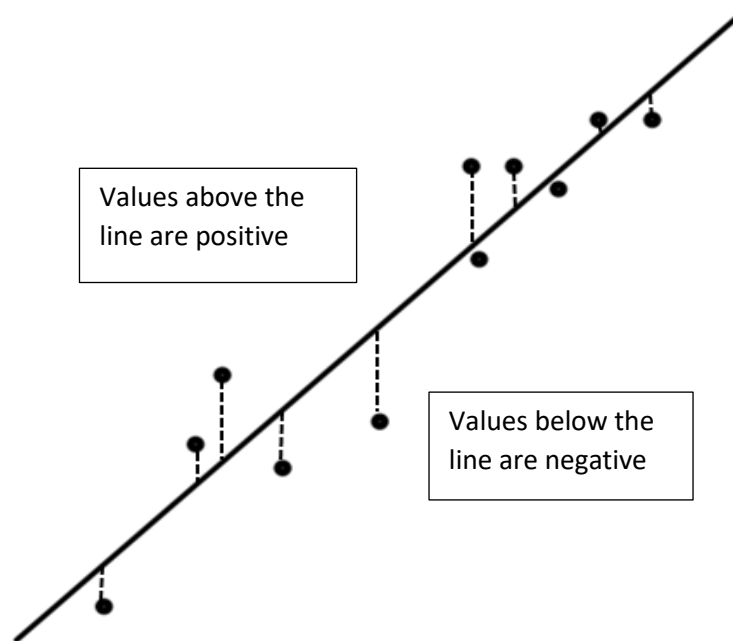
The vertical difference between the data points and the predicted regression line is known as the **residuals**. These values are squared to remove the negative numbers and then summed to give **SS$_R$**. This is effectively the error of the model or the **'goodness of fit'**, obviously the smaller the value the less error in the model.

JASP 0.16.1 - Dr Mark Goss-Sampson
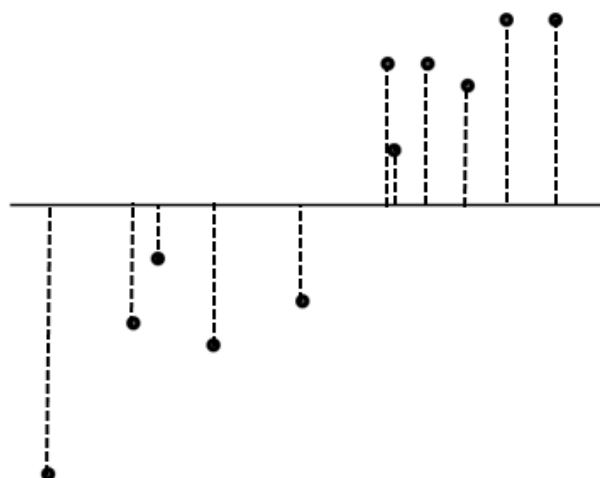
The vertical difference between the data points and the mean of the outcome variable can be calculated. These values are squared to remove the negative numbers and then summed to give the **total** sum of the squares SS_T. This shows how good the mean value is as a model of the outcome scores.
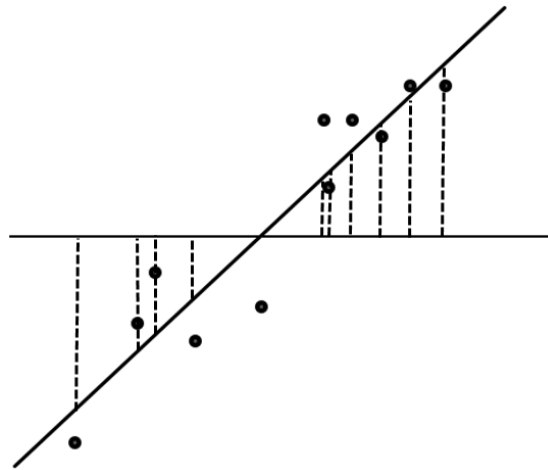
The vertical difference between the mean of the outcome variable and the predicted regression line is now determined. Again these values are squared to remove the negative numbers and then summed to give the **model** sum of squares (SS_M). This indicates how better the model is compared to just using the mean of the outcome variable. SS_T is the total sum of the squares.

So, the larger the SS_M the better the model is at predicting the outcome compared to the mean value alone. If this is accompanied by a small SS_R the model also has a small error.

R^2 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).

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

In regression, the model is assessed by the F statistic based on the improvement in the prediction of the model SS_M and the residual error SS_R. The larger the F value the better the model.

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

JASP 0.16.1 - Dr Mark Goss-Sampson

SIMPLE REGRESSION

Regression tests the null hypothesis (H₀) 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 ☑ Model fit
- ☐ From 5000 bootstraps ☐ R squared change
- ☐ Confidence intervals 95.0 % ☐ Descriptives
- ☐ Covariance matrix ☐ Part and partial correlations
- ☐ Vovk-Sellke maximum p-ratio ☐ Collinearity diagnostics

Residuals

- ☐ Statistics
- ☐ Durbin-Watson
- ☐ Casewise diagnostics
  - ○ Standard residual > 3
  - ○ Cook's distance > 1
  - ○ All

UNDERSTANDING THE OUTPUT

You will now get the following outputs:

Model Summary

Model	R	R²	Adjusted R²	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^2 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.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 - $p=.002$. This tells us that the model is a significantly better predictor of kicking distance than the mean distance.
Report as $F(1, 11) = 17.53$, $p=.002$.

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:

Distance = 57.105 + (6.452 * 60) = 454.6 m

FURTHER ASSUMPTION CHECKS
In Plots checks, tick the following two options:

▼ Plots

Residuals Plots
☐ Residuals vs. dependent
☐ Residuals vs. covariates
☑ Residuals vs. predicted
☐ Residuals vs. histogram
☑ Standardized residuals
☑ Q-Q plot standardized residuals
☐ Partial plots

JASP 0.16.1 - Dr Mark Goss-Sampson

This will result in two graphs:
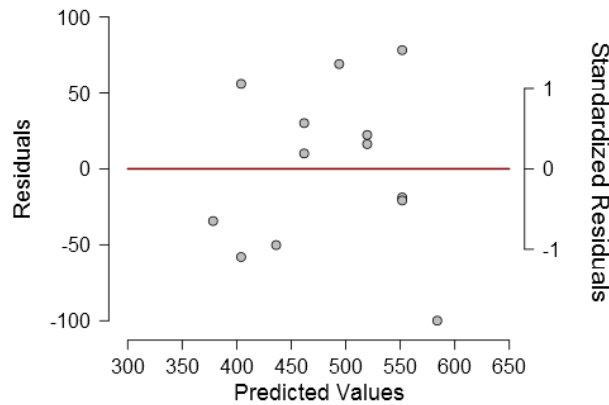
This graph shows a balanced random distribution of the residuals around the baseline suggesting that the assumption of homoscedasticity has not been violated. (See Exploring data integrity in JASP for further details.
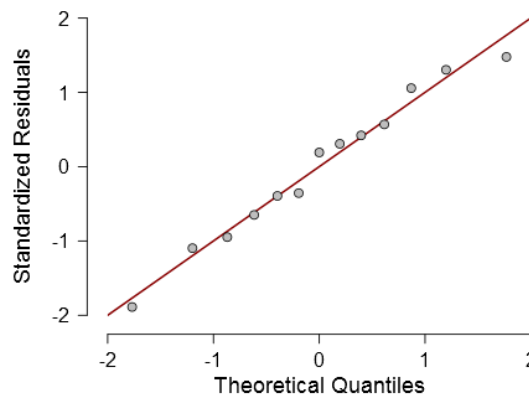
The Q-Q plot shows that the standardized residuals fit nicely along the diagonal suggesting that both assumptions or normality and linearity have also not been violated.

REPORTING THE RESULTS

A simple linear regression was used to predict rugby kicking distance from right leg strength. Leg strength was shown to explain a significant amount of the variance in the kicking distance: $F(1, 11) = 17.53$, $p=.002$, $R^2 = 0.614$. The regression coefficient (b = 6.452) allows the kicking distance to be predicted using the following regression equation:

Distance = 57.105 + (6.452 * Right leg strength)

JASP 0.16.1 - Dr Mark Goss-Sampson

MULTIPLE REGRESSION

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

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

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

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

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

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.

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.

Backward: All predictors are initially entered into 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.

Forward: The predictor with the highest simple correlation with the outcome variable is entered first. Subsequent predictors are selected based on 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.
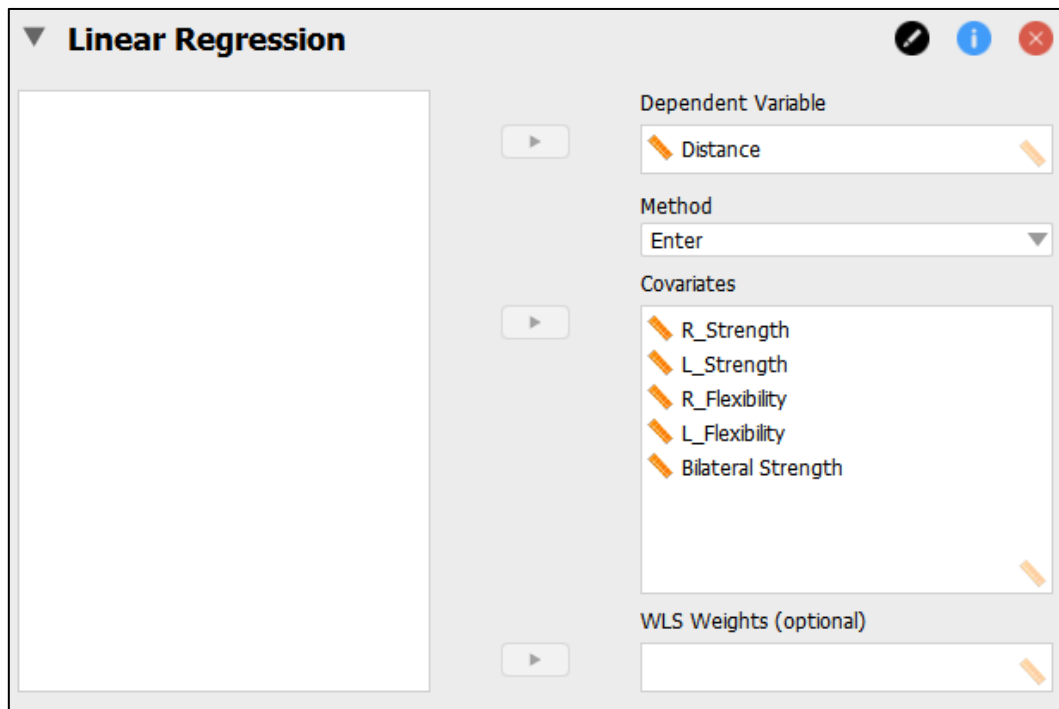
In the Variable section leave the Method as **Enter**. **Tick** the following options in the Statistics options, Estimates, Model fit, Collinearity diagnostics and Durbin-Watson.
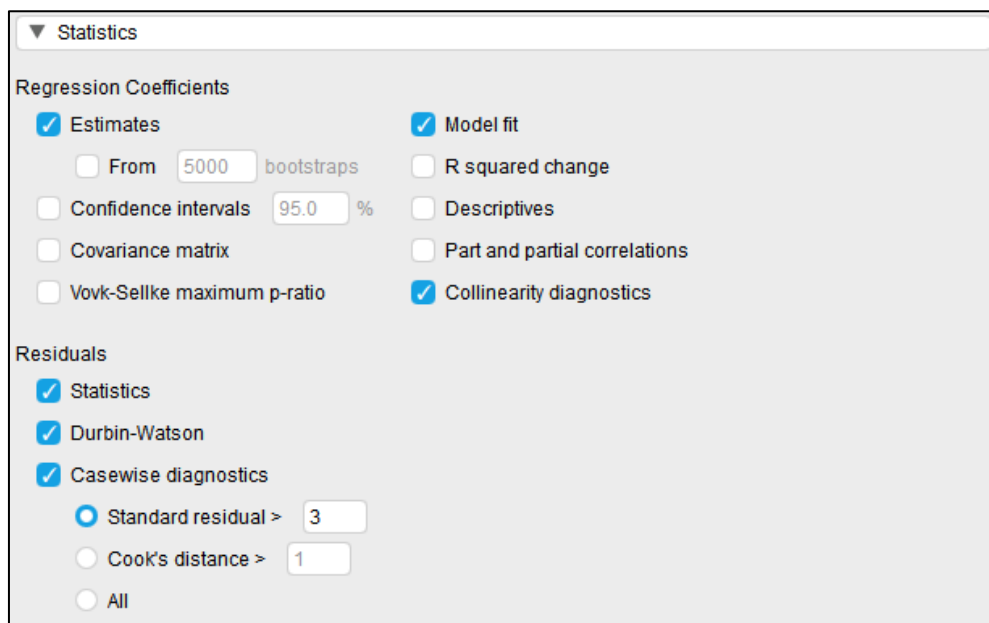
JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

You will now get the following outputs:

Model Summary - Distance

Model	R	R²	Adjusted R²	RMSE	Durbin-Watson Autocorrelation	Durbin-Watson Statistic	Durbin-Watson p
H_0	0.000	0.000	0.000	85.250	0.588	0.762	0.010
H_1	0.902	0.814	0.681	48.132	0.287	1.328	0.133

This provides information on a model based on the H_0 (no predictors) and the alternative H^1. The adjusted R^2 (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
H_1	Regression	70994.078	5	14198.816	6.129	0.017
	Residual	16216.845	7	2316.692		
	Total	87210.923	12			

Note. The intercept model is omitted, as no meaningful information can be shown.

The ANOVA table shows the F-statistic to be significant at $p=0.017$ 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	Collinearity Statistics VIF
H_0	(Intercept)	486.077	23.644		20.558	< .001		
H_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 both the H_0 and H^1 models 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 is 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).

The casewise diagnostics table is empty! This is good news. This will highlight any cases (rows) that have residuals that are 3 or more standard deviations away from the mean. These cases with the largest errors may well be outliers. Too many outliers will have an impact on the model and should be dealt with in the usual way (see Exploring Data Integrity).

Casewise Diagnostics

Case Number	Std. Residual	Distance	Predicted Value	Residual	Cook's Distance
.	.	.	.	.	.

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

The outputs are as follows:

Model Summary - Distance

Model	R	R²	Adjusted R²	RMSE	Durbin-Watson Autocorrelation	Durbin-Watson Statistic	Durbin-Watson p
1	0.902	0.814	0.681	48.132	0.287	1.328	0.133
2	0.902	0.813	0.720	45.146	0.303	1.296	0.088
3	0.897	0.805	0.740	43.505	0.284	1.313	0.144
4	0.884	0.782	0.738	43.618	0.132	1.676	0.439

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

JASP 0.16.1 - Dr Mark Goss-Sampson

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	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

Distance = 46.251 + (3.914 * R_Strength) + (2.009 * Bilateral Strength)

TESTING FURTHER ASSUMPTIONS.

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

JASP 0.16.1 - Dr Mark Goss-Sampson

Residuals vs. Predicted
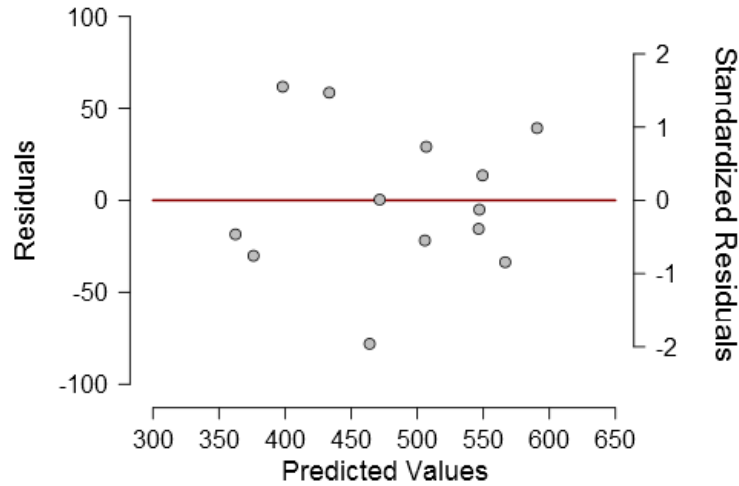
Q-Q Plot Standardized Residuals ▼
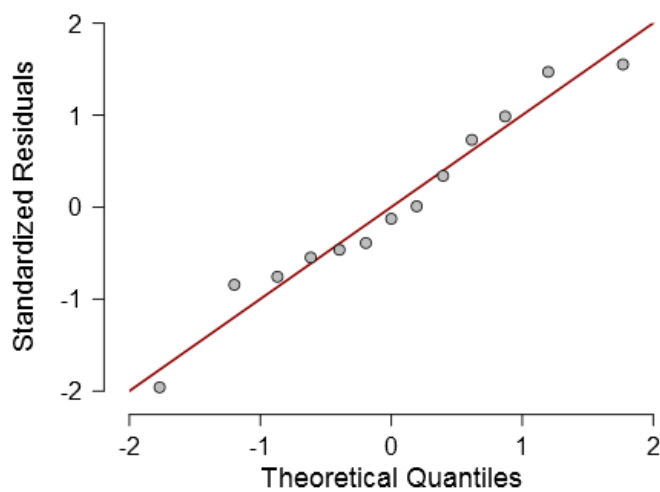
The balanced distribution of the residuals around the baseline suggests that the assumption of homoscedasticity has not been violated.

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

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.

REPORTING THE RESULTS

Multiple linear regression using backward data entry shows that right leg and bilateral strength can significantly explain 78.2% in the variance in kicking distance $F(2,10) = 17.92$, $p<.001$, $R^2 = 0.782$ Two significant regression coefficients where found, right leg strength ($b_1= 3.91$, $t=2.59$, $p=.027$) and bilateral strength ($b_2 = 2.009$, $t =2.77$, $p=0.02$). Kicking distance could be predicted using a regression equation of

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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 a binary yes or no outcome.

The null hypothesis tested is that there is no relationship between the outcome and the 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.
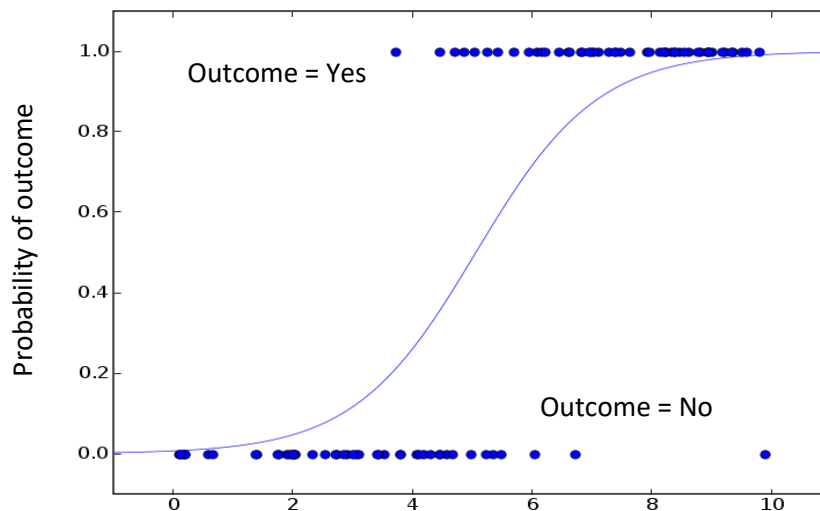
ASSUMPTIONS FOR BINARY LOGISTIC REGRESSION

- The dependent variable must be binary i.e. yes or no, male or female, good or bad.
- One or more independent (predictor variables) 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.

Four **pseudo R²** values are calculated in JASP, McFadden, Nagelkerke, Tjur and Cox & Snell. These are analogous to R^2 in linear regression and all give different values. What constitutes a good R^2 value varies, however, they are useful when comparing different models for the same data. The model with the largest R^2 statistic is considered to be the best.

The Wald test is used to determine statistical significance for each of the independent variables.

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 (2nd heart attack) into the Dependent variable, add the stress levels to Covariates and Exercise prescription to Factors. Leave the data entry method as Enter.
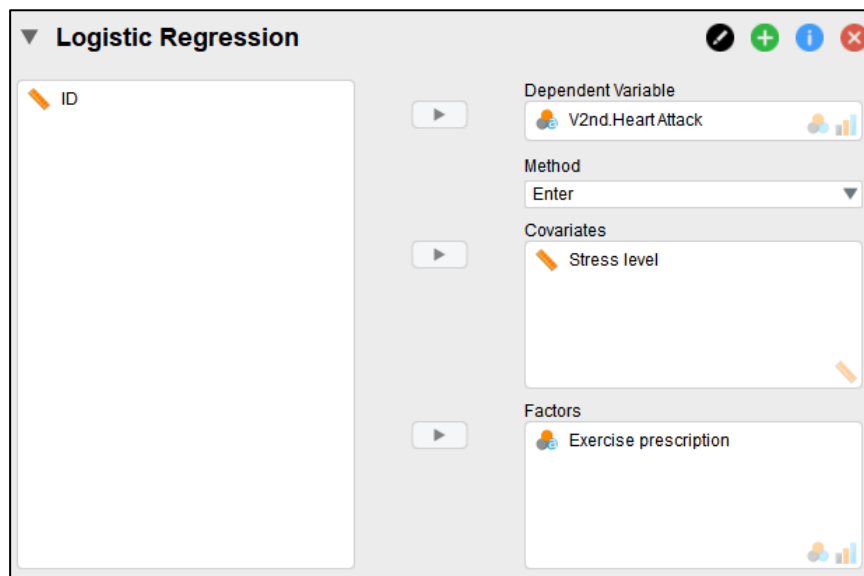
In the Statistics options tick Estimates, Odds ratios, Confusion matrix, Sensitivity and Specificity.

JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

The initial output should comprise 4 tables.

Model summary

Model	Deviance	AIC	BIC	df	X²	p	McFadden R²	Nagelkerke R²	Tjur R²	Cox & Snell R²
H₀	55.452	57.452	59.141	39						
H₁	34.195	40.195	45.261	37	21.257	< .001	0.383	0.550	0.126	0.412

The model summary shows that H1 (with the lowest AIC and BIC scores) suggests a significant relationship (X^2(37) =21.257, $p<.001$) between the outcome (2nd heart attack) and the predictor variables (exercise prescription and stress levels).

McFadden's R^2 = 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 | Wald Test | | |
					Wald Statistic	df	p
(Intercept)	−4.368	2.550	0.013	−1.713	2.933	1	0.087
Stress level	0.089	0.041	1.093	2.159	4.662	1	0.031
Exercise prescription (Yes)	−2.043	0.890	0.130	−2.295	5.268	1	0.022

Note. V2nd.Heart Attack level 'Yes' coded as class 1.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 2nd heart attack if undergoing an exercise intervention.

Confusion matrix

	Predicted	
Observed	No	Yes
No	15.000	5.000
Yes	5.000	15.000

Performance metrics

	Value
Sensitivity	0.750
Specificity	0.750

The confusion matrix shows where agreement (yes/yes) or disagreement (yes/no) occurred between the predicted and observed 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
☑ Conditional estimates plots
    Confidence interval 95.0 %
    ☑ Show data points

Residual Type
◉ Deviance
○ Pearson

Residual Plots
☐ Predicted - residual plot
☐ Predictor - residual plots
☐ Squared Pearson residuals plot

Performance Plots
☐ ROC plot
    Cutoff step 0.2
    ☐ Add cutoff labels
☐ PR plot
    Cutoff step 0.2
    ☐ Add cutoff labels
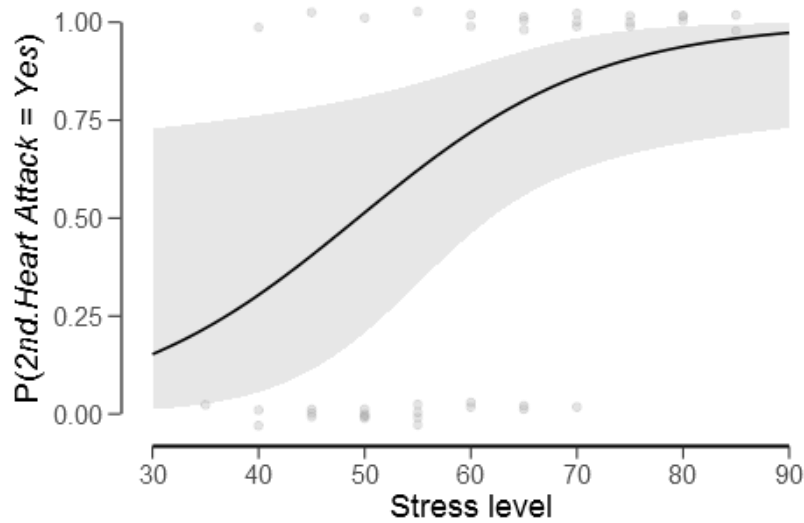
As stress levels increase the probability of having a second heart attack increases.
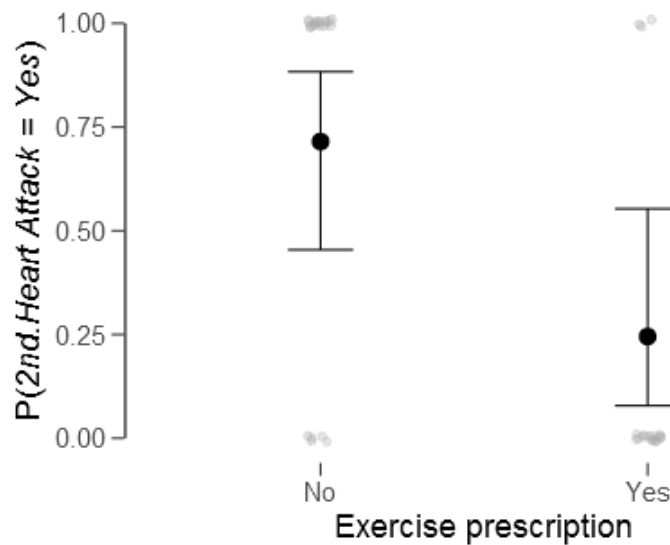
No exercise intervention increases the probability of a 2nd heart attack while it is reduced when it had been put in place.

REPORTING THE RESULTS

Logistic regression was performed to ascertain the effects of stress and exercise intervention on the likelihood that participants have a 2nd heart attack. The logistic regression model was statistically significant, $\chi2$ (37) = 21.257, $p< .001$. The model correctly classified 75.0% of cases. Increasing stress was associated with an increased likelihood of a 2nd 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 2nd heart attack to 13%.

JASP 0.16.1 - Dr Mark Goss-Sampson

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 dataset's 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. 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 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 an appropriate research method design.

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

JASP 0.16.1 - Dr Mark Goss-Sampson

CONTRASTS

Contrasts are 'a priori' tests (i.e. planned comparisons before any data were collected). As an example, researchers may want to compare the effects of some new drugs to the currently prescribed ones. These should only be a small set of comparisons in an attempt to reduce familywise error. The choice must be based on the scientific questions being asked, and chosen during the experimental design. Hence the term planned comparisons. Therefore they are looking at specified mean differences and therefore can be used if the ANOVA F test is insignificant.

JASP provides 6 planned contrast enabling different types of comparisons:

Deviation: the mean of each level of the independent variable is compared to the overall mean (the mean when all the levels are taken together).

Simple: the mean of each level is compared to the mean of a specified level, for example with the mean of the control group.

Difference: the mean of each level is compared to the mean of the previous levels.

Helmert: the mean of each level is compared to the mean of the subsequent levels.

Repeated: By selecting this contrast, the mean of each level is compared to the mean of the following level.2

Polynomial: tests polynomial trends in the data.

POST HOC TESTING

Post hoc tests are tests that were decided upon after the data have been collected. They can only be carried out if the ANOVA F test is significant.

JASP provides 4 types of post hoc testing

Standard

Games-Howell – used when you are unsure about the equality of group variances

Dunnett's – used to compare all the groups to one group i.e. the control group

Dunn – a non-parametric post hoc test used for testing small sub-sets of pairs.

JASP also provides 5 types of corrections 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. Does not assume independence of the comparisons.

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

Tukey – is 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.

Sidak – is similar to Bonferroni but assumes that each comparison is independent of the others. Slightly more powerful than Bonferroni.

EFFECT SIZE

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

Eta squared (η^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 (η_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 (ω^2) – Normally, statistical bias gets very small as sample size increases, but for small samples (n<30) ω^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 diets.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 Descriptive statistics and ***ω^2*** as the effect size;

In Assumptions checks, tick all options:

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

UNDERSTANDING THE OUTPUT

ANOVA - Weight loss kg

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p	ω^2
None	Diet	92.369	2.000	46.184	10.826	< .001	0.214
	Residuals	294.371	69.000	4.266			
Brown-Forsythe	Diet	92.369	2.000	46.184	10.826	< .001	0.214
	Residuals	294.371	64.352	4.574			
Welch	Diet	92.369	2.000	46.184	11.449	< .001	0.214
	Residuals	294.371	44.987	6.544			

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 the homogeneity of variance is not significant. However, if Levene's test shows a significant difference in variance, the Brown-Forsythe or Welch correction should be reported.
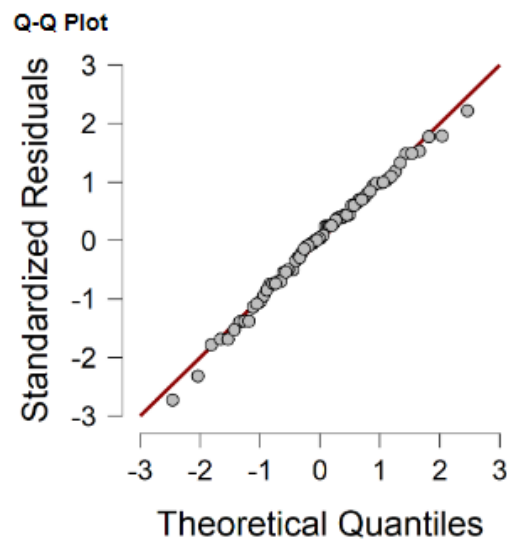
Q-Q Plot

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.

CONTRAST EXAMPLE

If for example, one planned to compare the effects of diets B and C to diet A. Click on the drop-down menu and select 'simple' next to diet. This will test the significance between the first category in the list with the remaining categories.

Simple Contrast - Diet

Comparison	Estimate	SE	df	t	p
Diet B - Diet A	0.404	0.596	69.000	0.678	0.500
Diet C - Diet A	2.579	0.596	69.000	4.326	< .001

As can be seen, only diet C is significantly different from diet A ($t(69)$ = 4.326, $p<.001$).

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 Standard type, use Tukey for the post hoc correction and tick flag significant comparisons.

JASP 0.16.1 - Dr Mark Goss-Sampson

Post Hoc Comparisons - Diet

		Mean Difference	SE	t	Cohen's d	p_{tukey}
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 in 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.

Also, in Descriptive Plots add the Factor 'Diet' to the horizontal axis and tick display error bars and in Raincloud plots do the same.

JASP 0.16.1 - Dr Mark Goss-Sampson
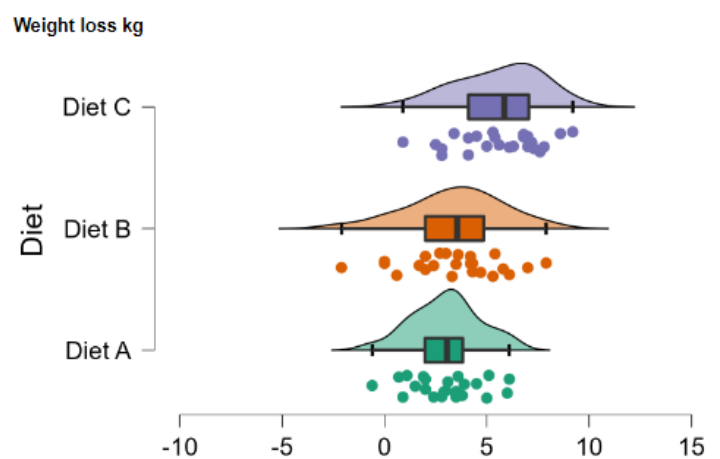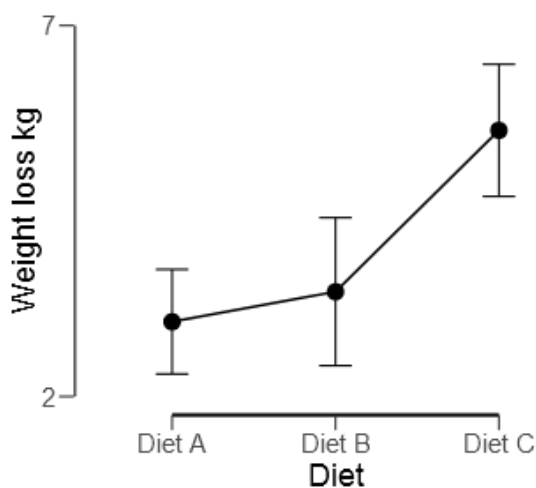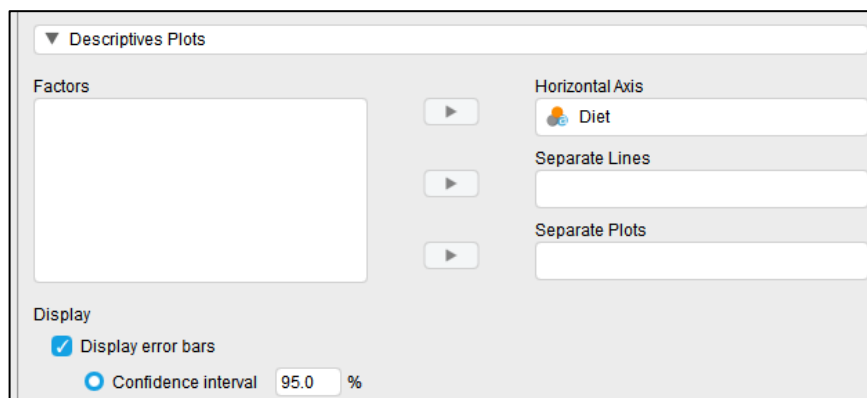
REPORTING THE RESULTS

Levene's test showed that there was equal variance for all three groups: F(2,69) = 1.3, p=.28*****.

Independent one way ANOVA showed a significant effect of the type of diet on weight loss after 10 weeks ($F_{(2, 69)}$ =46.184, $p<.001$, The ω^2 = 0.214 suggesting that this is a large effect size.

Post hoc testing using using pairwise comparisons of the estimated marginal means with Tukey's correction revealed that diet C resulted in significantly greater weight loss (*M* = 5.59, *SD* = 2.1 kg) than diet A (*M* = 3.0, *SD* = 2.36 kg, $p<.001$) or diet B (*M* = 3.41, *SD* = 2.36 kg, p=.001). There were no significant differences in weight loss between diets A and B (p=.777).

* If Levene's test shows a significant difference in variance, the Brown-Forsythe or Welch correction should be reported.

KRUSKAL-WALLIS – NON-PARAMETRIC ANOVA

If your data fails parametric assumption tests or is nominal, 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, the Kruskal-Wallis H test (also known as the "one-way ANOVA on ranks") is an omnibus test that 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 the 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

Q-Q Plot

JASP 0.16.1 - Dr Mark Goss-Sampson
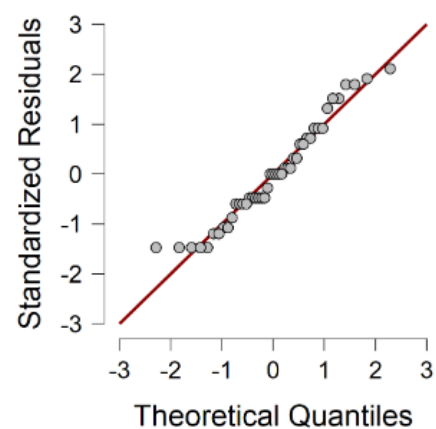
Although the ANOVA indicates a significant result, the data has not met the assumptions of homogeneity of variance as seen by the significant Levene's test and only shows linearity in the middle of the Q-Q plot and curves off at the extremities indicating more extreme values. Added to the fact that the dependent variable is based on subjective pain scores suggest the use of a non-parametric alternative.

Return to the statistics options and open the Nonparametrics option at the bottom. For the Kruskal-Wallis test Move the Treatment variable to the box on the right. In Post Hoc tests move treatment to the right box and tick Dunn's post hoc test and flag significant comparisons.
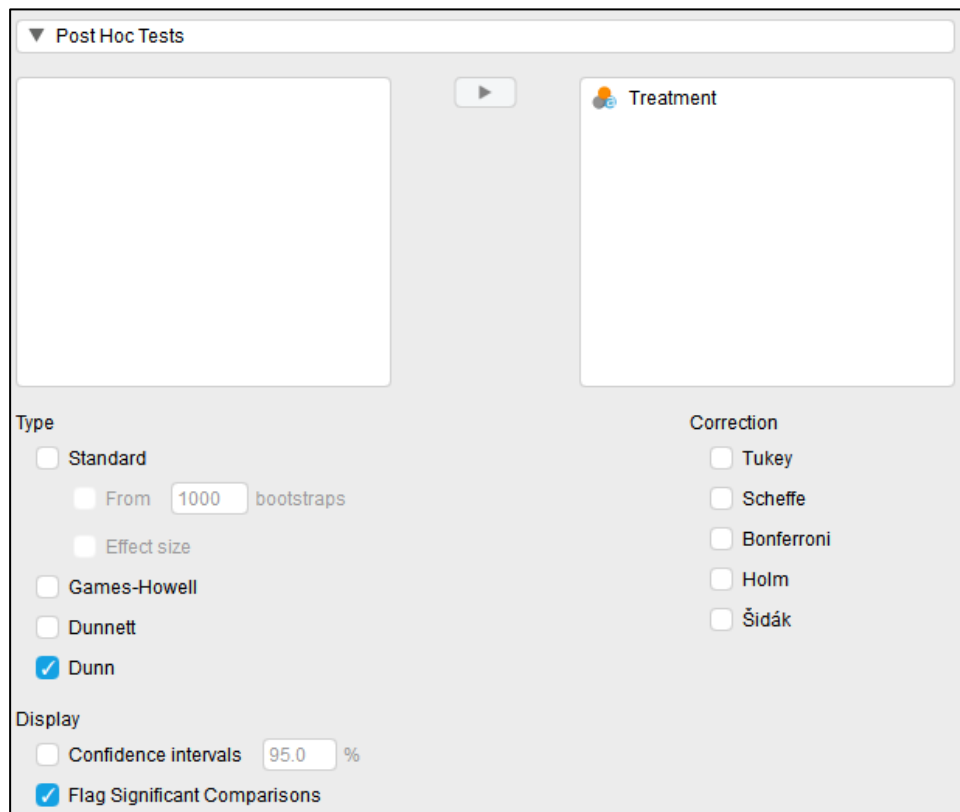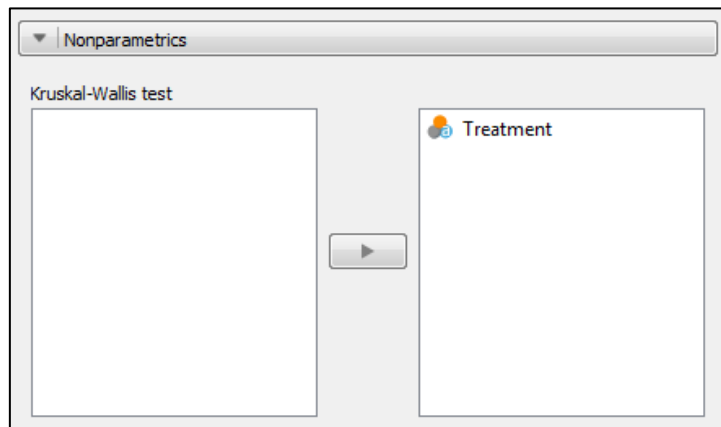
UNDERSTANDING THE OUTPUT

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

The Dunn's post hoc test provides its p-value as well as those for Bonferroni and Holm's Bonferroni

Dunn

Dunn's Post Hoc Comparisons - Treatment

Comparison	z	W_i	W_j	p	p_{bonf}	p_{holm}
Control - Cryo+Compression	4.317	34.600	14.200	< .001***	< .001***	< .001***
Control - 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

** p < .01, *** p < .001

correction. As can be seen, both treatment conditions are significantly different from the controls but not from each other.

Descriptive Statistics

	Pain Score		
	Control	Cryo+Compression	Cryotherapy
Median	7.000	3.000	3.000
MAD	2.000	1.000	1.000
IQR	4.000	1.500	1.500

JASP 0.16.1 - Dr Mark Goss-Sampson
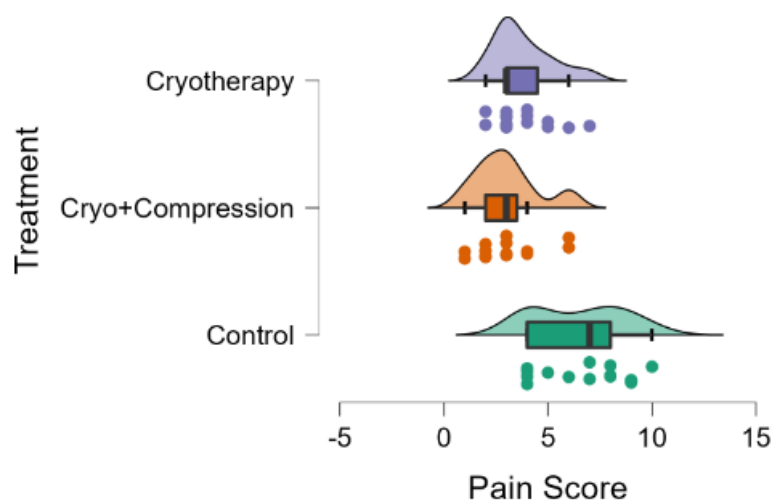
REPORTING THE RESULTS

Pain scores were significantly affected by treatment modality $H(2) = 19.693$, $p<.001$. Pairwise comparisons using Dunn's post hoc showed that both cryotherapy (*Mdn* = 3) and cryotherapy with compression (*Mdn* = 3) significantly reduces pain scores ($p=.002$ and $p<.001$ respectively) compared to the control group (*Mdn* = 7). There was no significant difference between cryotherapy and cryotherapy with compression ($p=.102$).

JASP 0.16.1 - Dr Mark Goss-Sampson

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.
- 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:

$$Variance (A-B) \approx Variance (A-C) \approx Variance (B-C)$$

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 (ε) corrections. A general rule of thumb is that if the ε 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 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 (η^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 (η_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 (ω^2) – Normally, statistical bias gets very small as sample size increases, but for small samples (n<30) ω^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.
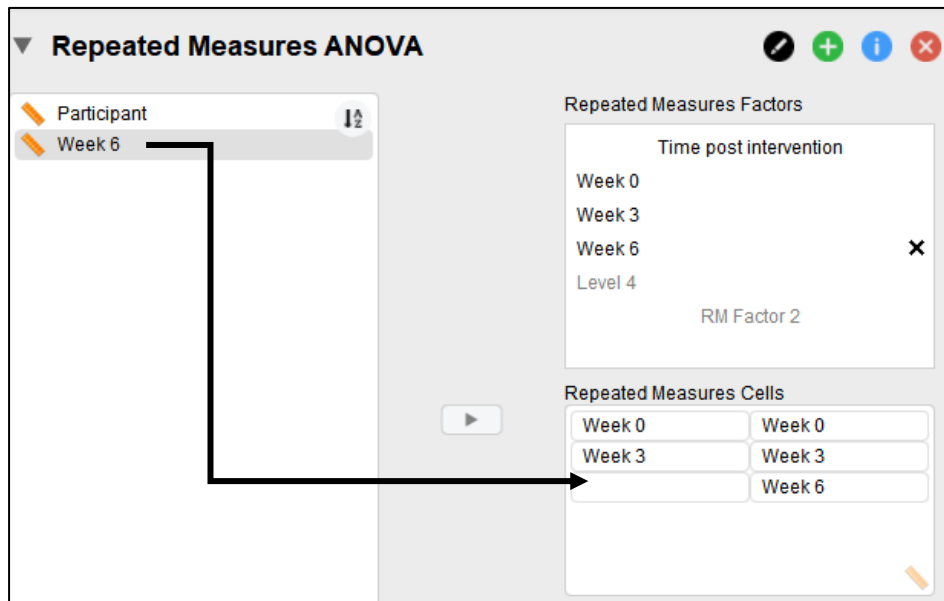
Tick Descriptive Statistics, Estimates of effect size and ω^2.
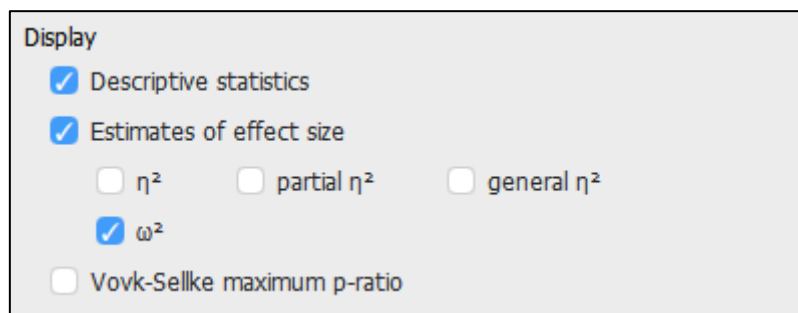
Under Assumption Checks tick Sphericity tests and all Sphericity correction options.
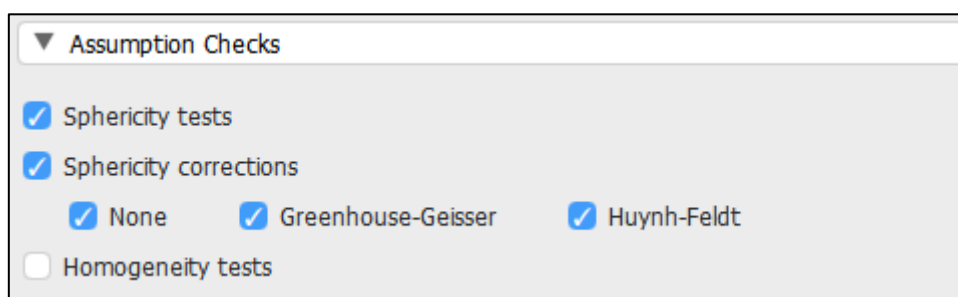
The output should consist of 4 tables. The third table, between-subject effects, can be ignored for this analysis.

JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
Time post intervention	None	4.320ª	2.000ª	2.160ª	212.321ª	< .001ª	0.058
	Greenhouse-Geisser	4.320ª	1.235ª	3.497ª	212.321ª	< .001ª	0.058
	Huynh-Feldt	4.320ª	1.284ª	3.365ª	212.321ª	< .001ª	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
ª 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 value of the mean square. 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 (ε) values are below 0.75. Therefore the ANOVA result should be reported based on the Greenhouse-Geisser correction:

Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X²	dfSphericity	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
Time post intervention	0.381	15.440	2	< .001	0.618	0.642	0.500

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	ω^2
Time post intervention	Greenhouse-Geisser	4.320ª	1.235ª	3.497ª	212.321ª	< .001ª	0.058
Residual	Greenhouse-Geisser	0.346	21.001	0.016			

Note. Type III Sum of Squares
ª 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, $p<.001$, ω^2 = 0.058.**

JASP 0.16.1 - Dr Mark Goss-Sampson

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 Holm 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 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.

JASP 0.16.1 - Dr Mark Goss-Sampson
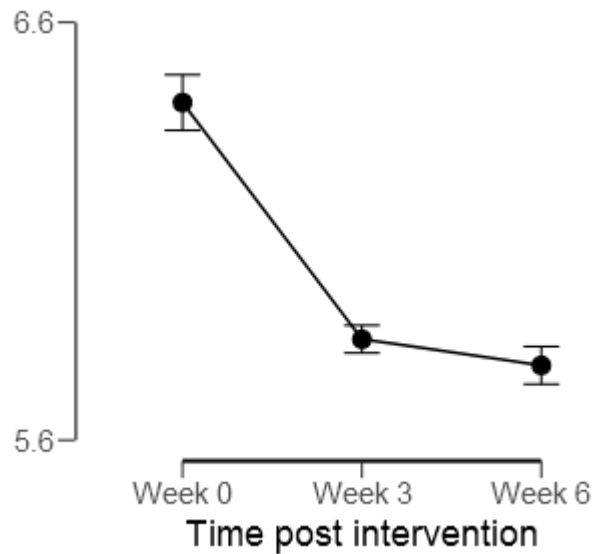
Post Hoc Comparisons - Time post intervention

		Mean Difference	SE	t	Cohen's d	p_{holm}
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	0.001

Note. Cohen's d does not correct for multiple comparisons.
Note. Bonferroni adjusted confidence intervals.

REPORTING THE RESULTS

Mauchly's test of sphericity showed a significant difference in variation between the group differences: $W(2) = 0.381$, $p<.001$, Therefore the Greenhouse-Geisser correction was used. This showed that cholesterol levels differed significantly between time points, $F(1.235, 21.0) =212.3$, $p<.001$. $\omega^2 = 0.058$ suggesting that this is a small to medium effect.

Post hoc testing using the Bonferroni correction revealed that cholesterol levels decreased significantly as time increased, weeks 0 – 3 ($M = 0.566$, $SE = 0.04$, $p<.001$), weeks 1 – 6 ($M = 0.63$ $SE = 0.04$, p=.004) and weeks 3-6 (M=0.06, $SE = 0.016$).

JASP 0.16.1 - Dr Mark Goss-Sampson

FRIEDMAN'S REPEATED MEASURES ANOVA

If parametric assumptions are violated or the data is ordinal you should consider using the non-parametric alternative, Friedman's test. Similar to the Kruskal-Wallis test, 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 that 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 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.
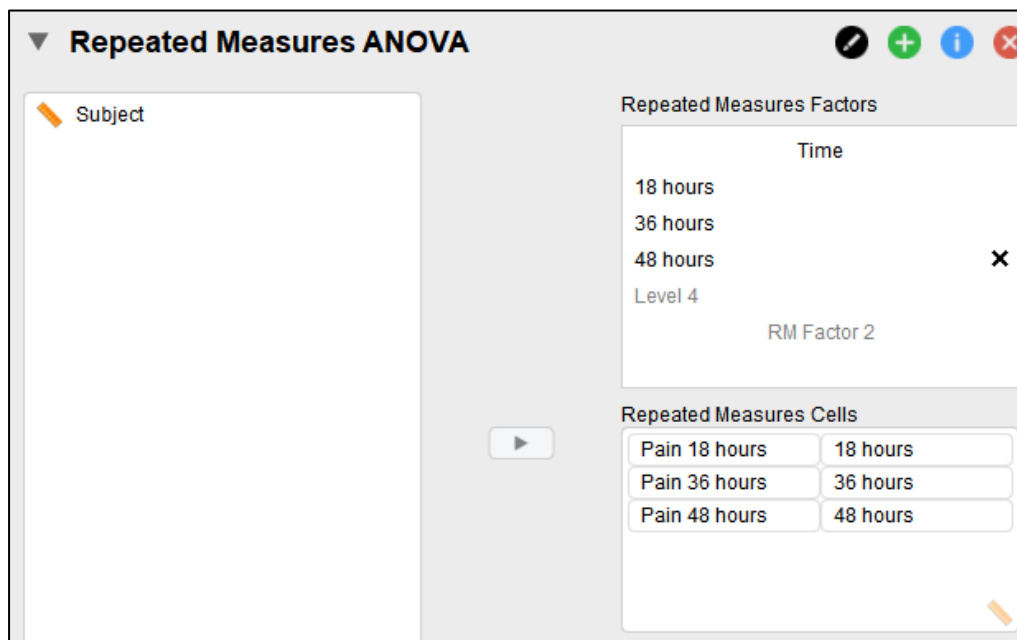
This will automatically produce the standard repeated measures within-subjects ANOVA table. To run the Friedman's test, expand the Nonparametrics tab, move Time to the RM factor box and tick Conover's post hoc tests.

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

Connover's Post Hoc Comparisons - Time

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
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.

Descriptive Statistics

	Pain 18 hours	Pain 36 hours	Pain 48 hours
Median	3.000	7.000	3.000
MAD	1.000	2.000	1.000
IQR	1.500	4.000	1.500

JASP 0.16.1 - Dr Mark Goss-Sampson
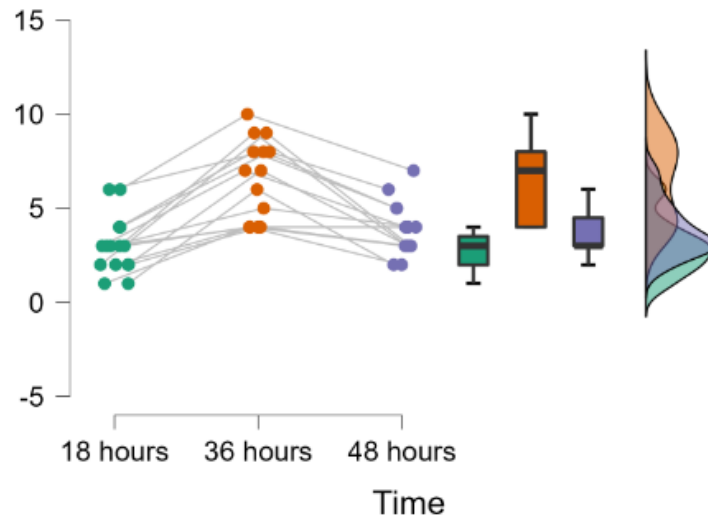
REPORTING THE RESULTS

Friedmann's test showed that time has a significant effect on subjective pain scores χ^2 (2) = 26.77, $p<.001$. Conover's pairwise post hoc comparisons showed that pain perception peaked at 36 hours (Mdn = 7) and was significantly higher than at 18 hours (Mdn = 3: T(28) = 15.2, $p<.001$) and then decreased by 48 hours (Mdn = 3: T(28) = 8.82 $p<.001$).

JASP 0.16.1 - Dr Mark Goss-Sampson

COMPARING INDEPENDENT GROUPS AND THE EFFECTS OF COVARIATES
ANCOVA

ANOVA can be used to compare the means of one variable (dependent) in two or more groups, whereas analysis of covariance (**ANCOVA**) sits between ANOVA and regression and compares the means of one (dependent) variable in two or more groups while taking into account the variability of other continuous variables (**COVARIATES**). ANCOVA checks for differences in 'adjusted' means (i.e. adjusted for the effects of the covariate). A covariate may not usually be part of the main research question but could influence the dependent variable and therefore needs to be adjusted or controlled for. As long as a good covariate is used ANCOVA will have improved statistical power and control over error.

Control for – to subtract statistically the effects of a variable (a control variable) to see what a relationship would be without it (Vogt 1977).

Hold constant – to "subtract" the effects of a variable from a complex relationship to study what the relationship would be if the variable were, in fact, a constant. Holding a variable constant essentially means assigning it an average value (Vogt 1977).

Statistical control – using statistical techniques to isolate or "subtract" variance in the dependent variable attributable to variables that are not the subject of the study (Vogt, 1999).

For example, when looking for a difference in weight loss between the three diets it would be appropriate to take into account the individuals' pre-trial bodyweight since heavier people may lose proportionately more weight.

JASP 0.16.1 - Dr Mark Goss-Sampson
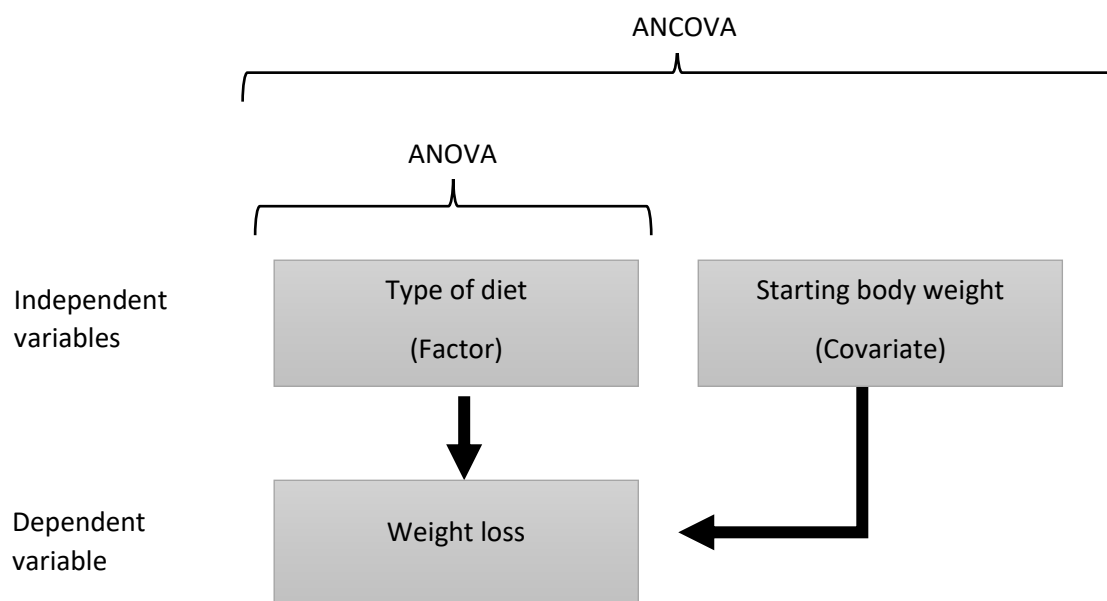
The null hypothesis tested is that there is no significant difference between the 'adjusted' means of all the groups.

ASSUMPTIONS

ANCOVA makes the same assumptions as the independent ANOVA makes. However, there are two further assumptions:

- The relationship between the dependent and covariate variables are linear.
- **Homogeneity of regression** i.e. the regression lines for each of the independent groups are parallel to each other.
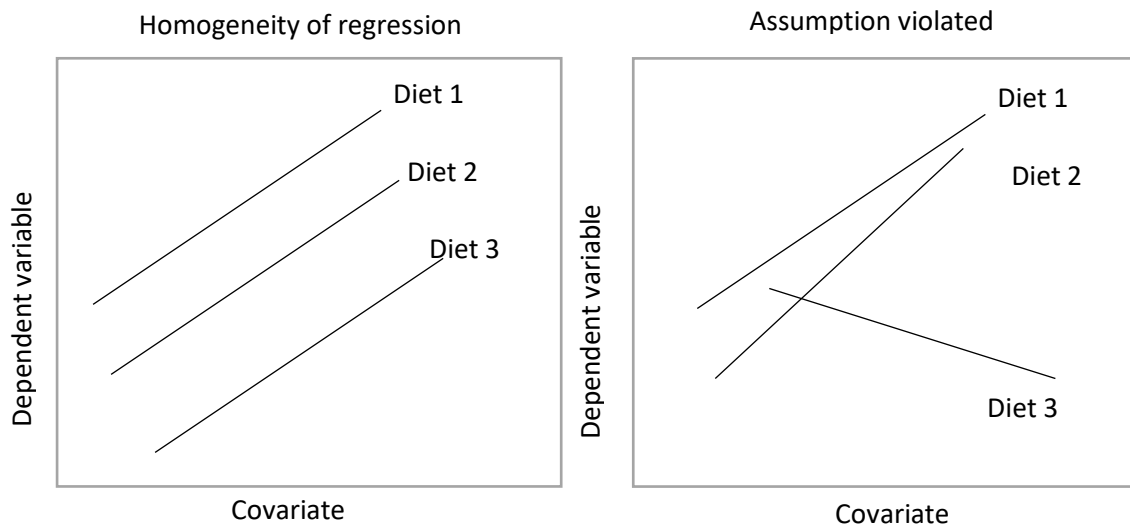
POST HOC TESTING

JASP provides 4 Types

- *Standard* – as above
- *Games-Howell* – used when you are unsure about the equality of group variances
- **Dunnett's** – used to compare all the groups to one group i.e. the control group
- *Dunn* – a non-parametric post hoc test used for testing small sub-sets of pairs.

JASP additionally provides 4 corrections 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 is a sequential Bonferroni method that is less conservative than the original Bonferroni test.
- *Tukey* – is 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 (η²)** - 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 (η$_p$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 (ω²)** – Normally, statistical bias gets very small as sample size increases, but for small samples (n<30) ω² 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 ANCOVA

Load **ANCOVA hangover.csv**. This dataset has been adapted from the one provided by Andy Field (2017). The morning after a Fresher's ball students were given either water, coffee or a Barocca to drink. Two hours later they reported how well they felt (from 0 – awful to 10 –very well). At the same time, data were collected on how drunk they were the night before (0-10).

Initially, run an ANOVA with wellness as the dependent variable and the type of drink as the fixed factor.

ANOVA - wellness

Cases	Sum of Squares	df	Mean Square	F	p	ω²
drink	4.267	2	2.133	1.714	0.199	0.045
Residual	33.600	27	1.244			

Note. Type III Sum of Squares

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.744	2	27	0.194

As can be seen from the results, homogeneity of variance has not been violated while the ANOVA shows that there is no significant difference in the wellness scores between any of the morning drinks. $F(2,27)=1.714$, $p=.199$. However, this may be related to how drunk the students were the night before!

Go to ANOVA > ANCOVA, put wellness as the dependent variable and the type of drink as the fixed factor. Now add drunkenness to the Covariate(s) box. In the first instance tick Descriptive statistics and ω² as the effect size;

JASP 0.16.1 - Dr Mark Goss-Sampson
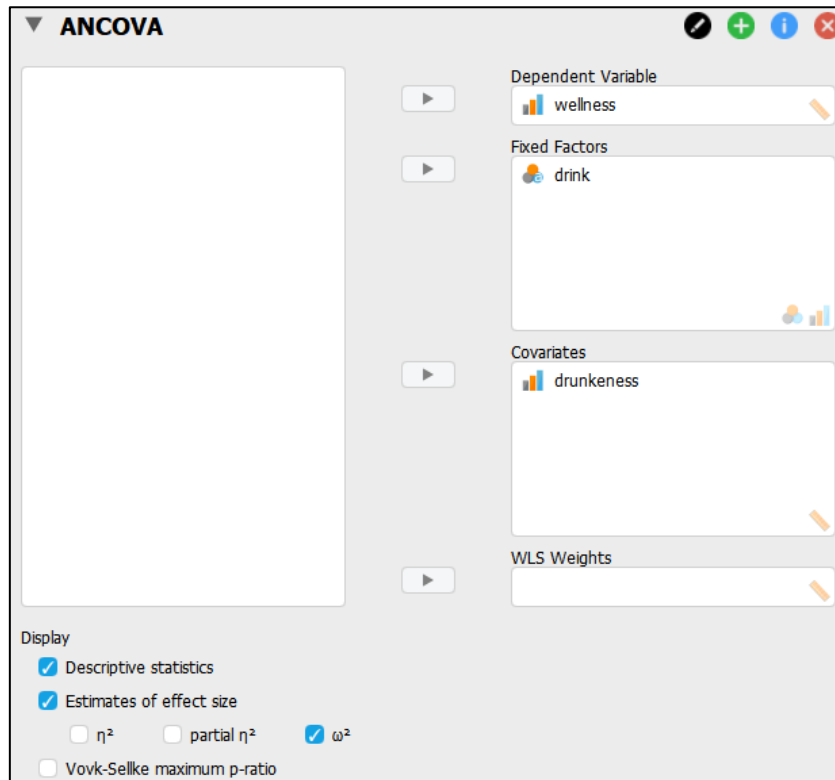
In Assumption Checks tick both options
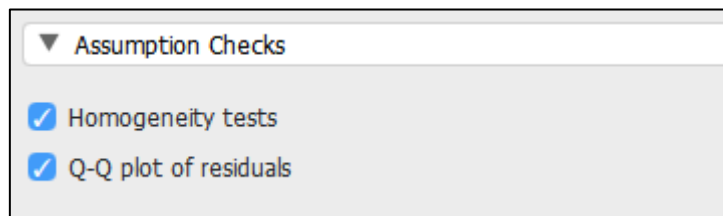
In Marginal Means move drink to the right

This should result in 4 tables and one Q-Q plot.

JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

ANCOVA - wellness

Cases	Sum of Squares	df	Mean Square	F	p	ω²
drink	8.503	2	4.252	7.470	0.003	0.173
drunkeness	18.801	1	18.801	33.032	< .001	0.427
Residual	14.799	26	0.569			

Note. Type III Sum of Squares

It can be seen that the covariate (drunkenness) significantly predicts wellness ($p<.001$). The effects of the type of drink on wellness, when adjusted for the effects of drunkenness are now significant ($p=.003$).
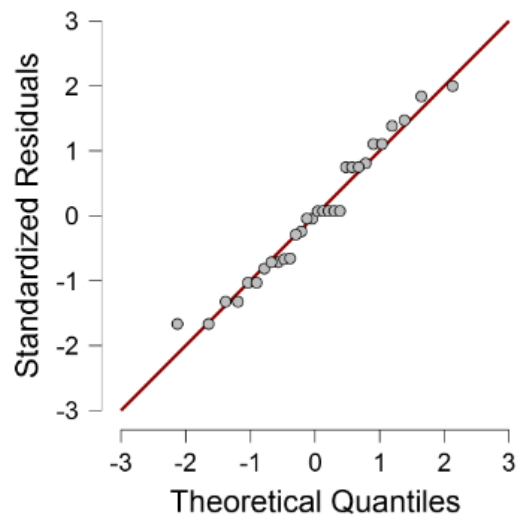
It can be seen that Levene's test is significant, unlike in ANOVA, no homogeneity of variance corrections (i.e. Welch) are provided. For ANCOVA this can be ignored. The Q-Q plot appears to be normal.

Descriptives - wellness

drink	Mean	SD	N
Barocca	6.000	1.414	10
Coffee	6.000	0.667	10
Water	5.200	1.135	10

The descriptive statistics show the unadjusted means for wellness in the three drink groups.

JASP 0.16.1 - Dr Mark Goss-Sampson

Marginal Means - drink

drink	Marginal Mean	SE	Lower CI	Upper CI
Barocca	6.571	0.258	6.040	7.103
Coffee	5.273	0.270	4.718	5.828
Water	5.356	0.240	4.862	5.849

The marginal means are now the wellness means having been adjusted for the effects of the covariate (drunkenness).

TESTING FURTHER ASSUMPTIONS

As previously mentioned the assumption of homogeneity of regression is important in ANCOVA. This can be tested by looking at the interaction between the type of drink and the drunkenness scores. Go to Model, drink and drunkenness will have been automatically added as individual Model terms. Now highlight both drink and drunkenness and add them both to Model terms.
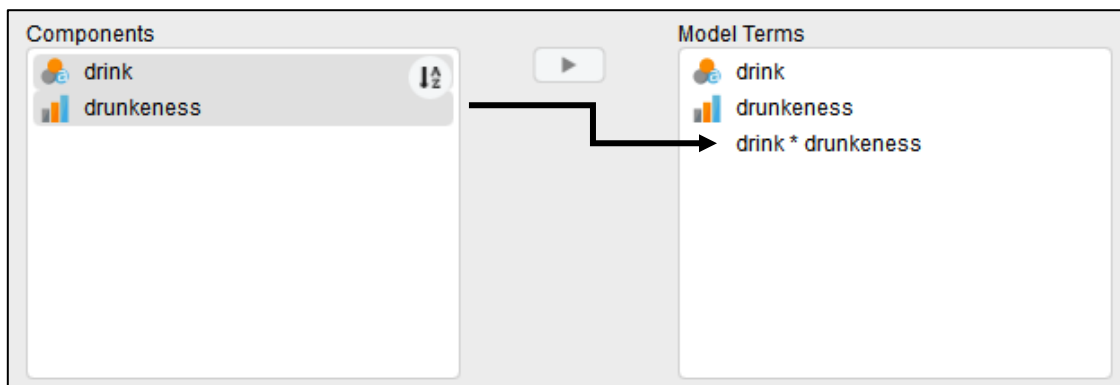
ANCOVA - wellness

Cases	Sum of Squares	df	Mean Square	F	p	ω^2
drink	2.290	2	1.145	1.876	0.175	0.047
drunkeness	4.865	1	4.865	7.971	0.009	0.189
drink * drunkeness	0.150	2	0.075	0.123	0.885	0.000
Residual	14.649	24	0.610			

Note. Type III Sum of Squares

The ANOVA table now has an extra row showing the interaction between the type of drink and drunkenness. This is not significant (*p*=.885), i.e. the relationships between drunkenness and wellness are the same in each drink group. If this is significant there will be concerns over the validity of the main ANCOVA analysis.

Having checked this, go back and remove the interaction term from the Model terms.

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

POST HOC TESTING

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

Post Hoc Comparisons - drink

		Mean Difference	SE	t	Cohen's d	p_{tukey}
Barocca	Coffee	1.298	0.406	3.198	1.174	0.010
	Water	1.215	0.345	3.523	0.948	0.004
Coffee	Water	−0.083	0.371	−0.224	−0.089	0.973

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

Post hoc testing shows that there is no significant difference between coffee and water on wellness. However, wellness scores were significantly higher after drinking a Barocca.

This can be seen from the Descriptive plots.
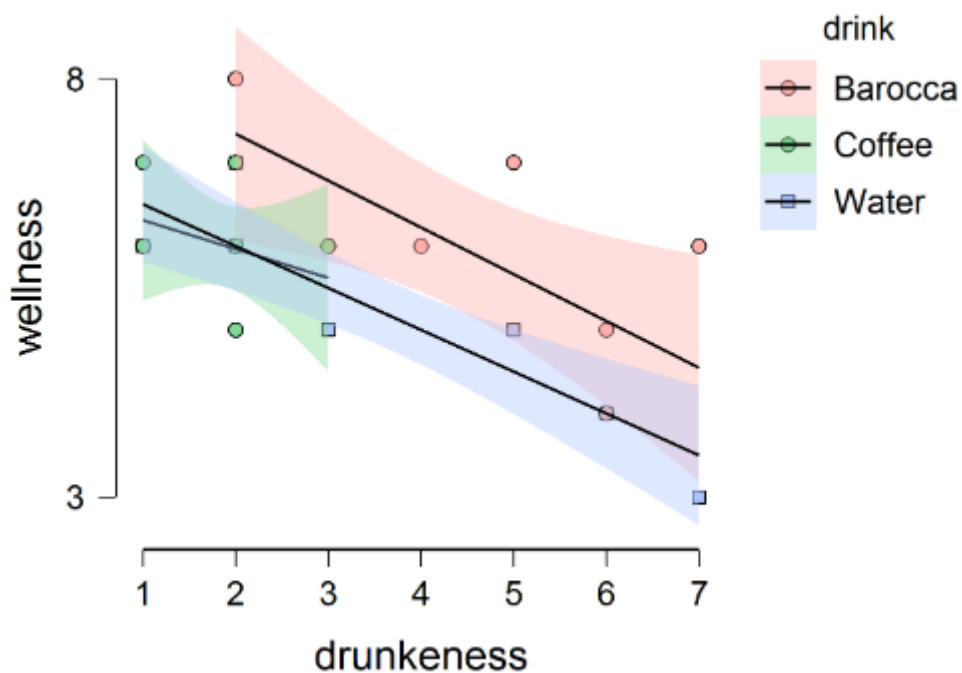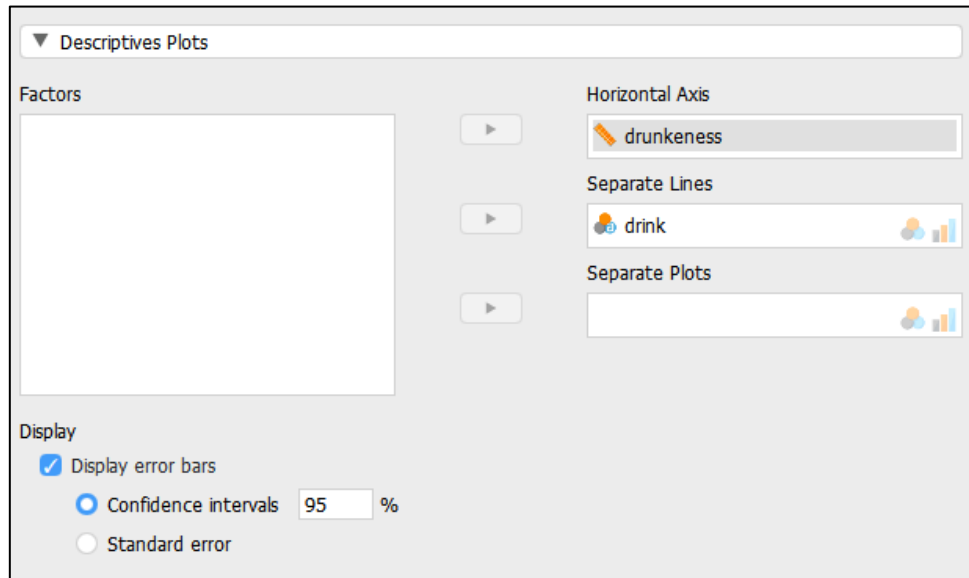
REPORTING THE RESULTS

The covariate, drunkenness, was significantly related to the morning after wellness, $F(1,26) = 33.03$, $p<.001$, $\omega^2 = 0.427$. There was also a significant effect of drink on wellness after controlling for drunkenness, $F(2, 26) = 7.47$, p=.003, $\omega^2 = 0.173$.

Post hoc testing using Tukey's correction revealed that drinking a Barocca resulted in significantly greater wellness compared to water ($p<.004$) or coffee ($p=.01$). There were no significant differences in wellness between water and coffee ($p=.973$).

JASP 0.16.1 - Dr Mark Goss-Sampson

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). I**nteraction** 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 that should either be addressed in the research design or can the 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 combinations of factors.
- There should be no significant outliers.

RUNNING TWO-WAY INDEPENDENT ANOVA

Open **2-way independent ANOVA.csv** in JASP. This comprises 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.

Tick Descriptive statistics and Estimates of effect size (ω^2).

In Descriptive plots, add the supplement to the horizontal axis and Gender to separate lines. In Additional Options,

JASP 0.16.1 - Dr Mark Goss-Sampson
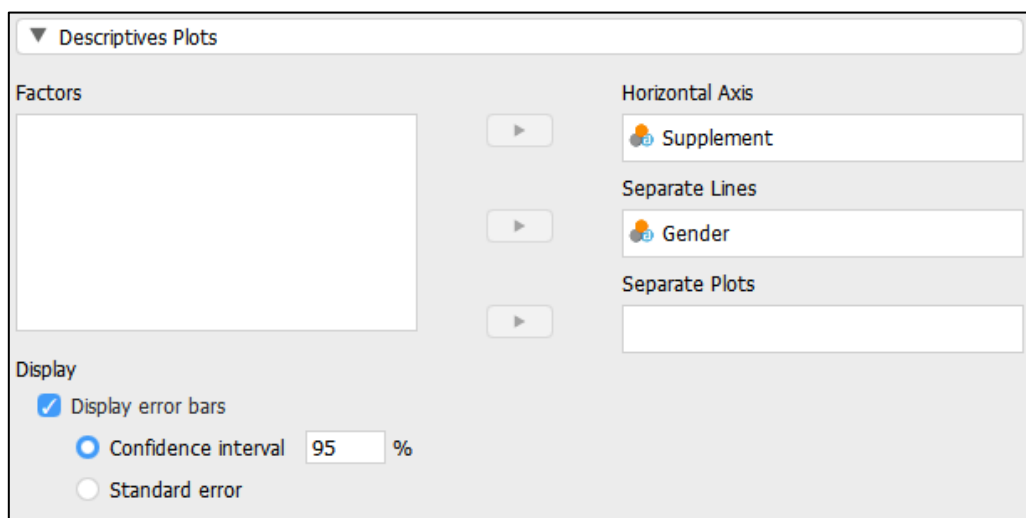
UNDERSTANDING THE OUTPUT

The output should comprise 2 tables and one plot.

ANOVA - Jump power

Cases	Sum of Squares	df	Mean Square	F	p	ω^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 are 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

JASP 0.16.1 - Dr Mark Goss-Sampson
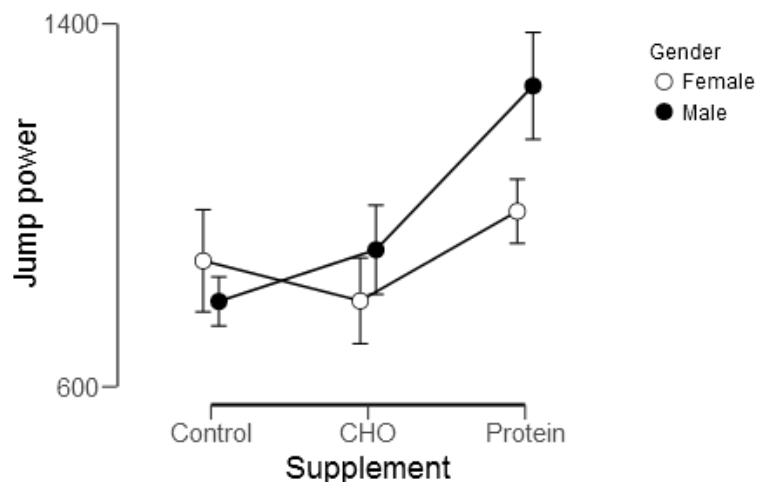
TESTING ASSUMPTIONS

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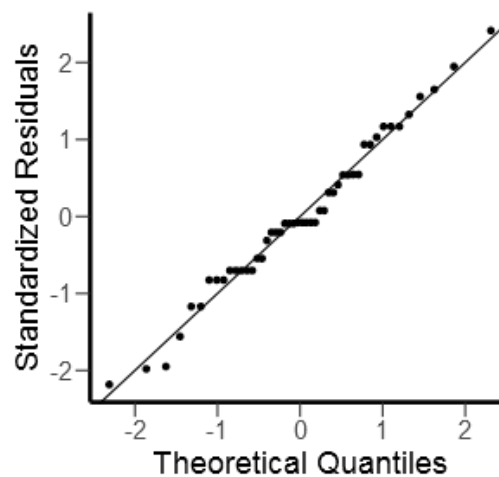
The Q-Q plot shows that the data appear to be normally distributed and linear. We can now accept the ANOVA result since none of these assumptions has been violated.

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

SIMPLE MAIN EFFECTS

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 limited pairwise comparisons.

JASP 0.16.1 - Dr Mark Goss-Sampson

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.

CONTRAST TESTING

There are two ways of testing for a difference between (combinations of) cells: post hoc tests and contrast analysis. JASP has a range of different contrast tests available, including custom contrasts. For example, we can contrast the three different supplements. Open up the Contrasts menu and next to Supplement click on the drop-down menu and select custom. This will add another series of options to this window.

In this window contrasts can be added, in this case, three contrasts can be defined:

This will result in the following table:

Custom Contrast - Supplement

Comparison	Estimate	SE	df	t	p
(1*CHO) - (1*Prtn.)	−279.315	39.561	42	−7.060	< .001
(1*CHO) - (1*Cntr.)	12.768	40.102	42	0.318	0.752
(1*Cntr.) - (1*Prtn.)	−292.083	38.853	42	−7.518	< .001

Comparing this table to the post hoc analysis below the estimates of the differences in marginal means are the same, as well as their standard errors and t-statistics. However, both the p-values and confidence intervals vary: the corrected p-values are typically higher, and the confidence intervals wider, for the post hoc analysis.

POST HOC TESTING

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

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

Standard

Post Hoc Comparisons - Supplement

		Mean Difference	SE	t	p_{tukey}
CHO	Control	12.77	40.10	0.32	0.95
	Protein	−279.32	39.56	−7.06	< .001***
Control	Protein	−292.08	38.85	−7.52	< .001***

*** p < .001
Note. P-value adjusted for comparing a family of 3
Note. Results are averaged over the levels of: Gender

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).

The post hoc comparisons for the interactions decomposes the results further.

Post Hoc Comparisons - Gender ⁕ Supplement

		Mean Difference	SE	t	p_{tukey}
Female, CHO	Male, CHO	−112.59	57.68	−1.95	0.39
	Female, Control	−88.21	57.68	−1.53	0.65
	Male, Control	1.16	57.68	0.02	1.00
	Female, Protein	−197.38	56.17	−3.51	0.01*
	Male, Protein	−473.84	57.68	−8.21	< .001***
Male, CHO	Female, Control	24.38	55.73	0.44	1.00
	Male, Control	113.75	55.73	2.04	0.34
	Female, Protein	−84.79	54.16	−1.57	0.62
	Male, Protein	−361.25	55.73	−6.48	< .001***
Female, Control	Male, Control	89.37	55.73	1.60	0.60
	Female, Protein	−109.17	54.16	−2.02	0.35
	Male, Protein	−385.63	55.73	−6.92	< .001***
Male, Control	Female, Protein	−198.54	54.16	−3.67	8.41e −3**
	Male, Protein	−475.00	55.73	−8.52	< .001***
Female, Protein	Male, Protein	−276.46	54.16	−5.10	< .001***

* p < .05, ** p < .01, *** p < .001
Note. P-value adjusted for comparing a family of 6

JASP 0.16.1 - Dr Mark Goss-Sampson

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, ω^2 = 0.058) and Supplement (F (2, 42) = 30.07, $p<.001$, ω^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, $p<.001$, ω^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 (*M* = 1263, *SD* = 140) compared to females (*M* = 987, *SD* = 92, F (1) =28.06, $p<.001$).

JASP 0.16.1 - Dr Mark Goss-Sampson

TWO-WAY REPEATED MEASURES ANOVA

Two-Way Repeated Measures ANOVA means that there are two factors in the experiment, for example, different treatments and different conditions. "Repeated measures" means that the same subject received more than one treatment and/or more than one condition.

Independent variable (Factor 2)	Independent variable (Factor 1) = time			
	Participant	Time 1	Time 2	Time 3
Condition 1	1	Dependent variable	Dependent variable	Dependent variable
	2	Dependent variable	Dependent variable	Dependent variable
	3	Dependent variable	Dependent variable	Dependent variable
Condition 2	1	Dependent variable	Dependent variable	Dependent variable
	2	Dependent variable	Dependent variable	Dependent variable
	3	Dependent variable	Dependent variable	Dependent variable

The factors are split into levels, therefore, in this case, Factor 1 has 3 repeated levels and Factor 2 has 2 repeated 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). I**nteraction** is where one factor influences the other factor.

The two-way repeated ANOVA is another omnibus test that is used to test the following main effect null hypotheses:

H01: the dependent variable scores are the same for each level in factor 1 (ignoring factor 2).

H02: the dependent variable scores are the same for each level in factor 2 (ignoring factor 1).

The null hypothesis for the interaction between the two factors is:

H03: the two factors are independent, or, that an interaction effect is not present.

ASSUMPTIONS

Like all other parametric tests, two-way repeated ANOVA makes a series of assumptions that should either be addressed in the research design or can the tested for.

- The independent variables (factors) should have at least two categorical related groups (levels).
- The dependent variable should be continuous and approximately normally distributed for all combinations of factors.

- Sphericity i.e. the variances of the differences between all combinations of related groups must be equal.
- There should be no significant outliers.

RUNNING TWO-WAY REPEATED MEASURES ANOVA

Open **2-way repeated ANOVA.csv** in JASP. This comprises 4 columns of data ("sit and reach" flexibility scores for two factors, Factor 1 with 2 levels (stretch and no stretch) and Factor 2 with 2 levels (warm-up and no warm-up). In Descriptive statistics check the data for significant outliers. Go to ANOVA > Repeated Measures ANOVA. Firstly each Factor and its levels should be defined, for RM Factor 1 – define this as Stretching and its levels as stretch and no stretch. Then define RM Factor 2 as Warm-up and its levels as warm-up and no warm-up. Then add the appropriate column of data to the assigned repeated measures cells.
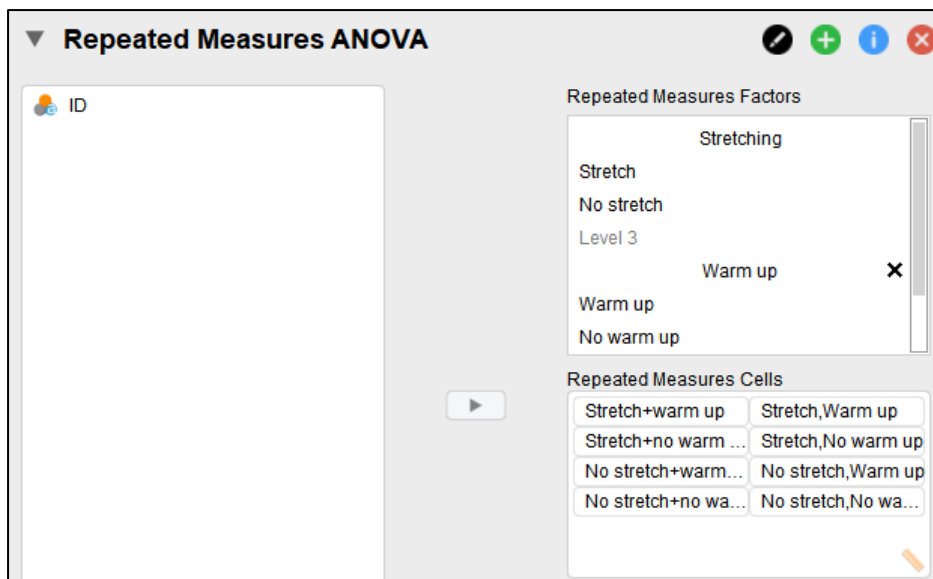
Also, tick Descriptive statistics and estimates of effect size - ω^2.
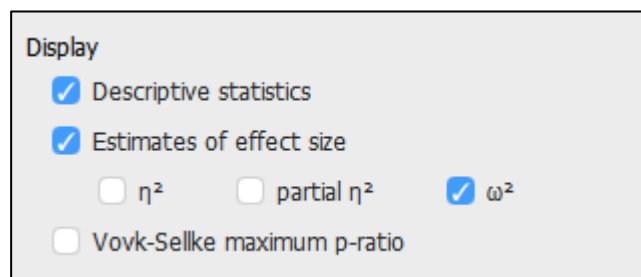
In Descriptive plots add the Stretching factor to the horizontal axis and the Warm-up factor to separate lines. Tick the display error bars option.

JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

The output should comprise 3 tables and one plot. The Between-Subjects Effects table can be ignored in this analysis.

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	ω²
Stretch	246.43	1	246.432	123.40	< .001	0.647
Residual	21.97	11	1.997			
Warm up	85.71	1	85.707	68.69	< .001	0.404
Residual	13.72	11	1.248			
Stretch ✳ Warm up	34.99	1	34.987	29.64	< .001	0.215
Residual	12.98	11	1.180			

Note. Type III Sum of Squares

The ANOVA within-subjects effects table shows that there are significant main effects for both stretch ($p<.001$) and warm-up ($p<.001$) on sit and reach distance. Both of these are associated with large effect sizes. There is also a significant interaction between stretch and warm-up ($p<.001$), this suggests that the effects of performing a stretch on sit and reach distance are different depending on whether or not a warm-up had been performed. These findings can be seen in both the descriptive data and plot.

JASP 0.16.1 - Dr Mark Goss-Sampson

Descriptives

Stretch	Warm up	Mean	SD	N
Stretch	Warm up	15.029	1.551	12
	No warm up	10.649	2.121	12
No stretch	Warm up	8.790	1.892	12
	No warm up	7.825	1.750	12
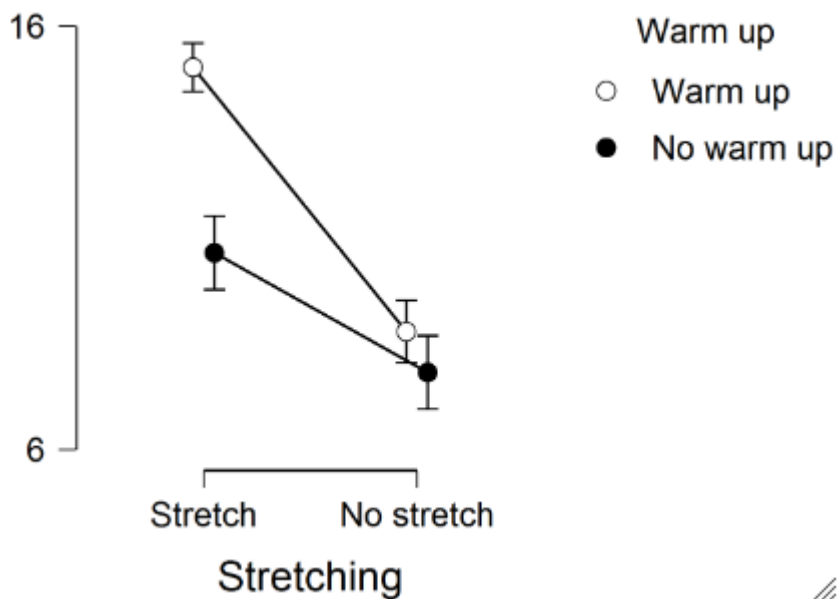
TESTING ASSUMPTIONS

In this case, there are no assumption checks. Sphericity can only be tested when there are at least three levels and homogeneity requires at least two unrelated data sets. If a factor has more than 2 levels Mauchly's test of Sphericity should also be run and the appropriate corrected F value used if necessary (See Repeated Measures ANOVA for details).

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

JASP 0.16.1 - Dr Mark Goss-Sampson

SIMPLE MAIN EFFECTS

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

▼ Simple Main Effects

Factors		Simple Effect Factor
	▸	Warm up
		Moderator Factor 1
	▸	Stretching
		Moderator Factor 2
	▸	
○ Pool error terms		

Simple Main Effects - Warm up

Level of Stretch	Sum of Squares	df	Mean Square	F	p
Stretch	115.106	1	115.106	112.673	< .001
No stretch	5.587	1	5.587	3.973	0.072

Note. Type III Sum of Squares

This table shows that when moderating for warm-up there is a significant difference ($p<.001$) in sit and reach performance when a stretch was also carried out but not without a stretch (p=.072).
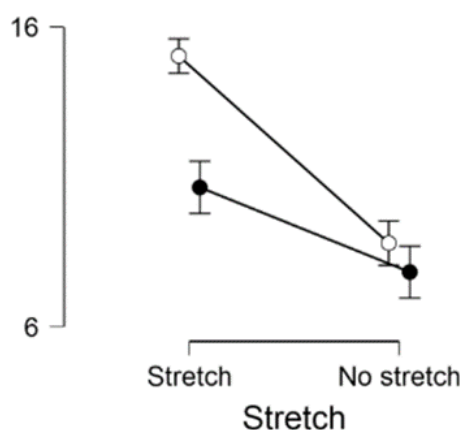
We can now moderate for stretch by changing the Simple Main Effects to use Stretch as the simple effect factor and warm up as the moderator factor. We can also replot the descriptives with a warm-up on the horizontal axis and stretch as separate lines.

JASP 0.16.1 - Dr Mark Goss-Sampson

Simple Main Effects - Stretch

Level of Warm up	Sum of Squares	df	Mean Square	F	p
Warm up	233.56	1	233.56	234.18	< .001
No warm up	47.86	1	47.86	21.95	< .001

Note. Type III Sum of Squares

In this case, when controlling for Stretch there were significant differences between both warm-up and no warm-up.
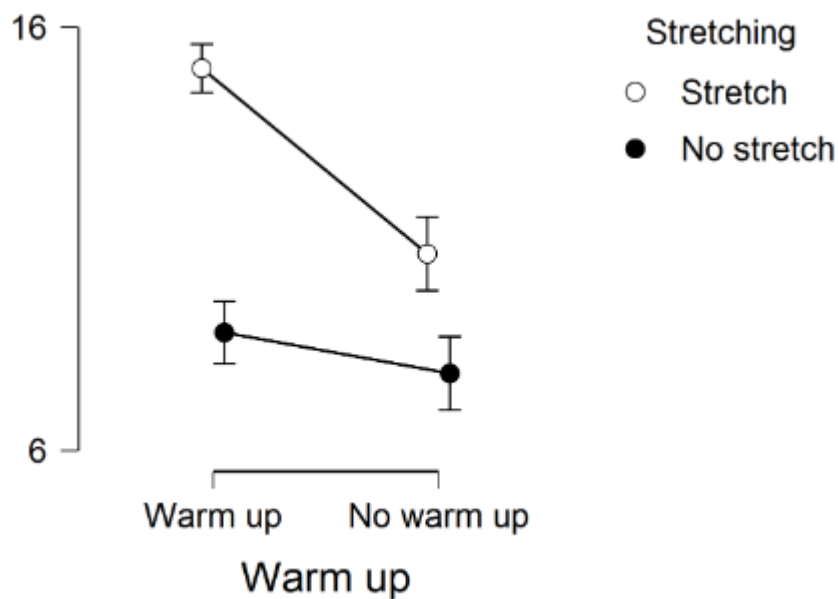
Both of these simple main effects can be visualised in their descriptive plots.

The data can be further visualised using Raincloud plots.

JASP 0.16.1 - Dr Mark Goss-Sampson

Raincloud plots

Jump power: Gender: Female
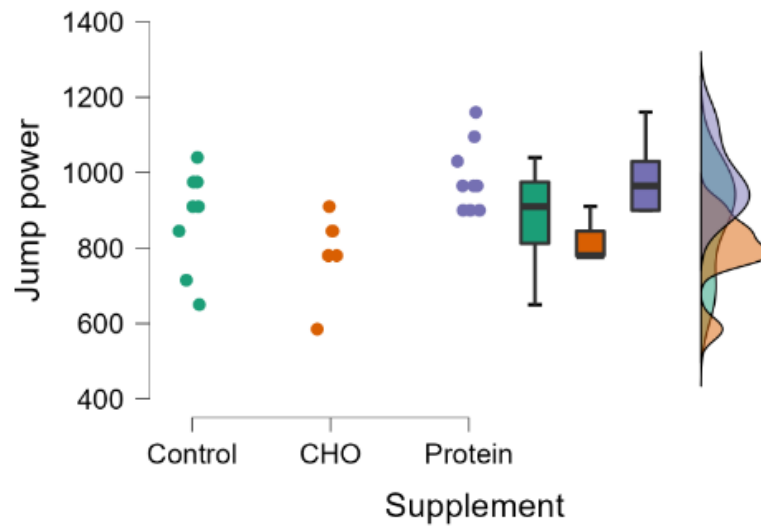
Jump power: Gender: Male
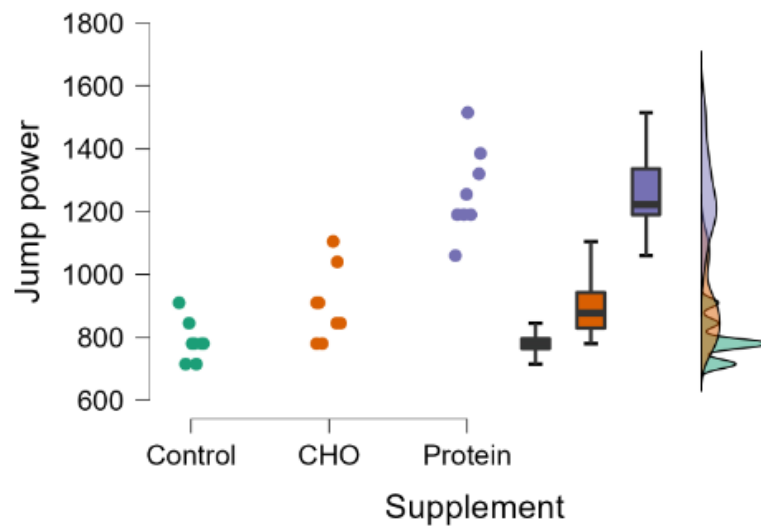
POST HOC TESTING

If the ANOVA is significant post hoc testing can now be carried out. In Post Hoc Tests add stretch, warm-up and the Stretching*warm-up interaction to the analysis box on the right, tick Effect size and, in this case, use Holm for the post hoc correction. Tick Flag significant comparisons.

JASP 0.16.1 - Dr Mark Goss-Sampson

Post hoc testing for the main effects confirms that there are significant differences in sit and reach distance when comparing the two levels of each factor. This is further decomposed in the Post hoc comparisons for the interaction.

Post Hoc Comparisons - Stretch

		Mean Difference	SE	t	Cohen's d	p_{holm}
Stretch	No stretch	4.532	0.408	11.11	3.207	< .001

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

Post Hoc Comparisons - Warm up

		Mean Difference	SE	t	Cohen's d	p_{holm}

Post Hoc Comparisons - Stretching ✱ Warm up

		Mean Difference	SE	t	p_{holm}
Stretch, Warm.up	No.Stretch, Warm.up	6.239	0.515	12.125	< .001***
	Stretch, No.warm.up	4.380	0.450	9.737	< .001***
	No.Stretch, No.warm.up	7.204	0.520	13.854	< .001***
No.Stretch, Warm.up	Stretch, No.warm.up	−1.859	0.520	−3.575	0.004**
	No.Stretch, No.warm.up	0.965	0.450	2.145	0.043*
Stretch, No.warm.up	No.Stretch, No.warm.up	2.824	0.515	5.488	< .001***

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

Note. P-value adjusted for comparing a family of 6

JASP 0.16.1 - Dr Mark Goss-Sampson

REPORTING THE RESULTS

A two-way ANOVA was used to examine the effect of stretch and warm-up type on sit and teach performance. There were significant main effects for stretch (F (1, 11) = 123.4, $p<.001$, $\omega^2 = 0.647$) and warm-up (F (1, 11) = 68.69, $p<.001$, $\omega^2 = 0.404$). There was a statistically significant interaction between the effects of stretch and warm up on sit and reach performance (F (1, 11) = 29.64, $p<.001$, $\omega^2 = 0.215$).

Simple main effects showed that sit and reach performance was significantly higher when both a stretch and warm-up had been done (F (1) =112.6, $p<.001$).

JASP 0.16.1 - Dr Mark Goss-Sampson

MIXED FACTOR ANOVA

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 (Factor 2)	Independent variable (Factor 1) = time or condition		
	Time/condition 1	**Time/condition 2**	**Time/condition 3**
Group 1	Dependent variable	Dependent variable	Dependent variable
Group 2	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. I**nteraction** 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 (Factor 2)	Independent variable (Factor 1) = time or condition		
	Time/condition 1	**Time/condition 2**	**Time/condition 3**
Group 1	All data	All data	All data
Group 2			

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

Independent variable (Factor 2)	Independent variable (Factor 1) = time or condition		
	Time/condition 1	**Time/condition 2**	**Time/condition 3**
Group 1	All data		*
Group 2	All data		

Simple main effects are effectively pairwise comparisons:

Independent variable (Factor 2)	Independent variable (Factor 1) = time or condition		
	Time/condition 1	**Time/condition 2**	**Time/condition 3**
Group 1	Data	Data	Data
Group 2	Data *	Data *	Data *

JASP 0.16.1 - Dr Mark Goss-Sampson

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 that should either be addressed in the research design or can the 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 and70%). 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:
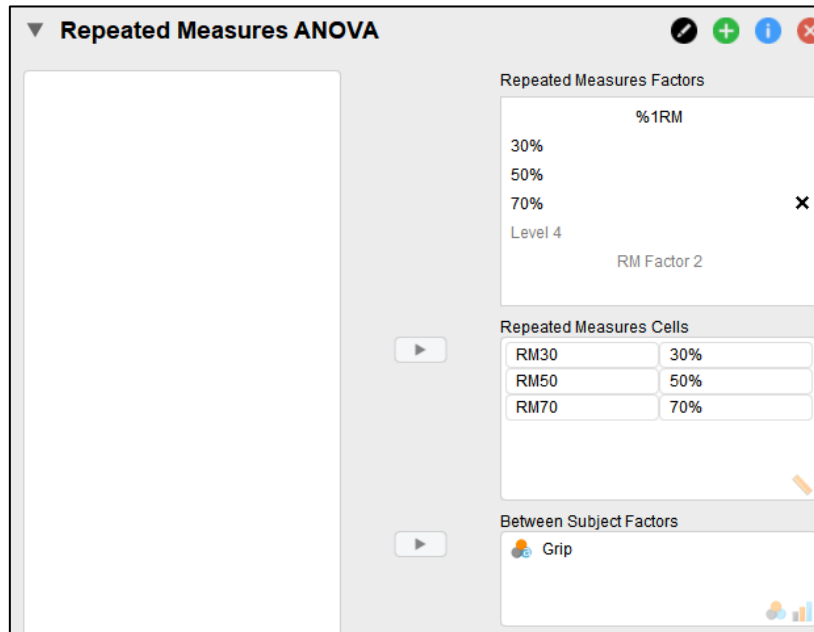
Additionally, tick Descriptive statistics and Estimates of effect size (ω2).
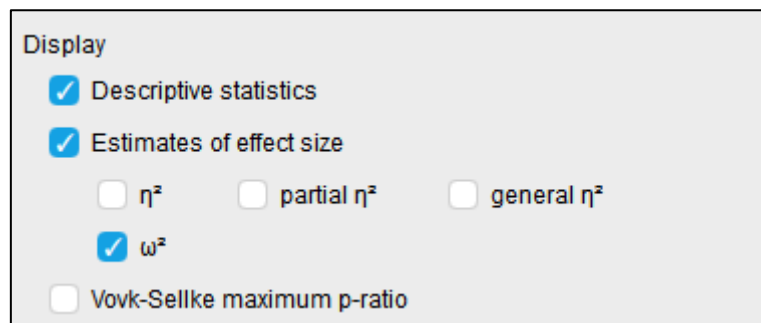
In Descriptive plots, move %1RM to the horizontal axis and Grip to separate lines. It is now possible to add a title for the vertical axis.
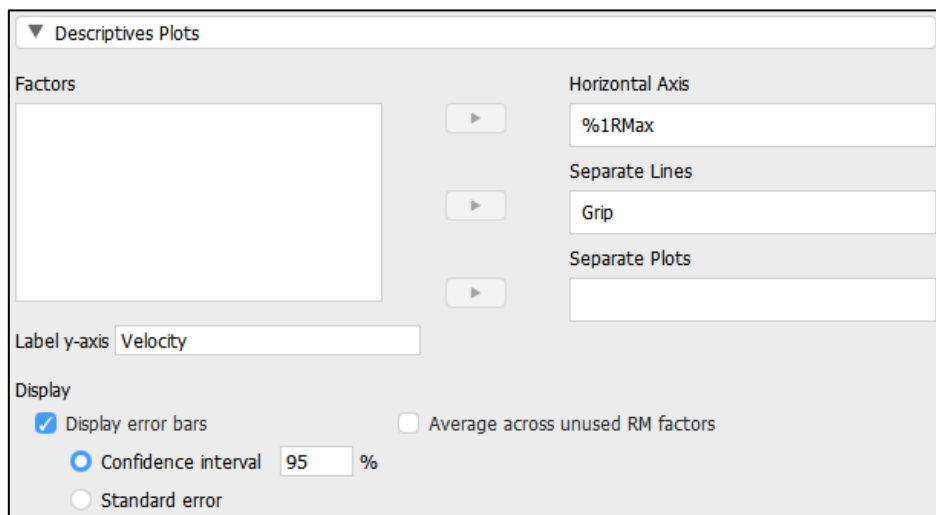
JASP 0.16.1 - Dr Mark Goss-Sampson

UNDERSTANDING THE OUTPUT

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	ω^2
%1RM	5.605ᵃ	2ᵃ	2.803ᵃ	115.450ᵃ	< .001ᵃ	0.744
%1RM ✳ Grip	0.583ᵃ	2ᵃ	0.291ᵃ	12.003ᵃ	< .001ᵃ	0.218
Residual	0.874	36	0.024			

Note. Type III Sum of Squares
ᵃ Mauchly's test of sphericity indicates that the assumption of sphericity is violated (p < .05).

The output should initially comprise 3 tables and 1 graph.

For the main effect for %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	ω^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 for 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
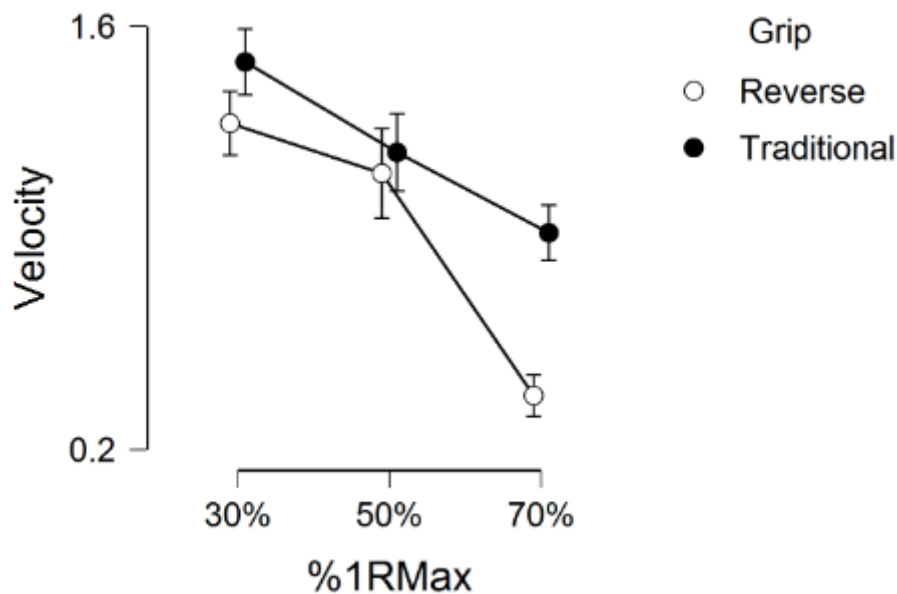
TESTING ASSUMPTIONS

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

Test of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
%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:

JASP 0.16.1 - Dr Mark Goss-Sampson

Within Subjects Effects ▼

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω²
%1RM	Greenhouse-Geisser	5.605ᵃ	1.480ᵃ	3.787ᵃ	115.450ᵃ	< .001ᵃ	0.744
%1RM * Grip	Greenhouse-Geisser	0.583ᵃ	1.480ᵃ	0.394ᵃ	12.003ᵃ	< .001ᵃ	0.218
Residual	Greenhouse-Geisser	0.874	26.639	0.033			

Note. Type III Sum of Squares

ᵃ 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 Holm for the post hoc correction. Only Bonferroni or Holm's corrections are available for repeated measures.

JASP 0.16.1 - Dr Mark Goss-Sampson
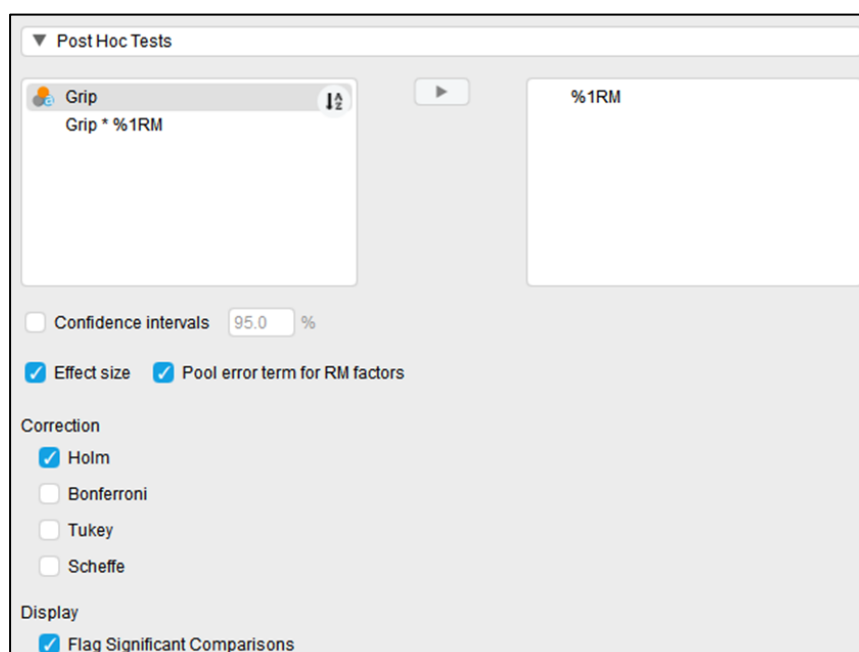
Post Hoc Comparisons - %1RMax

		Mean Difference	SE	t	Cohen's d	p_{bonf}
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 the load increases.
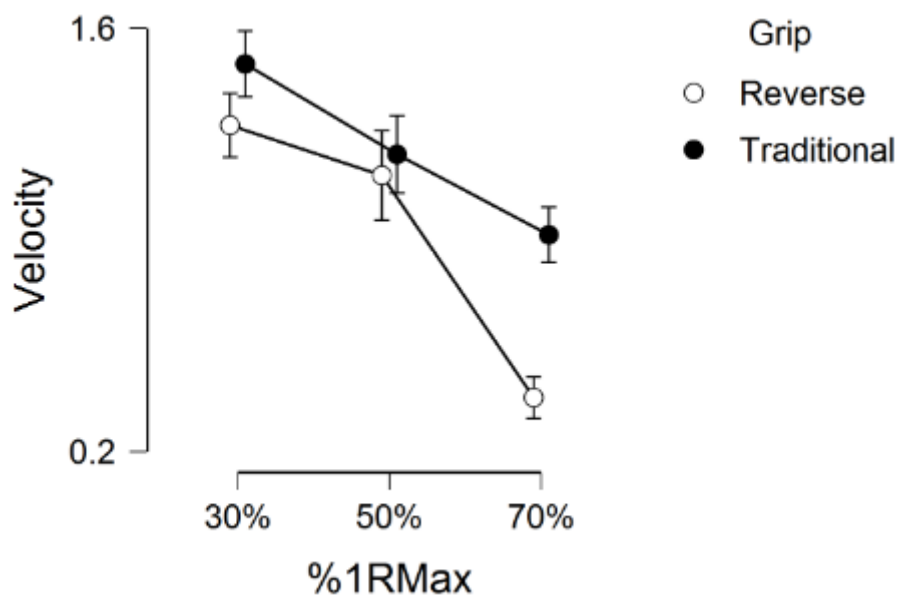
Finally, In Simple main effects add Grip to the Simple effect factor and %1RM to Moderator factor 1

JASP 0.16.1 - Dr Mark Goss-Sampson

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 (χ^2) test for independence (also known as Pearson's χ^2 test or the χ^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 χ^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 that 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:

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

Most parametric models are based on population means and standard deviations. The χ^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.

Equal Distribution	Male	Female	Row Total
Tall	25	25	50
Short	25	25	50
Column Total	50	50	

Observed Distribution	Male	Female	Row Total
Tall	57	24	**81**
Short	14	5	**19**
Column Total	**71**	**29**	

The model based on expected values can be calculated by:

Model (expected)   = ($row\ total\ x\ column\ total$)/100

Model – tall male   = (81 x 71) /100 = 57.5
Model – tall female   = (81 x 29) /100 = 23.5
Model – small male   = (19 x 71) /100 = 13.5
Model – small female   = (19 x 29) /100 = 5.5

These values can then be added to the contingency table:

JASP 0.16.1 - Dr Mark Goss-Sampson

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

the $\chi 2$ statistic is derived from $\sum \dfrac{(\textbf{observed} - \textbf{expected})^2}{\textbf{expected}}$

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	Abrahim, 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 chi square.csv** in JASP, add survived to rows as the dependent variable and sex into columns as the independent variable.
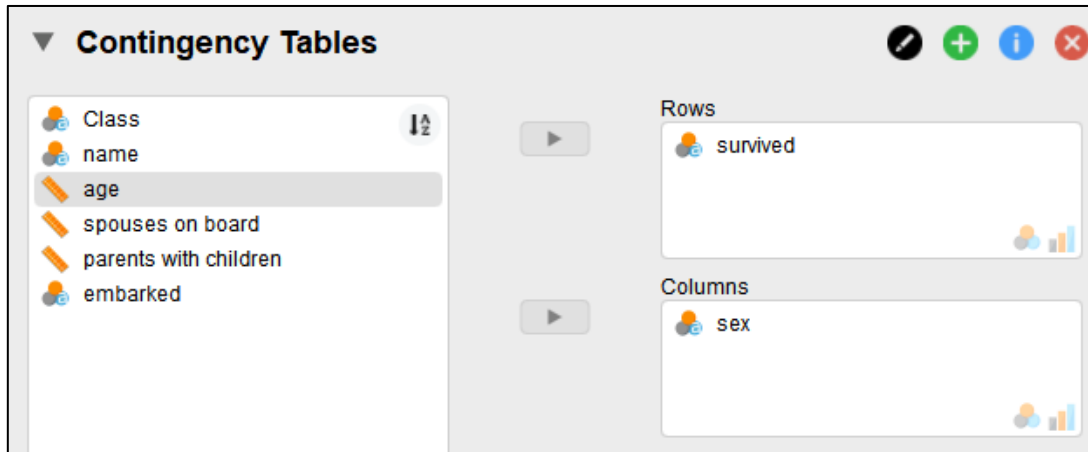
In statistics, tick all the following options:
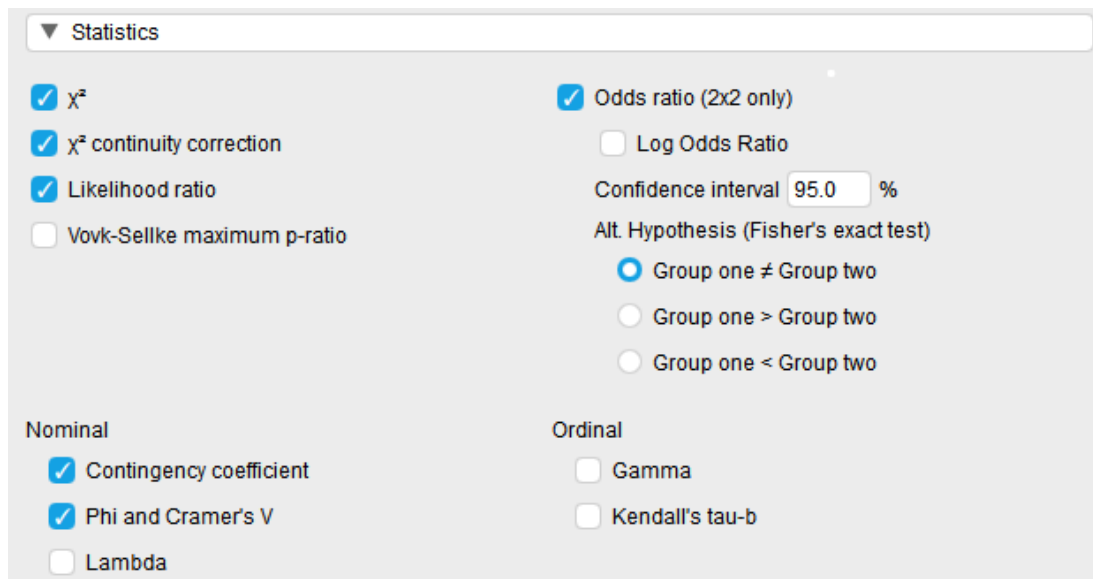
In Cells, tick the following:

JASP 0.16.1 - Dr Mark Goss-Sampson
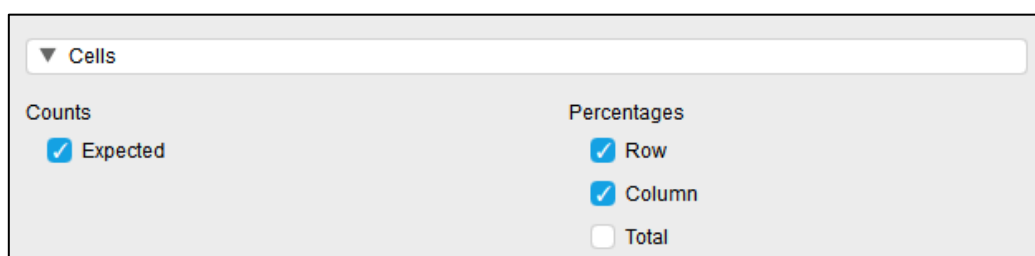
UNDERSTANDING THE OUTPUT

First, look at the Contingency table output.

Contingency Tables

survived		sex female	sex male	Total
		female	male	Total
No	Count	127.00	682.00	809.00
	Expected count	288.00	521.00	809.00
	% within row	15.70 %	84.30 %	100.00 %
	% within column	27.25 %	80.90 %	61.80 %
Yes	Count	339.00	161.00	500.00
	Expected count	178.00	322.00	500.00
	% within row	67.80 %	32.20 %	100.00 %
	% within column	72.75 %	19.10 %	38.20 %
Total	Count	466.00	843.00	1309.00
	Expected count	466.00	843.00	1309.00
	% within row	35.60 %	64.40 %	100.00 %
	% within column	100.00 %	100.00 %	100.00 %

Remember that χ^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
X²	365.9	1	< .001
X² continuity correction	363.6	1	< .001
Likelihood ratio	372.9	1	< .001
N	1309		

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

χ^2 continuity correction can be used to prevent the 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 the maximum-likelihood theory. For large samples, it is identical to Pearson $\chi2$. 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 has a large effect size.

Nominal

	Value
Contingency coefficient	0.47
Phi-coefficient	0.53
Cramer's V	0.53

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 [4]	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).

Odds Ratio

	Odds Ratio	95% Confidence Intervals		p
		Lower	Upper	
Odds ratio	0.088	0.068	0.115	
Fisher's exact test	0.089	0.067	0.116	< .001

For ease of understanding take the reciprocal of the OR (1/0.088) i.e. 11.36. This suggests that male passengers had 11.36 times more chance of dying than females.

---

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

JASP 0.16.1 - Dr Mark Goss-Sampson

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

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

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

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

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

JASP 0.16.1 - Dr Mark Goss-Sampson

Contingency Tables

survived		sex		
		female	male	Total
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 $p<.001$.

JASP 0.16.1 - Dr Mark Goss-Sampson

META-ANALYSIS IN JASP

Background

A meta-analysis is a statistical analysis that integrates results from multiple studies, providing a single numerical value of the overall treatment effect for that group of studies. The difference between statistical data analysis and meta-analysis is shown below. Effectively each study becomes a "participant" in the meta-analysis.

Statistical analysis		Meta-analysis	
Participant 1	Individual data	Study 1	Study data
Participant 2	Individual data	Study 2	Study data
Participant 3	Individual data	Study 3	Study data
Participant 4	Individual data	Study 4	Study data
Participant 5	Individual data	Study 5	Study data
Overall group data & statistics		Overall study group data & statistics	

Effect size and calculations

To perform a Meta-analysis in JASP the overall effect size (ES) and standard error (SE) of the study needs to be provided. An ES is a dimensionless estimate (no units) that indicates both direction and magnitude of the effect/outcome. The standard error measures the dispersion of different sample means taken from the same population and can be estimated from a single sample standard deviation. Some studies will provide this information, although many do not. However, they will provide results such as:

- Central tendencies and dispersion
- T or F statistics
- P-values
- Correlation coefficients
- Chi-square

All of these can be converted to an ES and the SE determined using David Wilson's "Meta-analysis effect size calculator"

https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php

For example, a study comparing a treatment to a control group may provide the variable mean, SD and n of each group. From this, the ES, a standardised mean difference (d) and estimated error variance (v) can be calculated:

JASP 0.16.1 - Dr Mark Goss-Sampson

Means, Standard Deviations, and Sample Sizes

	Mean	SD	N
Treatment	6.25	1.4	10
Control	4.95	1.7	10
	Calculate	Reset	
d =	0.8348		
95% C.I. =	-0.0791	1.7487	
v =	0.2174		

The 95% confidence intervals (CI) can be used to calculate the SE using the following equation:

SE = (upper CI limit – lower CI limit) / 3.92.   *NB: 3.92 = 2 x SD =95%*

SE = (1.7487- (-0.0791) / 3.92

SE = 0.466

A quicker way is to find the square root of the estimated error variance (v)

SE = √0.2174 = 0.466

Therefore, for this study the overall ES = 0.835 and SE = 0.466.

These data can be put into Excel and then used to perform a meta-analysis in JASP. The overall ES derived from the meta-analysis is calculated by combining the effect sizes of the included studies.

To interpret the meta-analysis output, one needs to understand the following concepts, heterogeneity, the model, effect size, and the forest plot.

Heterogeneity:

Heterogeneity describes any variability that may exist between the different studies. It is the opposite of homogeneity, which means that the population/data/results are the same. There are 3 types:

- Clinical: Differences in participants, interventions, or outcomes
- Methodological: Differences in study design, risk of bias
- Statistical: Variation in intervention effects or results

If there is no variability, then data can be described as homogenous. Meta-analysis is concerned with statistical heterogeneity. Statistical heterogeneity is used to describe variability among data/studies and occurs when the treatment effect estimates of a set of data/studies vary among one another. Studies with methodological flaws and small studies may overestimate treatment effects and can contribute to statistical heterogeneity.
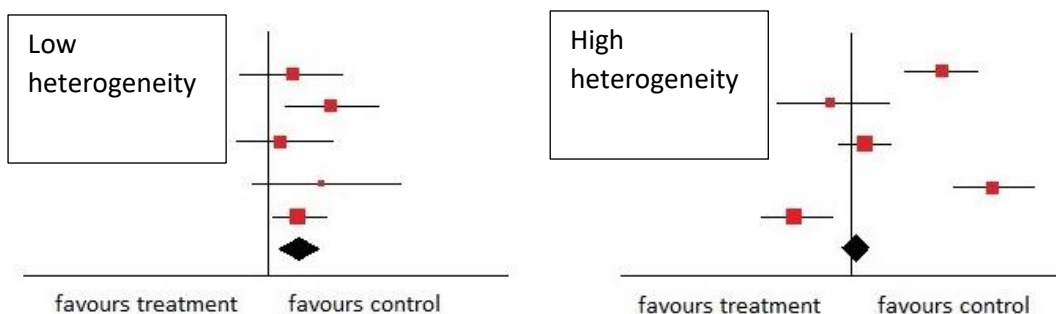
The diagram above[5] shows examples of forest plots exhibiting low and high heterogeneity. In the low condition, all the studies are generally lined up to the right of the vertical axis and all confidence intervals are overlapping. In the high condition, the studies are spread over either side of the vertical decision line and there is little overlap of the confidence intervals. Apart from visual observation of the level of heterogeneity meta-analysis provides quantitative statistical methods to measure heterogeneity.

Tests for heterogeneity

Q Cochran Q test - values are given for 2 tests outlined below, (these tests have low statistical power when only a few studies are included in the meta-analysis):

'Omnibus test of the Model Coefficients' tests the null hypothesis that all the ES are zero.

'Test for residual heterogeneity' tests the null hypothesis that all the effect sizes in the studies are all equal (homogeneity of effect sizes).

τ^2 Tau square is an estimate of the total amount of heterogeneity. This is interpreted as systematic unexplained differences between the observed effects of the separate studies. It is not affected by the number of studies but it is often hard to interpret how relevant the value is from a practical standpoint.

I^2 This measures the extent of heterogeneity not caused by sampling error. If there is high heterogeneity possible sub-group analysis could be done. If the value is very low, there is no point in doing further subgroup analyses. I^2 is not sensitive to changes in the number of studies and is therefore used extensively in medical and psychological research, especially since there is a **"rule of thumb"** to interpret it. A rough guide for interpretations has been suggested as follows[6]:

- 0% to 40%: might not be important.
- 30% to 60%: may represent moderate heterogeneity.
- 50% to 90%: may represent substantial heterogeneity.
- 75% to 100%: considerable heterogeneity.

H^2 The between-study variance is determined by equating the Q statistic to its expected value. H^2 has a value of 1 in the case of homogeneity and heterogeneity is assumed to be present when $H^2>1$.

Meta-analysis models

There are two models commonly used in meta-analysis and each makes different assumptions relating to the observed differences among the studies

---

[5] https://s4be.cochrane.org/blog/2018/11/29/what-is-heterogeneity/
[6] Cochrane Handbook for Systematic Reviews of Interventions

JASP 0.16.1 - Dr Mark Goss-Sampson

Fixed effects model: this assumes that all studies share a common true ES i.e. the data is homogeneous. All factors that could influence the ES are the same in all the study samples and therefore very little heterogeneity. Between-study differences are assumed to be due to chance and thus not incorporated into the model. Therefore, each study included in the meta-analysis is estimating the same population treatment effect, which, in theory, represents the true population treatment effect. Each study is weighted where more weight is given to studies with large samples sizes i.e. more information.

This model answers the following question **"What is the best estimate of the population effect size?"**

Random effects model: this assumes a distribution of the treatment effect for some populations. i.e. that the different studies are estimating different, yet related, intervention effects. Therefore, heterogeneity cannot be explained because it is due to chance. This model assigns a more balanced weighting between studies.

This model answers the question **"What is the average treatment effect".**

> *It is therefore important to check for significant heterogeneity to select the correct model for the meta-analysis.*

Random effects model selection.

JASP provides 8 random effects estimators to estimate the 4 indices of heterogeneity (see tests for heterogeneity above). All use a slightly different approach resulting in different pooled ES estimates and CIs. To date, the most often used estimator in medical research is the DerSimonian-Laird estimator however, recent studies have shown better estimates of between-study variance using Maximum-Likelihood, Empirical Bayes and Sidak-Jonkman methods.

The Forest Plot.

This plot provides a visual summary of the meta-analysis findings. Graphically it represents the ES and 95% CI for each study and an estimate of the overall ES and 95% CI for all the studies that were included. As previously mentioned, it can be used to visually assess the level of heterogeneity.

JASP 0.16.1 - Dr Mark Goss-Sampson
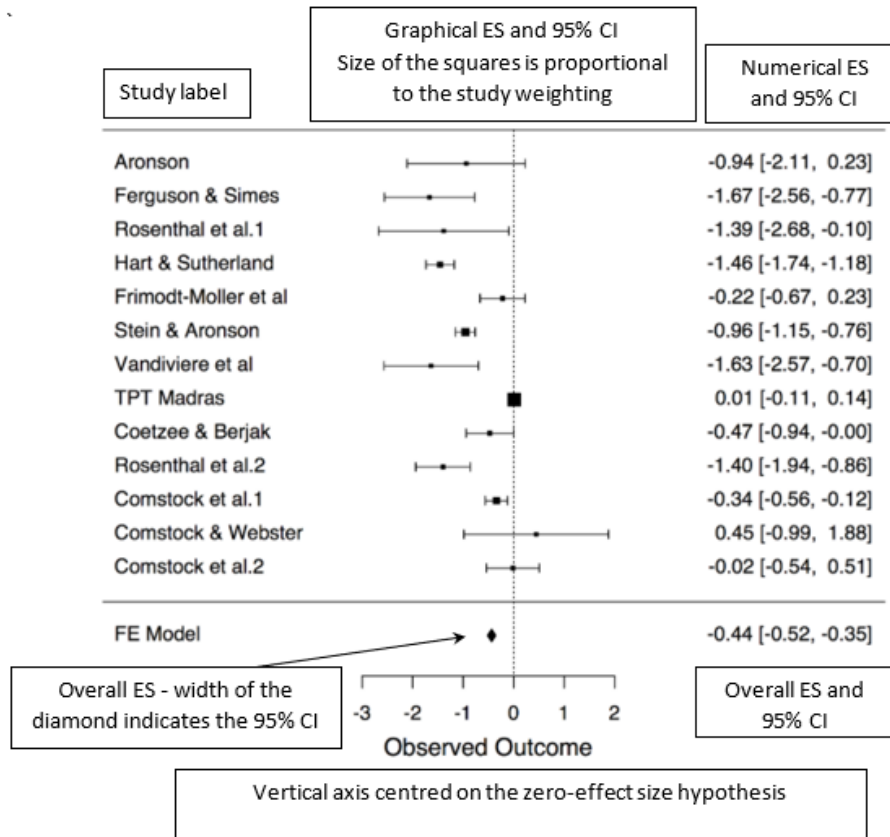
As a rule of thumb, if all the individual study CIs cover the final combined ES and its 95% CI (diamond) then there is low heterogeneity (or high homogeneity). In this case, 6 studies do not intersect with the diamond.

Funnel plot

Looking like an inverted funnel, this is a scatter plot of the intervention effect estimates from individual studies against each study's size or precision. The intervention effect should increase as the size of the study increases. Studies with low power will cluster at the bottom while higher power studies will be near the top. Ideally, the data points should have a symmetrical distribution. Funnel plot asymmetry indicates that the meta-analysis results may be biased. Potential reasons for bias include:

- Publication bias (the tendency of authors/publishers to only publish studies with significant results)
- True heterogeneity
- Artefact (wrong choice of effect measure)
- Data irregularities (methodological design, data analysis etc)

Funnel plot asymmetry can be analysed statistically using either a Rank-test or Eggers test.

JASP 0.16.1 - Dr Mark Goss-Sampson

Funnel Plot
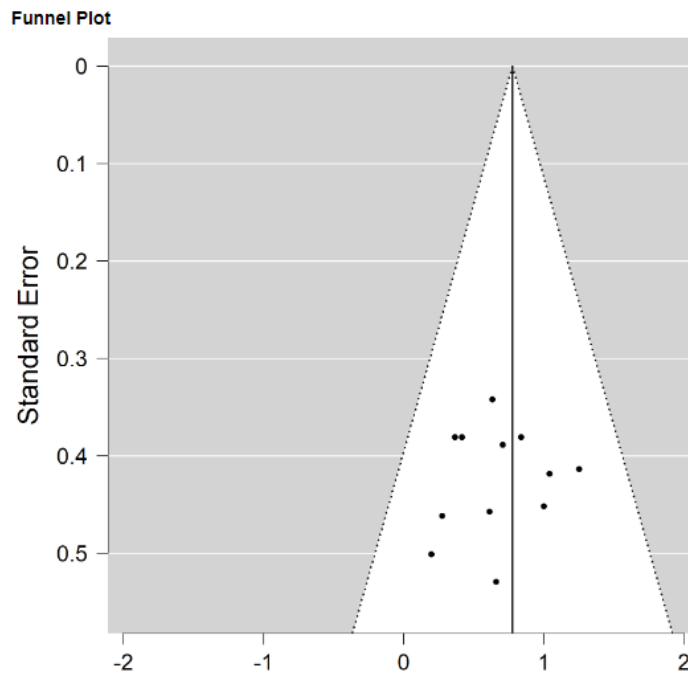
Running a meta-analysis in JASP

If you have not added the Meta-analysis module, click the blue cross at the top right corner of JASP and tick 'Meta-Analysis', this will add it to the main menu ribbon. Go to 'Meta-Analysis' and click 'Classical Meta-Analysis'.

Open meta analysis.csv. This contains the effect sizes and SEM from 14 studies for changes in DOMS pain scores in groups receiving no therapy vs cryotherapy. To start, add ES to 'Effect Size', SEM to 'Effect Size Standard Error' and Author to 'Study Labels'. Change the 'Method' to 'Fixed Effects'.
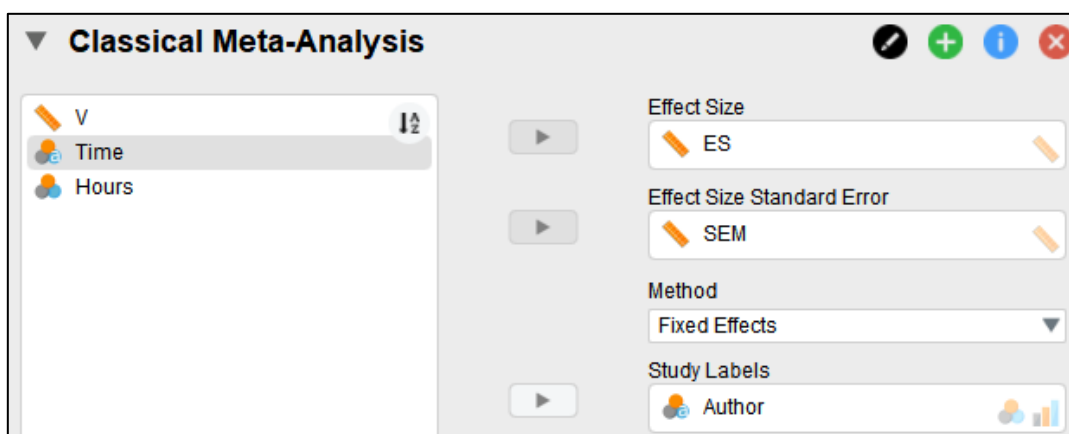
Initial output

This should produce two tables:

JASP 0.16.1 - Dr Mark Goss-Sampson

Fixed and Random Effects

	Q	df	p
Omnibus test of Model Coefficients	33.565	1	< .001
Test of Residual Heterogeneity	30.762	13	0.004

Note. p-values are approximate.

Coefficients

	Estimate	Standard Error	z	p
intercept	0.645	0.111	5.794	< .001

Note. Wald test.

The significance of the ES is shown in the 'Fixed and Random Effects' table:

Omnibus test of model coefficients tests the H_0 that all the coefficients in the second table are all zero i.e. the interventions have no significant effect. As can be seen, this is significant and H_0 can be rejected, thus the intervention has a significant effect.

The test of residual heterogeneity tests the H_0 that all the effect sizes in all the studies are equal. If this is not significant a fixed-effects model should be used. However, in this case, it is significant and therefore a random-effects model would be more appropriate.

The coefficients table provides an estimate (Wald test) of the overall effect size of the combined study and its significance. This confirms the result of the Omnibus test in the first table.

Now we know that a random-effects model is more appropriate to this dataset, return to the options and change the 'Method' to 'Restricted ML (Maximum Likelihood)'. This will then add another table showing the different estimates of residual homogeneity:

Residual Heterogeneity Estimates

	Estimate
τ^2	0.242
τ	0.492
I^2 (%)	58.150
H^2	2.390

Both τ^2 and I^2 show excess variance (heterogeneity) between the studies thus supporting the use of a random-effects model.

Further options.

Return to the Statistics options and in model fit, tick 'Forest Plot', 'Funnel Plot' and 'Rank Test for Funnel Plot Asymmetry':

The Forest plot shows the weighted effect sizes (the size of the squares reflects the weight of each study) and CIs used to determine the combined ES (diamond).

It can be easily seen that the overall effects of cryotherapy have a significant positive effect on DOMS scores (ES = 0.66).

Forest plot

Study	Effect Size [95% CI]
Davis	1.25 [0.44, 2.06]
Jordan	0.20 [-0.78, 1.18]
French	0.66 [-0.38, 1.70]
Joy	0.61 [-0.28, 1.51]
King	2.20 [1.29, 3.11]
Stanley	0.63 [-0.04, 1.30]
Dawson	-0.77 [-1.53, -0.01]
Ingrid	1.00 [0.11, 1.89]
Johnson	0.37 [-0.38, 1.11]
Poree	0.28 [-0.63, 1.18]
Robeson	0.71 [-0.05, 1.47]
Valiant	1.04 [0.22, 1.86]
Thomas	0.42 [-0.33, 1.16]
Nutfield	0.84 [0.09, 1.58]
RE Model	0.66 [0.32, 1.00]

Observed Outcome
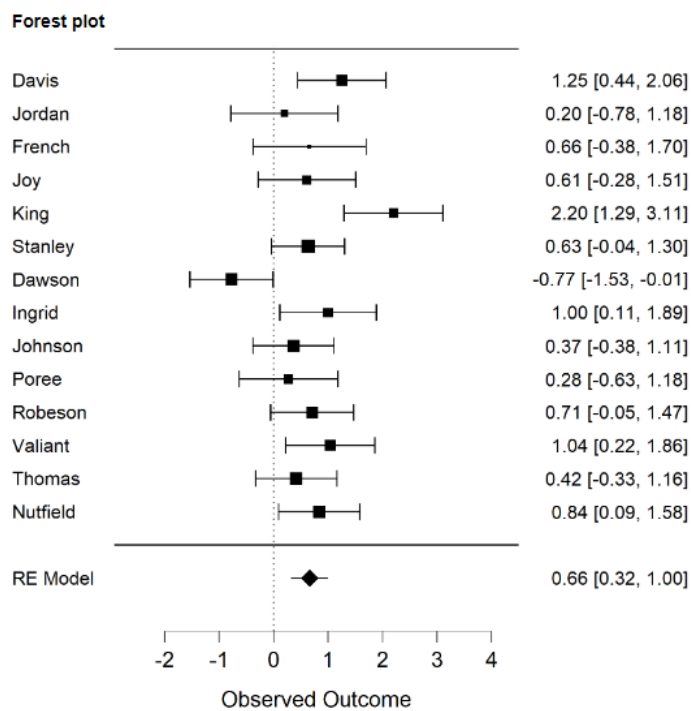
The funnel plot shows that the observed effects sizes appear to be symmetrically distributed around the vertical axis (based on the overall effect size estimate, in this case, 0.66) and lie within the 95% confidence triangle. Asymmetry is often reported as being indicative of publication bias. This plot is accompanied by the 'Rank Correlation Test' for funnel plot asymmetry which in this case is non-significant ($p=.912$).

JASP 0.16.1 - Dr Mark Goss-Sampson
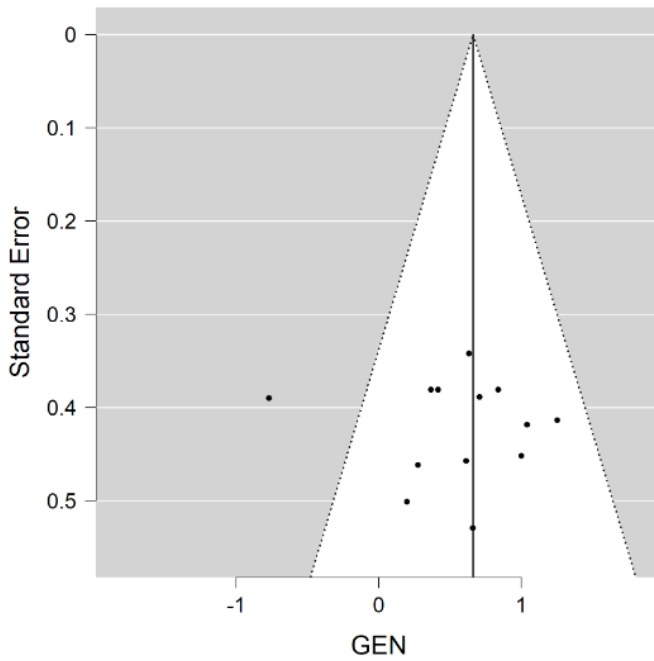
Funnel Plot

Rank correlation test for Funnel plot asymmetry

	Kendall's τ	p
Rank test	0.022	0.912

JASP 0.16.1 - Dr Mark Goss-Sampson

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 correlations

	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

JASP 0.16.1 - Dr Mark Goss-Sampson

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

JASP 0.16.1 - Dr Mark Goss-Sampson

One-way Independent ANOVA

Design example:

Independent variable	Group 1	Group 2	Group 3	Group...n
Dependent variable	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)

Levels

(Related groups)

	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 Repeated measures ANOVA

Design example:

Factor 1 Interventions	Level 1 i.e. intervention 1			Level 2 i.e. intervention 2		
Factor 2 Time	Level 1 i.e time 1	Level 2 i.e time 2	Level 3 i.e time 3	Level 1 i.e time 1	Level 2 i.e time 2	Level 3 i.e time 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 levels 1-n   **Factor 2 levels 1-n**
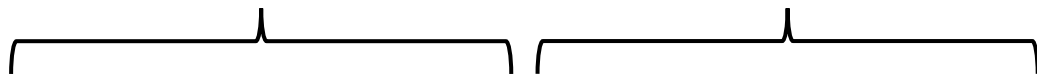
	A	B	C	D	E
1	Subject	Factor 1 level 1	Factor 1 level 2	Factor 2 level 1	Factor 2 level 2
2	A	7.38	6.52	9.27	14.32
3	B	7.71	10.83	11.48	16.38
4	C	6.19	10.42	9.77	15.45
5	D	9.27	11.78	15.45	16.96
6	E	11.41	9.52	11.65	15.64
7	F	5.29	5.82	9.22	13.01
8	G	8.54	9.43	10.92	17.35
9	H	7.89	8.43	8.26	12.57
10	I	5.49	6.64	11.39	14.02
11	J	9.26	9.36	13.03	16.24
12	K	6.9	7.09	9.02	14.7
13	L	8.57	9.64	8.33	13.71

JASP 0.16.1 - Dr Mark Goss-Sampson

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

(Categorical)

Factor 2 levels

(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 based on 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 has n 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 α, 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 is purely by chance.

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

JASP 0.16.1 - Dr Mark Goss-Sampson
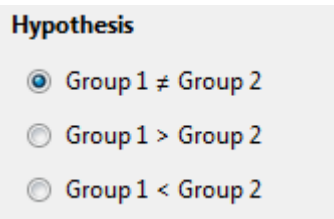
Type I and II errors

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)	**Type I error** An innocent person goes to prison	Correct decision
	Not guilty (H_0)	Correct decision	**Type II error** 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 **α**, 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
Between means	Cohen's d	<0.2	0.2	0.5	0.8
Correlation	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
Multiple Regression	Multiple correlation coefficient (R)	<0.10	0.1	0.3	0.5
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
Chi-squared	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:
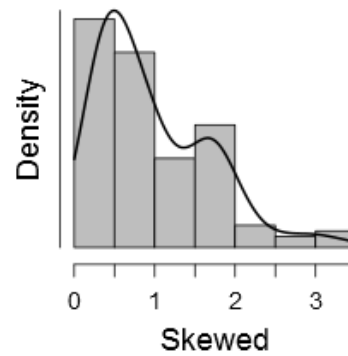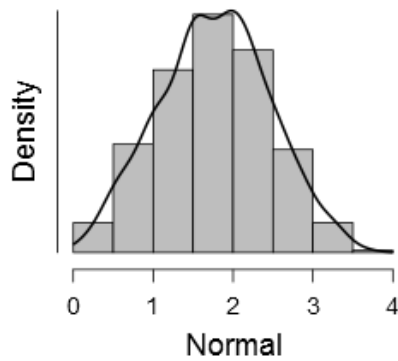
Sometimes transforming the data can rectify this but not always. It is also common to collect ordinal data (i.e. Likert scale ratings) for which terms such as mean and standard deviation are meaningless. As such there are no parameters associated with ordinal (**non-parametric**) data. The non-parametric counterparts include median values and quartiles.

In both of the cases described non-parametric statistical tests are available. There are equivalents to the most common classical parametric tests. These tests 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.

JASP 0.16.1 - Dr Mark Goss-Sampson

WHICH TEST SHOULD I USE?

Comparing one sample to a known or hypothesized population mean.

```
                        ┌──────────────┐
                        │  Data type   │
                        └──────────────┘
         ┌──────────────────┼───────────────────────────┐
         ▼                  ▼                            ▼
  ┌────────────┐     ┌────────────┐              ┌────────────┐
  │ Continuous │     │  Ordinal   │              │  Nominal   │
  └────────────┘     └────────────┘              └────────────┘
         │                  │              ┌──────────────┴──────────────┐
         ▼                  ▼              ▼                             ▼
  ┌────────────┐     ┌────────────┐  ┌──────────────┐          ┌──────────────┐
  │ One-sample │     │ One-sample │  │ 2 categories │          │ >2 categories│
  │   t-test   │     │ median test│  └──────────────┘          └──────────────┘
  └────────────┘     └────────────┘         │                         │
                                            ▼                         ▼
                     ┌──────────────┐ ┌──────────────┐     ┌──────────────────┐
                     │ Currently not│ │ Binomial test│     │ Multinomial test │
                     │ available in │ └──────────────┘     │  or Chi-square   │
                     │     JASP     │                      │ 'goodness of fit'│
                     └──────────────┘                      └──────────────────┘
```
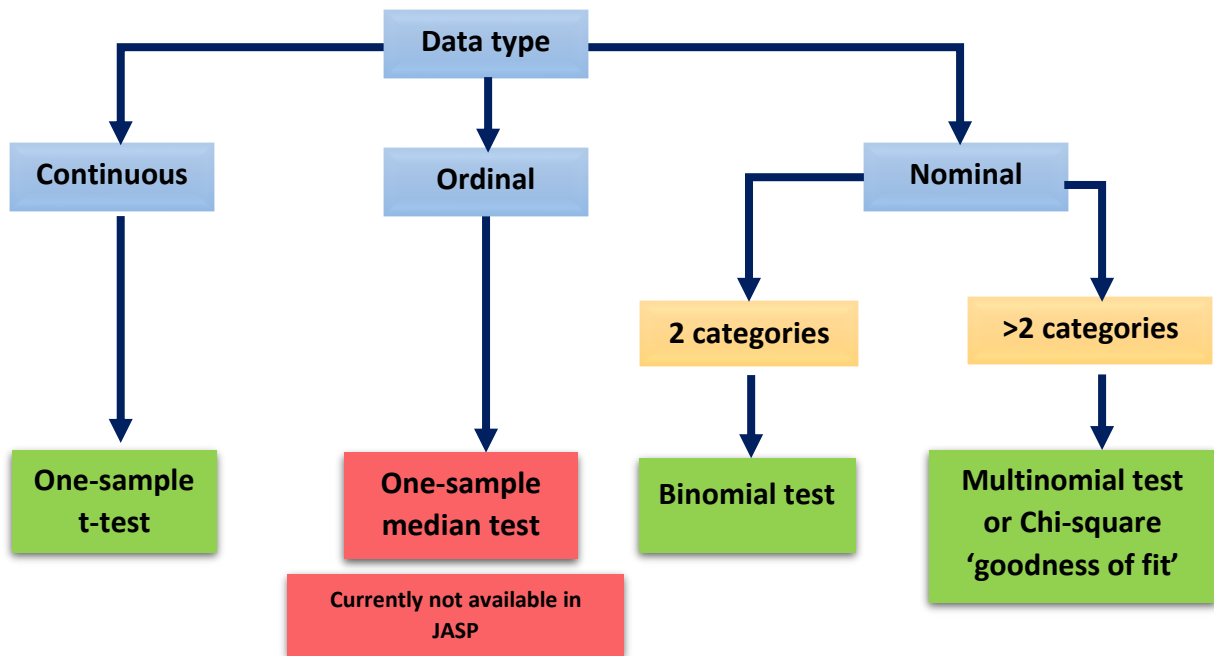
Testing relationships between two or more variables

```
                        ┌──────────────┐
                        │  Data type   │
                        └──────────────┘
         ┌──────────────────┼───────────────────────────┐
         ▼                  ▼                            ▼
  ┌────────────┐     ┌────────────┐              ┌────────────┐
  │ Continuous │     │  Ordinal   │              │  Nominal   │
  └────────────┘     └────────────┘              └────────────┘
         │                  │                           │
         ▼                  │                           │
  ┌────────────────┐        │                           │
  │ Are parametric │        │                           │
  │ assumptions met?│       │                           │
  └────────────────┘        │                           │
      │        │            │                           │
      ▼        ▼            │                           │
  ┌─────┐  ┌─────┐          │                           │
  │ Yes │  │ No  │          │                           │
  └─────┘  └─────┘          │                           │
      │        └────────┐   │                           │
      ▼                 ▼   ▼                            ▼
  ┌────────────┐  ┌──────────────┐            ┌──────────────┐
  │ Pearson's  │  │ Spearman's or│            │  Chi-square  │
  │correlation │  │ Kendall's tau│            │ contingency  │
  └────────────┘  └──────────────┘            │    tables    │
                                              └──────────────┘
```

JASP 0.16.1 - Dr Mark Goss-Sampson

Predicting outcomes

```
                              Data type
                                 │
        ┌────────────────────────┼────────────────────────┐
        ▼                        ▼                        ▼
   Continuous                 Ordinal                  Nominal
        │                        │                        │
        ▼                        ▼                        ▼
  More than one            Ordinal                   Logistic
  predictor variable?      regression                regression
        │                        │
   ┌────┴────┐          Currently not available in
   ▼         ▼                   JASP
   No        Yes
   │          │
   ▼          ▼
 Simple    Multiple
 regression regression
```
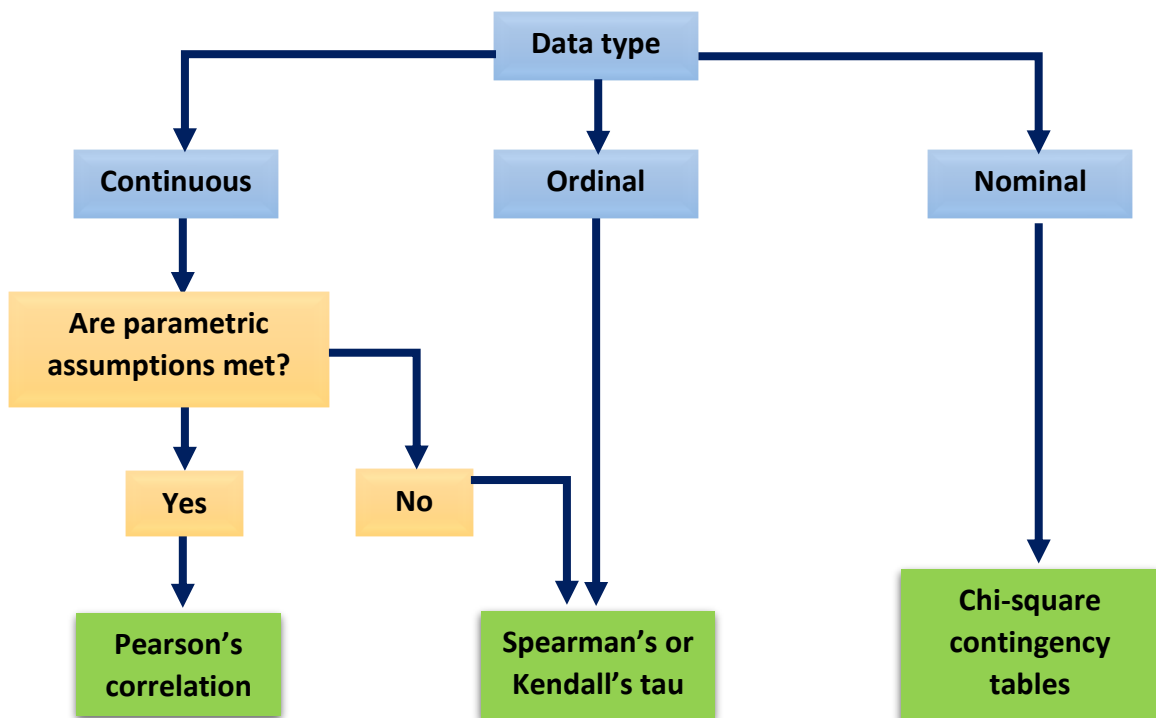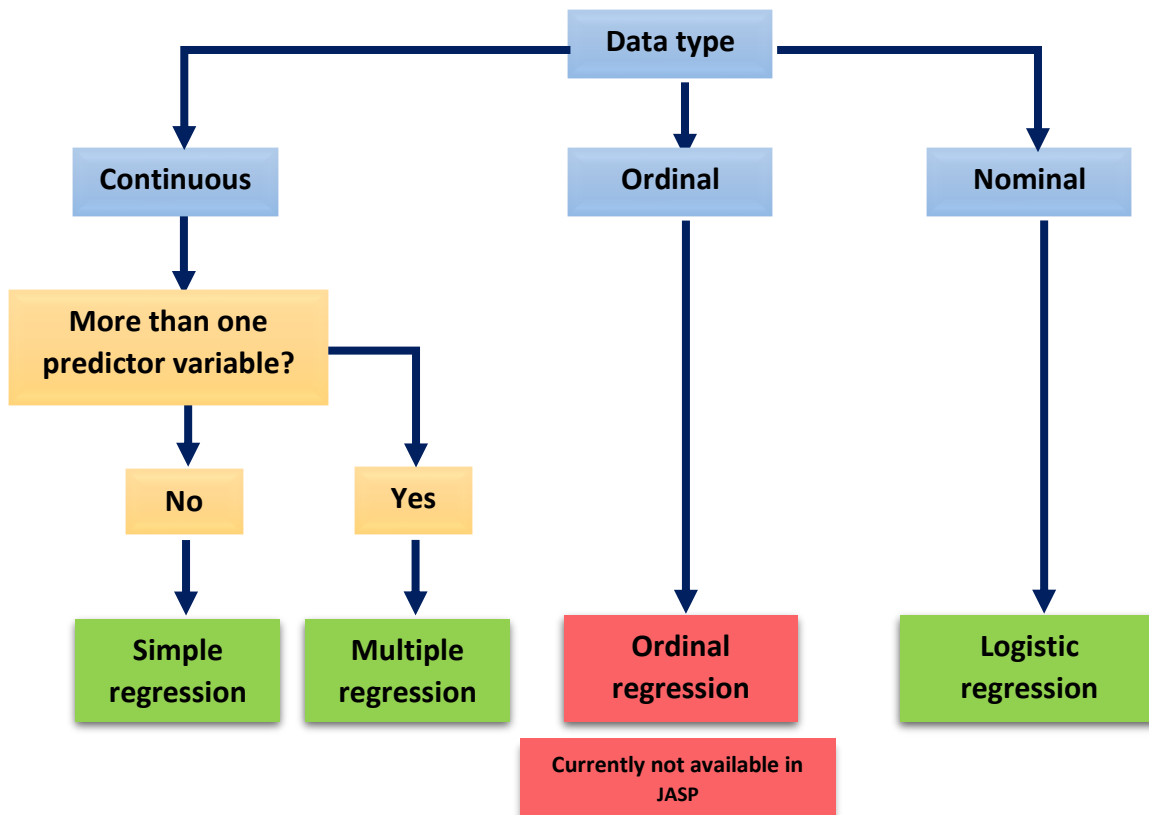
Testing for differences between two independent groups

```
                              Data type
                                 │
        ┌────────────────────────┼────────────────────────┐
        ▼                        ▼                        ▼
   Continuous                 Ordinal                  Nominal
        │                        │                        │
        ▼                        │                        │
  Are parametric               │                        │
  assumptions met?             │                        │
        │                        │                        │
   ┌────┴────┐                   │                        │
   ▼         ▼                   ▼                        ▼
   Yes       No          Mann-Whitney U          Chi-square or
   │          │              test               Fischer's Exact
   ▼          │                                      test
 Independent  │
 t-test       │
```
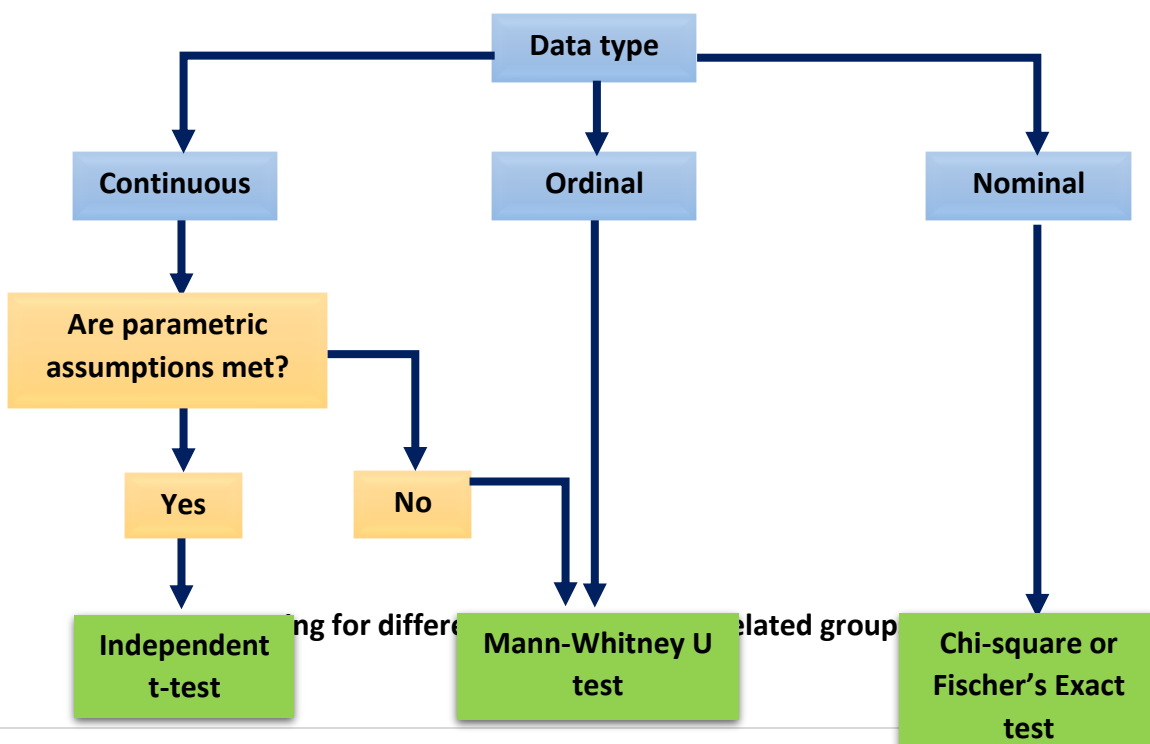
~~ing for differe~~ ~~elated group~~
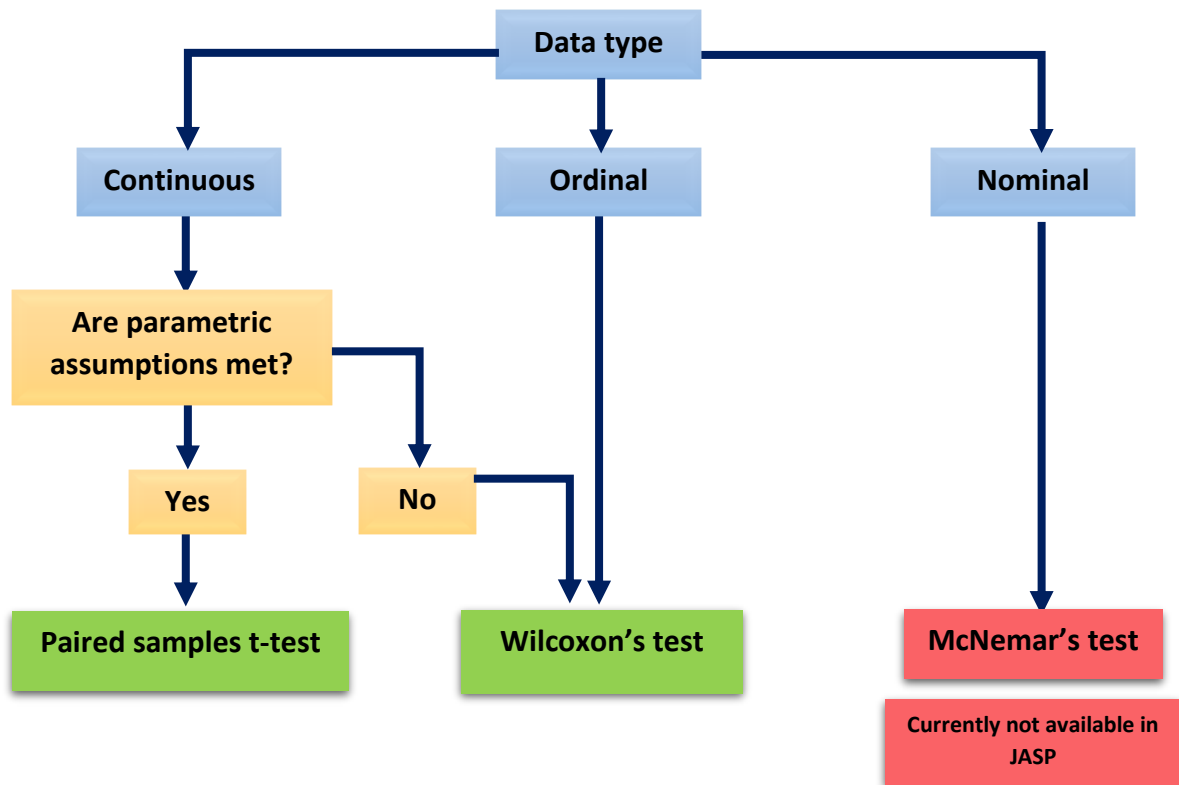
Testing for differences between three or more independent groups

JASP 0.16.1 - Dr Mark Goss-Sampson
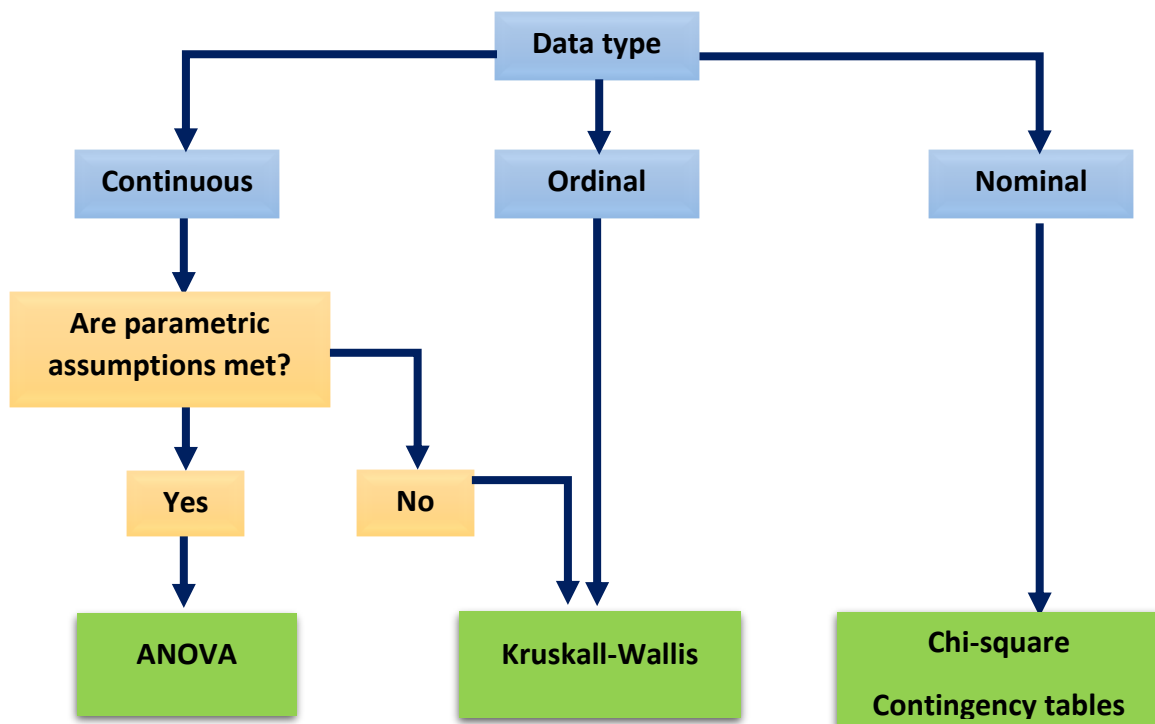
Testing for differences between three or more related groups

Data type
- Continuous
  - Are parametric assumptions met?
    - Yes → **RMANOVA**
    - No → **Friedman test**
- Ordinal → **Friedman test**
- Nominal → **Repeated measures logistic regression**
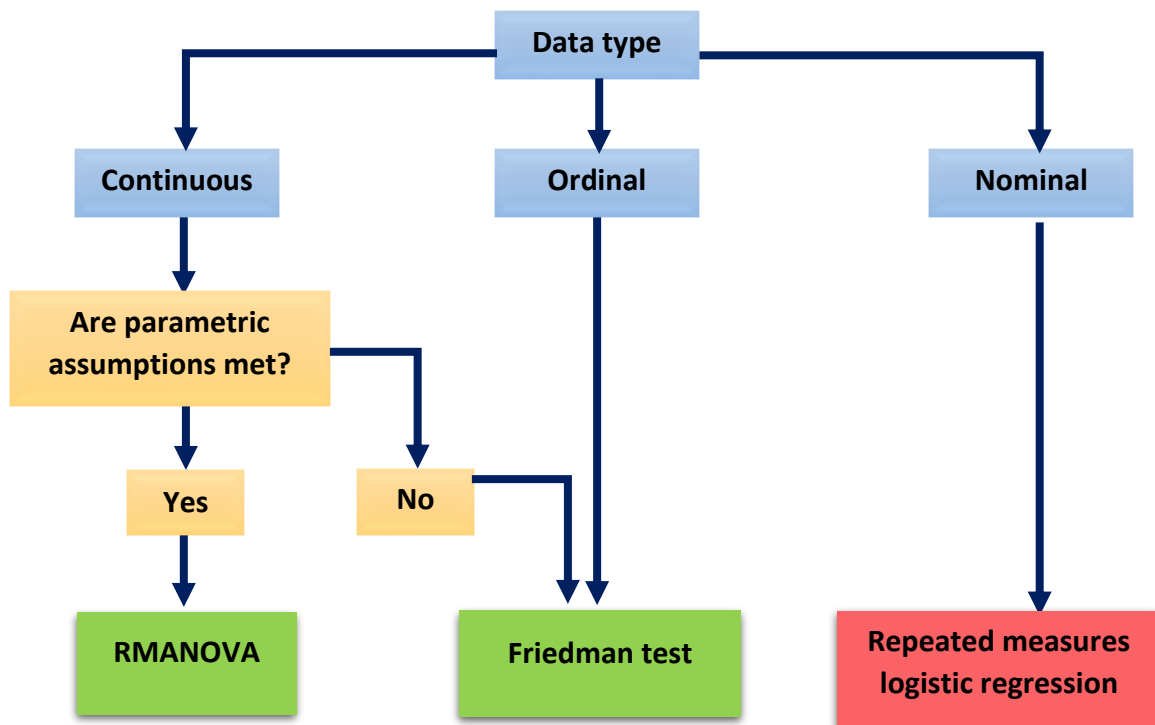
Test for interactions between 2 or more independent variables

Data type
- Continuous
  - Are parametric assumptions met?
    - Yes → **Two-way ANOVA**
    - No → **Ordered logistic regression**
- Ordinal → **Ordered logistic regression**
- Nominal → **Factorial logistic regression**