STAT193 Victoria University of Wellington

Help with iNZight v3.1.0

Summary of documents 1-10

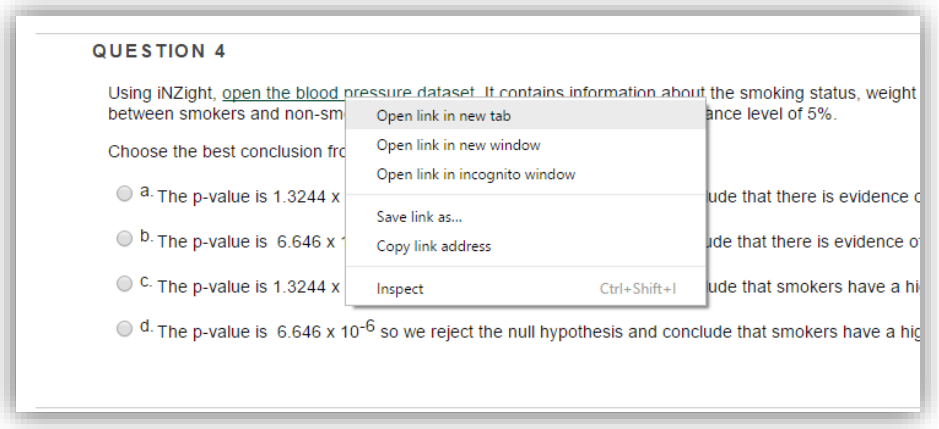
#1 – Saving spreadsheets, importing data, example (SURF) data

This is the first of 10 'Help with iNZight' documents to cover what you'll need for STAT193.

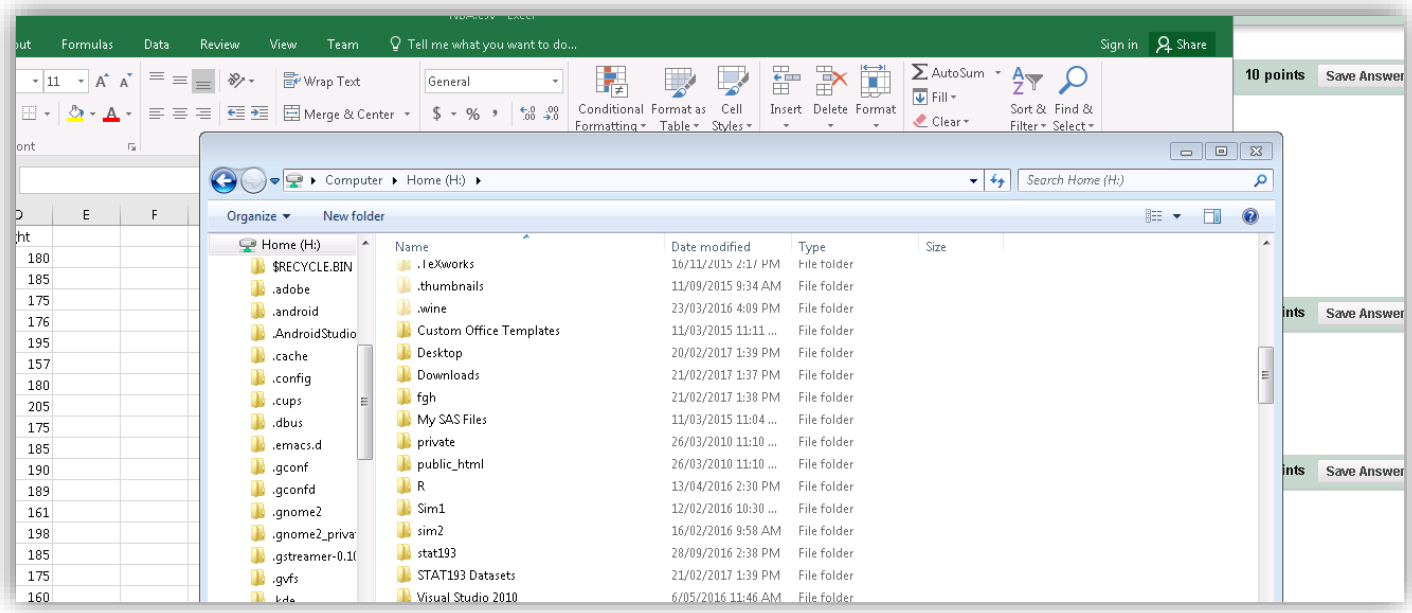
a. Open iNZight (version 3.1.0). If you are asked whether you want to create a new directory click 'yes'.

b. All datasets (spreadsheets) which you open will need to be saved before you can import them to iNZight. They actually need to be saved as csv (comma separated variable) files but if you are working on a tutorial quiz or a skills test the datasets will already be csv files.

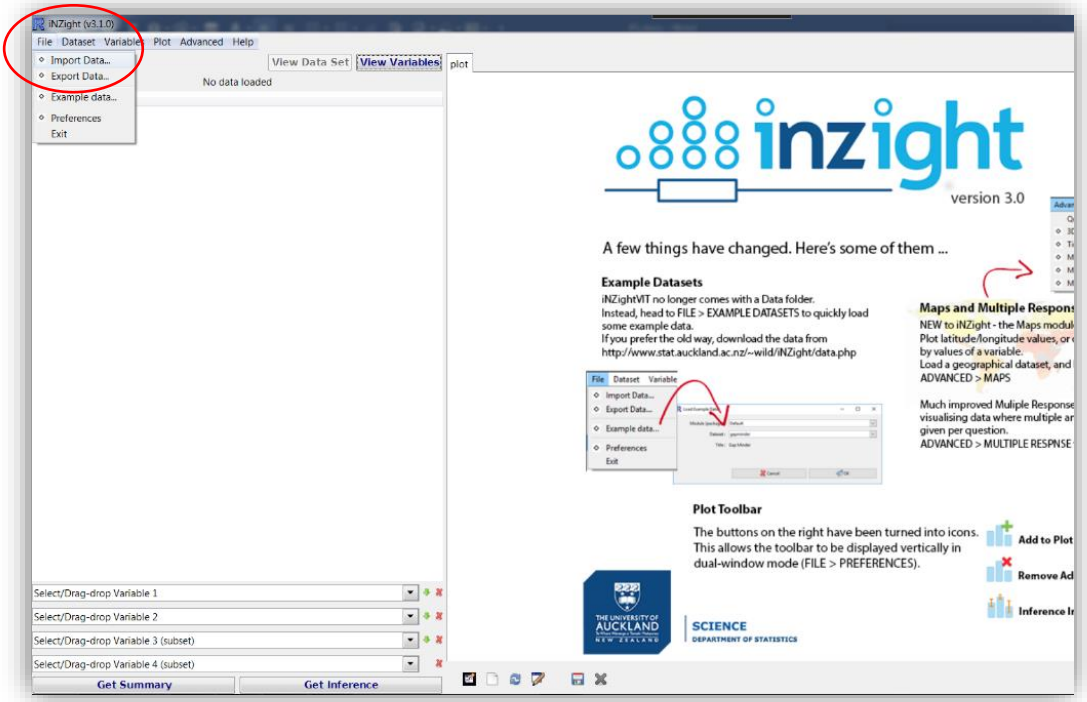
Right click on the dataset to open and select **Open link in new tab**. Then click on the downloaded dataset to open it:



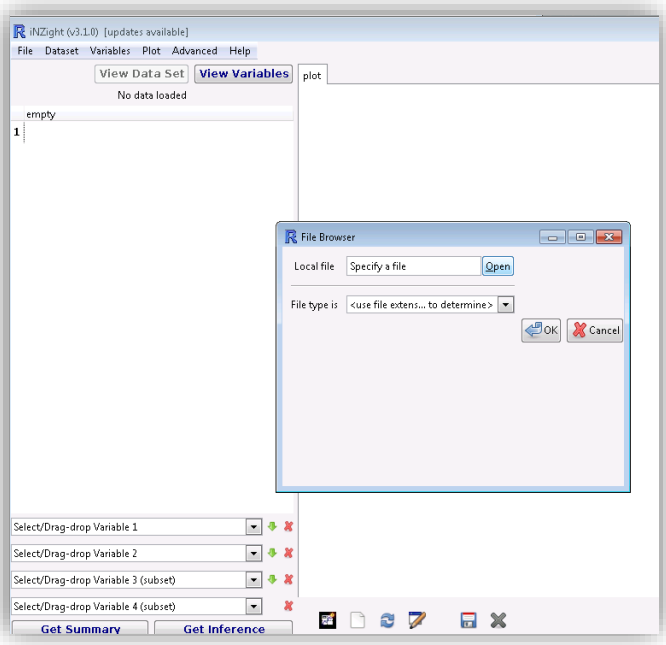
c. Save the dataset to your H-drive. We suggest you create a new folder which you label ‘STAT193 datasets’ in which to save all the datasets used in quizzes:

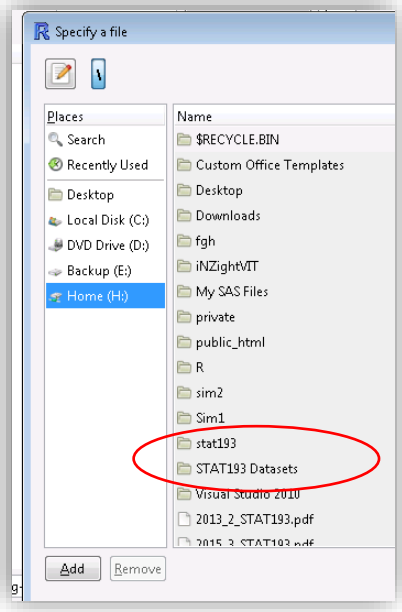


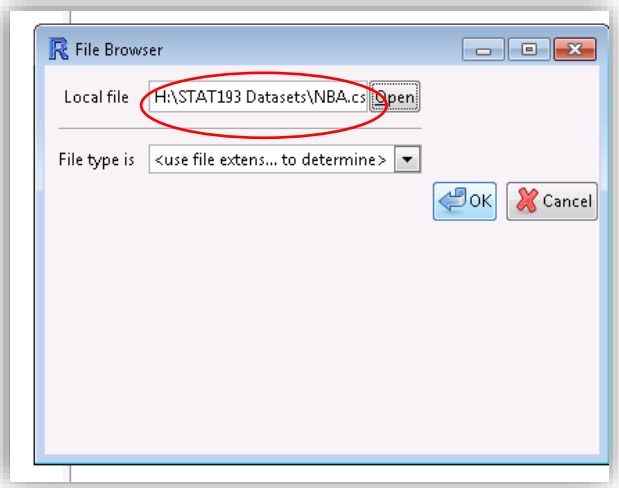
d. Now import the data to iNZight. Click on File>Import Data:



e. Under Local file find and select the dataset to import:

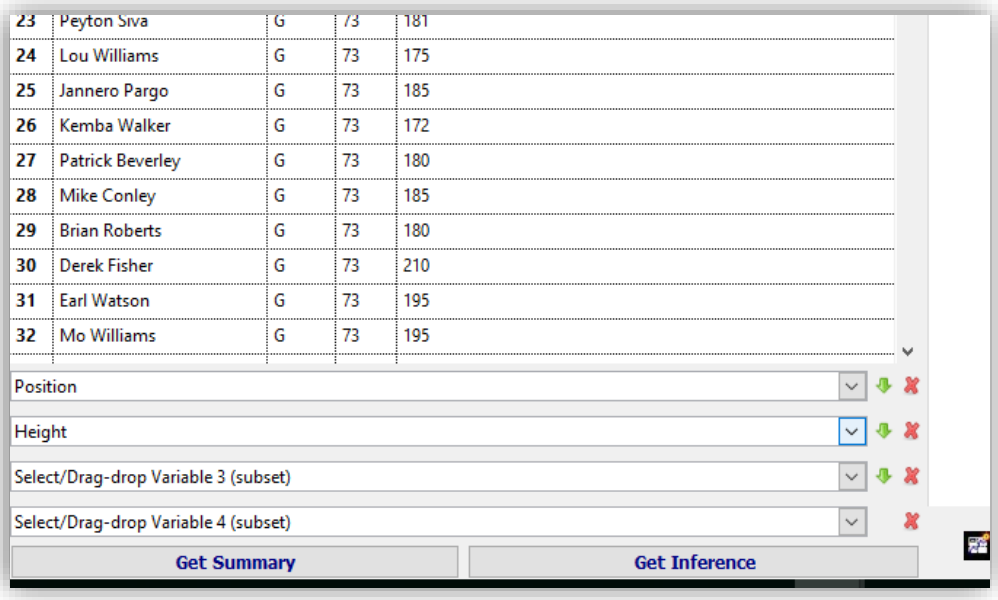


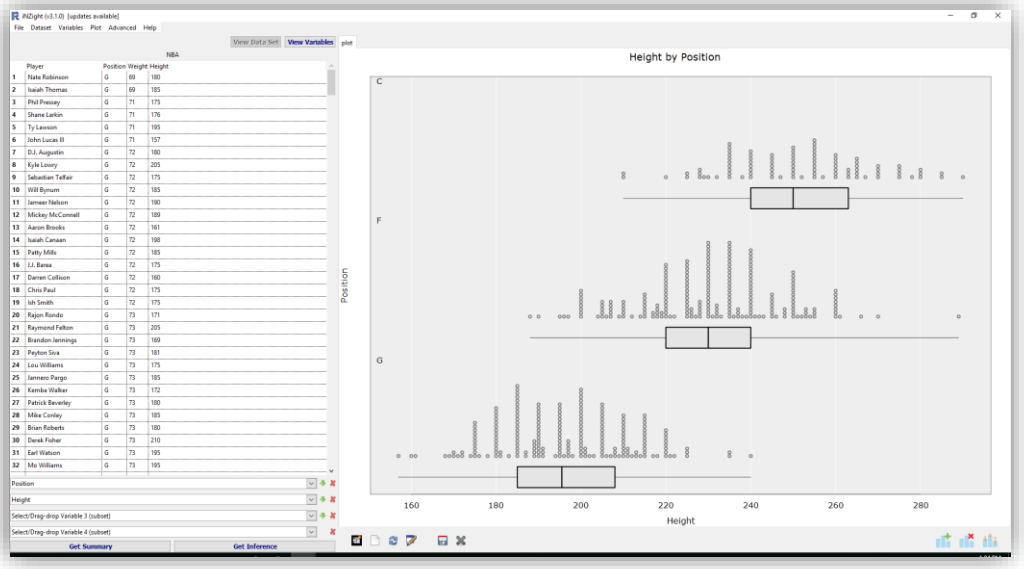




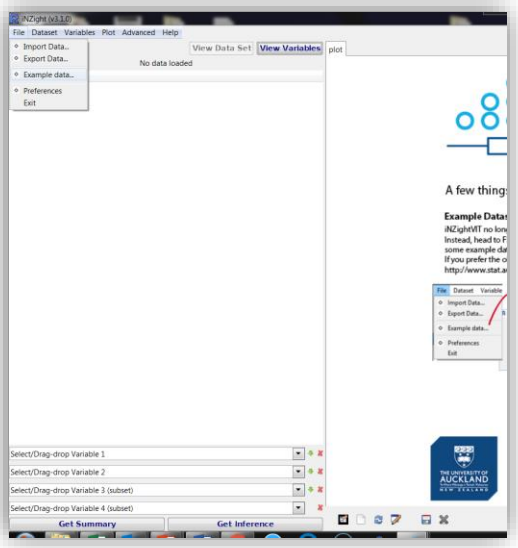
f. Click OK on the Specify a file window and again on the File Browser, the dataset will be imported:

g. You can now select the variable(s) you wish to work with under Variable 1 (and Variable 2):

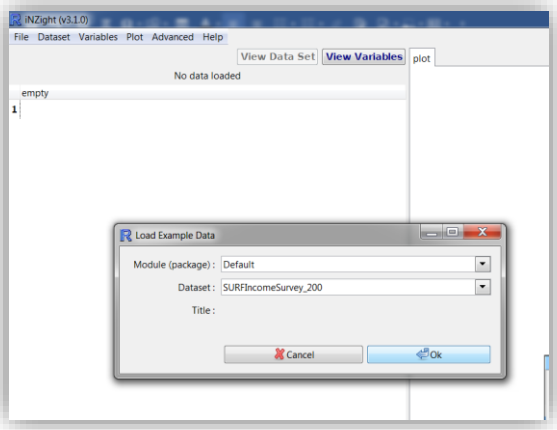


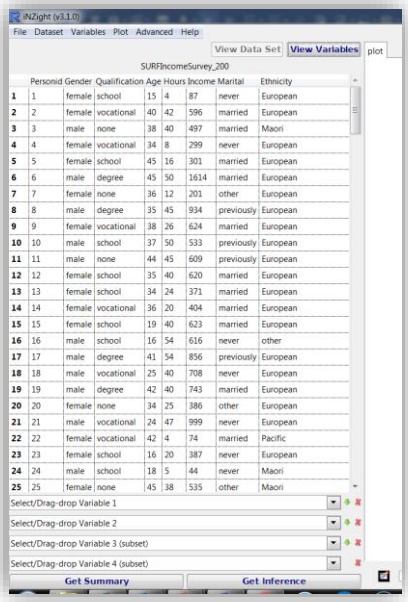


h. To use the Example data that iNZight stores, specifically the SURF dataset, click on File>Example data:



i. Select Default then SURFIncomeSurvey200, OK:

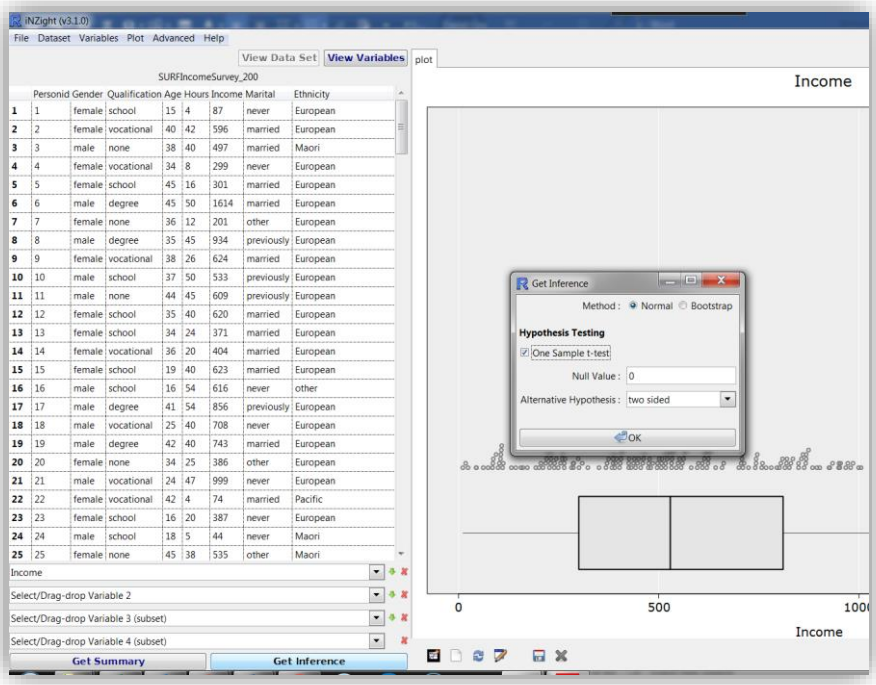




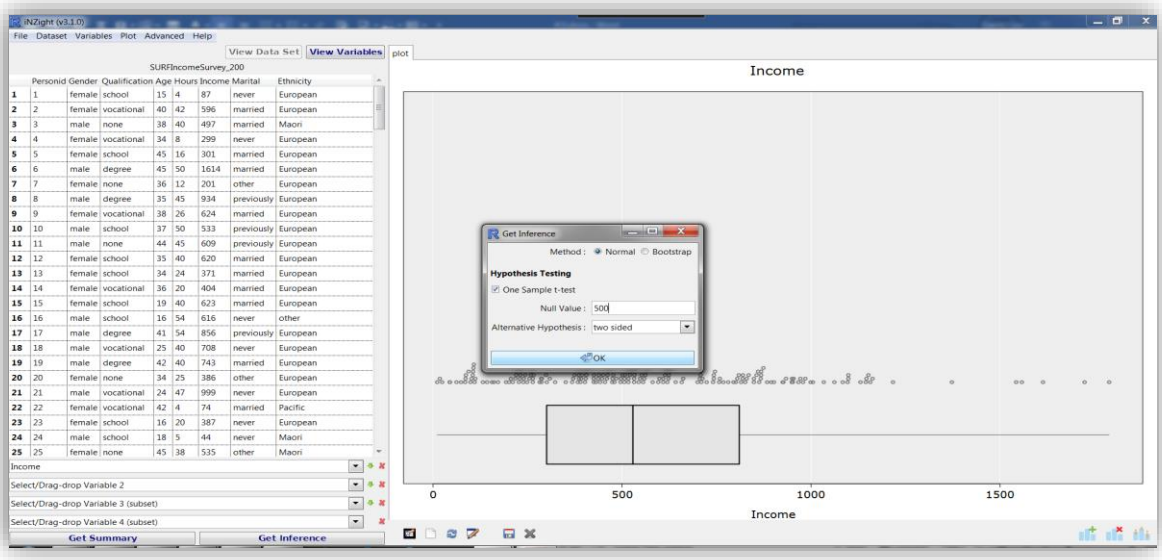
j. Select variables as required

#2 – t-test for a mean

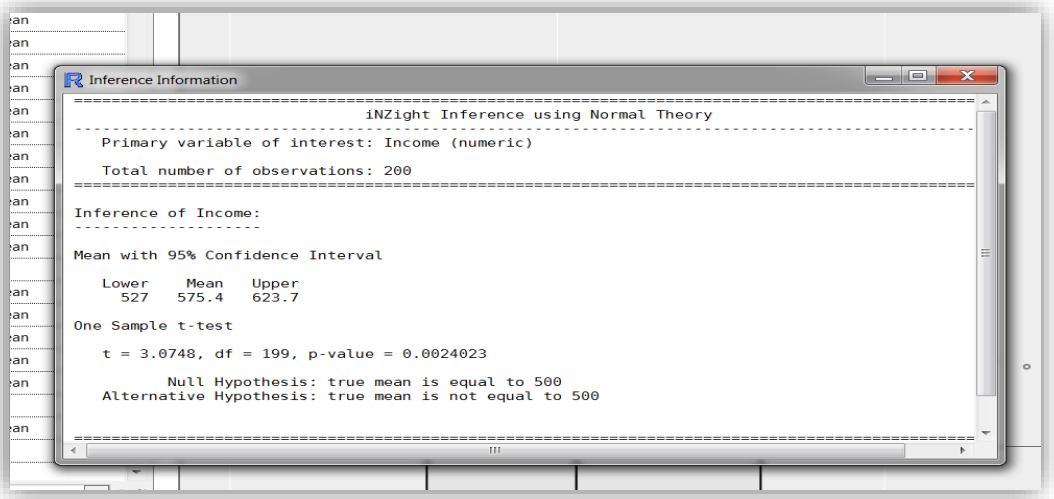
a. Select Get Inference and tick the box ‘One Sample t-test’:



b. Choose the required Null Value (the hypothesised mean under the null hypothesis). Now choose the type of alternative hypothesis, two-sided or one-sided.

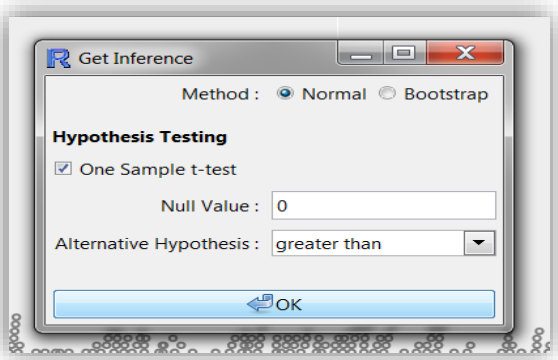


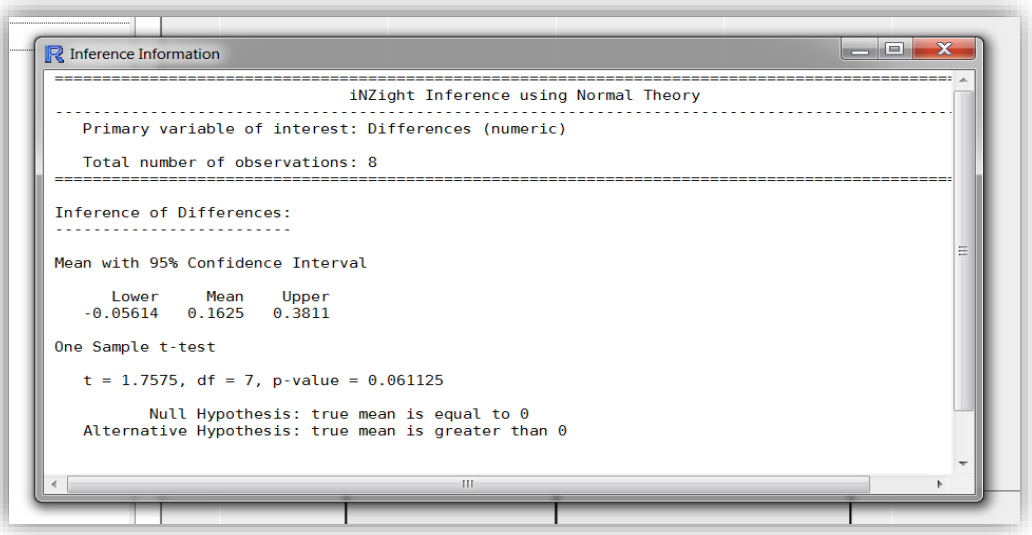
c. Click OK to display the inference:



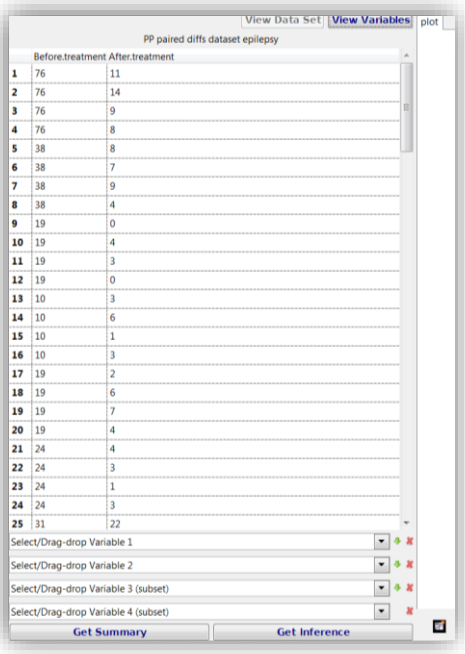
#3 – t-test for paired differences

a. Select Get Inference and tick the box 'One Sample t-test'. For this example we choose the usual null hypothesis of no change between the scores so that *the difference is zero*.

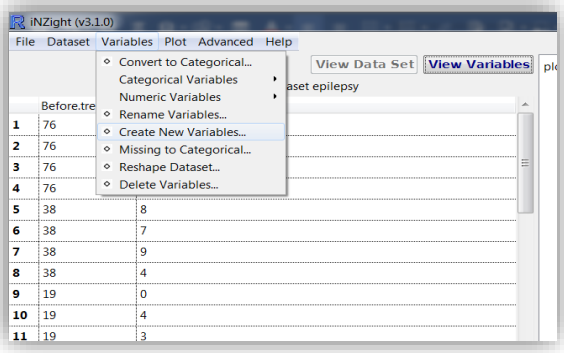


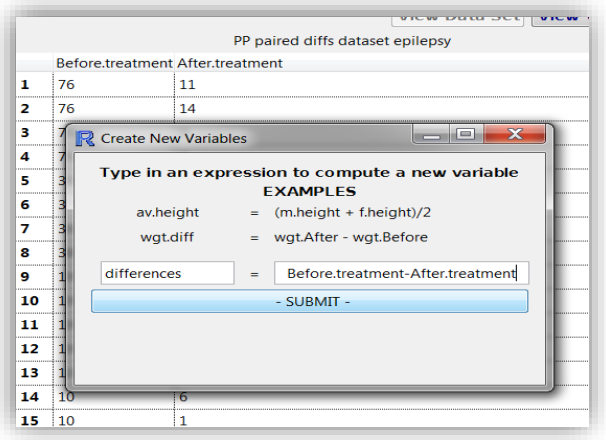


b. When the differences between the two sets of scores have not previously been calculated, use iNZight to do this calculation. We will create a new variable which will be the differences. Import the two columns of scores:

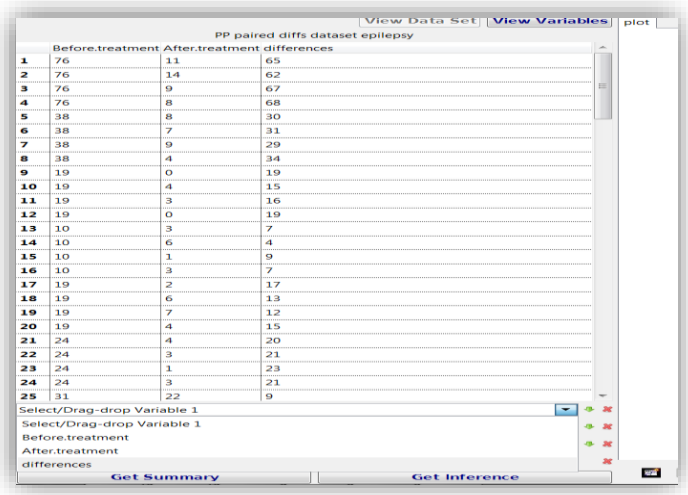


c. Go to Variables then Create New Variable. Type in a name for the variable such as ‘differences’ then in the right hand box type in the name of one variable *minus* the other. You must use the names of the variables exactly as iNZight has shown them, it is case sensitive:





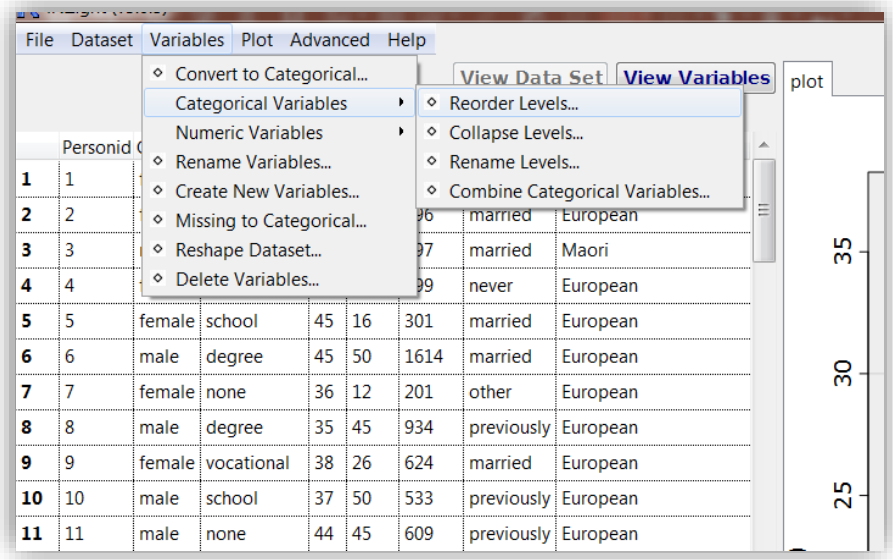
d. The variable you have created will be shown as an extra column, select it from the drop-down box as Variable 1:



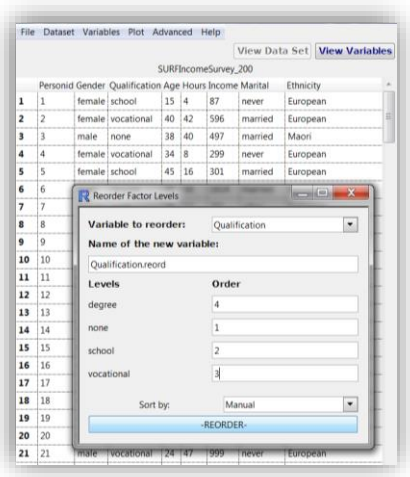
e. Proceed to Get Inference as above.

#4 – Reordering and renaming levels of a categorical variable, adding to plots, switching variables, saving plots

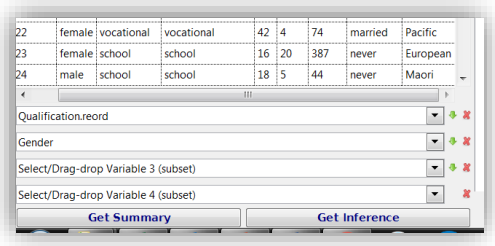
a. A categorical variable has different 'levels' (categories). If you wish to reorder these, for instance with an ordinal variable, go to Variables>Categorical Variables>Reorder Levels:



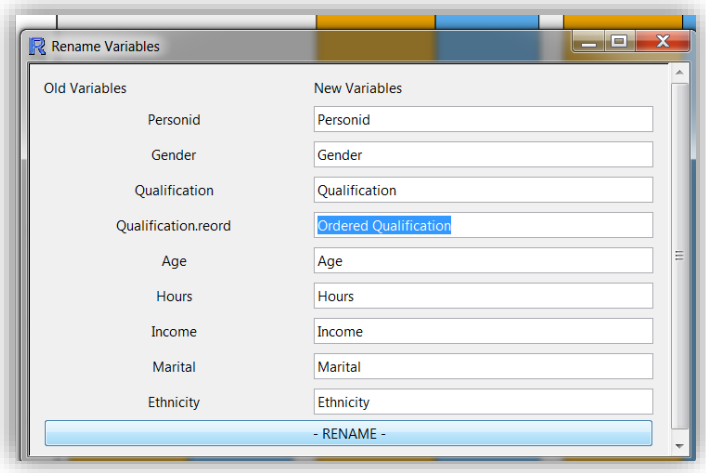
b. Select the variable to reorder, change the order numbers as required, click REORDER:



c. Select the reordered variable from the dropdown box:

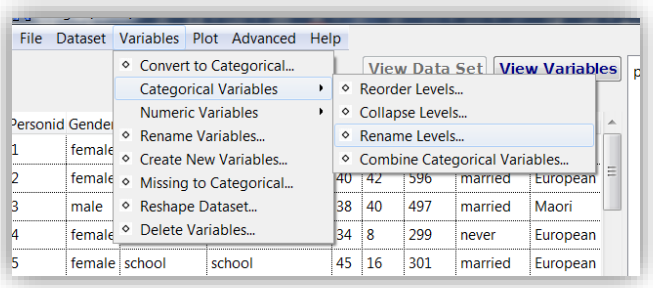


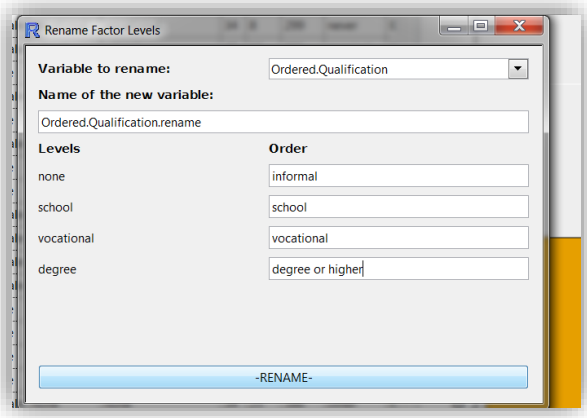
d. To rename a complete variable, simply go to Variables>Rename Variables, change as required, click RENAME:



Select the renamed variable from the dropdown box.

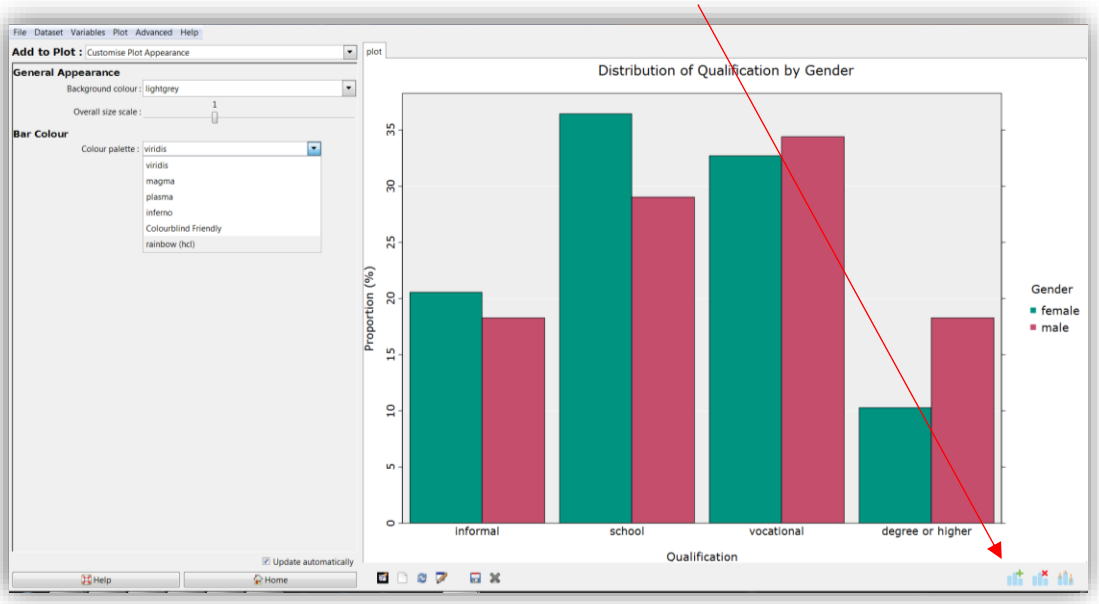
e. To rename a level of a categorical variable, go to Variables>Categorical Variables>Rename Levels. Select the appropriate variable, rename the level(s) as required, click RENAME:



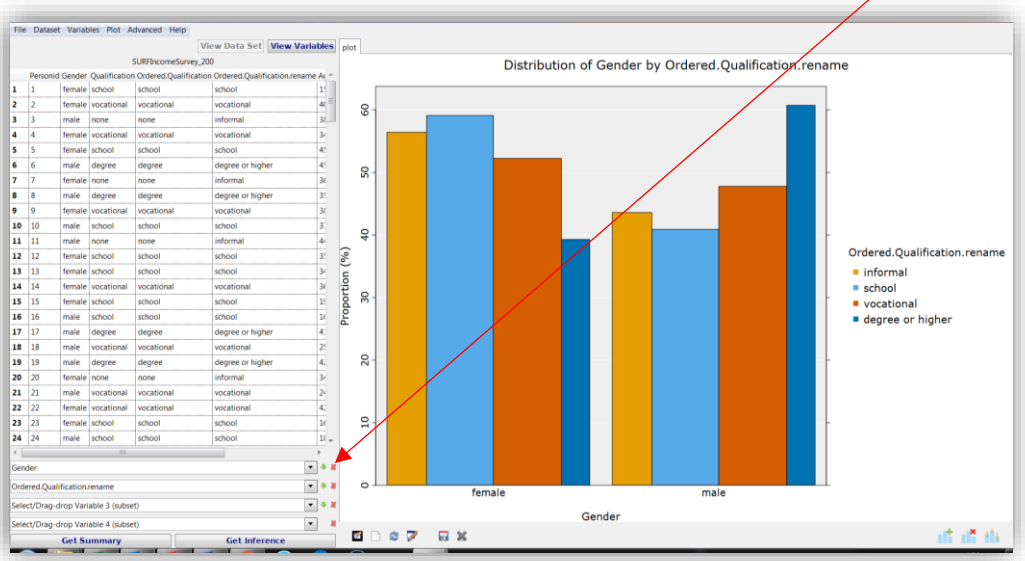


Select the renamed variable from the dropdown box.

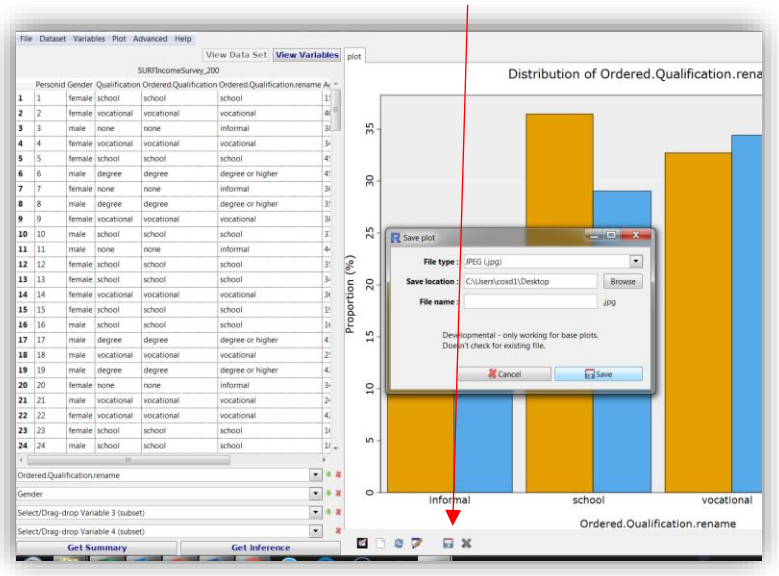
f. To change colours, title or axis labels, go to Add to Plot:



g. To switch/reverse the way the variables are displayed, click the arrow next to the Variable 1 dropdown box:

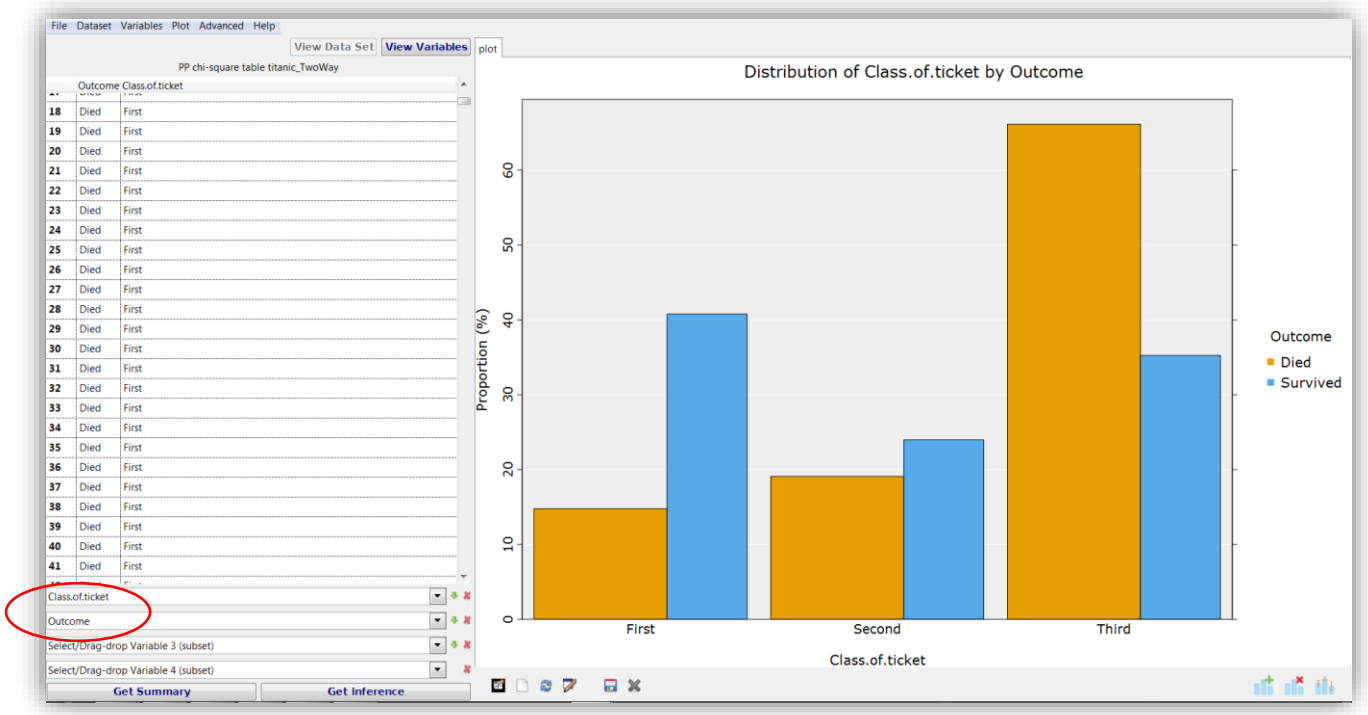


h. To save your graph, click Save Plot. Choose a location for your file:

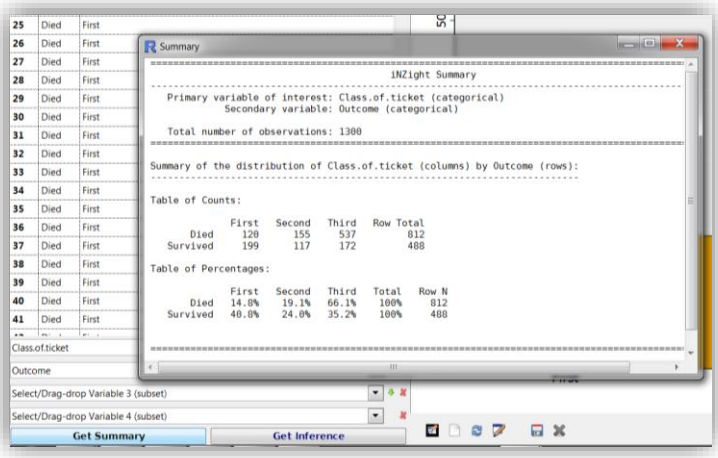


#5 – Chi-square test from data and frequency table

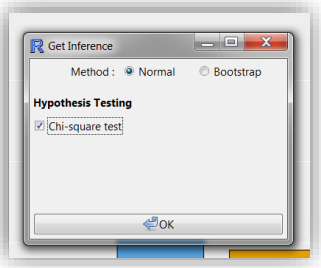
a. Import the data. Select the variables of interest (you may wish to rename the variables):

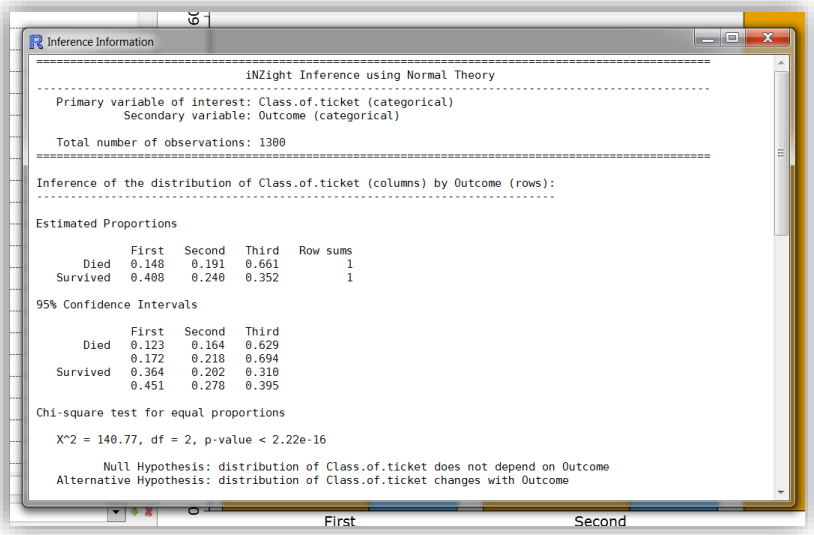


b. We are interested in whether or not there is an association between the two variables. We can first get a summary of the counts and the percentages using Get Summary. The Table of Counts is a contingency table. The Table of Percentages is given as percentages by row, in this case percentages of the outcome (died or survived):

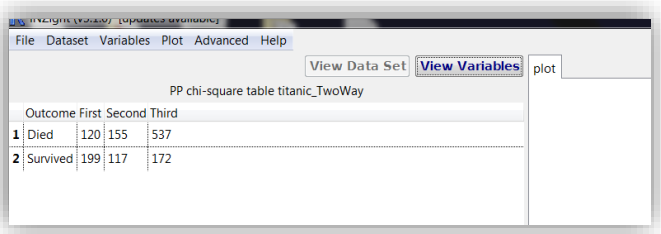


Under Get Inference we obtain the chi-square test statistic, the number of degrees of freedom and the p-value based on the null hypothesis of no association.

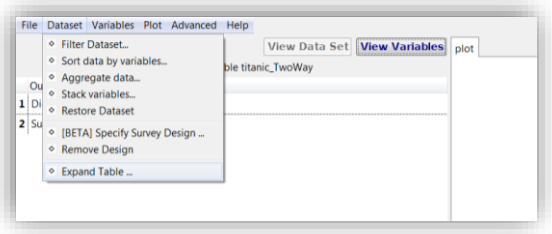


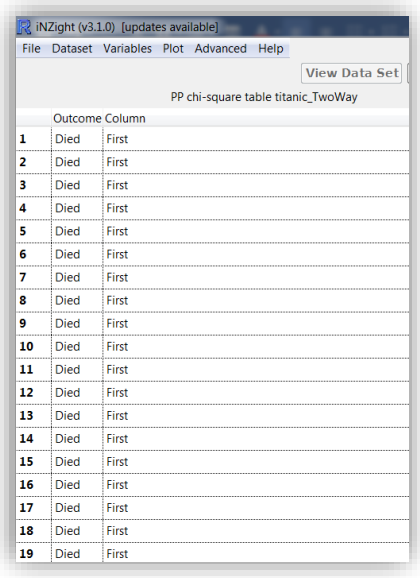


c. If the data you import are in the form of a contingency table…



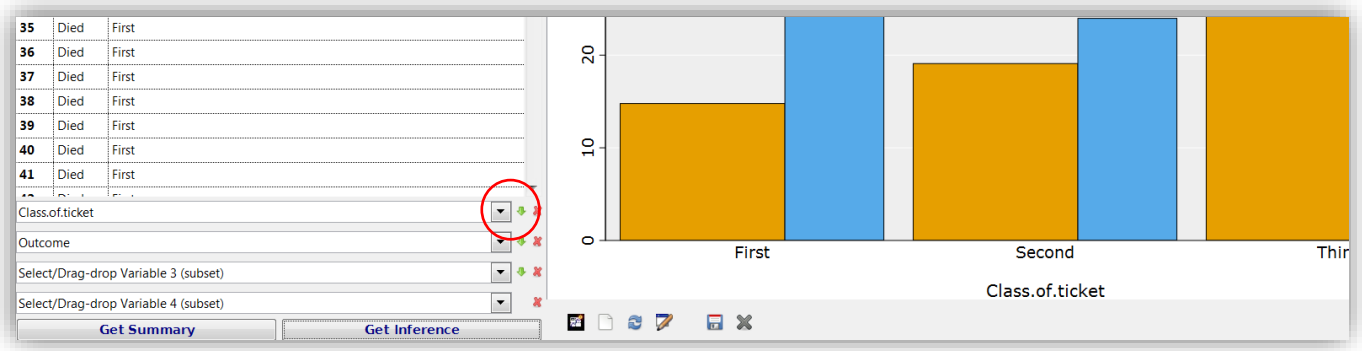
…expand the table by clicking on Dataset > Expand Table > OK:



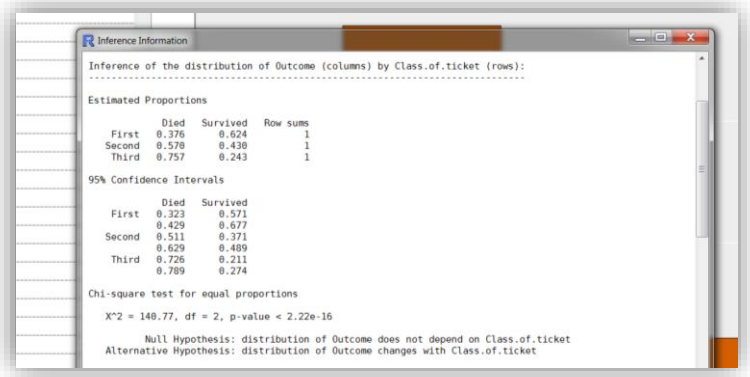


Then proceed as above.

Switch the variables with the green arrow to the right of the Variable 1 dropdown box:



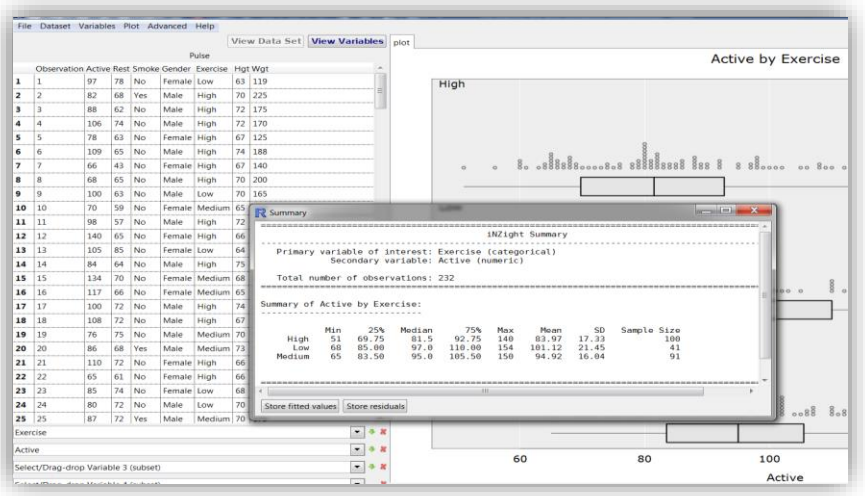
d. Then get the inference again:



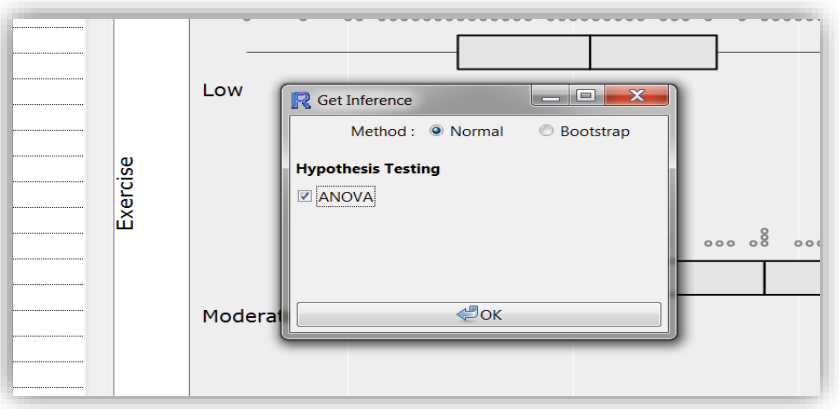
The proportions are now given out of each class.

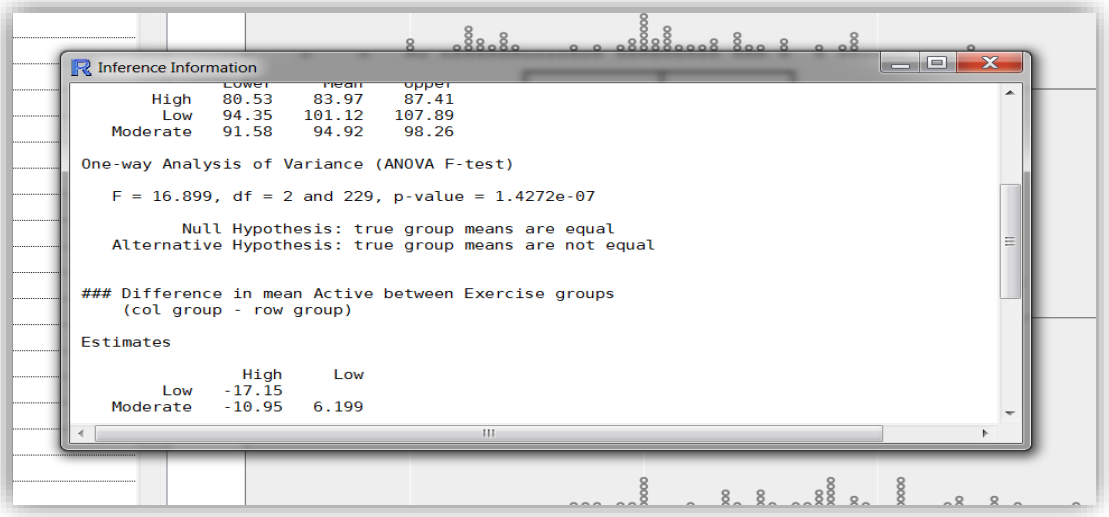
#6 – ANOVA test and confidence intervals

a. Import the data and select the two variables of interest (in either order). We can find a summary of the means, medians, quartiles and standard deviations for each level of the categorical variable under Get Summary:



b. To perform the ANOVA test go to Get Inference, select 'Normal' (rather than 'Bootstrap') and ANOVA, OK:

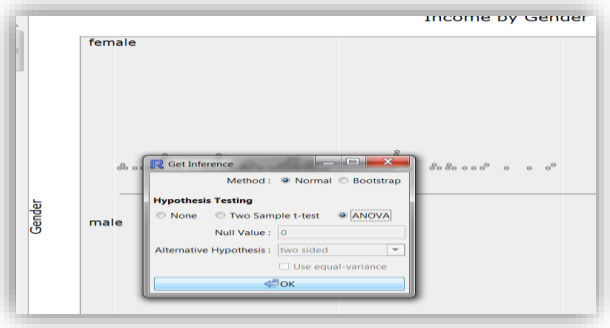




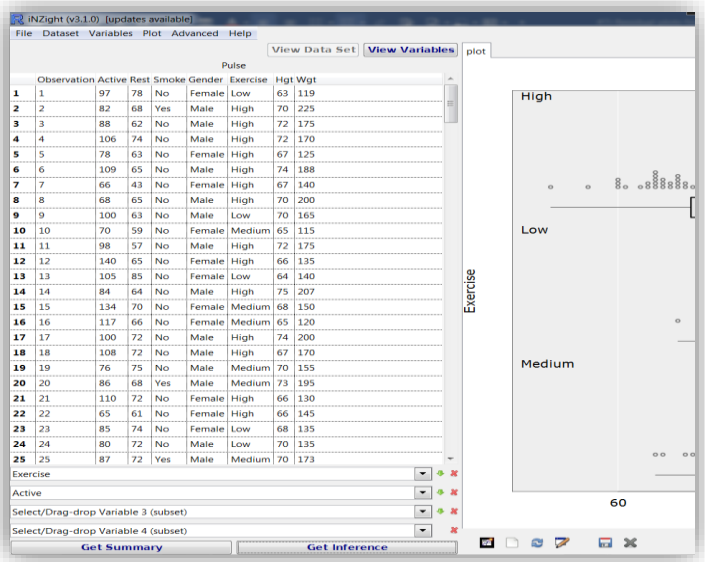
In this example the test statistic is large and the p-value much smaller than alpha. We reject the null hypothesis and conclude that the population means for our different levels of exercise are not all equal.

c. The Inference Information also contains **95% confidence intervals** for the difference between two particular group means, as well as p-values when the null hypothesis is that those group means are equal

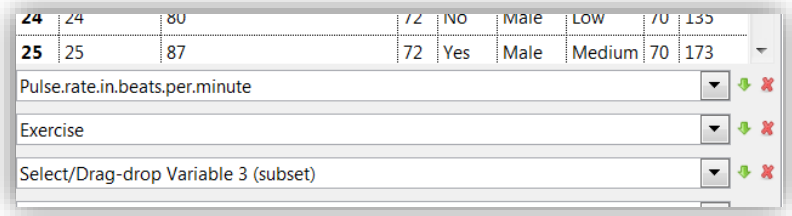
d. When there are only two levels of the categorical variable, iNZight will give the options of an ANOVA test or a t-test. As far as the p-value goes, it doesn't matter which one we choose (unless we specifically want to perform a one-sided test that one population mean is greater than the other – in this case use a t-test: see Help with iNZight #8):



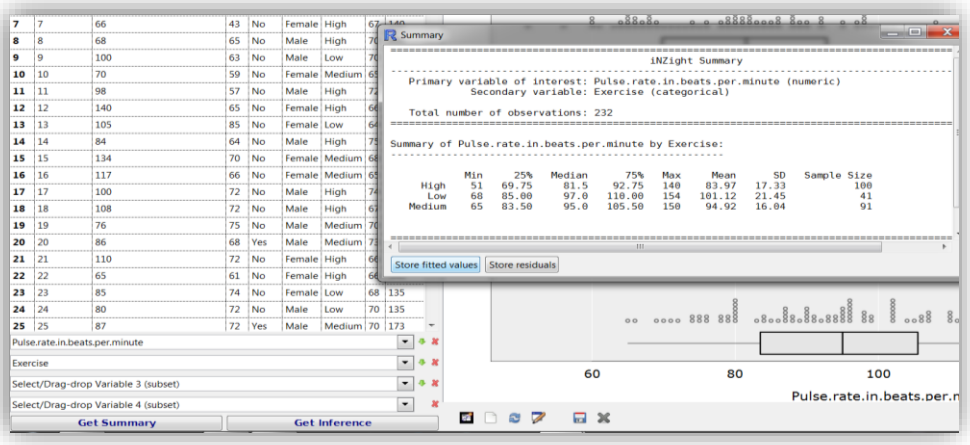
#7 – Residual plots for ANOVA

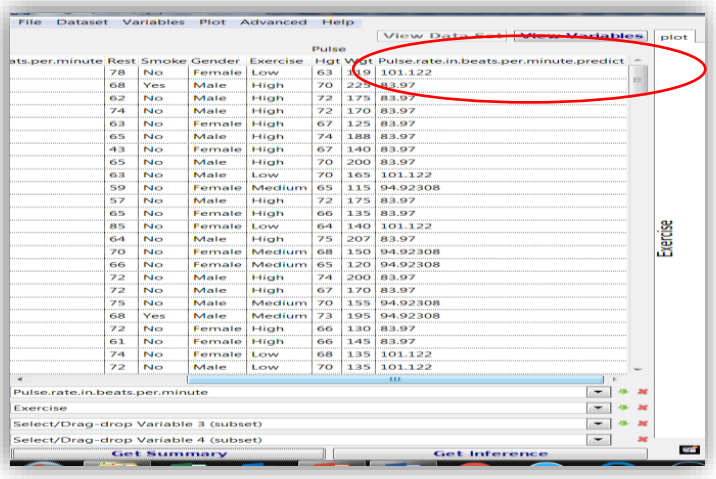


a. We'll rename the 'Active' variable to make it clearer what it measures:



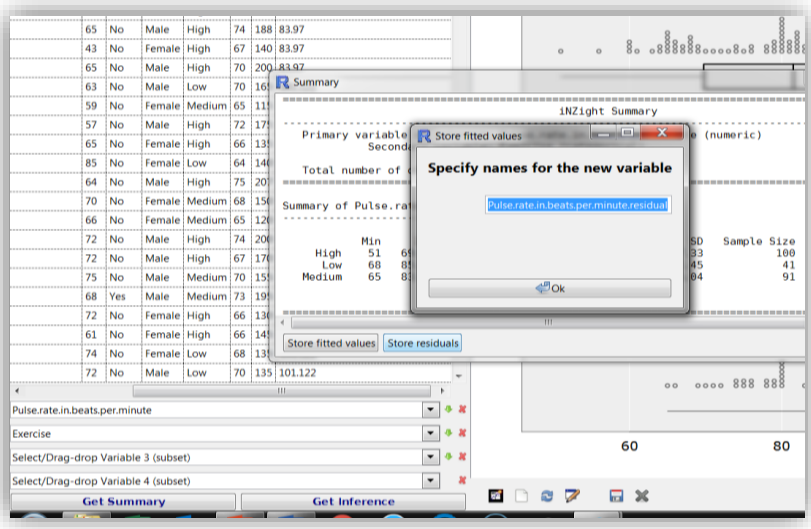
b. Before finding the residuals, we'll find the predicted, or 'fitted', values of Pulse rate for each level of exercise. These are in fact the group means of an ANOVA test. Go to Get Summary and select Store fitted values, OK. The predicted values using the sample data are listed as a new column on the right:





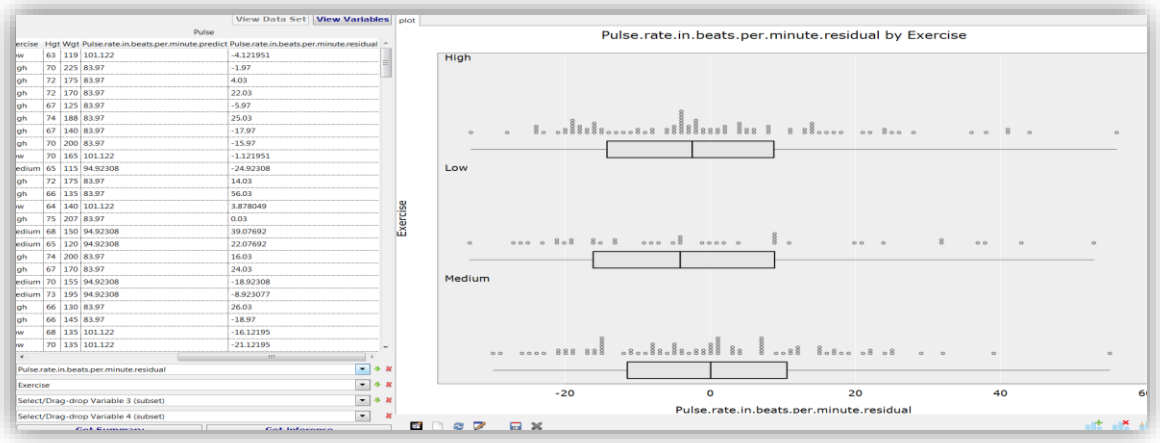
Because there are three levels of exercise, there are three distinct predicted values (group means).

c. The residuals are calculated by taking each value of Pulse rate and subtracting the corresponding predicted value. To do this, select Get Summary again, then Store residuals, OK:



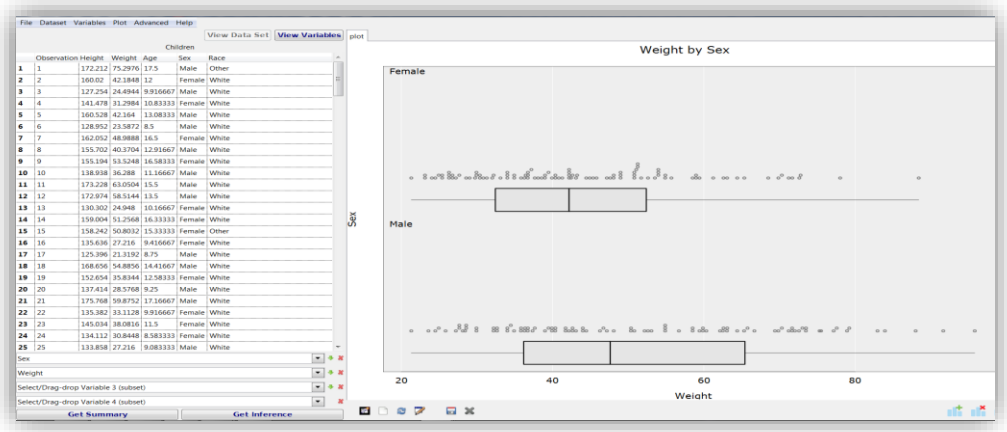
d. The residuals are again listed in a new right hand column

e. To show the residual plot, select this variable of Pulse rate residuals instead of just Pulse rate:

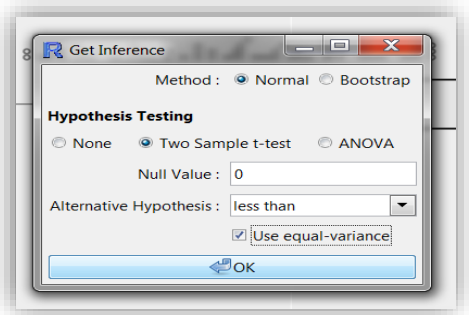


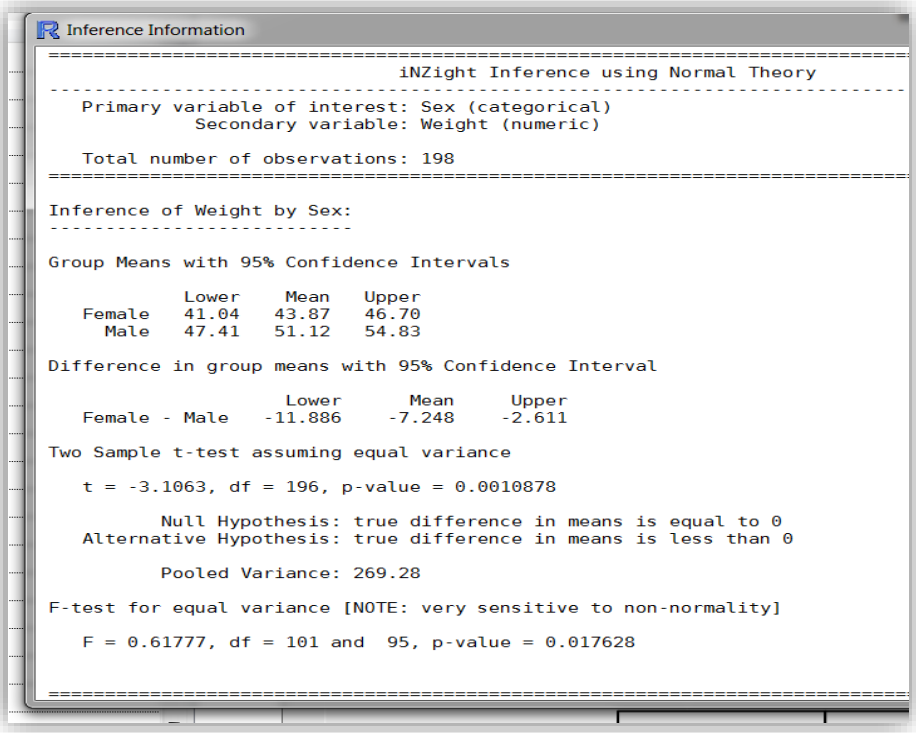
(1) To check the condition of equal variances look at the spread of the boxplots – we are especially interested in the central 50% of each sample. The above boxplots do look sufficiently similar in their variances.
(2) To check the condition of Normality we can look at the symmetry of the boxplots, especially whether there are any clear outliers which would make the data less symmetrical. The above boxplots do look sufficiently similar symmetrical.
(3) The third condition of random sampling could if necessary be investigated by looking at the sampling design of the study/experiment.

#8 – t-test for difference of 2 means and confidence intervals



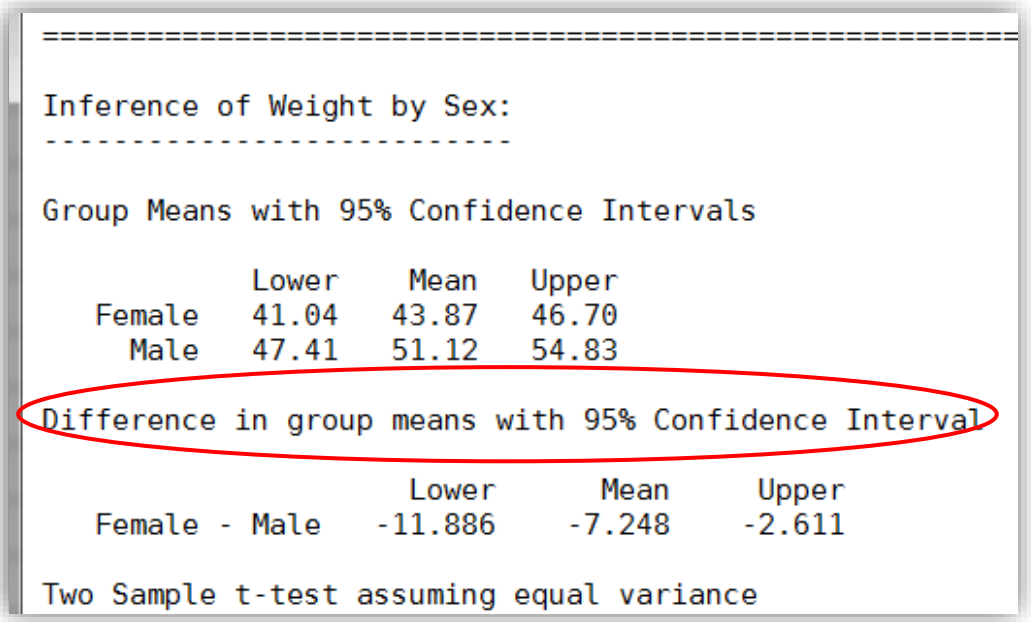
a. Select Get Inference then Two Sample t-test, the appropriate alternative hypothesis and tick Use equal variance (because that assumption appears to be met). For the Null Value, as with the paired difference t-test, our null hypothesis states that there is no difference between the means so we use the value of zero. Click OK:





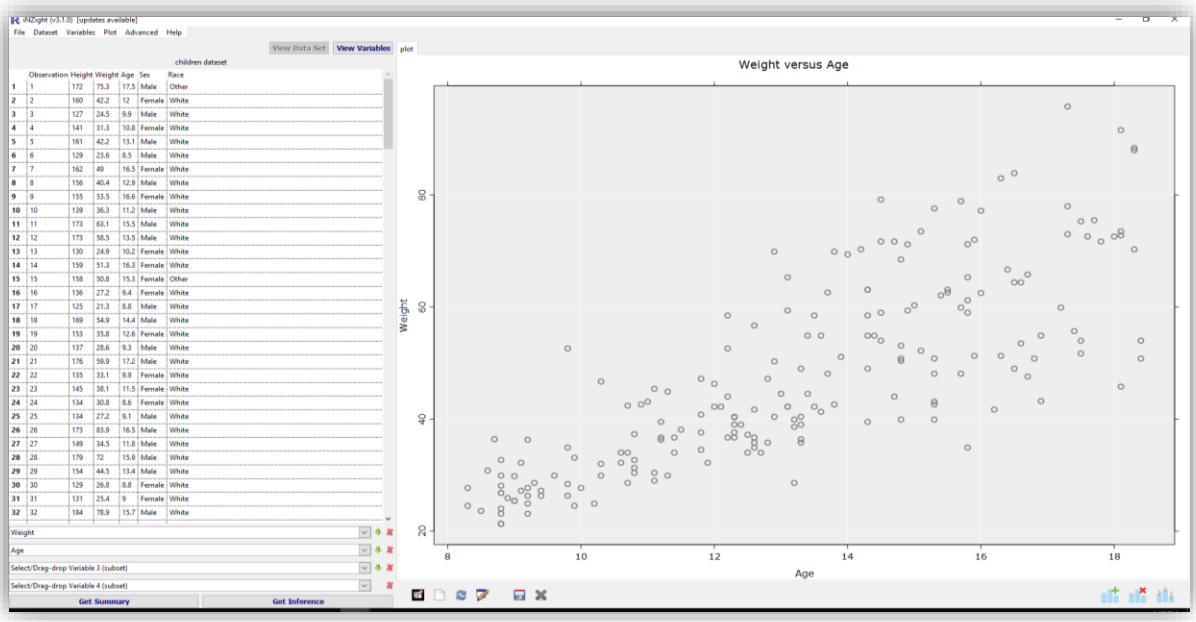
STAT193 does not use the 'F-test for equal variance' so we can ignore the last two lines of the output.

b. The Inference output also gives 95% confidence intervals for the group mean weights and a 95% confidence interval for the difference in the mean weights.

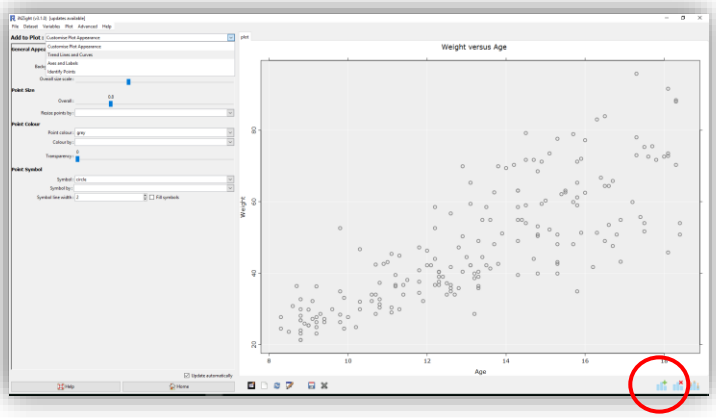


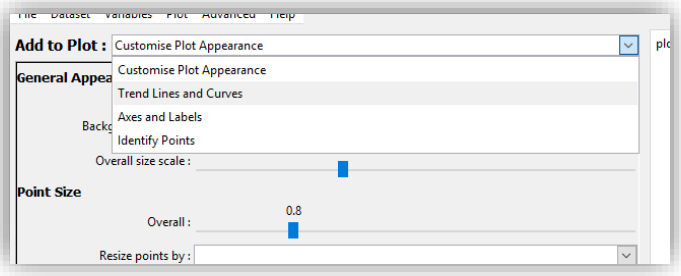
#9 – Scatterplots, Pearson's r, equation of regression line, residuals

a. When we are interested in the possible association between two numerical variables, the graph we will display is a scatterplot. iNZight requires the response variable as Variable 1:

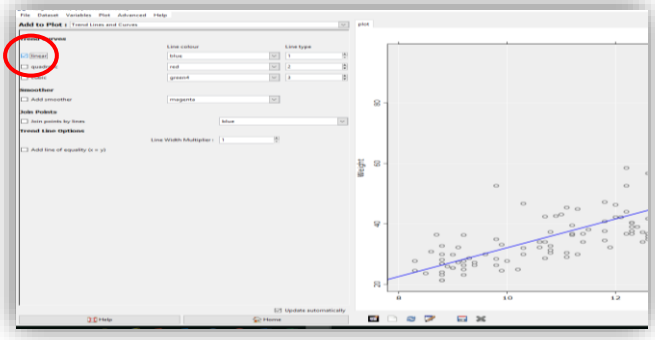


b. Before doing any analysis we must insert a Trend Line (regression line). Go to Add to Plot then Customise Plot Appearance and select Trend Lines and Curves:

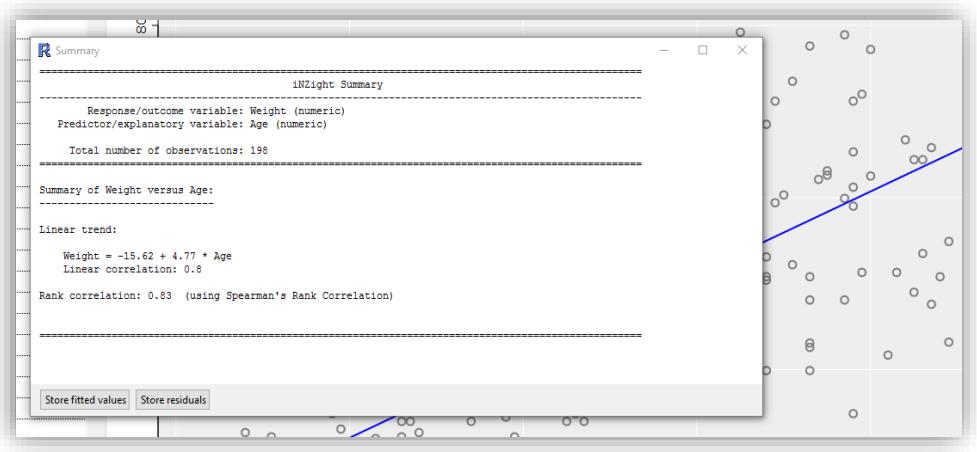




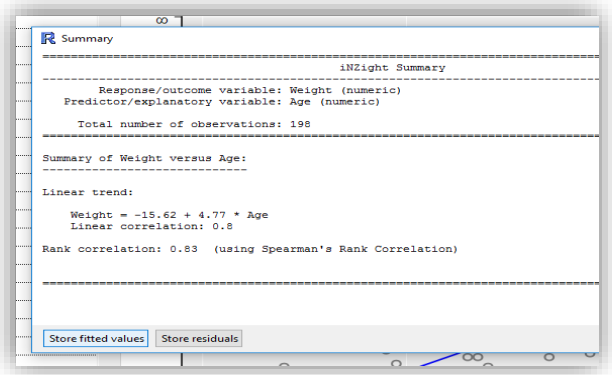
c. Select ‘linear’ then return to the home page:



d. Get Summary to display ***r,*** the correlation coefficient, and the equation of the trend line written in terms of the relevant variables (age and weight):

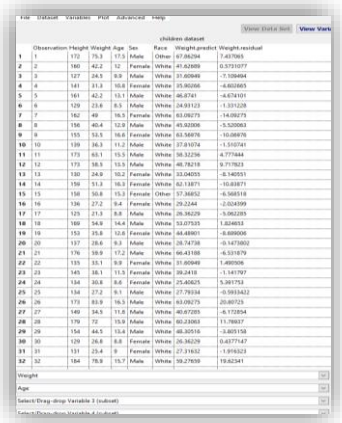


e. To display the predicted weight values from the equation, Get Summary then Store fitted values, OK:

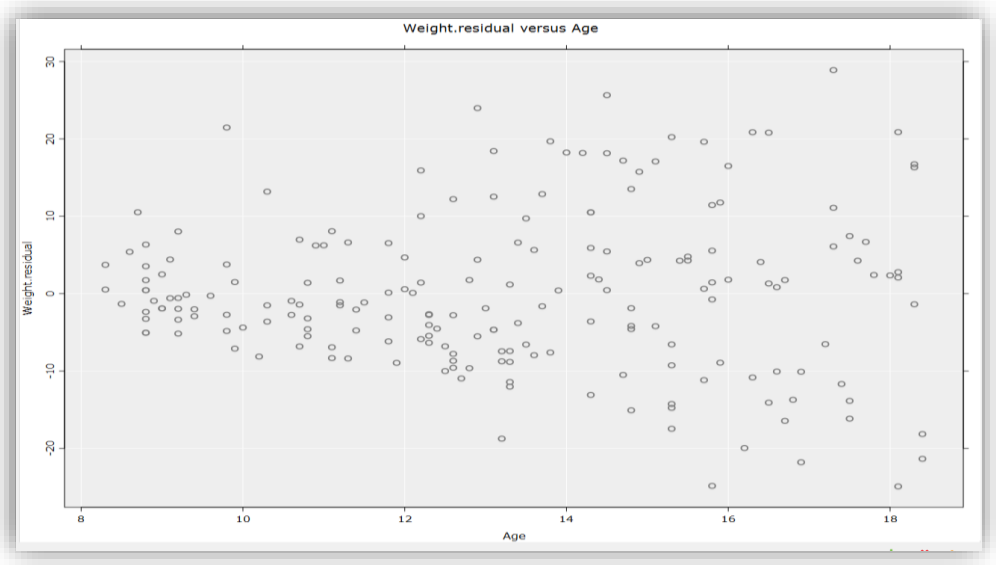


f. As with the predicted values and residuals for an ANOVA test, a new column of these fitted/predicted values is inserted to the right of the data:

g. As with ANOVA, the residuals are calculated by taking each value of weight and subtracting the corresponding predicted value. To display them, **don't** close the Summary window (otherwise you'll have to add a trend line on the scatterplot again!) and select Store residuals, OK. Again the column of residuals is inserted on the right:



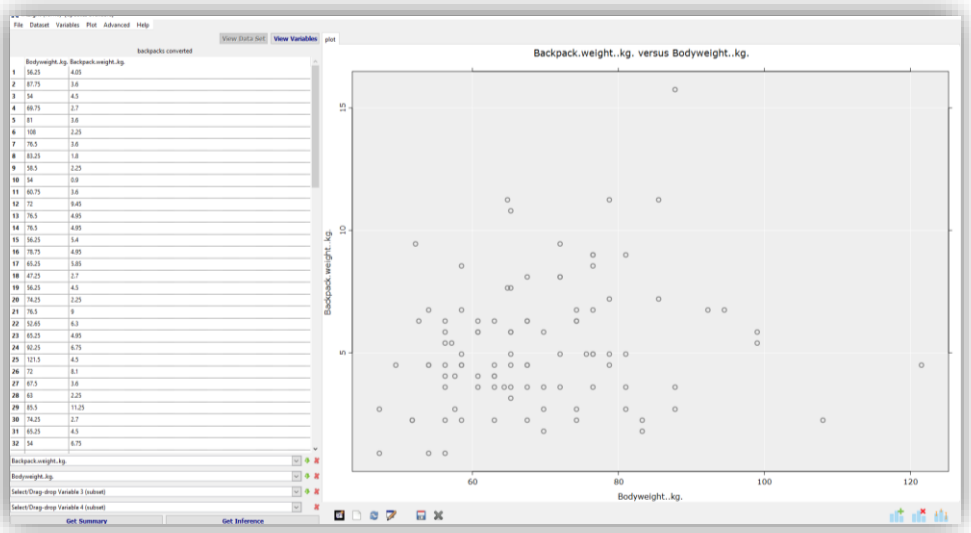
h. To show the residual plot, select as Variable 1 the residuals for the weight instead of just the weight:



#10 Hypothesis test and confidence intervals for gradient of regression line

a. We have chosen two numerical variables and we measure the strength of their association. One of the necessary assumptions of linear regression is that a linear relationship exists. If there actually is **no** association between the variables (so that they are independent), then r would approximately equal zero and furthermore the gradient of the regression line would approximately equal zero.

So whenever the gradient of the regression line does not equal zero then an association is suggested. We can do a hypothesis test on whether or not the gradient of the line equals zero, as well as finding a confidence interval for that gradient.

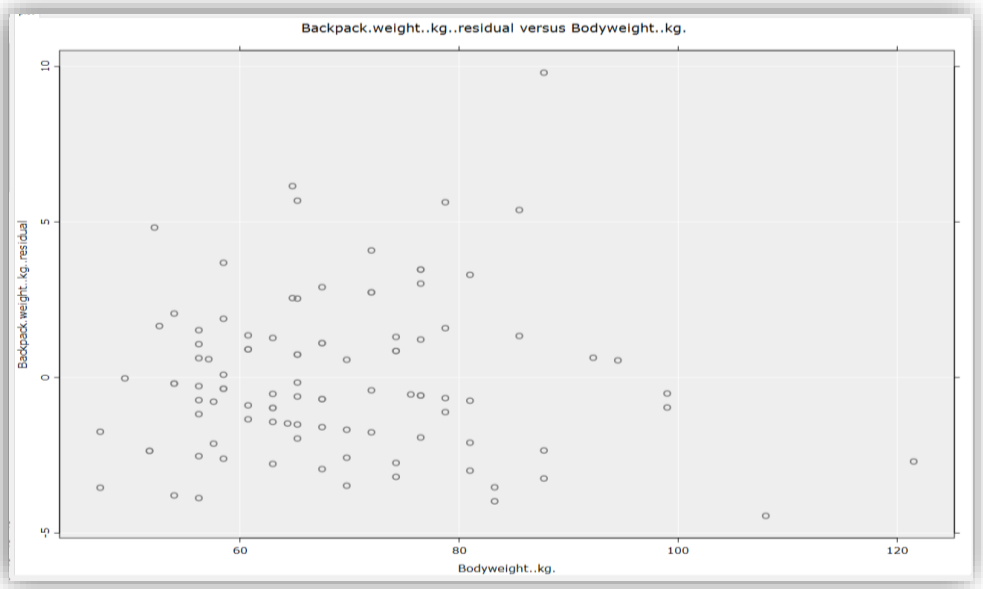


b. From the graph it appears that there is a weak linear correlation. To formally test whether there is a relationship, iNZight requires us to first add a trend line (Help with iNZight #9):

c. Under Get Summary a value of r is given as well as the equation of the fitted regression line from the data:

```
Summary of Backpack.weight..kg. versus Bodyweight..kg.:
-------------------------------------------------------

Linear trend:

    Backpack.weight..kg. = 2.69 + 0.03713 * Bodyweight..kg.
    Linear correlation: 0.19

Rank correlation: 0.21  (using Spearman's Rank Correlation)


=======================================================================
```

d. At this point we should examine the residuals (#9i) to check that the assumptions of linear regression are met, specifically that of a linear (rather than curved) relationship:



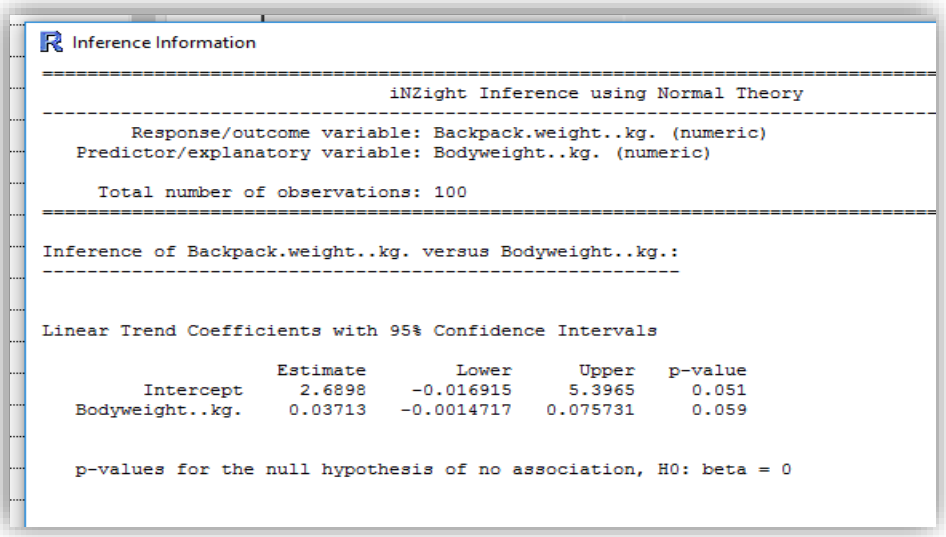
We will test whether in the equation of the linear relationship between the X and Y variables in the population i.e. $Y=\alpha+\beta X+\varepsilon$ the gradient β is equal to zero or not.

e. The null and alternative hypotheses are as follows:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

where β is the gradient of the linear relationship between bodyweight and backpack weight in the population. Choose a significance level of $\alpha = 0.01$ in this case, as we want strong evidence before we will reject H_0. Under Get Inference obtain the following:



We are interested in the second row of numbers, which gives a) the point estimate of the gradient b) the 95% confidence interval for β, the gradient for the population and c) the p-value based on the null hypothesis of 'no association'.