Hong Kong Baptist University

Department of Computer Science

COMP 7990 Principles and Practices of data analytics (2022-23)

Lab 2: Analyzing Data using Jamovi

Introduction

Data analytics is the science of analyzing raw data with the purpose of drawing conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. In this lab session, we will use a software package called jamovi to apply some statistical algorithms to derive insights from the data.

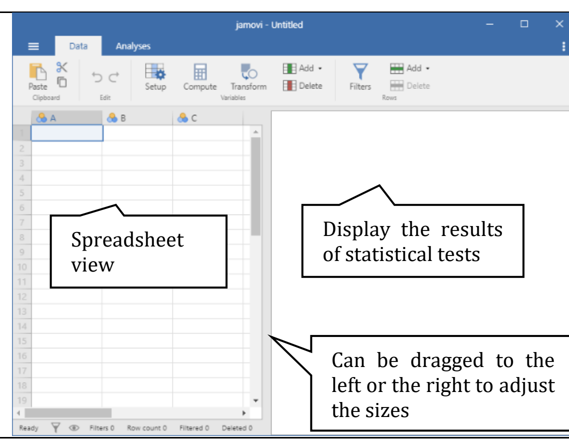
jamovi is a freeware used for interactive statistical analysis. It can take data from file formats that are commonly used for structured data such as CSV, SPSS, SAS, etc., and use them to generate tabulated reports, charts, and plots of distributions, descriptive statistics, and to conduct statistical analyses. jamovi can be obtained from https://www.jamovi.org/download.html.

Learning Outcome

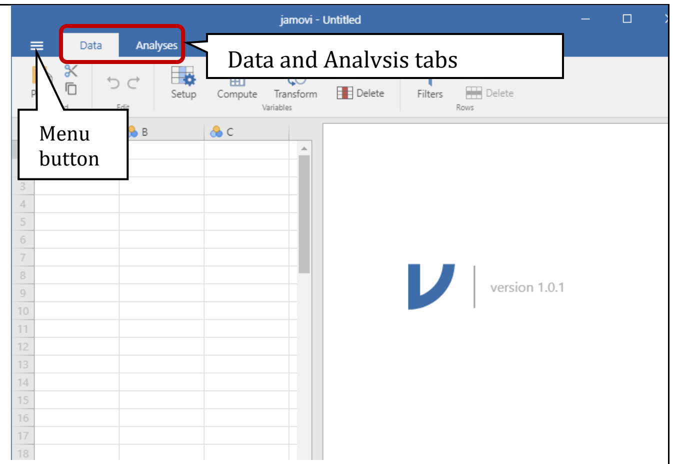
By finishing this lab session, you should be able to

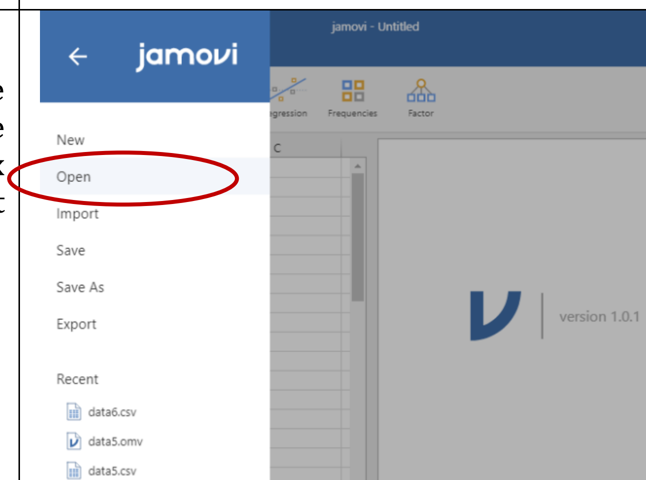
- Understand the basic functions of jamovi and perform basic statistical analyses
- Perform descriptive statistics and graphics, and basic inferential statistics for comparisons and correlations

Getting started with jamovi

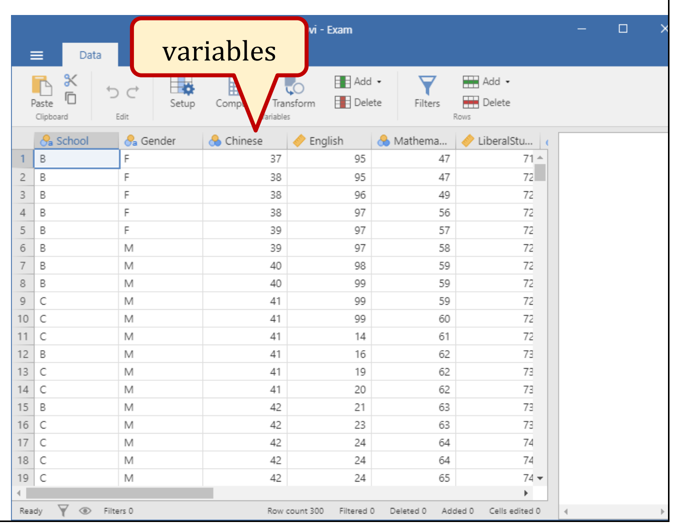
1. Open the **jamovi** program **LHS**: spreadsheet view **RHS**: show the results of statistical tests	

<table>
<tr>
<td>
2. Next to the menu button , there are two tabs: Data and Analysis

• Data: Create or update the data in various ways
• Analysis: Perform statistical analyses on the data
</td>
<td>


</td>
</tr>
<tr>
<td>
3. Download the datafile Lab2-Jamovi.zip
4. Click the menu button and choose Open to open the file Exam.csv. The file contains samples of secondary six examination results from three different schools.
</td>
<td>


</td>
</tr>
<tr>
<td>
5. Data is represented in the spreadsheet. Each column represents a variable and each row represents a case or participant. One can change the values, delete or add variables in the spreadsheet.

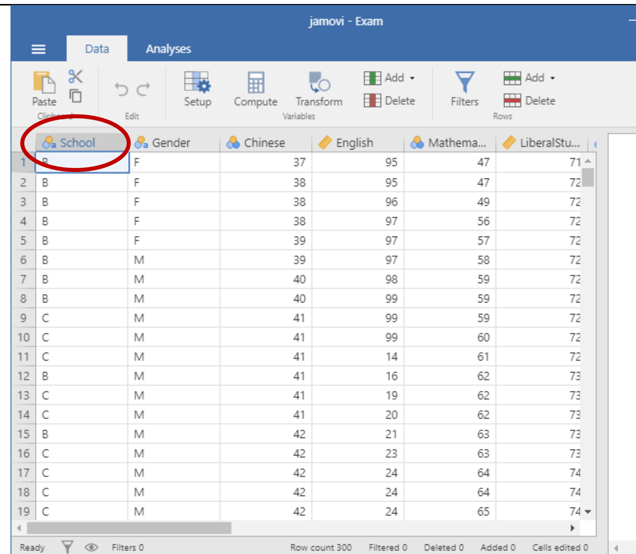
Cases or participants
</td>
<td>


</td>
</tr>
</table>

6. **Double click** the **School** variable to show the **variable pane.**

There are two types of variables:

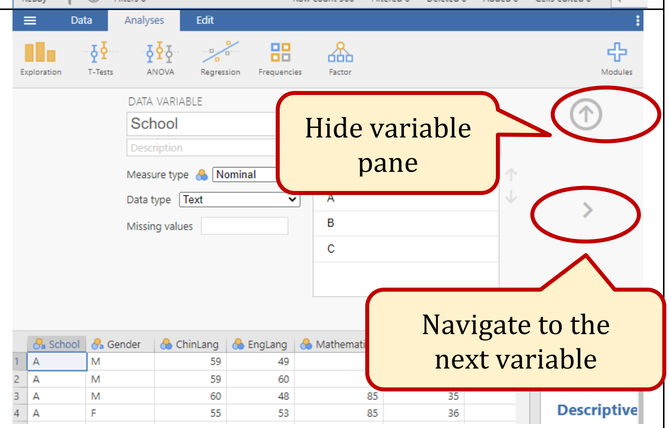
- **Data Variables**: data either loaded from a data file or type in by the user
- **Computed variables**: Those that take value by performing a computation on *other* variables

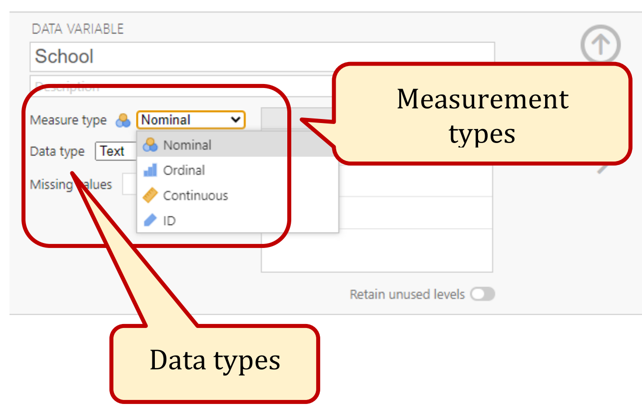


7. **Data variables** can be one of the three data types:
- **Integer**
- **Decimal**
- **Text**

and one of the following four **measure types**:

- **Nominal**: Categorial variables which are text labels. For example, the column called Gender with the value Male and Female would be nominal.
- **Ordinal**: Like Nominal variables, except the values have a specific order. An example is a Likert scale from 'strongly agree' to 'strongly disagree'.
- **Continuous**: Variables which exist on a continuous scale. An example might be height or weight. This is also referred to as 'Interval' or 'Ratio scale'.
- **ID:** ID variable type is unique to jamovi. It is for variables that contain identifiers that you would almost never want to analyze.





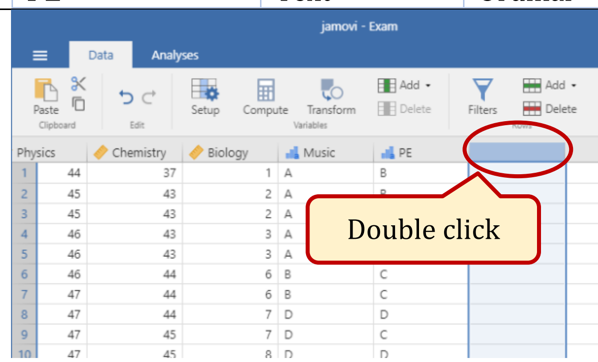
Variable	Data Type	Measure Type

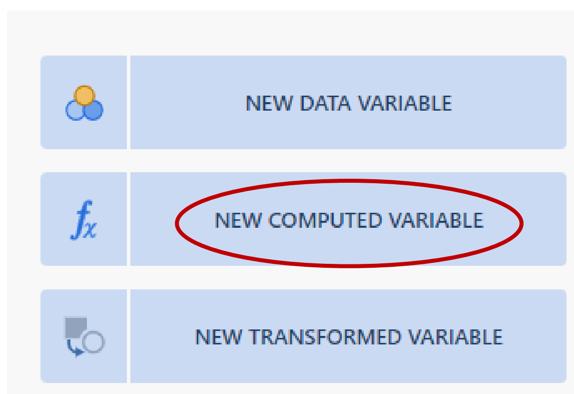
8. When opening a data file, jamovi will infer the variable type from the data in each column. You may manually specify the **data type** and **measure type** if they are incorrect.
9. **Update** the **data types** and **measure types** according to the table on the right.

School	Text	Nominal
Gender	Text	Nominal
ChinLang	Integer	Continuous
EngLang	Integer	Continuous
Mathematics	Integer	Continuous
LiberalStudies	Integer	Continuous
Physics	Integer	Continuous
Chemistry	Integer	Continuous
Biology	Integer	Continuous
History	Integer	Continuous
Geography	Integer	Continuous
Music	Text	Ordinal
PE	Text	Ordinal

10. Double click the empty variable next to the PE.

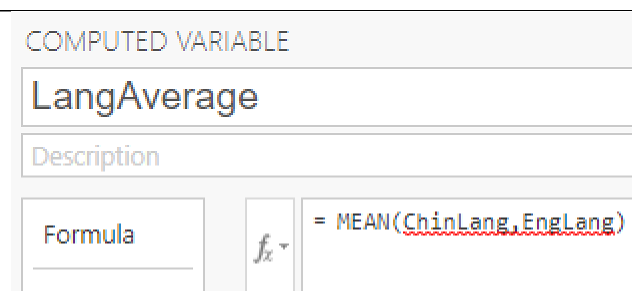
11. Click **NEW COMPUTED VARIABLE** to create a computed variable.



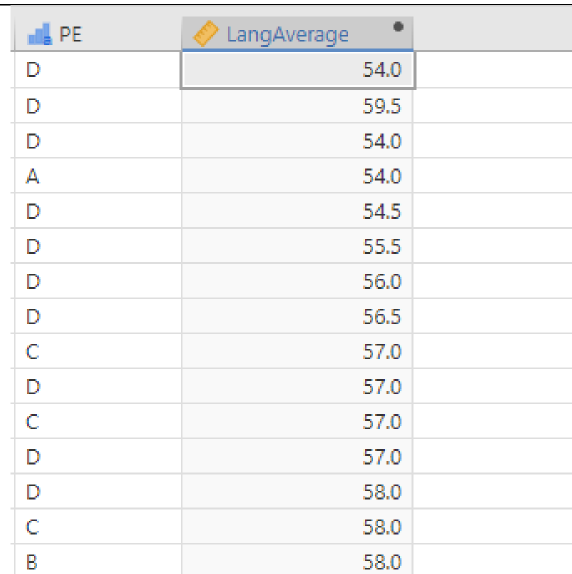


12. Type **'LangAverage'** as the computed variable name.

13. Click the **insert function button** (fx) and choose 'MEAN' under the functions. Under the variables, click **'ChinLang'**, type in a comma, and then click **'EngLang'**. Hide the variable pane after you have done.



14. The **LangAverage** column is added into the data table.

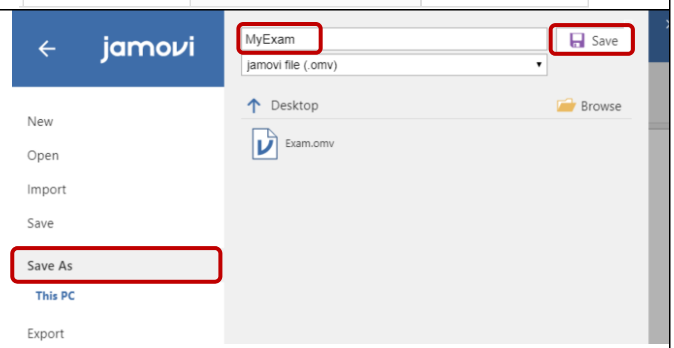


PE	LangAverage
D	54.0
D	59.5
D	54.0
A	54.0
D	54.5
D	55.5
D	56.0
D	56.5
C	57.0
D	57.0
C	57.0
D	57.0
D	58.0
C	58.0
B	58.0

15. Click the **menu button** and choose **Save As**.

16. Name the file **MyExam**.

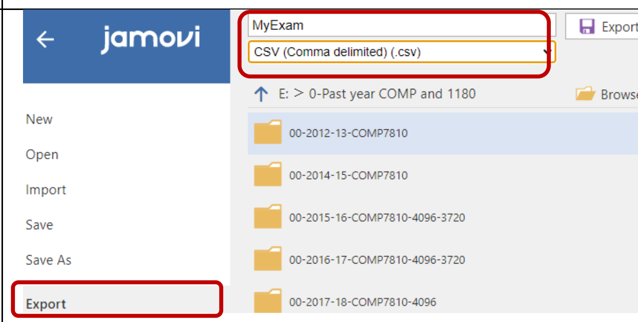
17. Click the **Save** button to save the data in jamovi (.omv) format. The omv files store data, variables and analyses.



18. Click the **Menu button** and choose **Export**.

19. Name the file **MyExam**. Choose **CSV(Comma delimited)** as the file type.

20. Click the **Export** button to save the data in CSV format.



Descriptive Statistics

Descriptive statistics helps to describe, show or summarize data in a compact, easily understood fashion. Typically, there are two general types of statistic that are used to describe data:

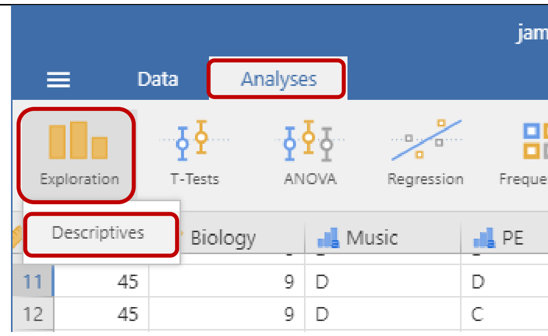
a) **Measures of central tendency**: Ways of describing the central position of a frequency distribution for a group of data. The <u>mean, median and mode</u> are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

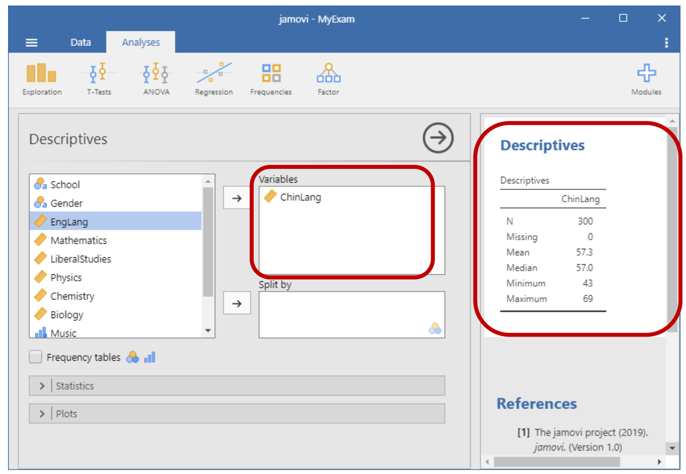
b) **Measures of dispersion/spread**: summarizing a group of data by describing how spread out the values are. Three measures of the spread of data: <u>range, standard deviation and variance.</u>

1. Open the **MyExam.omv** in jamovi if it is not already open.

2. Select **Analyses** → **Exploration** → **Descriptives**

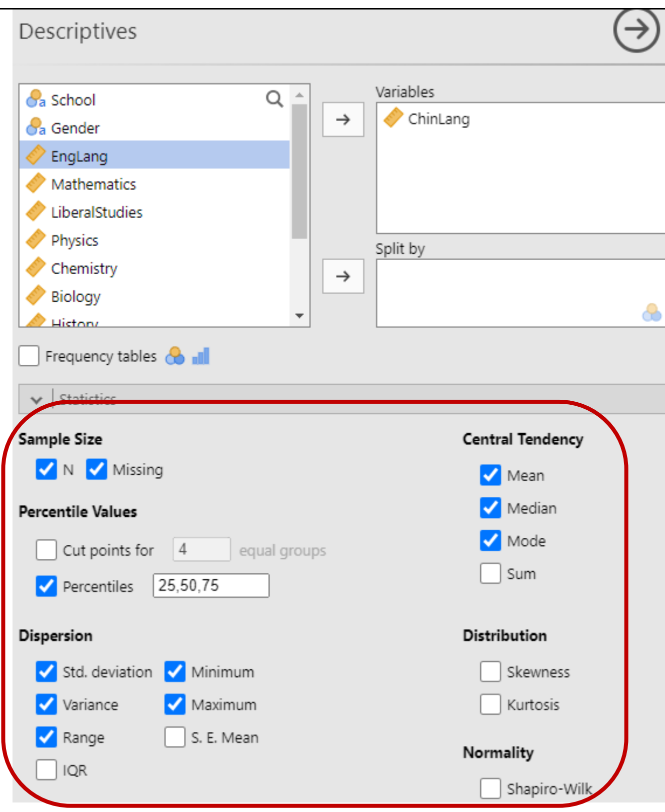
3. Move the **ChinLang** into the **Variables** box. **ChinLang** variable will automatically be summarized in a table on the right.





4. Expand the **Statistics** option and **checked** the following boxes:

   - **N:** Total number of samples.
   - **Missing:** Number of samples having a missing value.
   - **Mean:** The average or the sum of the values divided by the number of values.
   - **Median:** The value which divides the data into 2 equal parts i.e. number of terms on right side of it is same as the number of terms on the left side of it when the data is arranged in either ascending or descending order.
   - **Mode:** The value that has the highest frequency.
   - **Percentiles:** Divide your data into quarters provided data is sorted in ascending order. There are three quartile values. The first quartile value is at 25th percentile. The second quartile is 50th percentile (the median) and the third quartile is 75th percentile.



- **Std. deviation:** The measurement of the average distance between each quantity and mean. That is, how data is spread out from the mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.
- **Variance:** The average of the squared differences from the mean. That is the square of standard deviation.
- **Range:** The difference between the lowest and highest value.
- **Minimum:** The lowest value
- **Maximum**: The highest value

The output is automatically updated.

Descriptives

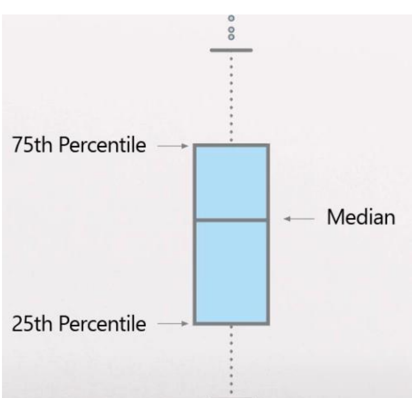
Descriptives

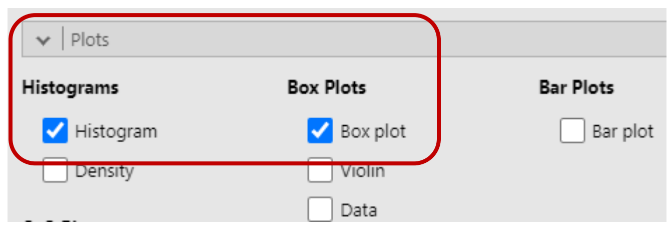
	ChinLang
N	300
Missing	0
Mean	57.3
Median	57.0
Mode	57.0
Standard deviation	4.89
Variance	23.9
Range	26
Minimum	43
Maximum	69
25th percentile	54.0
50th percentile	57.0
75th percentile	61.0

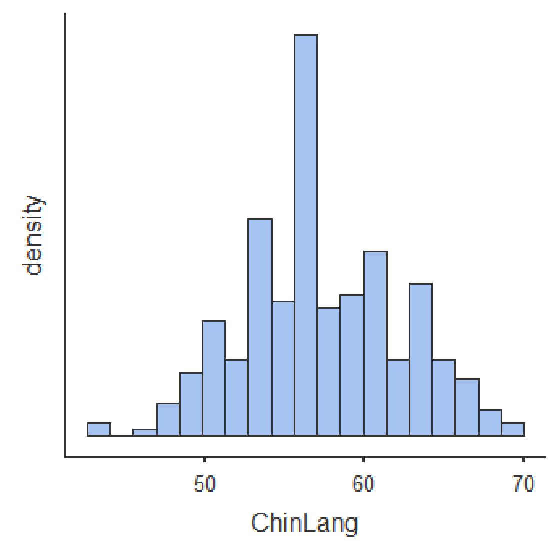
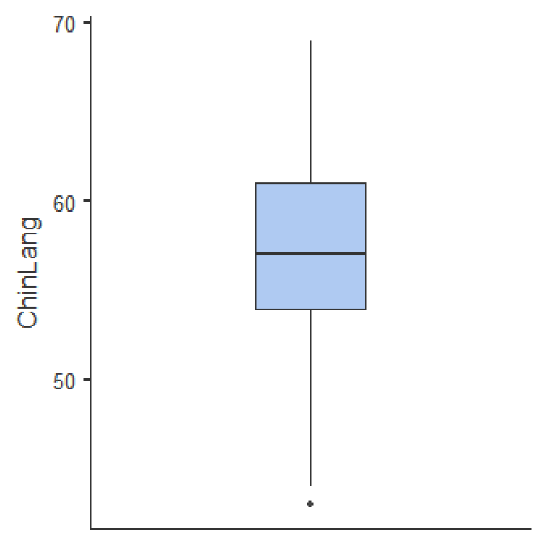
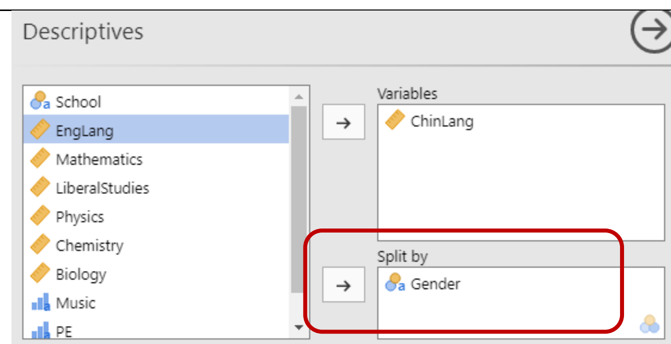
5. Expand the **Plots** option and **checked** the following boxes:

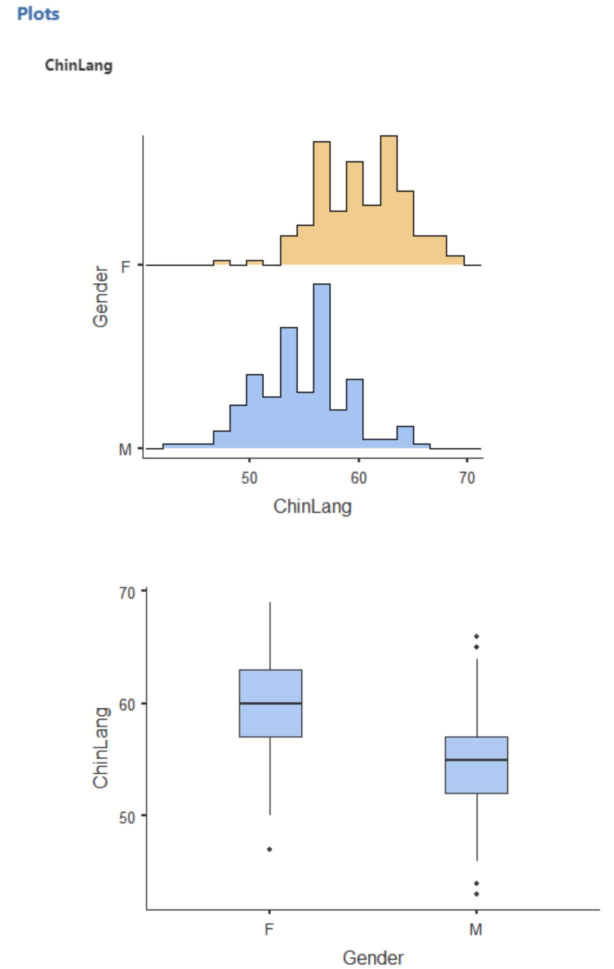
- **Histogram:** A chart that shows the frequency distribution of a variable. In a histogram, values are divided into bins and then count the number of observations that fall within each bin.

- **Box plot:** It provides a simple visual depiction of *median*, *interquartile range*, and the range of the data. The thick line in the middle of the box is the *median*. The box itself spans the range *from the 25th percentile to the 75th percentile* *Outliers* are plotted as a dot. IQR means interquartile range.





	Plots **ChinLang**  
6. Pull the **Gender** into the **Split by** box. Then, the statistics of ChinLang will be grouped by gender. 7. The **Descriptives table** and the **Plots** are updated automatically.	



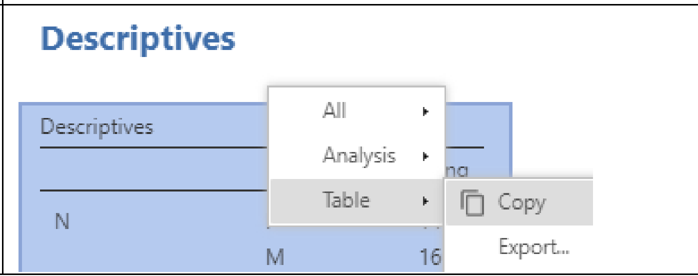
Descriptives

Descriptives

	Gender	ChinLang
N	F	140
	M	160
Missing	F	0
	M	0
Mean	F	60.2
	M	54.8
Median	F	60.0
	M	55.0
Mode	F	57.0
	M	56.0
Standard deviation	F	4.06
	M	4.10
Variance	F	16.5
	M	16.8
Range	F	22
	M	23
Minimum	F	47
	M	43
Maximum	F	69
	M	66
25th percentile	F	57.0
	M	52.0
50th percentile	F	60.0
	M	55.0
75th percentile	F	63.0
	M	57.0

8. You can right-click the table to copy it to another file such as Word and Excel.

9. **Save** the file **MyExam.omv**



Analyzing Differences among Groups

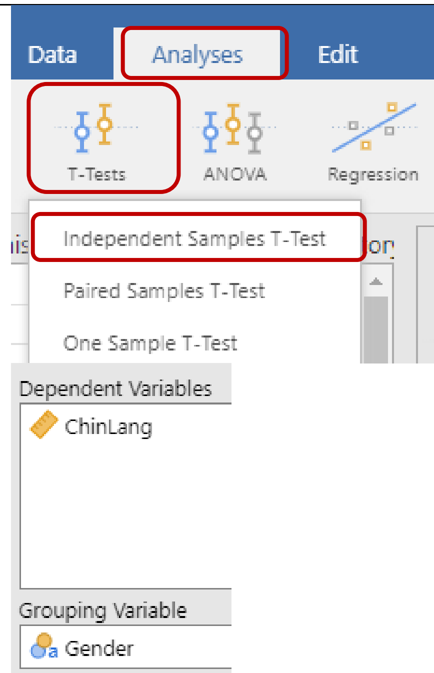
Plotting the data is a good way to get a feel for differences between groups, but statistics can provide us with two more pieces of information: a **confidence interval** for the difference between the group means and a measure of the probability that an effect is due to chance (statistical significance). We will use some statistical techniques such as **t-test** and **ANOVA** to do our analysis.

Note: α=0.05 is applied throughout the exercises. (If alpha equals 0.05, then the confidence level is 0.95)

Independent Samples T-Test

The **Independent Samples T-Test** compares the **means** of **two independent groups** and determine whether there is a statistically significant difference between the means in two unrelated groups. Also known as **Student's T-Test**.

1. Open the **MyExam.omv**
2. Select **Analyses** → **T-Tests** → **Independent Samples T-Test**
3. To compare the Chinese Language scores between female and male students. Pull the **ChinLang** into the **Dependent Variables** box and the **Gender** into the **Grouping Variable**.

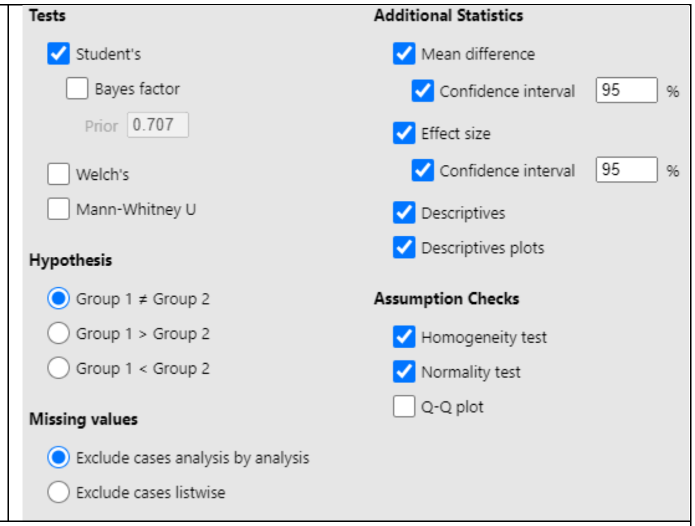


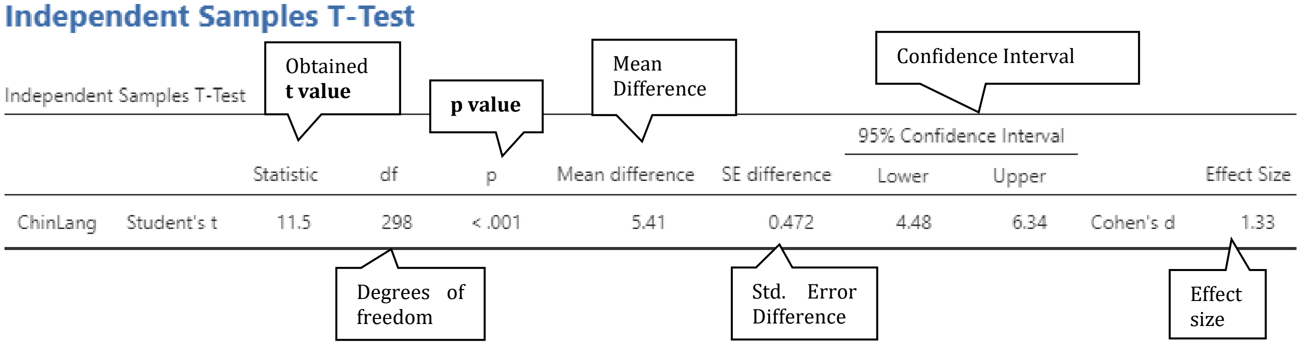
Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p
ChinLang	Student's t	11.5	298	< .001

4. Additional options for the analysis.
   - Under **Additional Statistics**, select **Mean Difference**, **Effect Size**, (with **Confidence interval 95%**), **Descriptives**, and **Descriptive plots**.
   - Under **Assumption Checks**, select **Normality test, Homogeneity test**
   - Notice that under **Hypothesis,** two-tailed t-test (**Group 1 ≠ Group 2**) is selected - we are hypothesizing that the mean Chinese language score of male students does not equal to that of female students.





Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference	95% Confidence Interval Lower	95% Confidence Interval Upper		Effect Size
ChinLang	Student's t	11.5	298	< .001	5.41	0.472	4.48	6.34	Cohen's d	1.33

Overall Test results:

- The test statistic, **t**, is 11.5.
- **df** = **degrees of freedom.** It indicates the sample size – 2 (two sample groups) = 298.
- p is the **p-value**. If this value is **smaller than 0.05**, the α ➔ there is support for our hypothesis (Group1's mean is not equal to Group2's mean, there is significant difference between the two means). If it is larger, then we reject our hypothesis. (https://www.socscistatistics.com/pvalues/tdistribution.aspx)
- **Mean difference** shows the mean difference between the two groups: **5.41**.
- **Confidence Interval** ➔ 95% of the time the true difference in means would lie between 4.48 and 6.34.
- **Cohen's d effect size**: how many standard deviation are between two group's mean. (Cohen's d is the most commonly used measure of effect size for a t-test.) 1.33 effect size means that two groups' means differ by 1.33 standard deviation. If two groups' means don't differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant. In this case, d is 1.33, it is a large effect.

d-value	Rough interpretation
About 0.2	small effect
About 0.5	moderate effect
About 0.8	large effect

5. The **Assumptions** section is **not** the t-test. It is used to make sure that assumption is met in your data for statistical tests.
   - **Normality Test (Shapiro-Wilk)** is to test whether the variable is normally distributed within each group. In this case, you can see that the W is 0.993 and the **p-value is 0.216 (> 0.05)**, it means that the <u>two groups are approximately normally distributed</u>.
   - **Homogeneity of Variances Test (Levene's)** is to test whether the variances of the two groups are equal. In this case, F is 0.187 and the **p-value is 0.666 (> 0.05),** it means that the variances are similar and the <u>assumption of equal variance was met.</u>

Assumptions

Normality Test (Shapiro-Wilk)

	W	p
ChinLang	0.993	0.216

Note. A low p-value suggests a violation of the assumption of normality

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
ChinLang	0.187	1	298	0.666

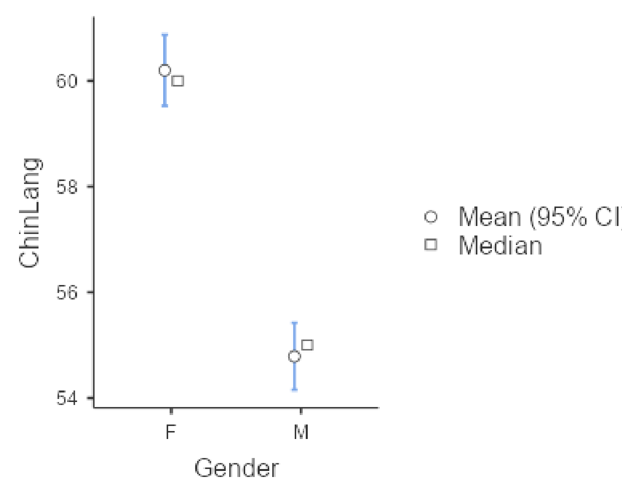
Note. A low p-value suggests a violation of the assumption of equal variances

6. **Group Descriptives** and the **Plots** show the descriptive statistics of each group and give us the general ideas about the central tendency and the dispersion of the groups.

Group Descriptives

	Group	N	Mean	Median	SD	SE
ChinLang	F	140	60.2	60.0	4.06	0.344
	M	160	54.8	55.0	4.10	0.324

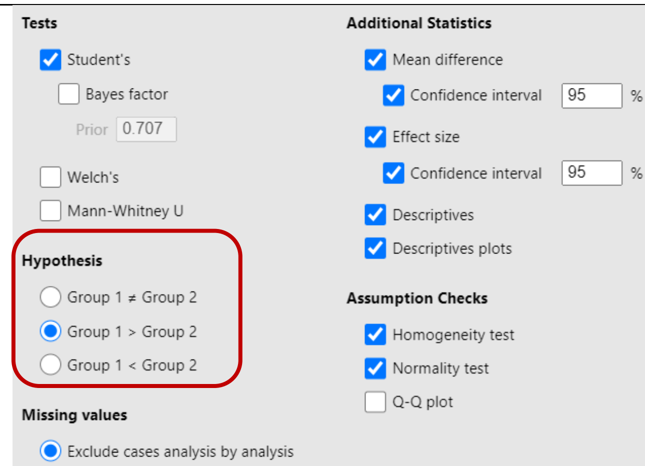
ChinLang



Overall Test results:

- An independent sample t-test was conducted to determine if there were significant differences in Chinese language score between male and female students.
- Chinese language scores for each level of gender were normally distributed, as assessed by Shapiro-Wilk test ($p > 0.05$).
- Homogeneity of variances was met, as assessed by Levene's Test for Equality of Variance ($p > 0.05$).
- On average, the Chinese language scores of female students (M=60.2, SD=4.06) were higher than the scores of male students (M=54.8, SD=4.10), a statistically significant difference, Mean difference = 5.41, 95%CI[4.48, 6.34], $t(298)=11.5$, $p<0.001$, $d=1.33$. These results support the hypothesis that the means for these two groups are not equal.

7. The **Hypothesis** can be changed to <u>one-tailed t-test</u> **Group 1 > Group 2** or **Group 1 < Group 2** to see whether the one group mean is higher or lower than the other group. In our data, female is listed first, so <u>Group 1 is female, and Group 2 is male</u>. We can verify it from the result below.



Independent Samples T-Test

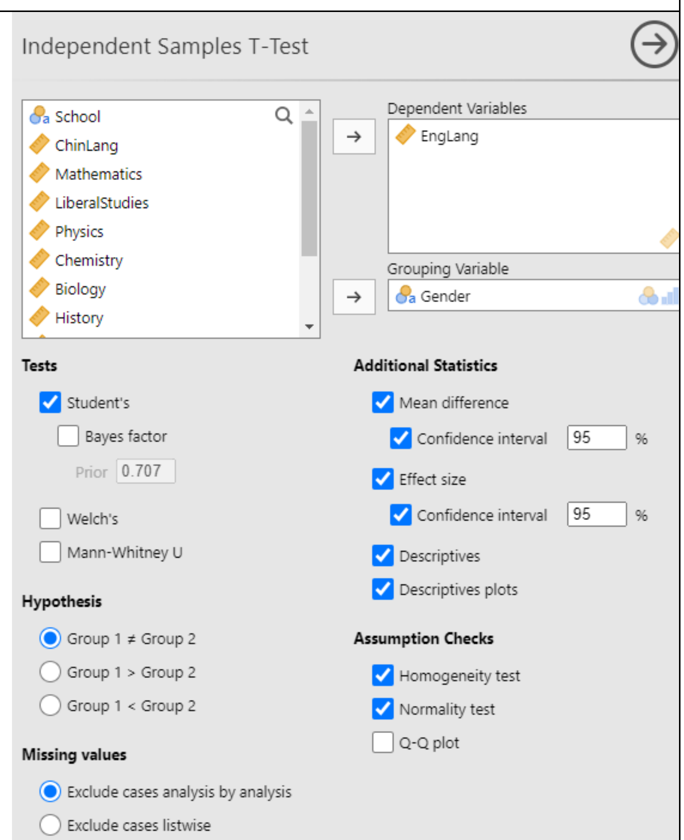
Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference	95% Confidence Interval Lower	95% Confidence Interval Upper		Effect Size
ChinLang	Student's t	11.5	298	< .001	5.41	0.472	4.63	Inf	Cohen's d	1.33

Note. H_a F > M

8. Let's repeat the test with **Hypothesis Group 1 ≠ Group 2** to compare the English Language scores between female and male groups.

9. Pull the **EngLang** into the **Dependent Variables** box and the **Gender** into the **Grouping Variable**.



10. There is a violation of the equal variances assumption. The **Levene's test** is having a p value less than 0.05.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference
EngLang	Student's t	14.5 [a]	298	< .001	9.84	0.680

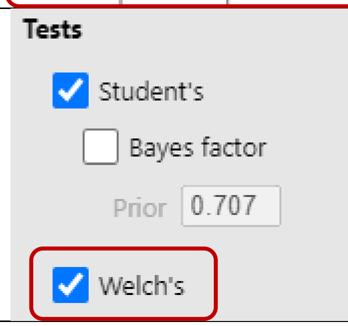
[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
EngLang	6.89	1	298	0.009

Note. A low p-value suggests a violation of the assumption of equal variances

11. **Levene's test** is violated, so **Welch's** test will be used. Check the **Welch's** box under the **Tests** section. Welch's t-test **does NOT require the assumption of equal variance between populations.**



Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference	95% Confidence Interval Lower	95% Confidence Interval Upper		Effect Size
EngLang	Student's t	14.5 [a]	298	< .001	9.84	0.680	8.50	11.2	Cohen's d	1.68
	Welch's t	14.7	295	< .001	9.84	0.670	8.52	11.2	Cohen's d	1.69

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Group Descriptives

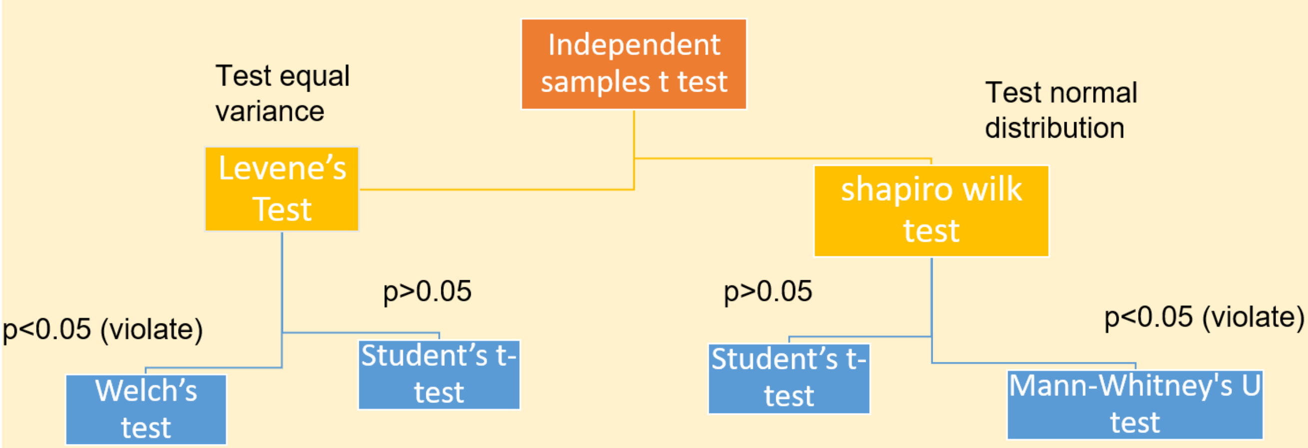
	Group	N	Mean	Median	SD	SE
EngLang	F	140	60.1	60.0	5.13	0.434
	M	160	50.3	50.0	6.45	0.510

Overall Test results:

- An independent sample t-test was conducted to determine if there were differences in English language score between male and female students.
- English language scores for each level of gender were normally distributed, as assessed by Shapiro-Wilk test (p > 0.05).
- Homogeneity of variances was violated, as assessed by Levene's Test for Equality of Variance (p =0.009), so the Welch's t-test was used.

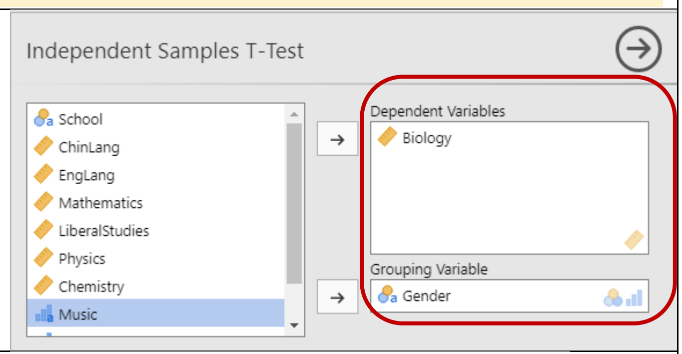
- On average, the English language scores of female students (M=60.1, SD=5.13) were higher than the scores of male students (M=50.3, SD=6.45), a statistically significant difference, Mean difference = 9.84, 95%CI[8.52, 11.2], t(295)=14.7, p<0.001, d=1.68. These results support the hypothesis that the two groups are not equal.

- Independent t-test assumes:
  1. the data from each group are from a normal distribution,
  2. the variances of these groups are equal. (*homogeneity of variance)*
  3. No outliers



12. Let us repeat the test with **Hypothesis Group 1 ≠ Group 2** to compare the Biology scores between female and male students.

13. Pull the **Biology** into the **Dependent Variables** box and the **Gender** into the **Grouping Variable**.



14. The **Assumptions** section shows there is a violation of the normality assumption. The **Shapiro-Wilk test** has a p value less than 0.05.

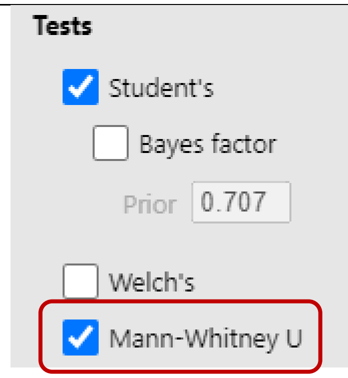
Assumptions

Normality Test (Shapiro-Wilk)

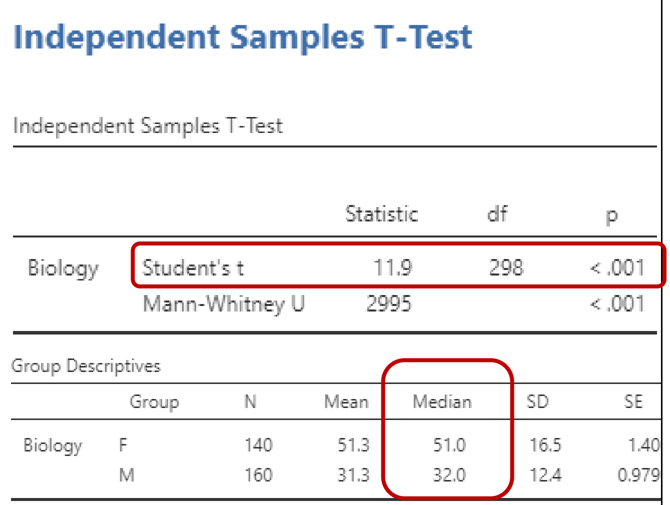
	W	p
Biology	0.930	< .001

Note. A low p-value suggests a violation of the assumption of normality

15. In this case, **check** the **Mann-Whitney U** box under the **Tests** section. **It does not assume normal distribution**. Mann-Whitney U-test does not compare mean score but median scores of the two samples.



16. The test statistic Mann-Whitney U is 2995. And the p value is <.001 which is less than 0.05.



Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p
Biology	Student's t	11.9	298	< .001
	Mann-Whitney U	2995		< .001

Group Descriptives

	Group	N	Mean	Median	SD	SE
Biology	F	140	51.3	51.0	16.5	1.40
	M	160	31.3	32.0	12.4	0.979

Overall Test results:

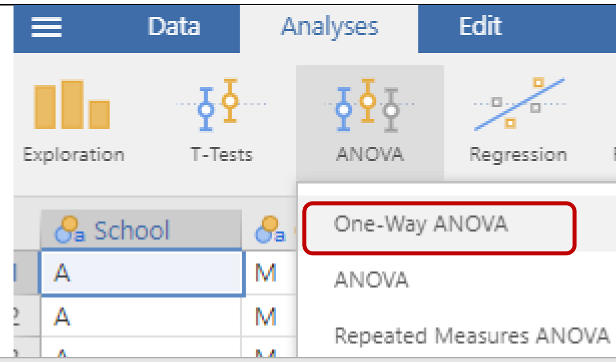
- An independent sample t-test was conducted to determine if there were differences in Biology score between male and female students.
- As assessed by Shapiro-Wilk test ($p < 0.001$), the normality assumption was violated, so Mann-Whitney U test was used instead.
- The Mann-Whitney U test showed that there was a significant difference ($U = 2995$, $p < 0.001$) in Biology scores between the female students compared to the male students.
- The median Biology score of female students was 51.0 compared to 32.0 for the male students. These results support the alternative hypothesis that the two groups are not equal.

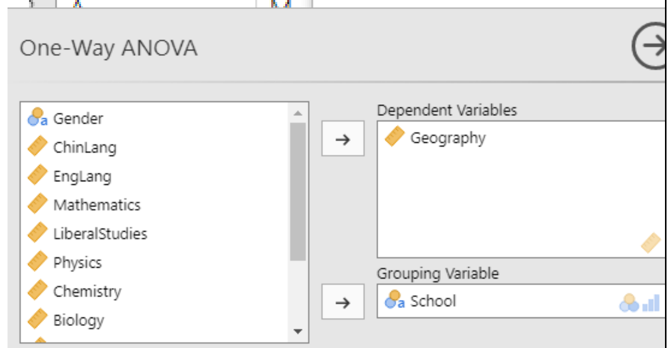
One-Way ANOVA

The **one-way analysis of variance (ANOVA)** is used to determine whether there are any statistically significant difference between the **means** of **two or more independent groups**. Although we tend to only see it used when there is a minimum of three, rather than two groups.

<table>
<tr>
<td>

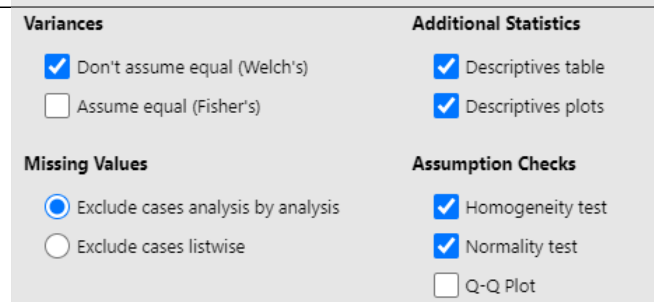
1. Open the **MyExam.omv** in jamovi if it is not opened.
2. Select **Analyses → ANOVA → One-Way ANOVA**
3. To compare the **Geography** scores among the three schools. Pull the **Geography** into the **Dependent Variables** box and the **School** into the **Grouping Variable.**

</td>
<td>




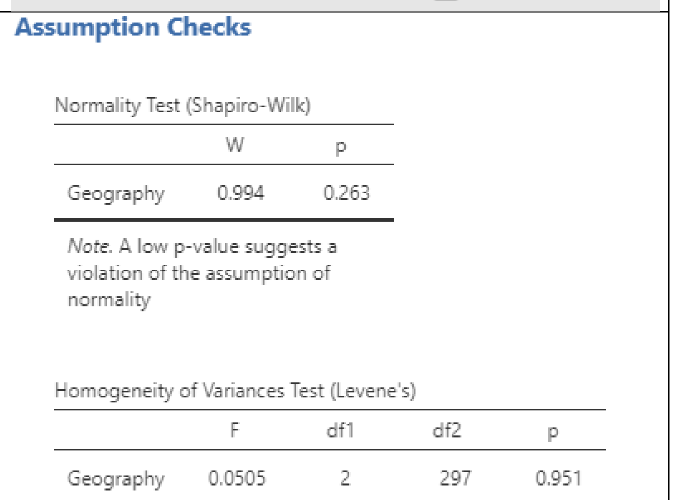
</td>
</tr>
<tr>
<td>

4. Under the **Additional Statistics**, select **Descriptives table** and **Descriptives plots.**
5. Under the **Assumption Checks**, select **Homogeneity test** and **Normality test.**

</td>
<td>


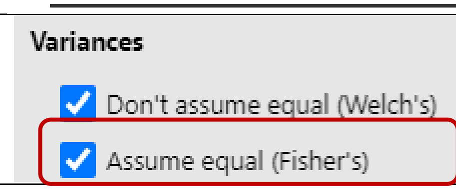
</td>
</tr>
<tr>
<td>

6. The **Assumption Checks** shows
   - No violation of normality (p=0.263 > 0.05)
   - No violation of equality of variances (p=0.951 >0.05)

</td>
<td>


</td>
</tr>
<tr>
<td>

7. With equal variances, select **Assume equal (Fisher's)** under the **Variances** section to perform Fisher's test

</td>
<td>


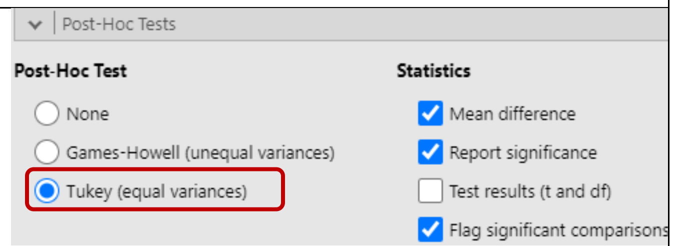
</td>
</tr>
</table>

8. From the Fisher's result, p value is 0.002, <0.05. Therefore, there is a statistically significant difference in the mean Geography scores among the three schools.

One-Way ANOVA

		F	df1	df2	p
Geography	Welch's	6.35	2	198	0.002
	Fisher's	6.44	2	297	0.002

9. Expand the **Post-Hoc Tests** option, under **Post-Hoc Test**, select **Tukey (equal variances)** to run the post-hoc test for groups with equal variances.



10. From the post-hoc test result, we can see there are significantly difference for the means in school A and C, as well as school A and B. ($p<0.05$)

11. Save the file **MyExam.omv**

Tukey Post-Hoc Test – Geography

		A	B	C
A	Mean difference	—	−1.45 *	−1.940 **
	p-value	—	0.028	0.002
B	Mean difference		—	−0.490
	p-value		—	0.659
C	Mean difference			—
	p-value			—

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

Group Descriptives

	School	N	Mean	SD	SE
Geography	A	100	60.8	3.99	0.399
	B	100	62.3	3.91	0.391
	C	100	62.8	4.02	0.402

Overall Test results:

- A one-way ANOVA test was conducted to determine if there were significantly differences in Geography score among three schools.
- Geography scores for each school were **normally distributed**, as assessed by **Shapiro-Wilk test** ($p > 0.05$).
- **Homogeneity of variances was met**, as assessed by **Levene's Test** for Equality of Variance ($p > 0.05$), so **Fisher's test** was used.
- There was a significant difference of the Geography scores at the $p < 0.05$ for the three schools. $F(2,297)=6.44$, $p=0.002$. The post hoc comparisons using Tukey test indicated that the mean Geography score for school B (M=62.3, SD=3.91) was not significantly different from school C (M=62.8, SD=4.02). However, the mean Geography score of school A (M=60.8, SD=3.99) was significantly different from both school B and C.

Paired Samples T-Test

The **paired-samples t-test** compares the **means** of **two related groups** to determine whether there is a statistically significant difference between these means. This test is also called the **paired t-test** or **dependent t-test**.

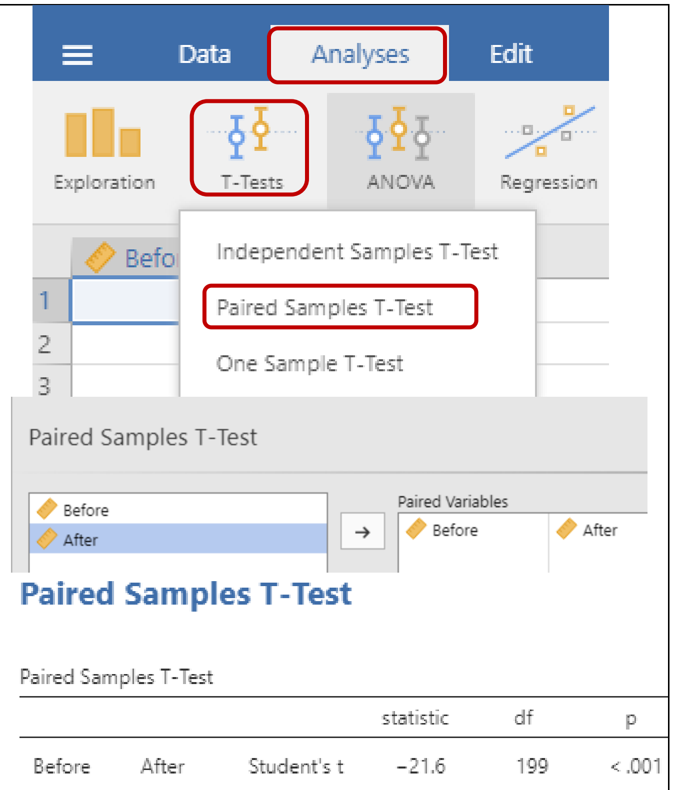
Examples of where this might occur are:

- Before-and-after observations on the same subjects. E.g. students' diagnostic test results before and after a module or course.
- A comparison of two different methods of measurement or two different treatments where the measurements or treatments are applied to the same subjects.

1. Open the **Enhancement.omv**. The file contains **Mathematics** scores of a group of students before and after taking the enhancement class. Let us compare the scores to see whether the enhancement class made a difference.

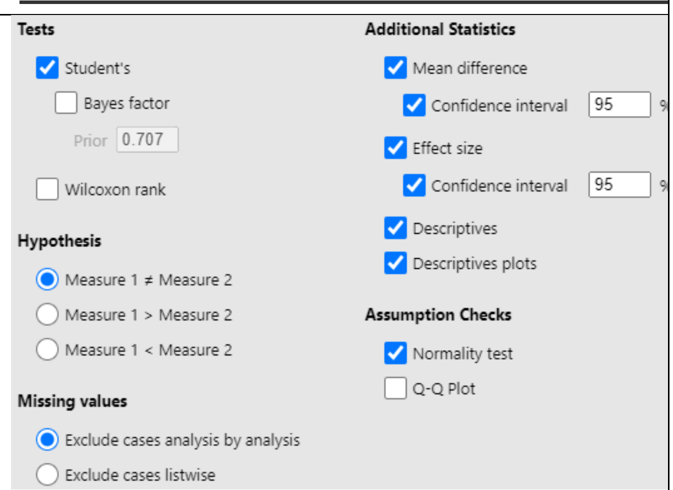
2. Select **Analyses → T-Tests → Paired Samples T-Test**

3. Pull the **Before** and **After** into the **Paired Variables** box.



4. Under the **Additional Statistics**, select **Mean difference, Effect size, Confidence interval, Descriptives** and **Descriptives plots**.

5. Under the **Assumption Checks**, select **Normality test**.



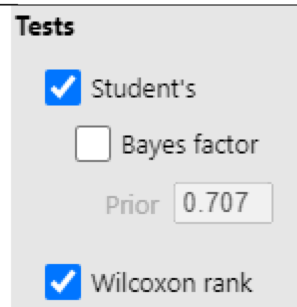
6. There is a violation of the normality assumption. The **Shapiro-Wilk** test has a **p value less than 0.05**.

Normality Test (Shapiro-Wilk)

			W	p
Before	-	After	0.331	< .001

Note. A low p-value suggests a violation of the assumption of normality

7. Therefore, **check** the **Wilcoxon rank** box under the **Tests** section. Use the **Wilcoxon W** instead of **Student's t**. Wilcoxon test does not compare mean score but median scores of the two samples.



Paired Samples T-Test

Paired Samples T-Test

			Statistic	df	p	Mean difference	SE difference	95% Confidence Interval Lower	95% Confidence Interval Upper
After	Before	Student's t	21.6	199	< .001	2.88	0.133	2.62	3.15
		Wilcoxon W	18755		< .001	3.00	0.133	3.00	3.00

8. From the **Wilcoxon W** result, p value is <0.001, which is below 0.05. Therefore, there is a **statistically significant difference in the median Mathematics score** before and after taking the enhancement class.

Descriptives

	N	Mean	Median	SD	SE
Before	200	76.7	76.5	5.40	0.382
After	200	79.6	79.5	5.60	0.396

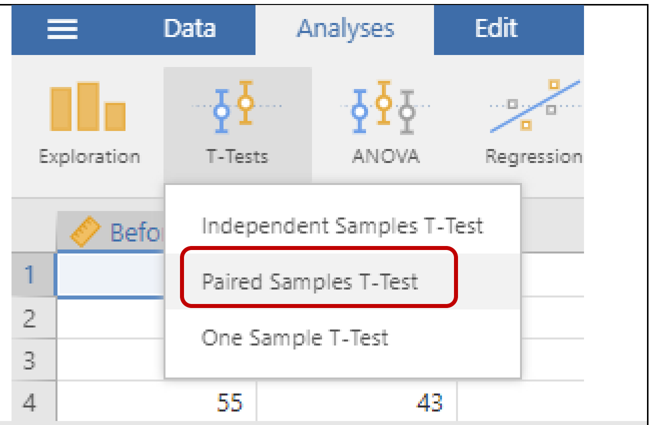
Overall Test results:

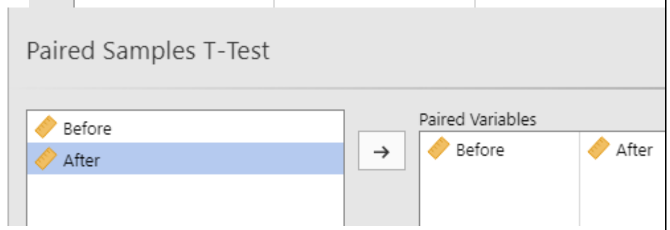
- A paired sample t-test was conducted to determine if there were statistically significant differences in Mathmatics scores before and after having the enhancement class.
- As assessed by Shapiro-Wilk test ($p < 0.001$), the normality assumption was violated, so Wilcoxon rank test was used instead.
- Wilcoxon rank test showed that there was a significant difference ($W = 1345$, $p < 0.001$) in Mathematics median scores before and after having the enhancement class.
- The median Mathematics score was 76.5 before the class, compared to 79.5 after the class. These results support that the enhancement class made a difference on the Mathematics scores.

9. Open the **Tutorial.omv**. The file contains **Mathematics** scores of a group of students before and after taking the tutorial class. Let us compare the scores to see whether the tutorial class made a difference.

10. Select **Analyses → T-Tests → Paired Samples T-Test**

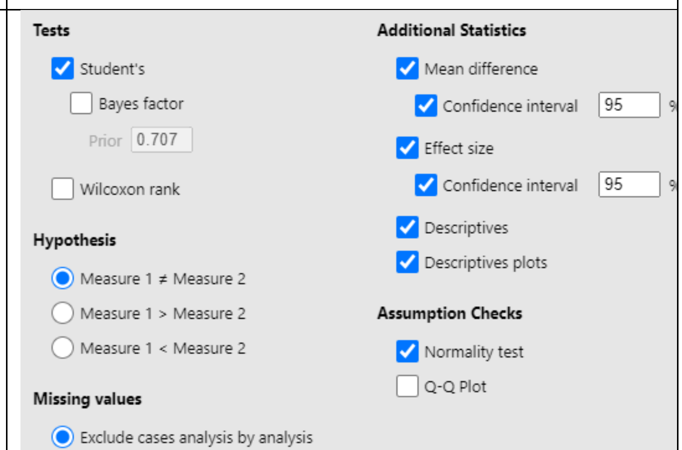
11. Pull the **Before** and **After** into the **Paired Variables** box.





12. Under the **Additional Statistics**, select **Mean difference, Effect size, Confidence interval, Descriptives** and **Descriptives plots**.

13. Under the **Assumption Checks**, select **Normality test**.



14. The test shows the scores are normally distributed. The **Shapiro-Wilk** test has a p value 0.710 (> 0.05).

Normality Test (Shapiro-Wilk)

			W	p
Before	-	After	0.995	0.710

Note. A low p-value suggests a violation of the assumption of normality

Paired Samples T-Test

Paired Samples T-Test

			statistic	df	p	Mean difference	SE difference	95% Confidence Interval Lower	95% Confidence Interval Upper		Effect Size
Before	After	Student's t	0.0800	199	0.936	0.0400	0.500	−0.945	1.03	Cohen's d	0.00566

15. From the Student's t result, p value is 0.936 (>0.05) Therefore, mean Mathematics scores before and after taking the tutorial class are similar.

Descriptives

	N	Mean	Median	SD	SE
Before	200	49.7	50.0	4.77	0.338
After	200	49.7	50.0	4.90	0.347

Overall Test results:

- A paired sample t-test was conducted to determine if there were statistically significant differences in Mathmatics mean scores before and after having the tutorial class.
- The scores were normally distributed, as assessed by Shapiro-Wilk test (p > 0.05).
- The t-test result shows there was no statistically significant difference (p=0.936 > 0.05) between the Mathematics mean scores before and after having the tutorial class. The tutorial class made no difference on the Mathematics mean scores.

Correlation and Regression

Correlation is a statistical technique that can show **whether and how strongly pairs of variables are related**. For example, height and weight are related; correlation can tell how much of the variation in people's weights is related to their heights. **Regression** is to **predict the value of an outcome variable** based on the known value of one or more predictor variables.

a) <u>Pearson Correlation Coefficient</u>
Pearson is the most widely used **correlation coefficient**. It measures the **linear association** between **continuous variables**. (interval or ratio scale data)

b) <u>Spearman's Correlation Coefficient</u>
Spearman's rank correlation coefficient can be defined as a **special case of Pearson coefficient** applied to **ranked variables**. Unlike Pearson, Spearman's correlation is not restricted to linear relationships. Instead, it measures **monotonic association** (only strictly increasing or decreasing, but not mixed) between two ranked variables. A ranking is a relationship between a set of items such that for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In other words, rather than comparing means and variances, Spearman's coefficient looks at the relative order of values for each variable. This makes it appropriate to use with both **continuous** and **discrete data**.

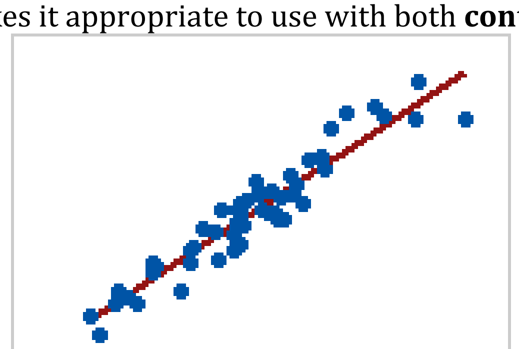


Figure 1: Linear relationship

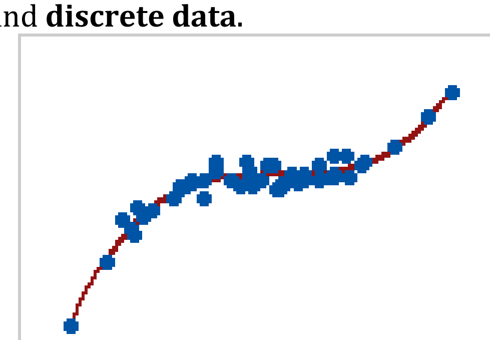
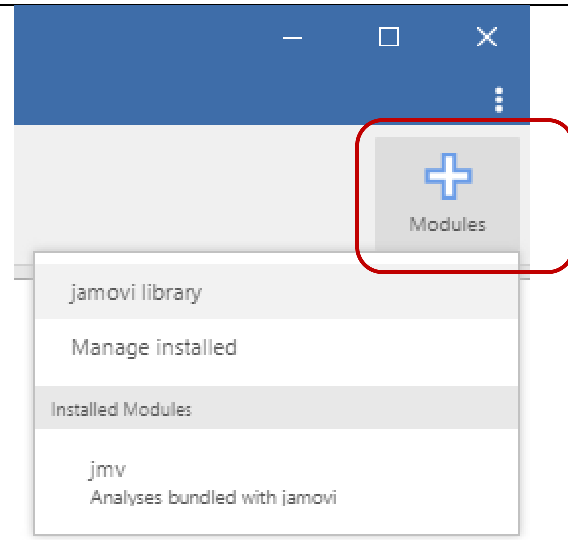
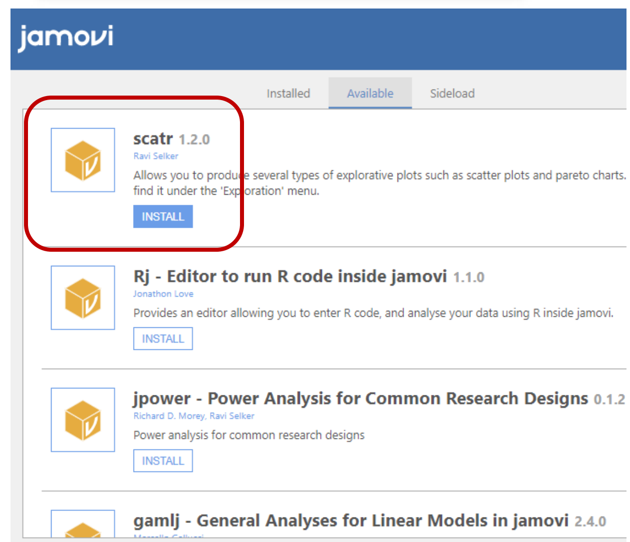


Figure 2: Monotonic relationship

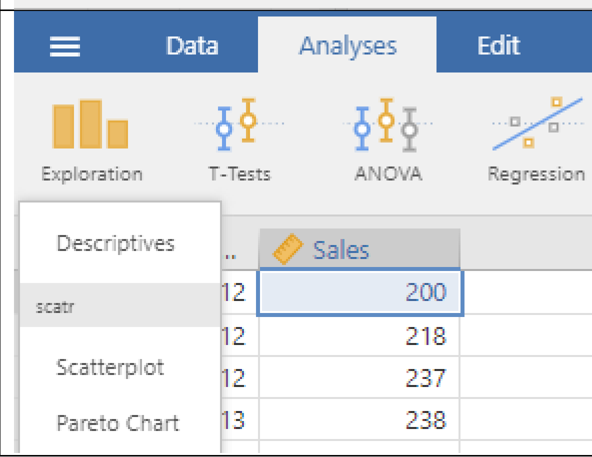
1. Open the **Ice-cream.omv**. The file recorded how much ice-cream was sold on a day in an ice-cream stall and the temperature on that day.

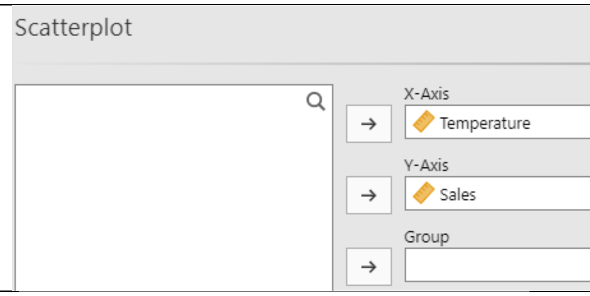
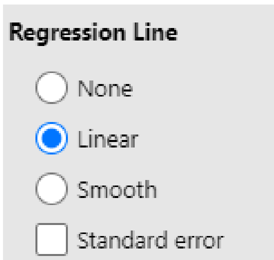
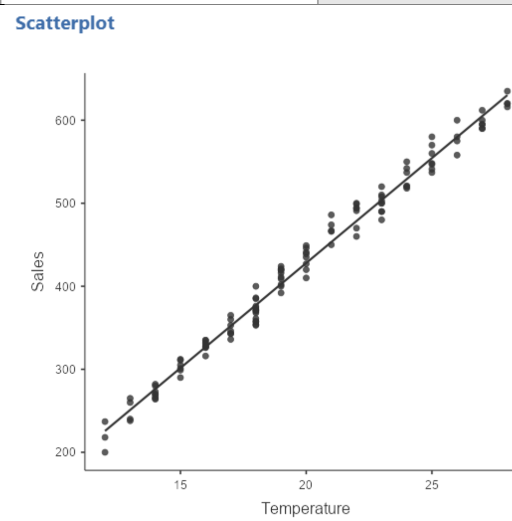
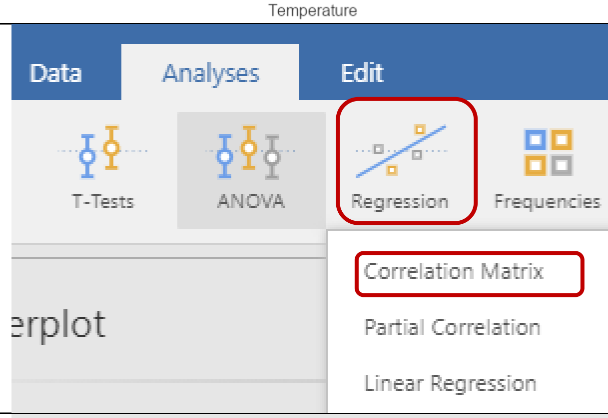
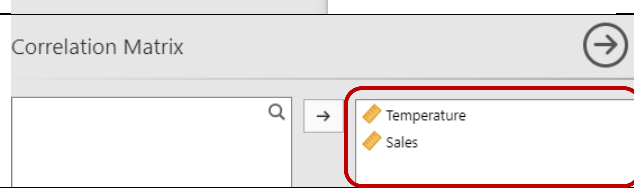
2. Click **Modules** button at the top right and Install **scatr** module.





3. Select **Analyses** → **Exploration** → **Scatterplot** to observe relationship between the ice-cream sales and the temperature.



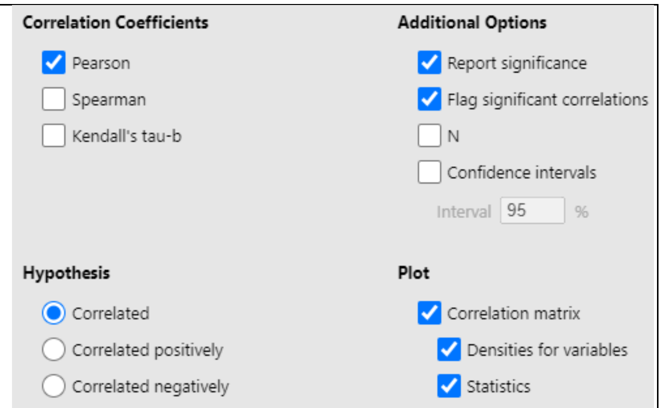
4. Pull the **Temperature** into the **X-Axis** and the **Sales** into **Y-Axis** box. 5. From the observation, the amount of the ice-cream sales increases as the temperature increases. There is a **positive correlation**.	
6. Select **Linear** under the **Regression Line**. 	
7. To find the correlation coefficient, select **Analyses → Regression → Correlation Matrix** (Correlation matrix is a way to examine linear relationships between two or more continuous variables)	
8. Pull the **Temperature** and **Sales** to the box on the right.	

9. Under **Correlation Coefficients**, select **Pearson**, as the relationship is linear.

10. Under **Additional Options**, select **Report significance** and **Flag significant correlations** to show the p value and highlight (by asterisk) the significant results respectively.

11. Under **Hypothesis**, select **Correlated** to test whether there is significant correlation. The two other options are to test whether there is positive correlation or negative correlation respectively.

12. Under the **Plot**, select the **Correlation matrix** with **Densities for variables** and **Statistics.**



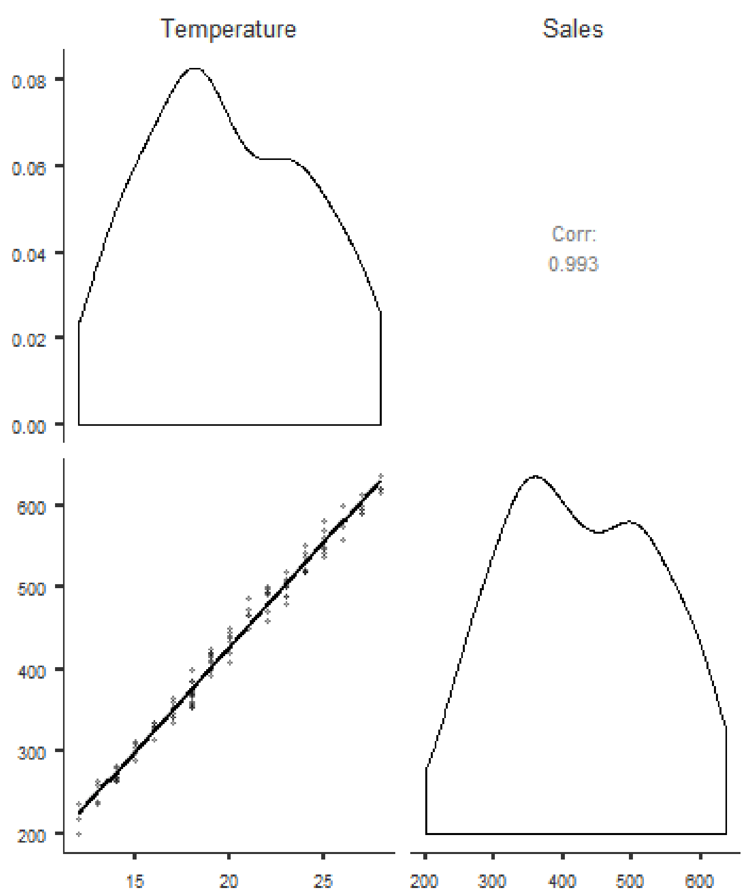
13. The **p value is less than 0.05**, so there is a significant correlation between the two variables. The correlation coefficient **r = 0.993 is positive**, this means that the two variables are **positively correlated**. The absolute value of r tells the strength of the correlation:
    - +1 ➔ perfect positive correlation
    - +0.6 ➔ strong positive correlation
    - +0.1 ➔ weak positive correlation
    - +0 ➔ No correlation
    - -0.1 ➔ weak negative correlation
    - -0.6 ➔ strong negative correlation
    - Near -1 ➔ perfect negative correlation

Correlation Matrix

		Temperature	Sales
Temperature	Pearson's r	—	
	p-value	—	
Sales	Pearson's r	0.993 ***	—
	p-value	< .001	—

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

Plot

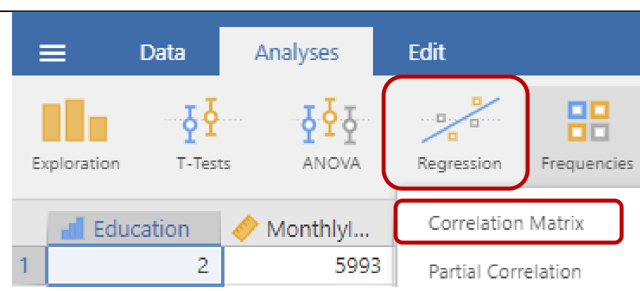


Overall Test results:

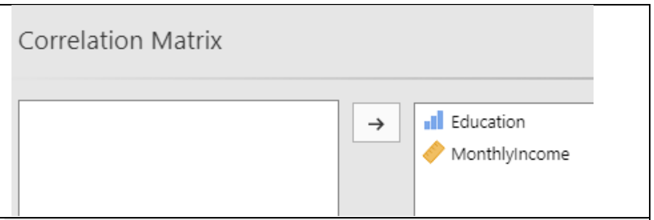
- Results of the Pearson correlation indicated that there was a very strong, significant positive association between temperature and ice-cream sales, ($r = 0.993$, $p < 0.001$).
- As temperature increased, the amount of ice-cream sales increased.

14. Open the **Income.omv**. The file contains the information about education level and the monthly income of the general public obtained from a survey. The education is divided into 5 levels: 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', and 5 'Doctor'.

15. Select **Analyses** → **Regression** → **Correlation Matrix**



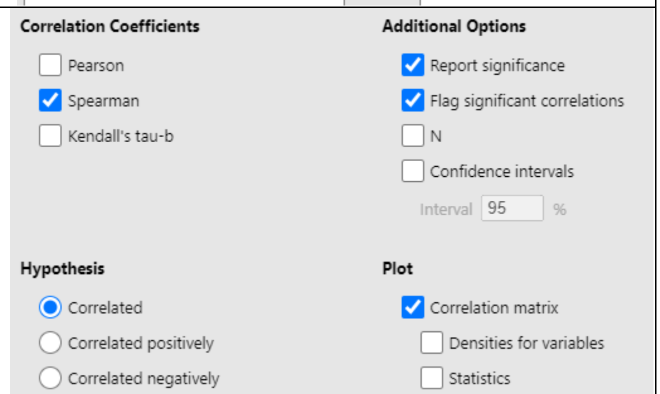
16. Pull the **Education** and **MonthlyIncome** to the box on the right.



17. Under **Correlation Coefficients**, uncheck the **Pearson**, and select **Spearman** correlation coefficient.

18. Under **Additional Options**, select **Flag significant correlations** to show the p value and highlight (by asterisk) the significant results respectively.

19. Under **Hypothesis**, select **Correlated** to test whether there is significant correlation.



20. The **p value is less than 0.05**, so there is a significant correlation between the two variables. The correlation coefficient **Spearman's rho = 0.241** shows that the two variables are **positively correlated** (but not very strong).

Correlation Matrix

Correlation Matrix

		Education	MonthlyIncome
Education	Spearman's rho	—	
	p-value	—	
MonthlyIncome	Spearman's rho	0.241 ***	—
	p-value	< .001	—

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

Overall Test results:

- Results of the Spearman correlation indicated that there was a weak, significant positive association between education level and monthly salary of a person, (rho = 0.241, p <0.001).
- This implies that a person has a higher education level is more likely to get a higher monthly salary.

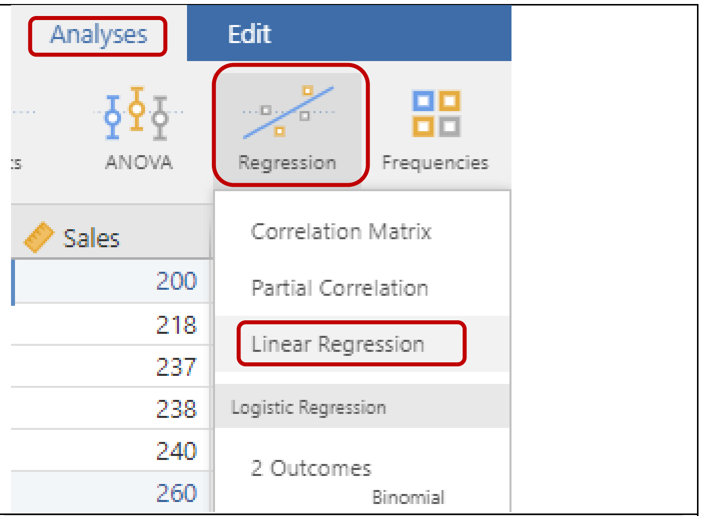
Linear Regression

Linear regression is used for finding linear relationship between outcome and one or more predictors. There are two types of linear regression- **Simple** and **Multiple**.

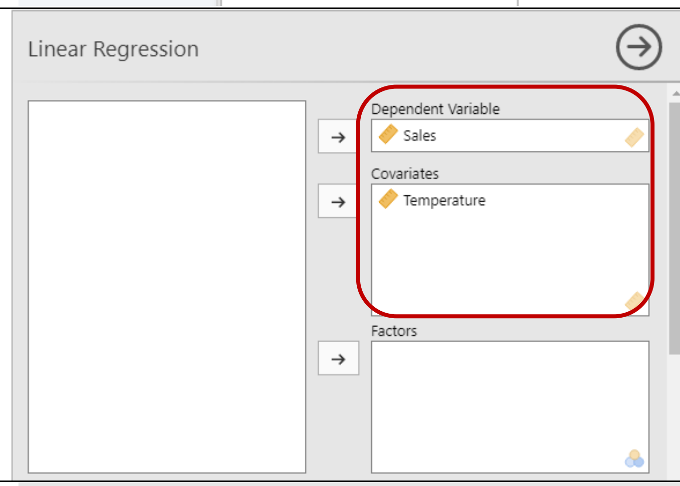
a) <u>**Simple linear regression**</u> - it is useful for finding relationship between **two continuous variables**. One is predictor and the other is outcome (dependent) variable. It is used for <u>predicting the value of the outcome variable</u> from the known predictor variable. It looks for statistical relationship in which the relationship between the variables is not perfect.

1. In the previous part, we found there is a strong positive correlation between the temperature and the ice-cream sales. Now, let us use simple linear regression to predict the ice-cream sales from temperature.

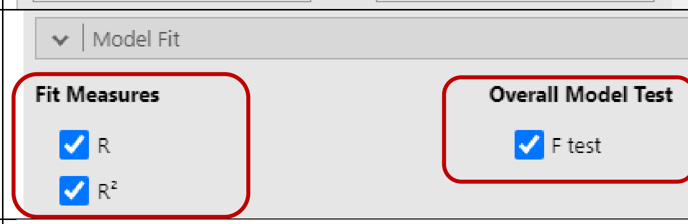
2. Open the **Ice-cream.omv.** Select **Analyses → Regression → Linear Regression**



3. Pull the **Sales** into the **Dependent Variable** box and the **Temperature** into the **Covariates** box. Note that the **Covariates** box is for continuous predictors where the **Factors** box is for categorical predictors.



4. Expand the **Model Fit** option, select the **R**, **R²** and the **F test**.



5. The **Model Fit Measures** table displays information about how the two variables relate to one another.
   - **R** column represents the correlation between the two variables
   - **R²** column indicates the proportion of the variation in the outcome variable that can be explained by the model.
   - **F test** tells us whether the model is a significant predictor of the outcome variables. As the p value is less than 0.05, the model significantly predicts the amount of ice-cream sales.

Model Fit Measures

			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.993	0.986	8432	1	116	< .001

6. The **Model Coefficients** table shows the only predictor temperature significantly contributes to the model as its p value is less than 0.05.

7. We can produce a statistical model to predict the dependent variable:

   ice-cream sales = -77.6 + 25.3(temperature)

   By inserting the temperature (in Celsius) of a day into the equation, we can predict the ice-cream sales on that day.

Model Coefficients - Sales

Predictor	Estimate	SE	t	p
Intercept	−77.6	5.591	−13.9	< .001
Temperature	25.3	0.275	91.8	< .001

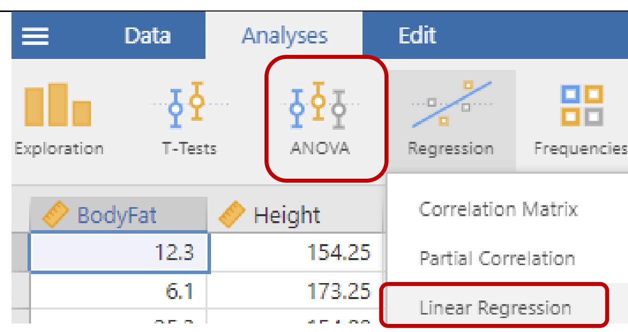
Overall Test results:

- A simple linear regression was carried out to predict the amount of ice-cream sales based on the temperature.
- A significant regression equation was found ($F(1,116) = 8432$, $p < 0.001$), with an R^2 of 0.968. The amount of ice-cream sales is equal to -77.6+25.3(temperature) dollars when the temperature is measured in degree Celsius.
- The amount of the ice-cream sales increased by $25.30 for each degree increase of temperature.

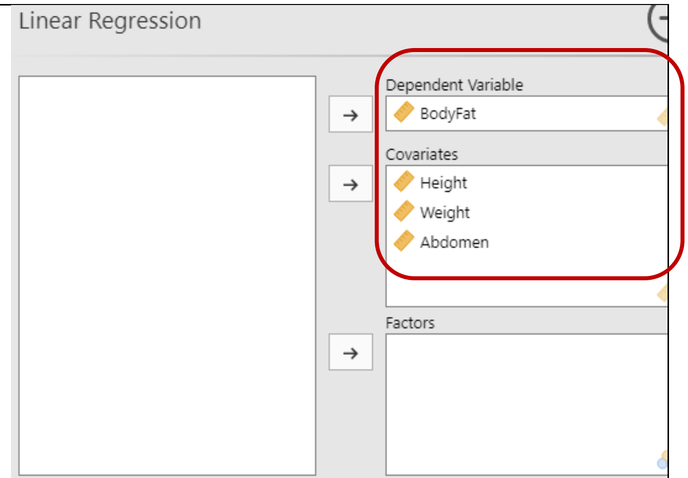
b) **Multiple linear regression**, also known as **multiple regression**, uses **several predictor variables** to predict the outcome variable. The goal of multiple linear regression is to model the **linear relationship** between the predicator variables and the outcome variable.

1. Open the **BodyFat.omv**. The file contains measurements of percent body fat, height(cm), weight(kg), abdomen circumference(cm) and age of 250 men. It is known that percent body fat is difficult and expensive to measure accurately. Let us build a model to predict the percent body fat from the other measurements.

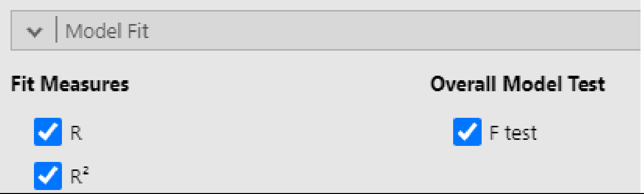
2. Select **Analyses → Regression → Linear Regression**



3. Pull the **BodyFat** into the **Dependent Variable** box and the other variables into the **Covariates** box. Note that the **Covariates** box is for continuous predictors where the **Factors** box is for categorical predictors.



4. Expand the **Model Fit** option, select the **R**, **R²** and the **F test**.



5. In **Model Coefficients** table, it shows that only **Height** and **Abdomen** are the significant predictors (p is <0.05), but it is **only true with the current selection of the predictors**. If we remove one of the predictors, the result will change.

Linear Regression

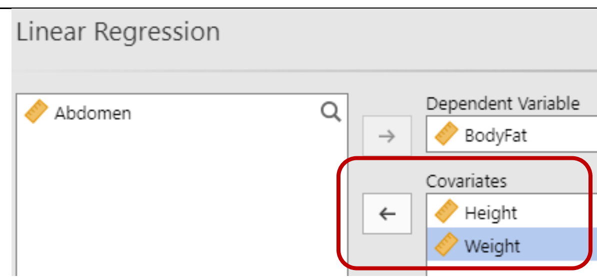
Model Fit Measures

			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.851	0.725	216	3	246	< .001

Model Coefficients - BodyFat

Predictor	Estimate	SE	t	p
Intercept	−31.483	11.5401	−2.73	0.007
Height	−0.101	0.0310	−3.25	0.001
Weight	−0.225	0.1583	−1.42	0.157
Abdomen	0.913	0.0716	12.74	< .001

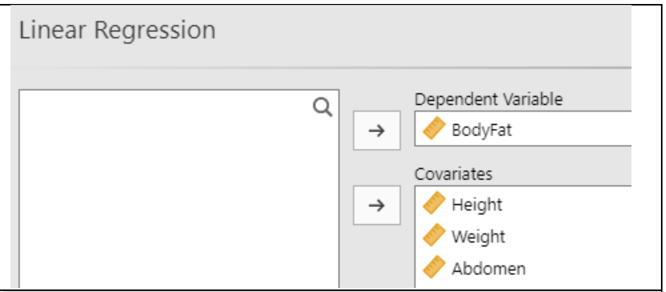
6. Remove **Abdomen** from the **Covariates** box. Now, height and weight are significant predictors.



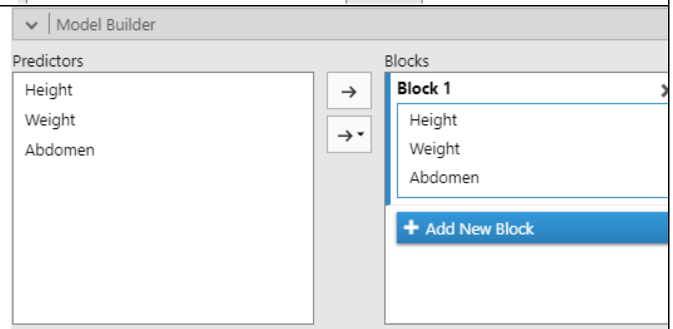
Model Coefficients

Predictor	Estimate	SE	t	p
Intercept	76.781	10.0412	7.65	< .001
Height	0.263	0.0154	17.14	< .001
Weight	−1.488	0.1587	−9.38	< .001

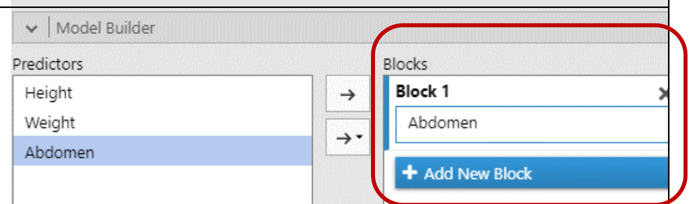
7. Add the **Abdomen** variable back to the **Covariates** box.



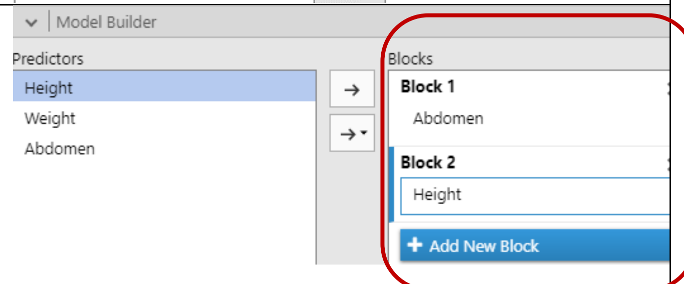
8. Expand the **Model Builder** option. Model builder allows us to build the model from “Blocks”. (“Blocks” are individual building blocks of the final model. The common practice is to add a variable in each block and compared the models statistically so we can decide which model is significantly better than the others.)



9. Remove all the predictors except **Abdomen** from the Block 1.



10. Click **+ Add New Block** to create **Block 2** and pull the **Height** into it. (now, there are only two blocks)



11. The **Model Fit Measures** table shows both models can predict the percent body fat significantly. (p value <0.001)

Model Fit Measures

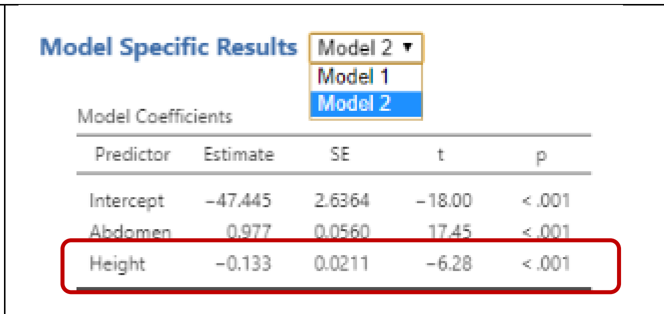
			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.824	0.678	523	1	248	< .001
2	0.850	0.723	322	2	247	< .001

12. The **Model Comparisons** table shows that from model 1 to 2, there is a statistically significant increase (p<0.001). In this case, model 2 is better as it gives higher R^2 value.

Model Comparisons

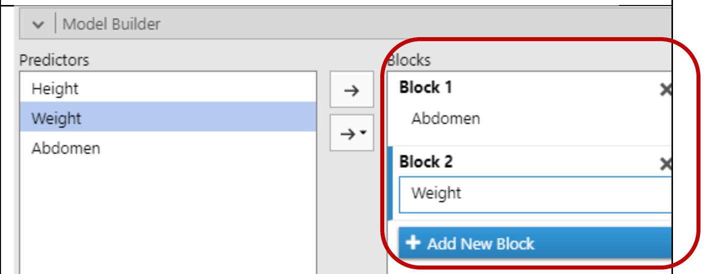
Comparison							
Model		Model	ΔR^2	F	df1	df2	p
1	-	2	0.0443	39.5	1	247	< .001

13. Use the **Model Specific Results** pull down box to switch between the models. In Model 2, the **Height** predictor have significant contribution to the model.

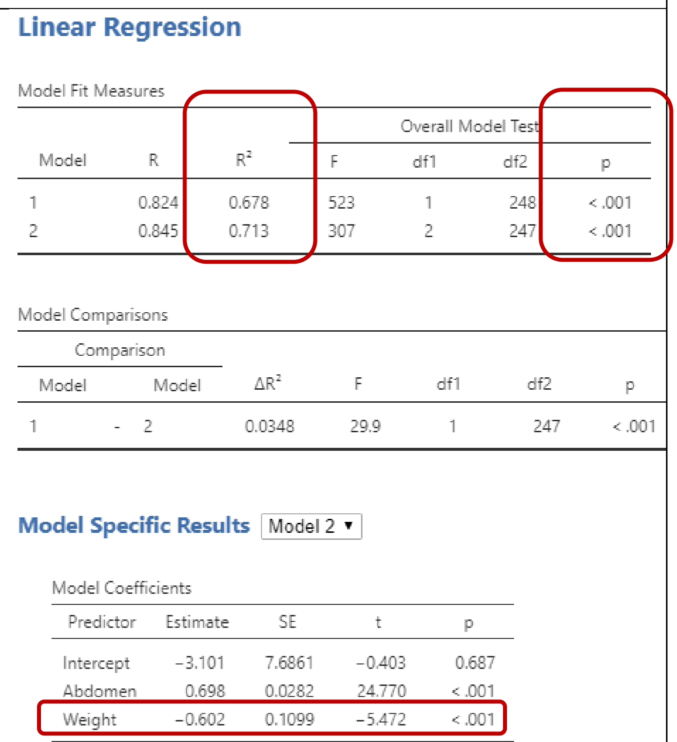


Predictor	Estimate	SE	t	p
Intercept	−47.445	2.6364	−18.00	< .001
Abdomen	0.977	0.0560	17.45	< .001
Height	−0.133	0.0211	−6.28	< .001

14. Let try another predictor. Remove the **Height** predictor from **Block 2** and pull **Weight** into it.



15. The Model 2 is better than Model 1. However, comparing this Model 2 with the Model 2 in step 11, the one with **Height** predictor. The previous model with **Height** predictor performs better than this one as it gives higher R^2 value (0.723).



			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.824	0.678	523	1	248	< .001
2	0.845	0.713	307	2	247	< .001

Model Comparisons

Comparison						
Model	Model	ΔR^2	F	df1	df2	p
1 -	2	0.0348	29.9	1	247	< .001

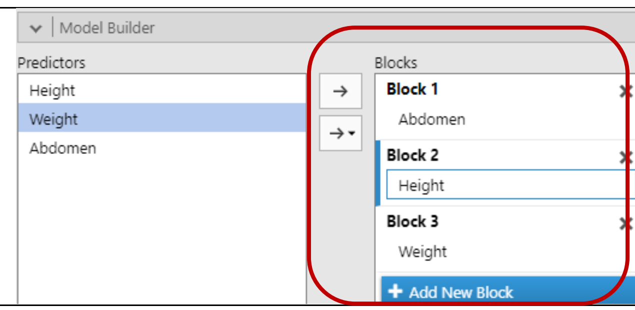
Model Specific Results Model 2

Model Coefficients

Predictor	Estimate	SE	t	p
Intercept	−3.101	7.6861	−0.403	0.687
Abdomen	0.698	0.0282	24.770	< .001
Weight	−0.602	0.1099	−5.472	< .001

16. Remove the **Weight** predictor from **Block 2** and pull **Height** into it.

17. Click **+ Add New Block** to create **Block 3** and pull the **Weight** into it.



18. The **Model Fit Measures** table shows all the three models can predict the percent body fat significantly.

19. The **Model Comparisons** table shows that there is no significant difference from model 2 to 3.

20. Select **Model 3** next to **Model Specific Results.** It shows that **Weight** predictor does not have significant contribution to the model (p =0.157 > 0.05).

Linear Regression

Model Fit Measures

			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.824	0.678	523	1	248	< .001
2	0.850	0.723	322	2	247	< .001
3	0.851	0.725	216	3	246	< .001

Model Comparisons

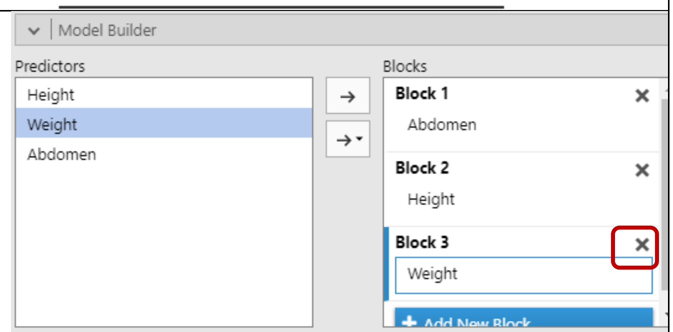
Comparison							
Model		Model	ΔR^2	F	df1	df2	p
1	-	2	0.04431	39.47	1	247	< .001
2	-	3	0.00226	2.02	1	246	0.157

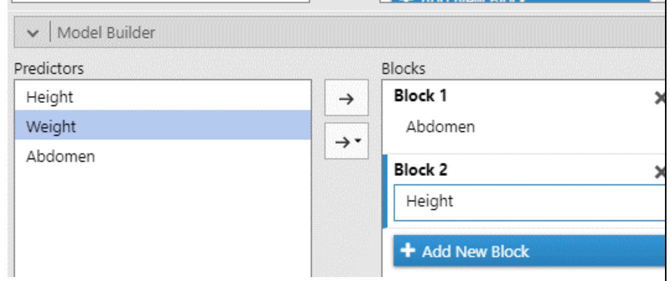
Model Specific Results Model 3

Model Coefficients - BodyFat

Predictor	Estimate	SE	t	p
Intercept	−31.483	11.5401	−2.73	0.007
Abdomen	0.913	0.0716	12.74	< .001
Height	−0.101	0.0310	−3.25	0.001
Weight	−0.225	0.1583	−1.42	0.157

21. Remove **Block 3** from the **Model Builder** by clicking the **cross** at the top right of **Block 3**.





8. We can produce a statistical model that allow us to predict values of the outcome variable based on the predictor variables:

Percent Body Fat = - 47.445
+ 0.977(abdomen circumference)
- 0.113(height)

where the abdomen circumference and height are measured in cm

Model Specific Results Model 2

Model Coefficients - BodyFat

Predictor	Estimate	SE	t	p
Intercept	−47.445	2.6364	−18.00	< .001
Abdomen	0.977	0.0560	17.45	< .001
Height	−0.133	0.0211	−6.28	< .001

Overall Test results:

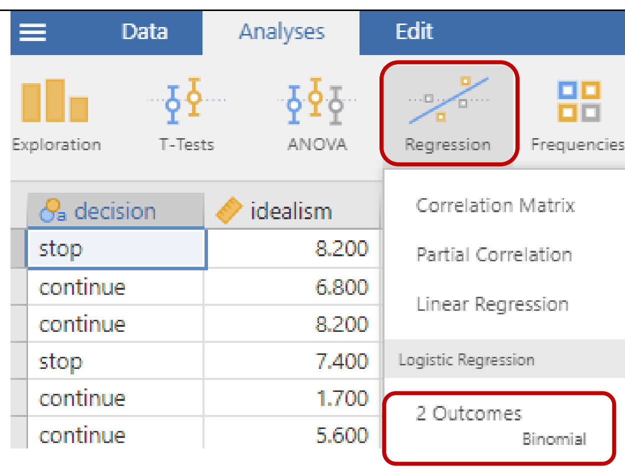
- A multiple linear regression was carried out to predict the percent body fat based on the abdomen circumference and the height of a person.
- A significant regression equation was found ($F(2,247) = 322$, $p < 0.001$), with an R^2 of 0.723. A person's predicted percent body fat is equal to -47.445 + 0.977(abdomen circumference) -0.113(height) where abdomen circumference and height are measured in centimeters.
- The percent body fat increased 0.977% for each cm of abdomen circumference and decreased 0.113% for each cm of height. Both abdomen circumference and height were significant predictors of percent body fat.

Logistic Regression

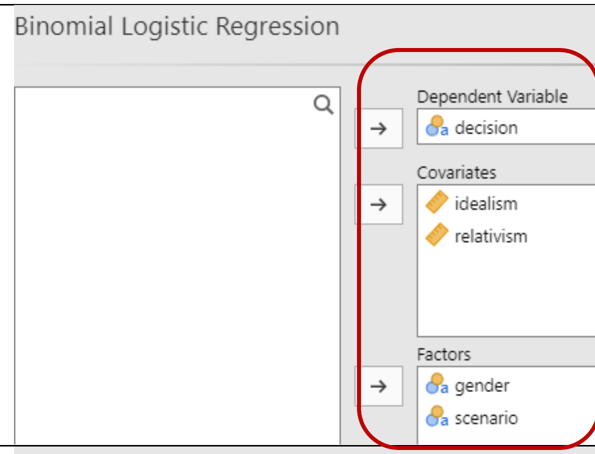
Logistic Regression is a regression technique that is used when we have a **categorical outcome**. This technique can be used to analyze and predict variables that are 'Discrete', 'Nominal' and 'Ordered'. Logistic regression solves the limitation of linear regression in which the outcome variable must be continuous. There are three types of logistic regression - **Binomial**, **Multinomial** and **Ordinal**. Binomial Logistic Regression (binary logistic regression) uses one or more predictor variables to predict **only dichotomous outcome** variable.

1. Open the **AnimalResearch.omv**. The file contains a sample of people's decisions on whether to continue or stop the animal test under five different scenarios. The dataset is obtained and modified from the morality of animal research by Wuensch, K. L., & Poteat, G. M[4][8]. The columns are the decision, idealism score, relativism score and gender of the person and the animal test scenario.

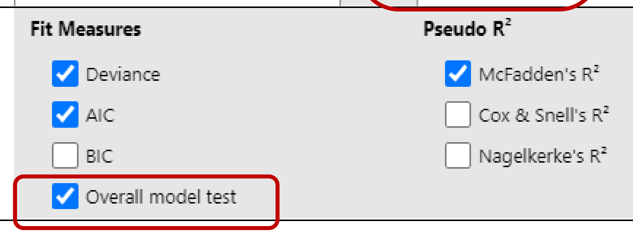
2. Select **Analyses → Regression → 2 Outcomes Binomial**



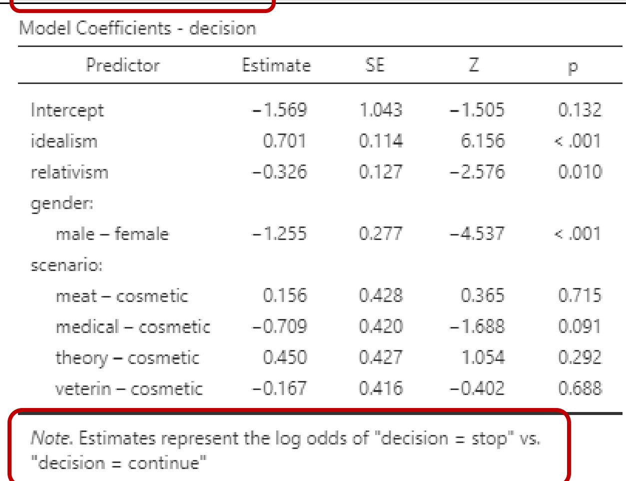
3. Pull the **decision** into the **Dependent Variable** box, the **idealism** and **relativism** into the **Covariates** box, and the **gender** and **scenario** into the **Factors** box.



4. Expand the **Model Fit** option, select the **Overall model test**.



5. The bottom of **Model Coefficients** table shows the *estimates represent the log odds of "decision = stop" vs. "decision = continue"*. (our comparisons for the decision variable is in reference to "continue") It could be changed in the **Reference Levels**.



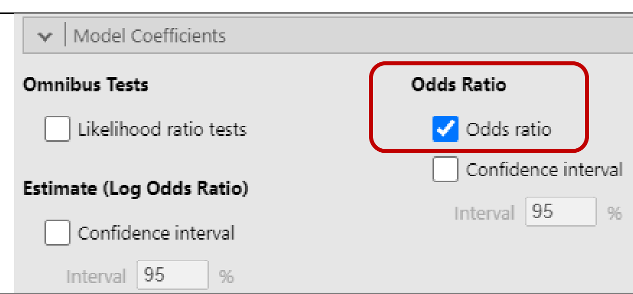
Model Coefficients - decision

Predictor	Estimate	SE	Z	p
Intercept	−1.569	1.043	−1.505	0.132
idealism	0.701	0.114	6.156	< .001
relativism	−0.326	0.127	−2.576	0.010
gender:				
male – female	−1.255	0.277	−4.537	< .001
scenario:				
meat – cosmetic	0.156	0.428	0.365	0.715
medical – cosmetic	−0.709	0.420	−1.688	0.091
theory – cosmetic	0.450	0.427	1.054	0.292
veterin – cosmetic	−0.167	0.416	−0.402	0.688

Note. Estimates represent the log odds of "decision = stop" vs. "decision = continue"

6. Expand the **Reference Levels** option. Change the **Reference Level** of **decision** to **stop**. Now we should see the note has changed to *estimates represent the log odds of "decision = continue" vs. "decision = stop"*. (our comparisons for the decision variable is in reference to "stop")

7. Expand the **Model Coefficients** option. Select **Odds ratio**.
8. The odds ratio tells *how each predictor affect the decision*. For example, for every unit increase in relativism, the odds of deciding to continue the animal research increases by a factor of 1.386.

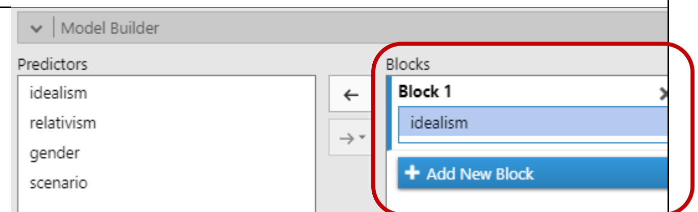


Model Coefficients

Predictor	Estimate	SE	Z	p	Odds ratio
Intercept	1.569	1.043	1.505	0.132	4.804
idealism	−0.701	0.114	−6.156	< .001	0.496
relativism	0.326	0.127	2.576	0.010	1.386
gender:					
male – female	1.255	0.277	4.537	< .001	3.508
scenario:					
meat – cosmetic	−0.156	0.428	−0.365	0.715	0.855
medical – cosmetic	0.709	0.420	1.688	0.091	2.033
theory – cosmetic	−0.450	0.427	−1.054	0.292	0.638
veterin – cosmetic	0.167	0.416	0.402	0.688	1.182

Note. Estimates represent the log odds of "decision = continue" vs. "decision = stop"

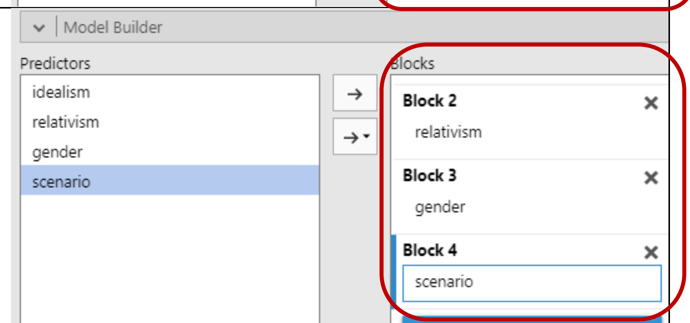
9. Expand the **Model Builder** option. Remove all the predictors except **idealism** from the Block 1.



10. Click **+ Add New Block** to create **Block 2** and pull the **relativism** into it.

11. Repeat the step to create **Block 3** and pull the **gender** into it.

12. Repeat the step to create **Block 4** and pull the **scenario** into it.



13. The **Model Fit Measures** table shows all the four models can predict the decision significantly.

14. The **Model Comparisons** table shows there is a significant difference from model 1 to 2 and model 2 to 3. In this case, model 3 is better as it gives higher R^2 value. There is no significant difference from model 3 to 4 even though the R^2 value of model 4 is the highest. So, model 3 is the best model.

Model Fit Measures

				Overall Model Test		
Model	Deviance	AIC	R^2_{McF}	χ^2	df	p
1	375	379	0.119	50.6	1	< .001
2	366	372	0.139	59.3	2	< .001
3	347	355	0.186	79.1	3	< .001
4	338	354	0.206	87.5	7	< .001

P<0.05

Model Comparisons

Comparison					
Model		Model	χ^2	df	p
1	-	2	8.71	1	0.003
2	-	3	19.77	1	< .001
3	-	4	8.44	4	0.077

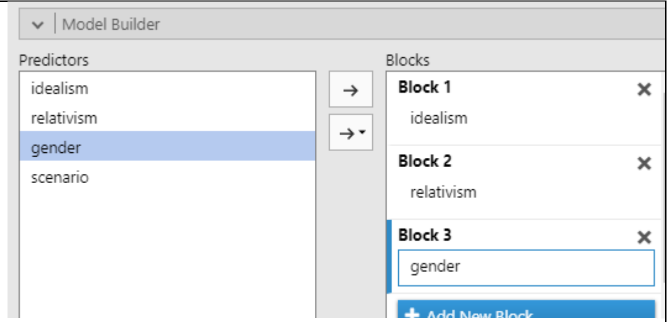
15. Remove **Block 4** from the **Blocks** under **Model Builder**. Now, only Block 1 to 3 are left in the Blocks.

16. We can produce a statistical model that allow us to predict values of the outcome variable based on the predictor variable:

    ln(p/(1-p)) = 1.488 - 0.689(idealism) + 0.343(relativism) + 1.171(gender)

    where gender = 1 for male, 0 for female

17. Odds ratio for gender is 3.225. Males were 3.225 times more likely to <u>continue</u> the animal test than females.



Model Specific Results Model 3

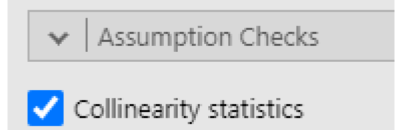
Model Coefficients

Predictor	Estimate	SE	Z	p	Odds ratio
Intercept	1.488	0.979	1.52	0.128	4.427
idealism	−0.689	0.112	−6.18	< .001	0.502
relativism	0.343	0.124	2.76	0.006	1.409
gender:					
male – female	1.171	0.268	4.37	< .001	3.225

Note. Estimates represent the log odds of "decision = continue" vs. "decision = stop"

18. To make sure the result is reliable and valid, expand **Assumption Checks** option, select **Collinearity statistics.**

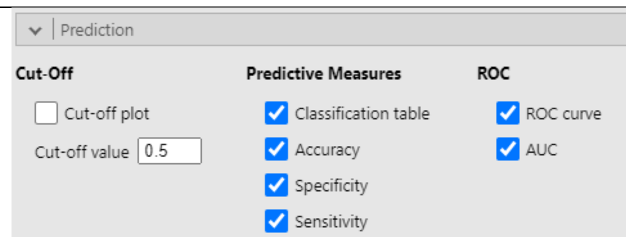
19. **Variance Inflation Factors** is used to test for multicollinearity. The <u>VIF</u> scores of the predictors are <u>below 10</u> and the <u>tolerance</u> scores are <u>above 0.2</u>. Therefore, the assumption of **no multicollinearity** has been met. (independent variables in a regression model are not correlated)



Collinearity Statistics

	VIF	Tolerance
idealism	1.02	0.976
relativism	1.02	0.980
gender	1.01	0.994

20. Expand the **Prediction** option. Under the **Predictive Measures**, select the **Classification table**, **Accuracy**, **Specificity** and **Sensitivity**. Under the **ROC**, select the **ROC curve** and **AUC.**



21. The **classification table** shows:
    - **True positive rate /sensitivity** P(correct|continue). [73/ (73+55)] = 57%
    - **True negative rate /specificity** P(correct|stop). [151/ (151+36)] =80.7%

22. The **Predictive Measures** shows the overall success rate (accuracy) is 71.1%.

Classification Table – decision

	Predicted		
Observed	stop	continue	% Correct
stop	151	36	80.7
continue	55	73	57.0

Note. The cut-off value is set to 0.5

23. **ROC curve is a plot of the values of sensitivity against one minus specificity.** A model with high discrimination ability will have high sensitivity and specificity simultaneously, leading to an ROC curve which goes close to the top left corner of the plot. A model with no discrimination ability will have an ROC curve which is the 45 degree diagonal line.

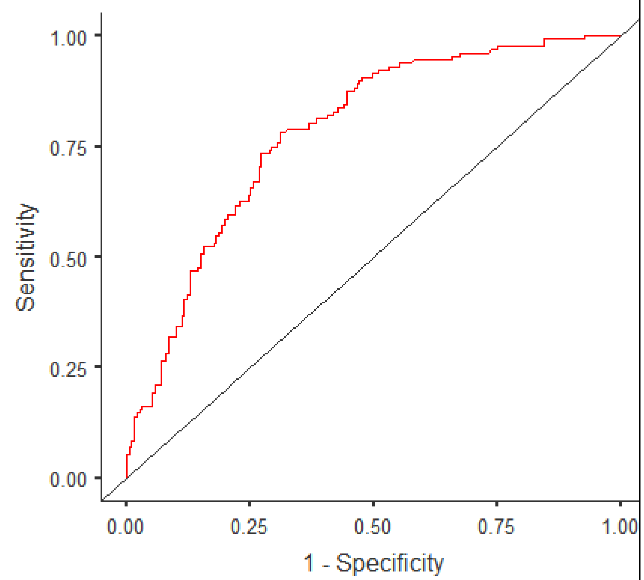
24. **AUC is the area under the ROC curve**. In this case, AUC is 0.782.

Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.711	0.807	0.570	0.782

Note. The cut-off value is set to 0.5

ROC Curve



Overall Test results:

- A binomial logistic regression was carried out to ascertain the effects of idealism, relativism and gender on the likelihood that participants decide to continue an animal test.
- The logistic regression model was statistically significant, $\chi2(3) = 79.1$, $p < 0.001$. The model explained 18.6% (McFadden's R^2) of the variance in decision and correctly classified 71.1% of cases.
- An **increase in relativism** score was associated with **an increase** in the likelihood of animal test
- An **increase in idealism** score was associated with **a decrease** in the likelihood of animal test [Model: 1.488 - 0.689(idealism) + 0.343(relativism) + 1.171(gender)]

Take home assignment

Open the file **lab2-assignment-ans.docx**, download the file **Assignment2-dataset.zip**. Complete it individually.

Submission

Submit the following files to buelearning website:

- lab2-assignment-ans.docx
- diet.omv
- weather.omv
- terminate.omv

References

[1] Area under the ROC curve – assessing discrimination in logistic regression – The Stats Geek. (2017, August 29). Retrieved from https://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/

[2] jamovi datalab.cc. (n.d.). Retrieved from https://datalab.cc/tools/jamovi

[3] Jamovi Tutorials · TysonBarrett.com. (n.d.). Retrieved from https://tysonbarrett.com/jamovi/

[4] *Logistics.sav*. (n.d.). Retrieved from http://core.ecu.edu/psyc/wuenschk/SPSS/Logistic.sav

[5] Navarro DJ and Foxcroft DR. (2019). *learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.70)*. doi:10.24384/hgc3-7p15

[6] user guide. (n.d.). Retrieved from https://www.jamovi.org/user-manual.html

[7] Using Jamovi: Correlation and Regression · TysonBarrett.com. (2018, March 28). Retrieved from https://tysonbarrett.com/jekyll/update/2018/03/28/jamovi_correlation_regression/

[8] Wuensch & Poteat, 1998. (n.d.). Retrieved from http://core.ecu.edu/psyc/wuenschk/Articles/JSB&P1998/JSB&P1998.htm

[9] ROC https://www.youtube.com/watch?v=egTNM8NSa2k

[10] Figure 1 source: https://support.minitab.com/en-us/minitab-express/1/scatterplot_linear_relationship.png

[11] Figure 2 source: https://support.minitab.com/en-us/minitab-express/1/scatterplot_cubic_relationship.png