

Statistical Test

Key terms [3]:

Null hypothesis (H_0) is a statement about the population & sample data used to decide whether to reject that statement or not. Typically, the statement is that there is no difference between groups or association between variables.

Alternative hypothesis (H_1) is often the research question and varies depending on whether the test is one or two tailed.

Significance level: The probability of rejecting the null hypothesis when it is true, (also known as a type 1 error). This is decided by the individual but is normally set at 5% (0.05) which means that there is a 1 in 20 chance of rejecting the null hypothesis when it is true.

Test statistic is a value calculated from a sample to decide whether to accept or reject the null hypothesis (H_0) and varies between tests. The test statistic compares differences between the samples or between observed and expected values when the null hypothesis is true.

***p*-value:** the probability of obtaining a test statistic at least as extreme as ours if the null is true and there really is no difference or association in the population of interest. The *p*-values are calculated using different probability distributions depending on the test. A significant result is when the *p*-value is less than the chosen level of significance (usually 0.05).

The normality test

Hypothesis

[<http://webspace.ship.edu/pgmarr/Geo441/Lectures/Lec%205%20-%20Normality%20Testing.pdf>]:

H_0 : The sample data is not significantly different than a normal population.

H_1 : The sample data is significantly different than a normal population.

Testing for the normality
[<http://webspace.ship.edu/pgmarr/Geo441/Lectures/Lec%205%20-%20Normality%20Testing.pdf>]:

If the p -value is less than the predefined significance level, the null hypothesis will be rejected. In other word, the data is not normally distributed if the p -value < 0.05 .

- p -value < 0.05 indicate that the data is NOT normally distributed (i.e., reject H_0).
- p -value > 0.05 indicate that the data is normally distributed (i.e., accept H_0).

The statistical tests for normality [4, 7]:

- The Shapiro–Wilk test is more appropriate method for small sample sizes (< 50 samples) although it can also be handling on larger sample size (2000 samples).
- The Kolmogorov–Smirnov test is used for $n \geq 50$.

Example of performing the normality test with SPSS

1. Click on the menu Analyze⇒Descriptive Statistics⇒Explore. The Explore dialogue box is shown in Figure 1.1.

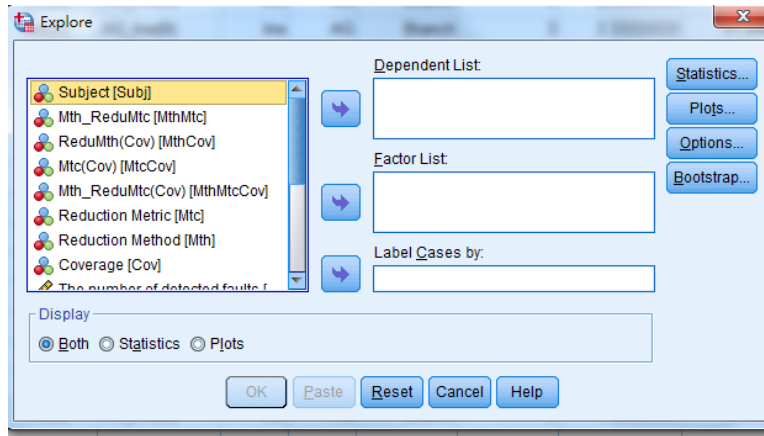


Figure 1.1 The Explore dialogue box of the normality test.

2. At the Explore dialogue box, take the variable in the left box that needs to be tested for normality to the Dependent List: box and the Factor List: box.

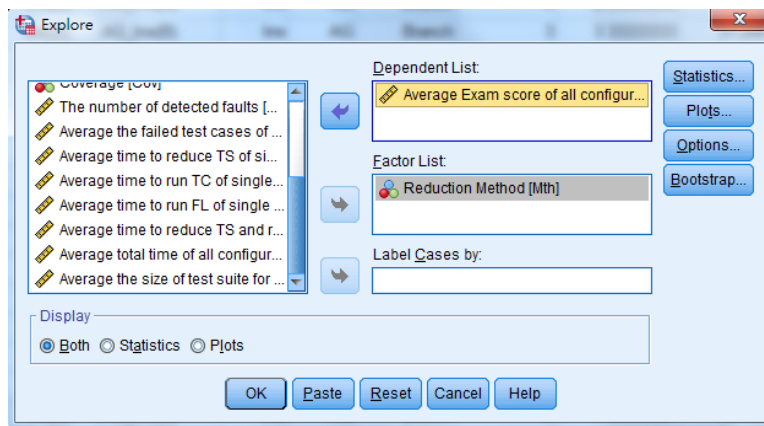


Figure 1.2 Transferring the variable into the Dependent List: box and the Factor List: box.

3. At the Explore dialogue box, click on the Statistics button. The Explore: Statistics dialogue box is shown below (keep the default options), and click the Continue button.

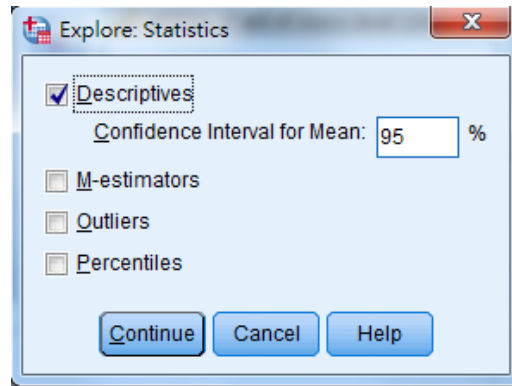


Figure 1.3

4. At the Explore dialogue box, click on the Plots: button. The Explore: Plots is shown below.

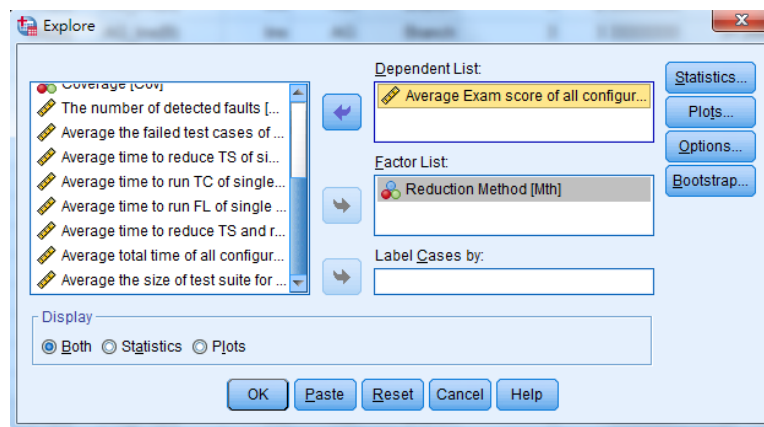


Figure 1.4

5. The Explore: Plot dialogue box, choose the options. In this example, we choose the Histogram as Descriptive and check the Normality plots with tests checkbox. Then click on the Continue button.

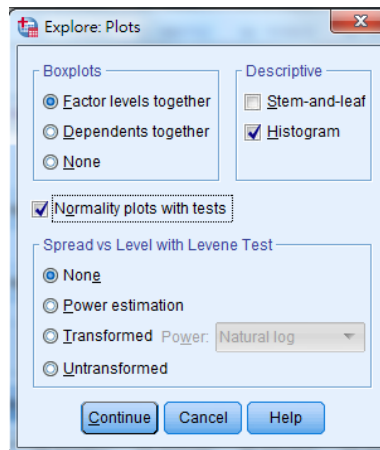


Figure 1.5

6. At the Explore dialogue box, click on the OK button.

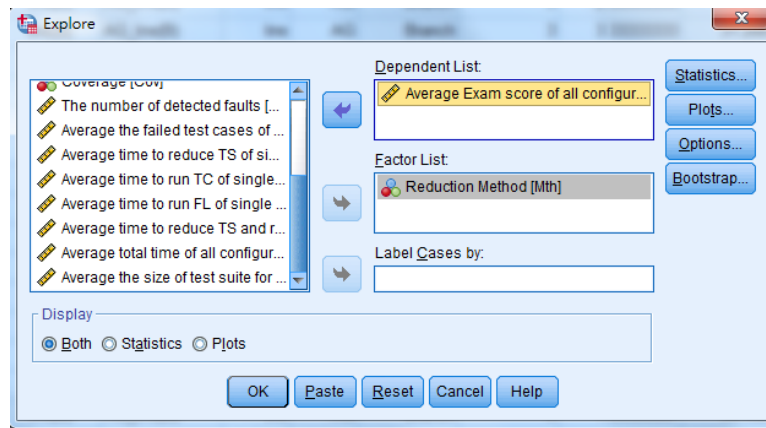


Figure 1.6

Case Processing Summary

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Average Exam score of all configurations (%)	AG	108000	100.0%	0	0.0%	108000	100.0%
	GRE	108000	100.0%	0	0.0%	108000	100.0%
	HGS	108000	100.0%	0	0.0%	108000	100.0%

Tests of Normality

		Kolmogorov-Smirnov ^a		
		Statistic	df	Sig.
Average Exam score of all configurations (%)	AG	.078	108000	.000
	GRE	.081	108000	.000
	HGS	.084	108000	.000

a. Lilliefors Significance Correction

Figure 1.7

Example output [4]

Considering the Sig. column of the Shapiro-Wilk Test, the distribution of the data for all three groups (i.e., the “Beginner”, “Intermediate” and “Advanced”) is normal (i.e., $p\text{-value} > 0.05$).

Tests of Normality							
Course		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Time	Beginner	.177	10	.200 [*]	.964	10	.827
	Intermediate	.166	10	.200 [*]	.969	10	.882
	Advanced	.151	10	.200 [*]	.965	10	.837

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Figure 1.8 The statistical test result of normality.

If the distribution of the data is normal, the parametric test will be adopted. Otherwise, use the non-parametric test.

Parametric test	Non-parametric test
One-way ANOVA	Kruskal-Wallis test
t-test	Mann-Whitney <i>U</i> test
Pearson	Spearman

Parametric test

2.1. One-way ANOVA (Analysis of variance)

The One-way ANOVA test is used to compare three or more independent groups [2].

ANOVA is based on two assumptions. Therefore, before we carry out ANOVA, we need to check that these are met:

- 1) The observations are random samples from normal distributions.
- 2) The populations have the same variance, σ^2 .

The hypotheses used are:

H_0 : all group means are equal.

H_1 : all group means are NOT equal at least one pair of group means.

When perform ANOVA:

- p -value > 0.05 indicate that all group means are equal (i.e., accept H_0).
- p -value < 0.05 indicate that there is a significant difference between at least one pair of group means (i.e., reject H_0).

If there exists significant difference, the Post-Hoc test will be further performed. The Post-Hoc test is the multiple comparison procedures are used to determine which groups are significantly different after obtaining a statistically significant result from an ANOVA [10]. Approaches for pairwise comparisons with ANOVA [11][<https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=option-one-way-anova>].

[10] <http://facstaff.uwa.edu/tdevaney/guides/posthocTests.pdf>

[11] http://web.pdx.edu/~newsomj/uvclass/ho_posthoc.pdf

The Multiple Comparisons shows which groups differed from each other. The Tukey post hoc test is generally the preferred test for conducting post hoc tests on a one-way ANOVA, but there are many others.

- LSD (Fisher's Least -Significant Difference)
- Tukey's HSD (Tukey's Honesty Significant Difference)

Example of performing the one-way ANOVA test with SPSS

In this example, we want to compare seven TSR techniques in terms of APFDc.

1. Click on the menu Analyze⇒Compare Means⇒One-Way ANOVA and the One-Way ANOVA dialogue is shown in Figure 2.1.

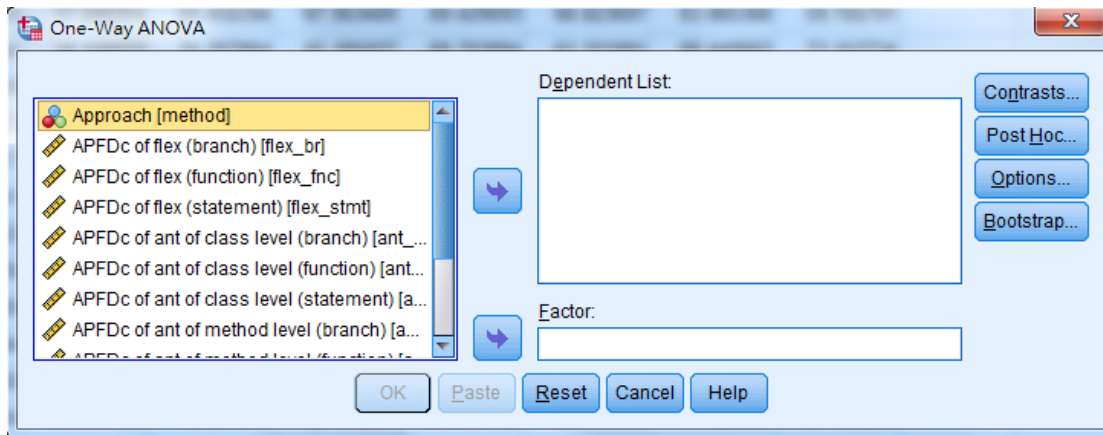


Figure 2.1 The One-Way ANOVA dialogue for the one-way ANOVA test.

2. At the One-Way ANOVA dialogue, transfer the dependent variable and the independent variable for testing to the Dependent List: box and Factors: box, respectively, and click the **OK** button.

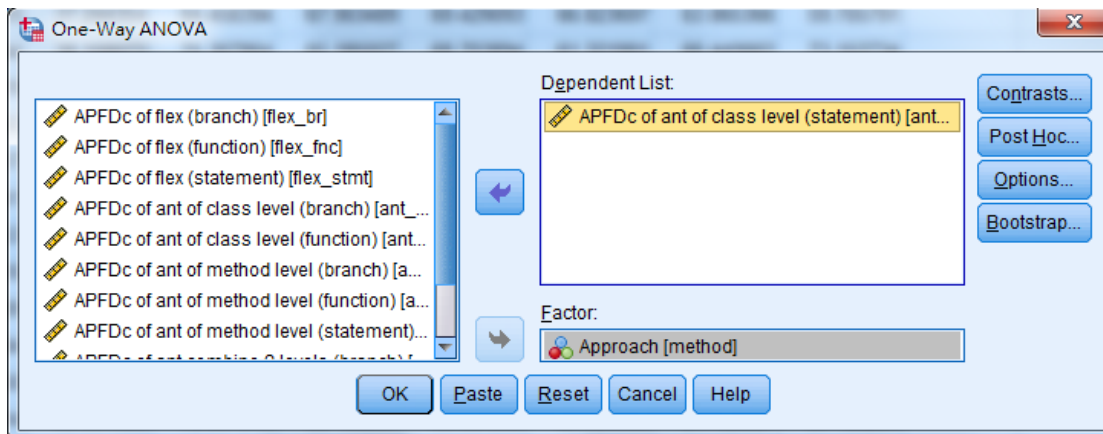


Figure 2.2 Choosing the dependent and independent variables.

3. The obtained result is depicted in Figure 2.3.

ANOVA

APFDc of ant of class level (statement)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	647773.646	6	107962.274	7217.025	.000
Within Groups	104610.996	6993	14.959		
Total	752384.642	6999			

Figure 2.3

From this figure, the p -value in the Sig. column is 0.000 which is less than 0.05. This indicates that there is a statistically significant difference in the average of APFDc between the different TSR techniques. Therefore, we further perform post hoc test.

4. At the One-Way ANOVA dialogue (i.e. Figure 2.2), click on the **Post Hoc...** button. In the One-Way ANOVA: Post Hoc Multiple Comparisons dialogue, choose Tukey (Tukey's HSD (Honesty Significant Different)) as shown in Figure 2.4 and click the **Continue** button.

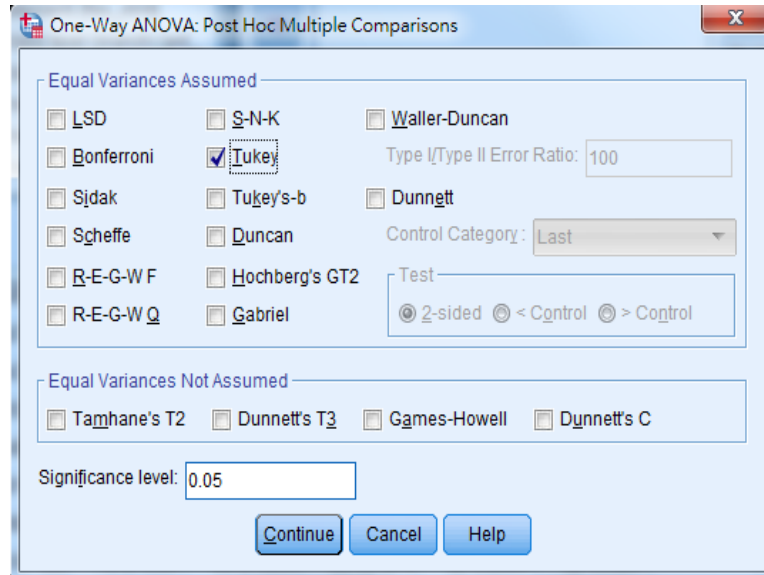


Figure 2.4 Selecting the Post-Hoc test.

5. The obtained results of Tukey HSD post hoc tests are illustrated in Figures 2.5 and 2.6.

Post Hoc Tests

Multiple Comparisons						
Dependent Variable: APFDc of ant of class level (statement)						
Tukey HSD						
(I) Approach	(J) Approach	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Ori	TG	-2.17463965*	.172970445	.000	-2.68476085	-1.66451845
	AG	-21.5679766*	.172970445	.000	-22.07809776	-21.05785536
	KP	-13.7762217*	.172970445	.000	-14.28634291	-13.26610051
	Liu	-18.4158256*	.172970445	.000	-18.92594683	-17.90570443
	TVA	-2.16588997*	.172970445	.000	-2.67601117	-1.65576877
	AVA	-25.2184910*	.172970445	.000	-25.72861224	-24.70836983
TG	Ori	2.174639651*	.172970445	.000	1.66451845	2.68476085
	AG	-19.3933369*	.172970445	.000	-19.90345811	-18.88321571
	KP	-11.6015821*	.172970445	.000	-12.11170326	-11.09146086
	Liu	-16.2411860*	.172970445	.000	-16.75130718	-15.73106477
	TVA	.008749678	.172970445	1.000	-.50137152	.51887088
	AVA	-23.0438514*	.172970445	.000	-23.55397258	-22.53373018
AG	Ori	21.5679766*	.172970445	.000	21.05785536	22.07809776

Figure 2.5

Considering Figure 2.5, each row shows the comparison between one out of seven TSR techniques and the others. In addition, the wildcard (*) in the 2nd column and the Sig. value in the 4th column indicate the difference between two groups is statistically significant. For example, at the 1st row, the Ori TSR technique's mean value is significant difference from the others (i.e., all p -values < 0.05). Considering the TG TSR technique, it's mean value is significant difference from the others except the TVA TSR technique (i.e., the p -values > 0.05).

Homogeneous Subsets

APFDc of ant of class level (statement)						
Tukey HSD ^a						
Approach	N	Subset for alpha = 0.05				
		1	2	3	4	5
Ori	1000	36.67011867				
TVA	1000		38.83600864			
TG	1000		38.84475832			
KP	1000			50.44634038		
Liu	1000				55.08594430	
AG	1000					58.23809523
AVA	1000					61.88860970
Sig.		1.000	1.000	1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 1000.000.

Figure 2.6

According to Figure 2.6, the first column shows the order of seven TSR techniques according to their mean values in ascending order. The second column indicates the number of APFDc in each TSR technique. The remaining columns identify the sets of TSR techniques (i.e., subsets 1 through

6) that are statistically significantly different from each other. If two TSR techniques appear in the same subset, then these TSR techniques are not significantly different. Additionally, a TSR technique (e.g., subset 5) that is ordered in one column is said to be statistically significantly different from TSR techniques listed in another subset. Therefore, based on the above output, AVA is the best TSR technique while Ori is the worst. In addition, TVA and TG are not significantly different from each other since they are both listed in the subset 2.

Non-parametric test

We want to compare 37 TSR techniques in terms of *EXAM* score.

3.1. The Kruskal-Wallis test or *H* test

Objective:	To assess the differences between several groups (i.e., more than two groups).
Hypotheses:	H ₀ : The median of all samples are equal. H ₁ : The median of all samples are different.

When perform ANOVA:

- $p\text{-value} > 0.05$ indicate that all group means are equal (i.e., accept H_0).
- $p\text{-value} < 0.05$ indicate that there is a significant difference between at least one pair of means (i.e., reject H_0).

Example of performing the Kruskal-Wallis test with SPSS

1. Click on the menu Analyze⇒Nonparametric Tests⇒Legacy Dialogs⇒K Independent Samples. The Tests for Several Independent Samples dialogue is depicted in Figure 2.1.

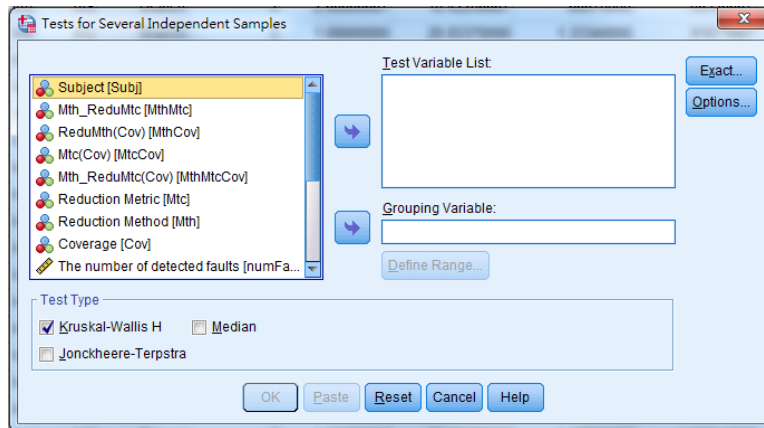


Figure 3.1

2. At the Tests for Several Independent Samples dialogue, select the tested variable (i.e., the *EXAM* score) to the Test Variable List: box. For the Grouping Variable: box, we will transfer the independent variable (i.e., TSR technique) to it. Then, click on **Define Range...** button.

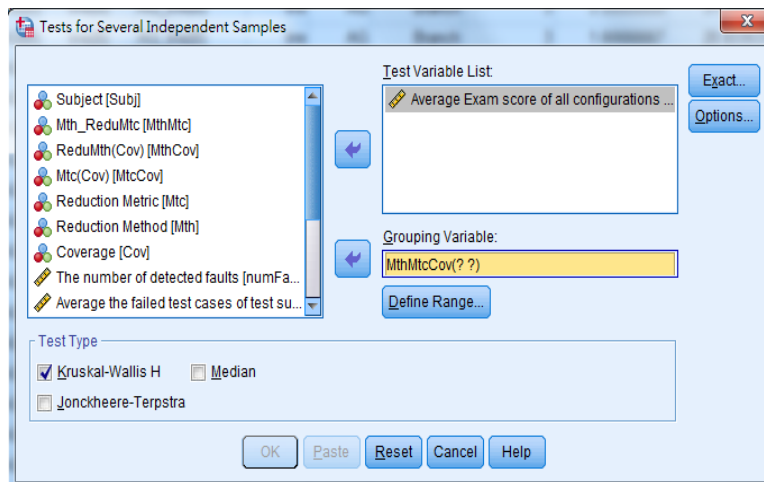


Figure 3.2

3. At the Several Independent Samples dialogue in Figure 2.3, we will define the range of variables, click on the **Continue** button, and click on the **OK** button in Figure 2.4.

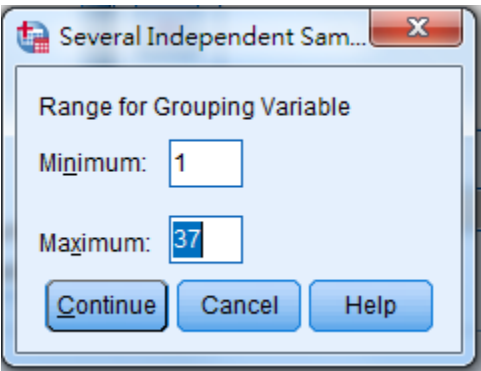


Figure 3.3 Defining the range of variable to be tested.

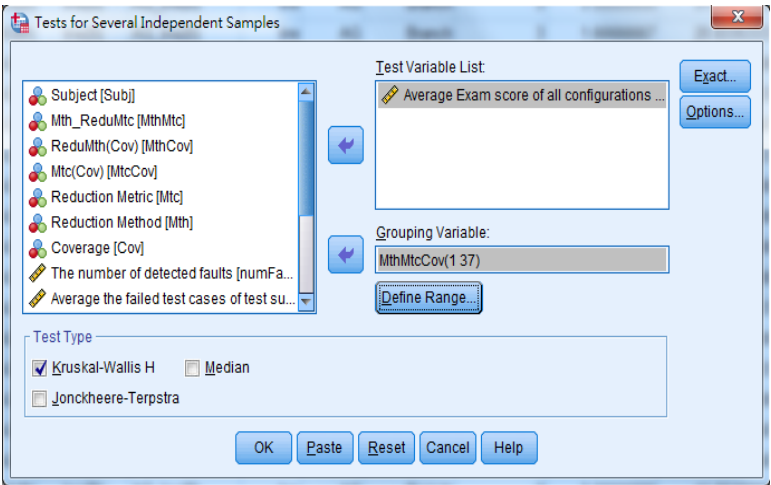


Figure 3.4 The Tests for Several Independent Samples dialogue after defined the range of variable.

4. The obtained result of the Kruskal-Wallis test is shown in Figure 2.5.

Test Statistics^{a,b}

	Average Exam score of all configurations (%)
Chi-Square	102446.663
df	36
Asymp. Sig.	.000

a. Kruskal Wallis Test
b. Grouping Variable: Mth_Redumtc(Cov)

Figure 3.5 The result of the Kruskal-Wallis test.

1. Click on the menu Analyze⇒Nonparametric Tests⇒Independent Samples

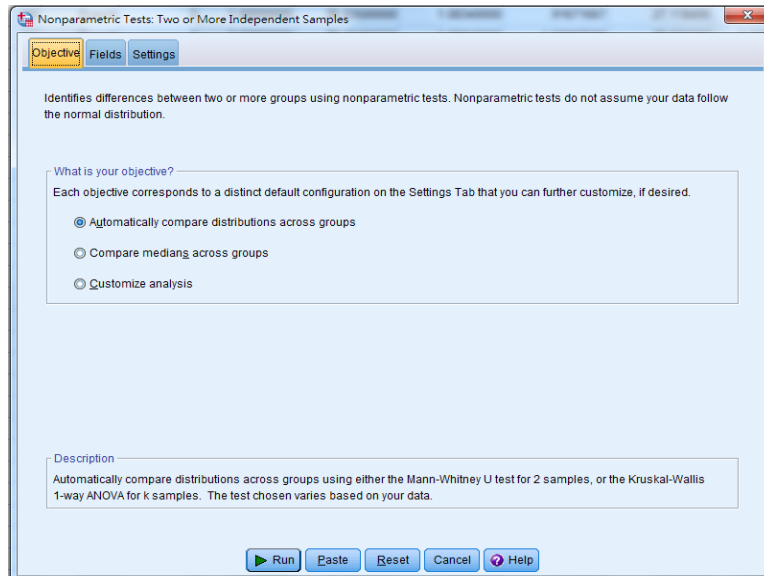



Figure 3.6

2. At the Fields tab, we transfer the tested variable (i.e., the *EXAM* score) to the Test Fields: box and the independent variable (i.e., TSR techniques) to the Groups: box. Finally, we click on the  button.

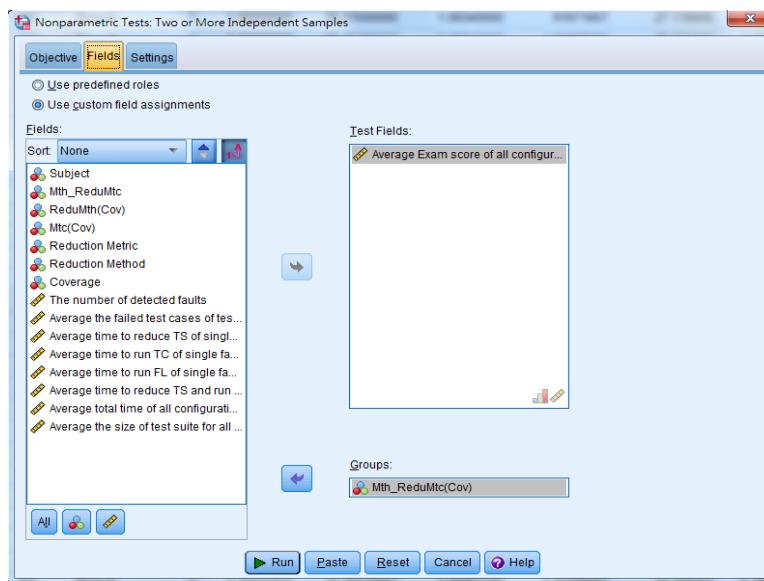


Figure 3.7

Considering the Settings tab in Figure 3.8, the statistical test will be chosen automatic based on to the data. If we would like to choose, we can select the Customize test option in Figure 3.9 to define the test by yourself

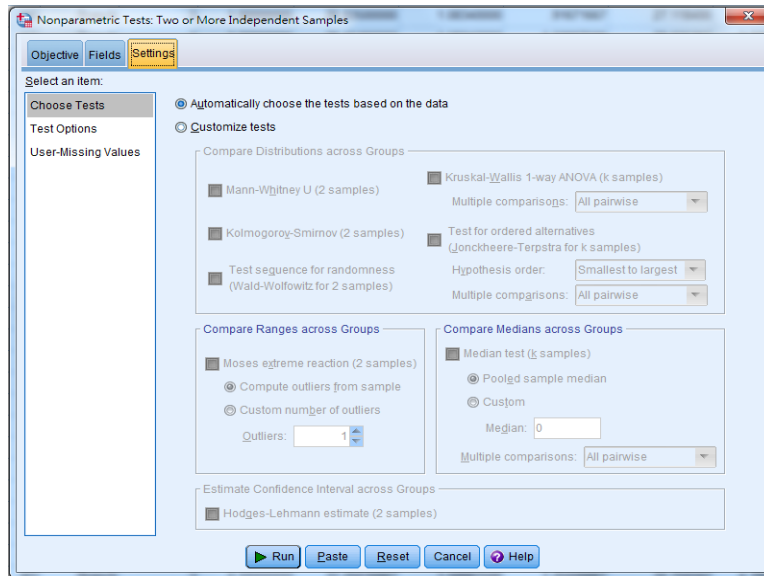


Figure 3.8

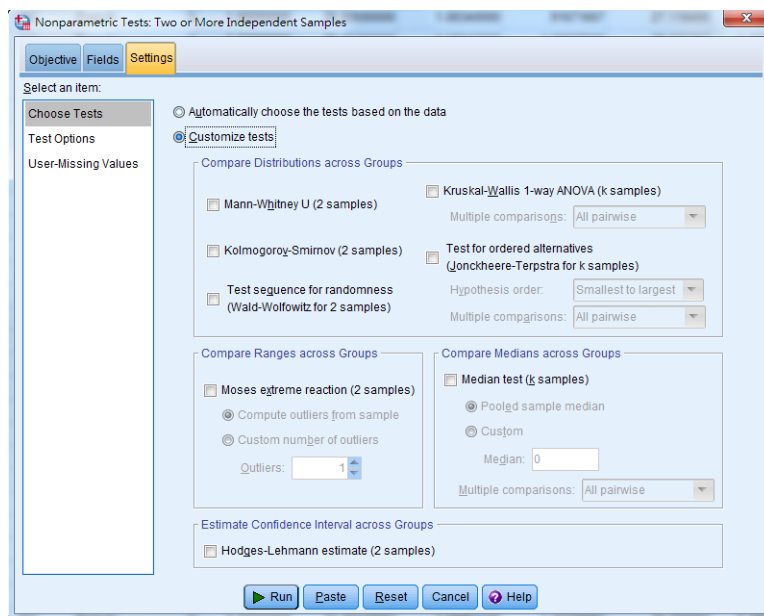


Figure 3.9 Customizing the test to perform the Kruskal-Wallis test

3. The obtained result

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Average Exam score of all configurations (%) is the same across categories of Mth_ReduCtc(Cov).	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 3.10

Double click on the table to show model viewer

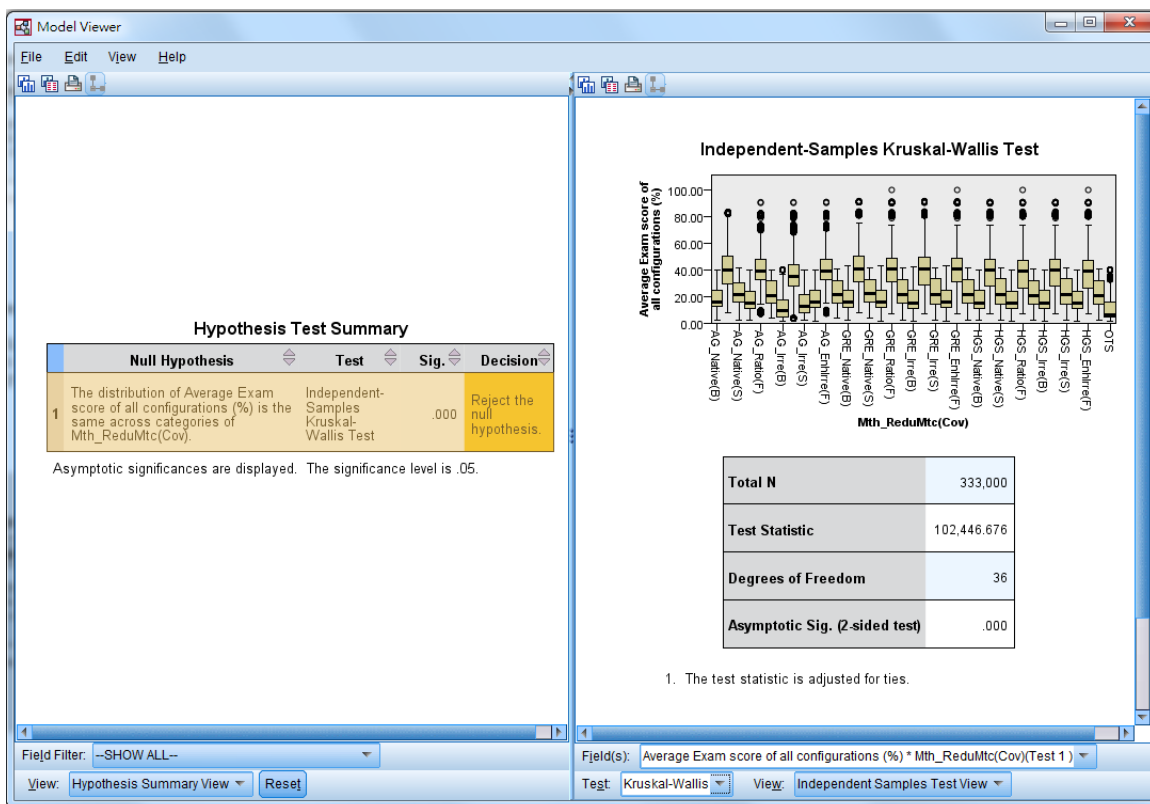


Figure 3.11

Choose View: ex. Pairwise comparison

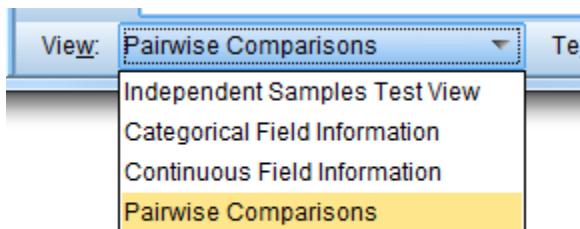


Figure 3.12
The obtained result

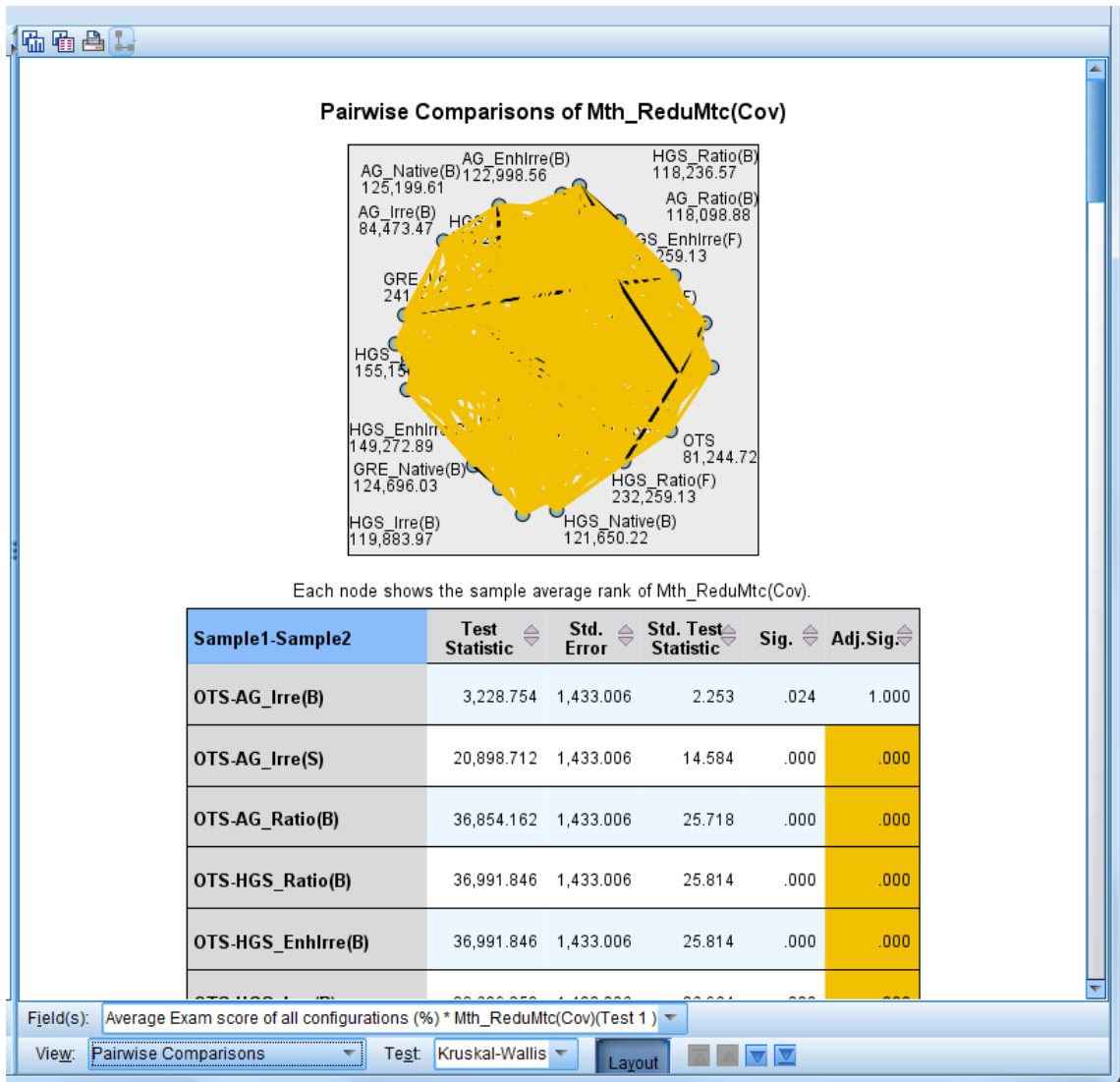


Figure 3.13

3.2. The Mann-Whitney U test

for pairwise comparisons at significant level 0.05

The Mann-Whitney U test is used to compare the difference between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed [12].

[12] <https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php>

The hypotheses used are:

H_0 : two samples come from the same population (i.e. that they both have the same median).

H_1 : two samples come from the DIFFERENT population (i.e. that they have the different median).

When perform Mann-Whitney U test:

- p -value < 0.05 indicate that two samples come from the different population (i.e., reject H_0).

Example of performing the Mann-Whitney U test with SPSS

1. Click on the menu Analyze⇒Nonparametric Tests⇒Legacy Dialogs⇒2 Independent Samples. Then, the Two-Independent-Samples Test dialogue is shown in Figure 3.14

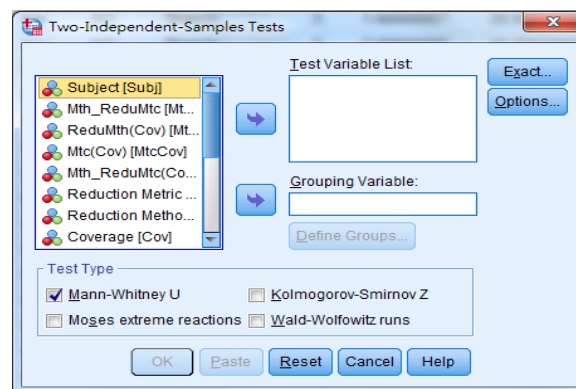


Figure 3.14 The Two-Independent-Samples Test dialogue.

2. At the Two-Independent-Samples Test dialogue, we transfer the *EXAM* score and the TSR technique to the Test Variable List: box and the Grouping Variable: box, respectively. Then, click on the **Define Groups...** button.

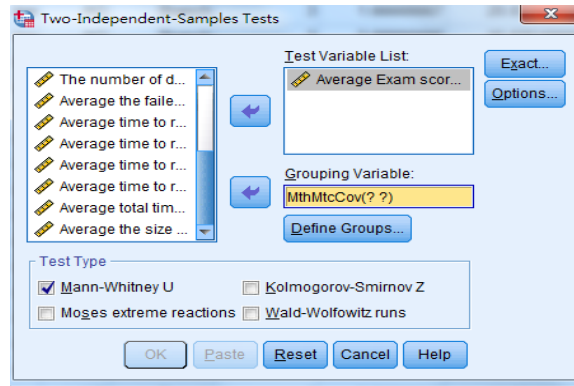


Figure 3.15 Transferring the variables to perform the test.

3. At the Two Independent Sample dialogue, for, we will define a pair of TSR techniques by identifying a group to the Group 1 box and Group 2 box, respectively, and then click on the **Continue** button. Finally, click on the **OK** button in Figure 3.17.

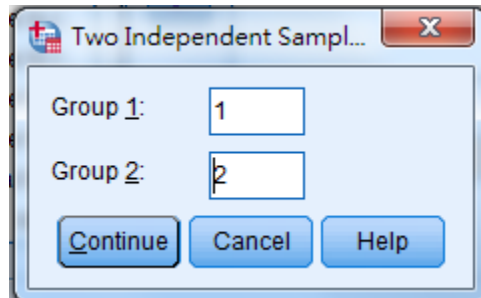


Figure 3.16 defining two groups to compare.

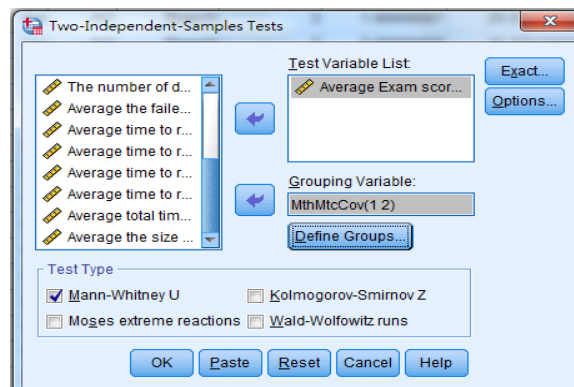


Figure 3.17

4. The Mann-Whitney U test's result is shown in Figure 3.18

Ranks				
	Mth_Redumtc(Cov)	N	Mean Rank	Sum of Ranks
Average Exam score of all configurations (%)	AG_Native(B)	9000	5960.34	53643025.00
	AG_Native(F)	9000	12040.66	108365975.0
	Total	18000		

Test Statistics ^a	
	Average Exam score of all configurations (%)
Mann-Whitney U	13138525.00
Wilcoxon W	53643025.00
Z	-.78.495
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Mth_Redumtc (Cov)

Figure 3.18 The result of Mann-Whitney U test.

3.3. The Vargha and Delaney test

- This metric measures the effect size (A_{I2}) between the values of two groups 1 and 2. More specifically,
- An effect size is a measure of the strength or magnitude of the effect of an independent variable on a dependent variable which helps assess whether a statistically significant result is meaningful.
- The value of A_{I2} ranges from 0 to 1.
- The Vargah-Delaney's effect size **Error! Reference source not found.** between groups X and Y can be computed as follows.

$$A_{I2} = (R_1/m_1 - (m_1 + 1)/2)/m_2,$$

where R_1 represents the rank sum of the Group 1 from the Mann-Whitney U test, m_1 represents the number of population in the Group 1, and m_2 represents the number of population in the Group 2

Example of performing the Vargha and Delaney test

From the output in Figure 3.18, we will compute the effect size between AG_Native(B) and AG_Native(F). Groups 1 and 2 are AG_Native(B) and AG_Native(F), respectively

$$A_{I2} = (R_1/m_1 - (m_1 + 1)/2)/m_2$$

$$A_{12} = (53643025/9000 - (9000+1)/2)/9000$$

$$A_{12} = 0.16$$

3.4. The Spearman's rank correlation coefficient (ρ)

Objective:	To determine the monotonic relationship between two variables [56, 57].
Hypotheses:	H ₀ : No relationship between two variables. H ₁ : There exists the relationship between two variables.

The result of this test is denoted by the Greek letter ρ (rho) or r_s . The value of ρ ranges between -1 (perfect negative correlation) and $+1$ (perfect positive correlation). The case of $\rho=0$ indicates that no correlation exists between these two variables. Additionally, the absolute value of ρ (i.e., $|\rho|$) can be used to represent the strength of the correlation between the two variables. More specifically, if the absolute value of ρ approaches to 1, the relationship between two variables is very strong; on the other hand, if its absolute value is close to 0, their relationship is very weak. The strength of correlation can be determined according to Table 4.2 **Error! Reference source not found..**

Table 3.1: The interpretation for the strength of correlation coefficient.

$ \rho $	Strength of Correlation
[0.00, 0.20)	very weak or independent
[0.20, 0.40)	weak
[0.40, 0.60)	moderate
[0.60, 0.80)	strong
[0.80, 1.00]	very strong

Example of performing the Spearman test with SPSS

1. Click on the menu Analyze⇒Correlate⇒Bivariate. The Bivariate Correlations dialogue with the default setting is illustrated in Figure 3.20

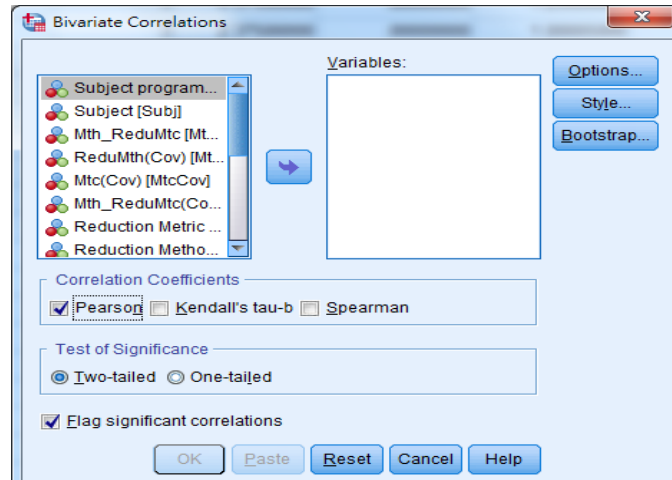


Figure 3.19 The Bivariate Correlations dialogue.

2. At the Bivariate Correlations dialogue, we choose two variables to analyze the relationship. In this example, we want to analyze the relationship between the EXAM score and the size of test suite. Thus, we transfer these two variables to the Variables: box, choose the Spearman test (i.e., these two are not normally distributed), and click on the **OK** button.

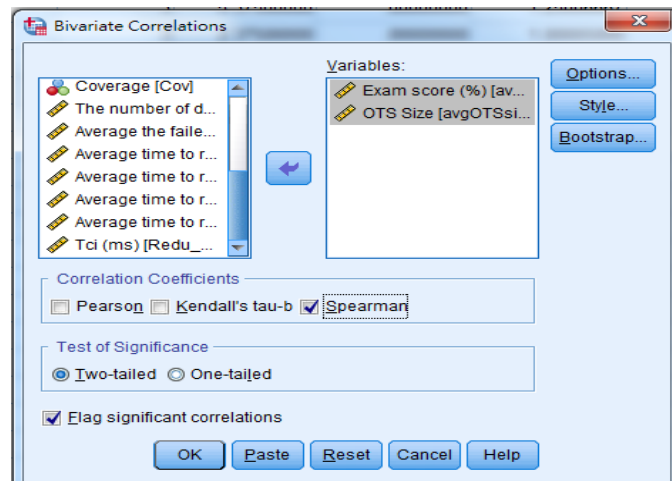


Figure 3.20 Choosing two variables to analyze the relationship.

3. The result of the Spearman test is shown in Figure 3.22.

Correlations

			Exam score (%)	OTS Size
Spearman's rho	Exam score (%)	Correlation Coefficient	1.000	-.270**
		Sig. (2-tailed)	.	.000
		N	9000	9000
	OTS Size	Correlation Coefficient	-.270**	1.000
		Sig. (2-tailed)	.000	.
		N	9000	9000

** . Correlation is significant at the 0.01 level (2-tailed).

Figure 3.21 The Spearman test's result

Reference

- [1] https://www.sheffield.ac.uk/polopoly_fs/1.579181!/file/stcp-marshallsamuels-NormalityS.pdf
- [2] https://www.lboro.ac.uk/media/media/schoolanddepartments/mlsc/downloads/1_5_OnewayANOVA.pdf
- [3] <https://www.statstutor.ac.uk/resources/uploaded/tutorsquickguidetostatistics.pdf>
- [4] <https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php>
- [5] <https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93>
- [6] <https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics-2.php>
- [7] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/>
- [8] <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics-2.php>
- [9] <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php>
- [59] A. Vargha and H. D. Delaney, "A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong," *Educational and Behavioral Statistics*, Vol. 25, No. 2, pp. 101-132, June 2000.