Chapter 9

Transformation of Data

9 Transformation of Data

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

Data is collected for every research conducted. The collected date needs to be analysed using the appropriate statistical test. Every data collected can be either normally distributed or not. Those that are normally distributed are subjected to parametric tests in their analyses while those that are not are intend subjected to non parametric test. Parametric test are said to be more reliable than the parametric test because less assumptions are made on then compared with non parametric. However whenever data is collected from an experiment or a study and it is not normally distributed but a parametric test is subjected to it, the results become misleading and therefore unreliable. In order to use a parametric test on data that is not normally distributed, it is converted or transformed into normally distributed data. This chapter explains and demonstrate how to transform data that is not normally distributed to normally distributed data in order to apply parametric tests to analyse them. The SPSS software has been used to demonstrate how to transform data in that regard.

Keywords

Normal Distribution, Parametric, Non Parametric, Statistical Tests, Data

9.1 Introduction

Data collection and analysis are an indispensable aspect of research which great skills and expertise is needed to handle. In order to make meaningful and reliable findings in a research study, the accuracy of the data and the appropriateness of the tools used in analyzing the data must be employed. Statistical tests are grouped broadly into parametric and non parametric tests. Each of these tests has their strengths and weaknesses. For benefits to be derived from the use of these tests for data analysis data, the research must adopt the use of the right tests i.e. either parametric or non parametric for the analysis.

Parametric test are said to be stronger compared to non parametric tests because the latter are based on less assumptions and are less quantitative in nature. The main assumption that parametric tests is based on is that they must be applied on normally distributed data. This chapter would not focus on the assumptions underpinning both tests but on when to use both tests.

Whenever data is collected from a research study and inferential statistics is expected to be used in the analysis of the data, such data must be tested to ascertain whether they are normally distributed or not before an appropriate parametric or non parametric test is used for their analysis. When the data is proven to be normally distributed, then the appropriate test for the analysis would be a parametric tests if not then a non parametric test should be selected for the analysis.

For data confirmed to be normally distributed, it must satisfy some requirements or conditions such as: it must have the mean, mode and median to be equal; kurtosis and skewness must range between -1.96 to 1.96; a histogram of such data must follow the shape of the normal curve; data must follow a normal Q-Q plot trend; and a box-plot of the data must be symmetrical. Any

data that does not follow or approximately satisfy these conditions cannot be said to be normally distributed.

However, when the normality of data is tested and it proves not to be normally distributed, various methods can be used to transform the data to that which is anormally distributed in order for a parametric test to be used for its analysis. It is only when the transformed data still does not conform to a normally distributed one that the researcher can go ahead to analyse it with an appropriate non parametric test. The process of converting data that is not normally distributed to that which is normally distributed is referred to as *transformation.*

This chapter uses SPSS software to demonstrate how to check whether data collected from a researcher study is normally distributed considering the conditions and requirements for normality as well as show how to transform data that is not normally distributed.

9.2 Testing the Data Normality

It should be noted that it is the data collected on the dependent variables that are tested to be either normally distributed or not. The acronym SPSS stands for Statistical Package for Social Scientist and it is an application used for data summarizing and analyzing. SPSS has two windows (views) for handling of data – Data and Variable views where the collected data can be labeled and entered. The data view is the window where data is entered whiles the variable view is where the levels of measurement, label and codes of variables whose data had been entered are specified.

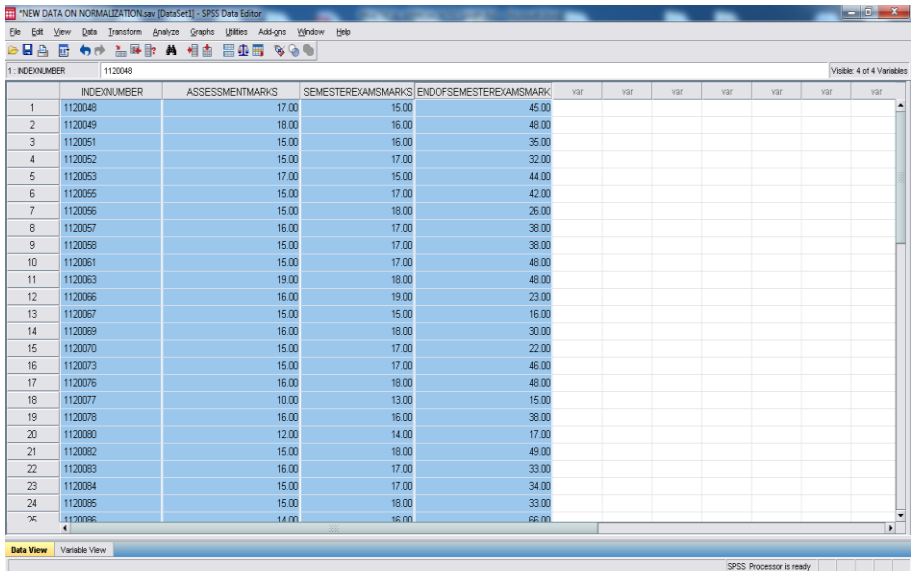


Figure 9.1 *Data view page.*

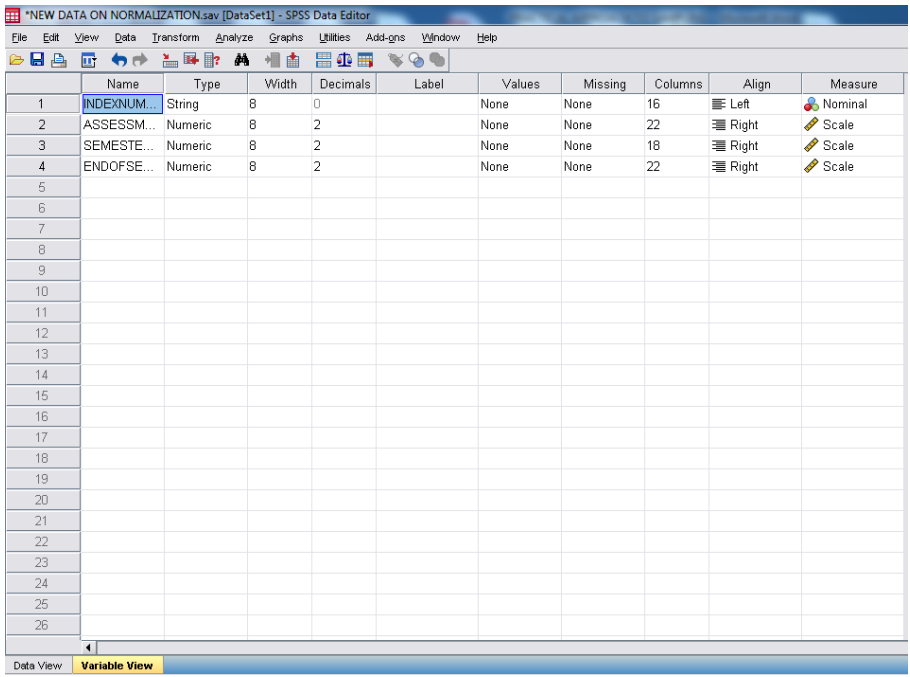


Figure 9.2 *Variable view page.*

The researcher can copy or import the data collected from Excel application into the Data view page, or better still if the data is not already typed, it can be typed directly into the cells in the Data view window as seen in Fig. 9.1. Once the data has been entered, the researcher must click on the variable view button of the page to open its window and then label appropriately the variables with their specifications – level of measurement, type of variable i.e either numeric or string etc. as shown in Figure 9.2.

Let us proceed to use the data below to demonstrate how to test for the normality of a collected data from a research study:

Table 9.1 *Data used for Illustration on the Test for Normality.*

Index No	Assessment Marks	Semester Exam Marks	End of Semester Exam Marks
1120048	17	15	45
1120049	18	16	48
1120051	15	16	35
1120052	15	17	32
1120053	17	15	44
1120055	15	17	42
1120056	15	18	26
1120057	16	17	38
1120058	15	17	38
1120061	15	17	48
1120063	19	18	48
1120066	16	19	23
1120067	15	15	16
1120069	16	18	30
1120070	15	17	22
1120073	15	17	46
1120076	16	18	48
1120077	10	13	15
1120078	16	16	38

Index No	Assessment Marks	Semester Exam Marks	End of Semester Exam Marks
1120080	12	14	17
1120082	15	18	49
1120083	16	17	33
1120084	15	17	34
1120085	15	18	33
1120086	14	16	66
1120087	19	17	28
1120090	15	18	34
1120091	16	16	44
1120094	15	18	46
1120097	17	18	53
1120098	15	17	46
1120099	15	18	26
1120103	15	18	35

After the data has been entered and label, the *analyze command button* at the top part is pressed, and cursor moved to descriptive statistics and then to *'explore'* as indicated in Figure 9.3.

The '*explore'* button is then clicked to open another window shown below where the cursor is placed on dependent variable whose normality is to be tested and then moved by clicking on the '*arrow'* to send into the dialogue box provided for the dependent list as shown in Figure 9.4. One can highlight on all the dependent variables in the box if he or she needs to test their normality and then move them all into the dialogue box provided by click on the '*arrow'*. But for this illustration let us use only the dependent variable "End of Semester Exams Mark".

Once the dependent variable whose normality is to be tested is moved into box of the dependent list, click on '*plots'* and then use the cursor to check the

boxes for ‘histogram’, ‘dependents together’ and ‘normality plots with tests’ after which one must click on the continue button as shown in Figure 9.5.

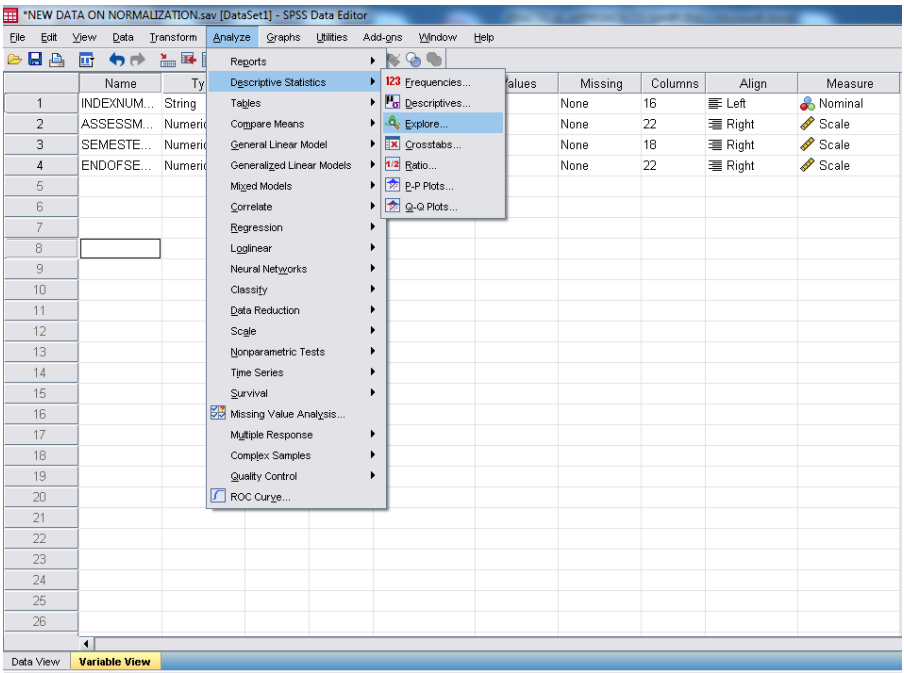
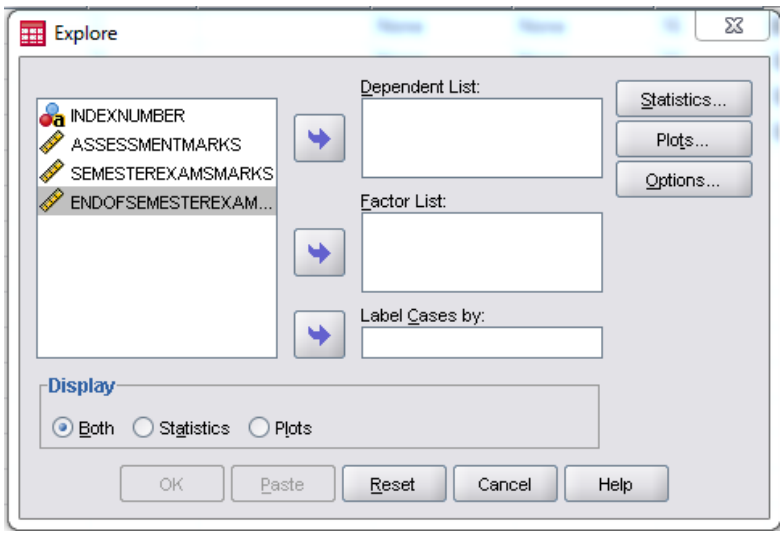


Figure 9.3 *Outcome of Clicking the Analyze Button.*



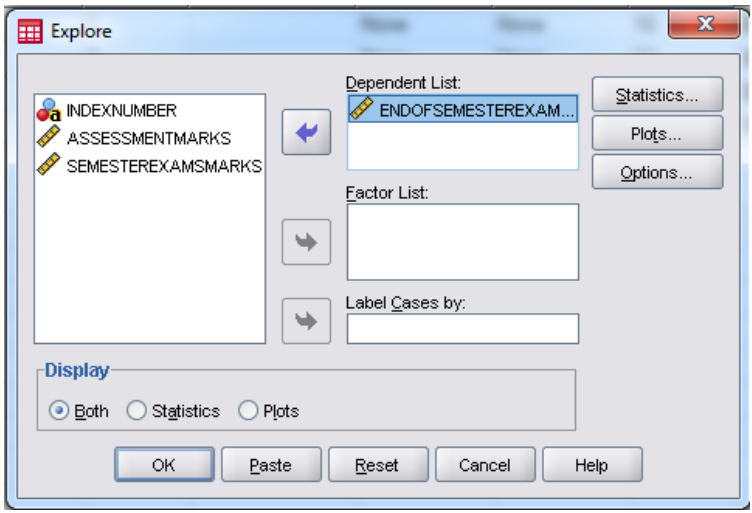


Figure 9.4 *Outcome of Clicking the Explore button.*

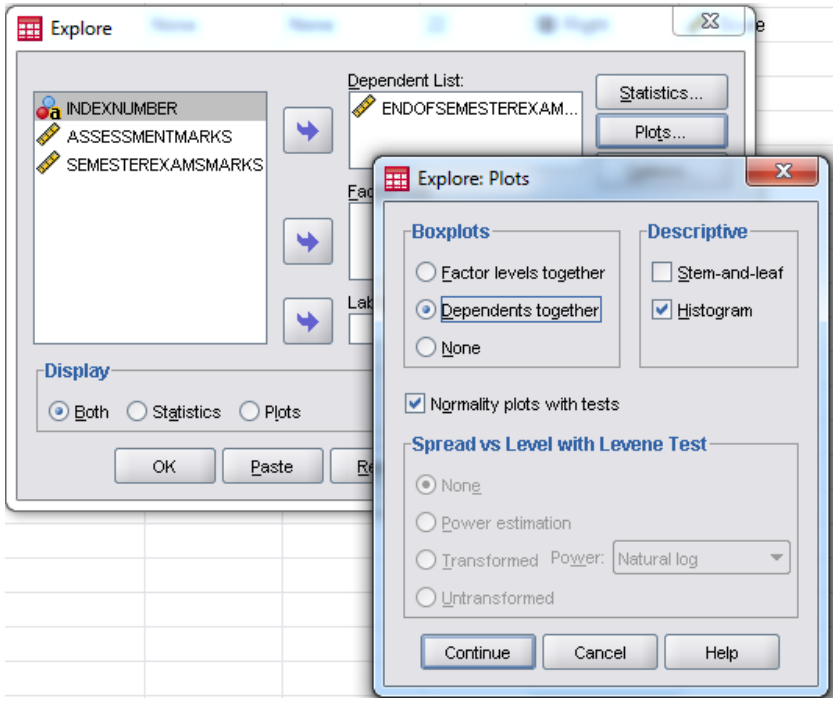


Figure 9.5 *Outcome of Clicking on the 'Plots' button.*

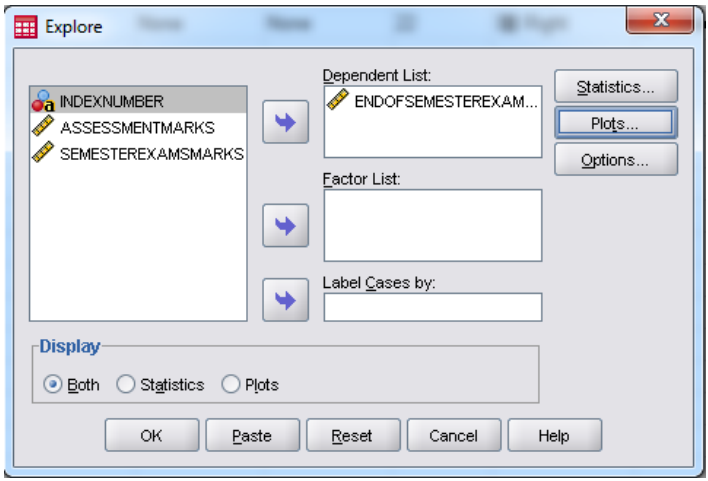


Figure 9.6 *Explore window with 'both' checked in the display box.*

By clicking on the *'continue'* button in the opened window sends one back to the explore window. The researcher can then click check *'both'* in the display box within the explore window (shown in Figure 9.5) and then proceed to click on *'Ok'* to obtain the output in an output window.

Now proceed to analyse the outcomes of the output box one-by-one juxtaposing them with the requirements for a data to be considered as normally distributed as follows:

Check the values for the mean, median and mode of the Data

Table 9.2.1 *Statistics.*

Mean		**37.1515**
Median		38.0000
Mode		48.00

From the table above obtained from the output of the SPSS application on the data being tested for, the mean and the median are not equal though close to each other. Since the output generated from the normality shown in the table

above reveals that the mean (37.15), median (38.0) and mode (48.00) of the data are not equal. Therefore the data is likely not to be normally distributed. It should be noted that the modal value of the data was obtained by clicking on the analyze – frequency – statistics -'check mode box' buttons in that logical sequence.

Compare the value of Skewness and Kurtosis for the Data

Table 9.2.2 *Descriptives.*

			Statistic	Std. Error
Endofsemesterexamsmark	Mean		37.1515	2.03153
	95% Confidence Interval for Mean	Lower Bound	33.0134	
		Upper Bound	41.2896	
	5% Trimmed Mean		37.0421	
	Median		38.0000	
	Variance		136.195	
	Std. Deviation		1.16703E1	
	Minimum		15.00	
	Maximum		66.00	
	Range		51.00	
	Interquartile Range		17.00	
	Skewness		-.021	.409
	Kurtosis		-.054	.798

For data to be considered as normally distributed, the z-scores for skewness and kurtosis must range between -1.96 to 1.96. In order to convert the skewness and kurtosis values to z-scores, the researcher must divide the kurtosis and skewness values by their respective standard error besides them as shown in the table above.

$$Z - scores\ for\ Skewness = \frac{-0.021}{0.409} = -0.05$$

$$Z - scores\ for\ Kurtosis = \frac{-0.054}{0.798} = -0.07$$

The values fall within the z-scores of -1.96 to 1.96 but though the data meets this requirement, this alone cannot indicate that the data is normally distributed or not.

The Normality test values

Table 9.2.3 *Tests of Normality.*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Endofsemesterexamsmark	.115	33	.200*	.968	33	.435
Lilliefors Significance Correction						
*. This is a lower bound of the true significance.						

For data to pass the Kolmogorov-Smirnov test and be considered as normally distributed, the p-value (indicated in the table above as sigma) must be less than the level of significance (α) which in this case is 5% (0.05). Since the p-value or sigma is 0.20 and greater than an (α) = 5% (0.05), then the data is not normally distributed. For Sharpiro-Wilk test of normality, a data can be considered as normally distributed only when the p-value (sigma) is greater than the level of significance (α) which in this case is 5% (0.05). For this data though the p-value (sigma) according to Shapiro-Wilk test (0.435) is greater than the level of significance (α) which in this case is 5% (0.05), the data is not normally distributed. It thus should be noted that the Shapiro-Wilk test is not always right and therefore cannot be used as the only conclusive test to determine the normality of data.

Examine the trend of Normal Q-Q Plot for the Data

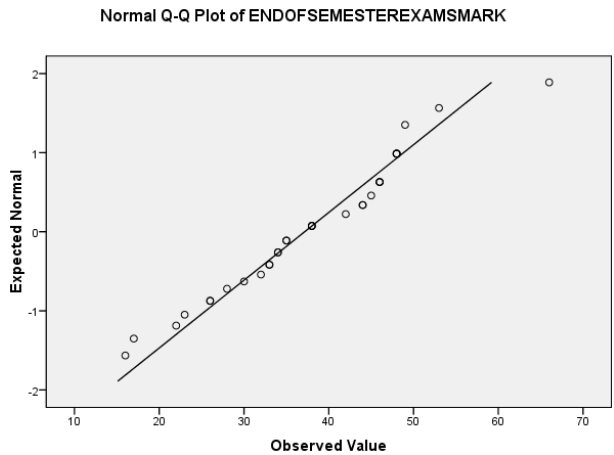


Figure 9.7.1 *Normal Q-Q Plot for the Data.*

For a normally distributed data, all or most of data points must lie exactly on or almost close to the trend line shown in a Normal Q-Q Plot. From the Normal Q-Q Plot for the data tested as shown above, most of the points do not fall on or close to the trend line and hence the data cannot be said to be normally distributed.

Examine the shape of the Histogram plot for the Data

When data is normally distributed, a drawn histogram for such data must look and have a shape of a normal curve with the data histogram formed being bilaterally symmetrical. However, the histogram for the data being tested as shown above does not show such characteristics and therefore cannot be said to be normally distributed.

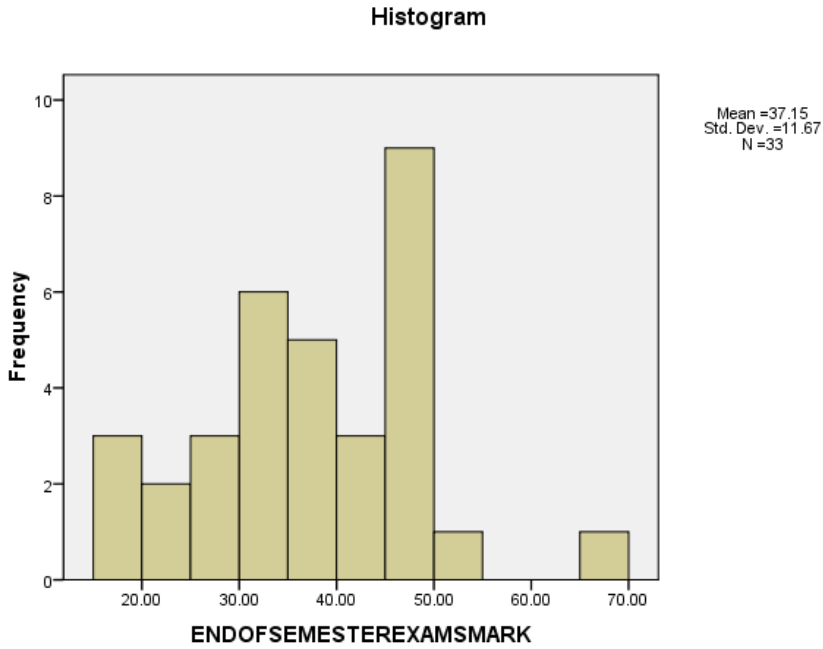


Figure 9.7.2 *Histogram for the Data.*

Examine the Box –Plot for the Data

The Box-Plot just like the histogram must show that the data have two bilateral symmetrical halves for it to be considered as normally distributed. The upper end to the middle portion or the mid section is more stretched as compared with the lower end to the mid section of the plot. This shows that the data is not normally distributed.

Thus since majority of the tests do not meet the conditions or the requirements for the data to be passed as normally distributed, the researcher can proceed to try and see whether the data can be transformed to become normally distributed before a parametric test can be used for its analysis. If all transformations applied do not succeed in converting the data to a normally

distributed one, then the researcher can select an appropriate non parametric test for its analysis.

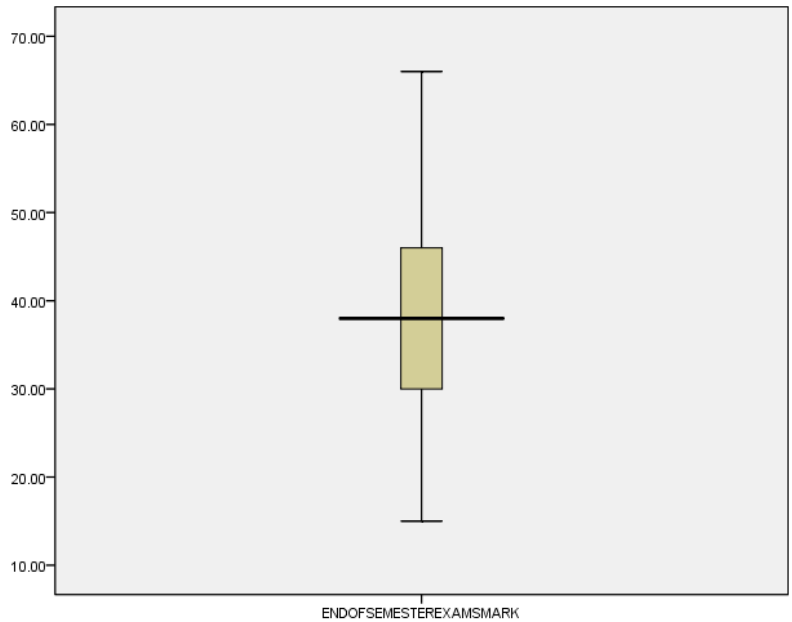


Figure 9.7.3 *Box-Plot for the Data.*

9.3 Transformation of Data

Since all the requirements for the normality test suggest that the data being considered is not normally distributed, the researcher must now try and transform the data. For one to transform the data using the SPSS application, the following described procedures must be used:

Description of Data Transformation Procedure

To transform the existing data, click on the *'transform'* button and then move to the *'compute'* variable button and click it to open a new window – compute variable window as shown in the Figure 9.8.1.

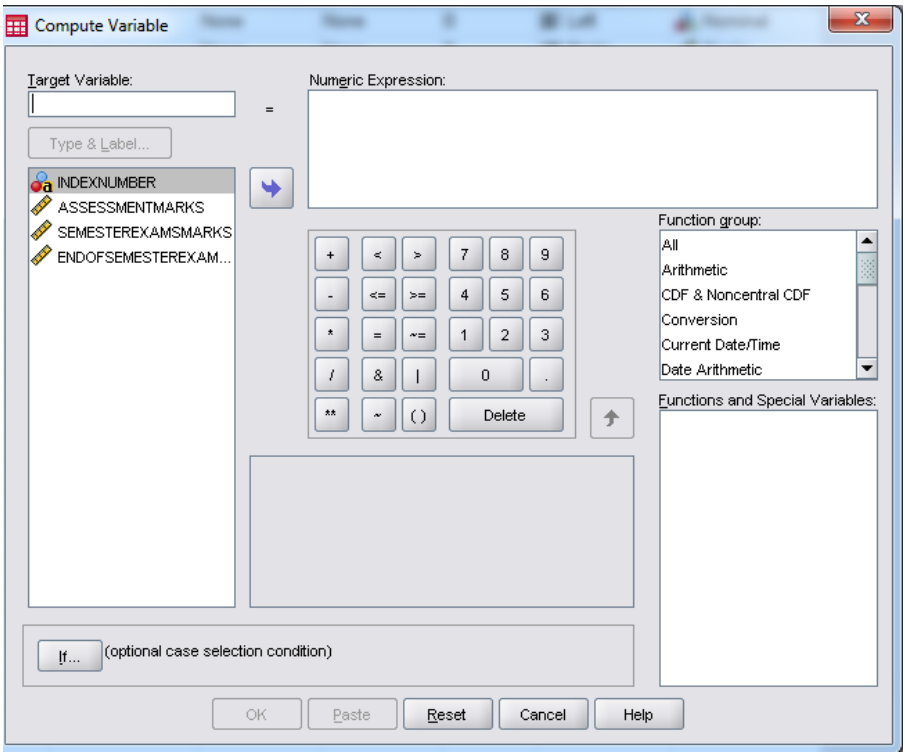


Figure 9.8.1 *Outcome of clicking on the 'compute variable' button.*

Type the variable to be transformed in the 'Target Variable' bar and then click on the 'type and label dialogue box' that appears to give the new label for the data to be transformed. One must then move the cursor to function group dialogue box and click on arithmetic, which will then enable the various 'functions and special variables box' from which one can select the function to use to perform the transformation. For this example, let us use the function 'log_{10}' as shown in Figure 9.8.2.

Click on the function 'log_{10}' and then move the cursor to the 'type and label' dialogue box and click on the target variable and then use the enabled 'arrow' to

move the target variable into the numeric expression dialogue box by clicking on it as shown in Figure 9.8.3.

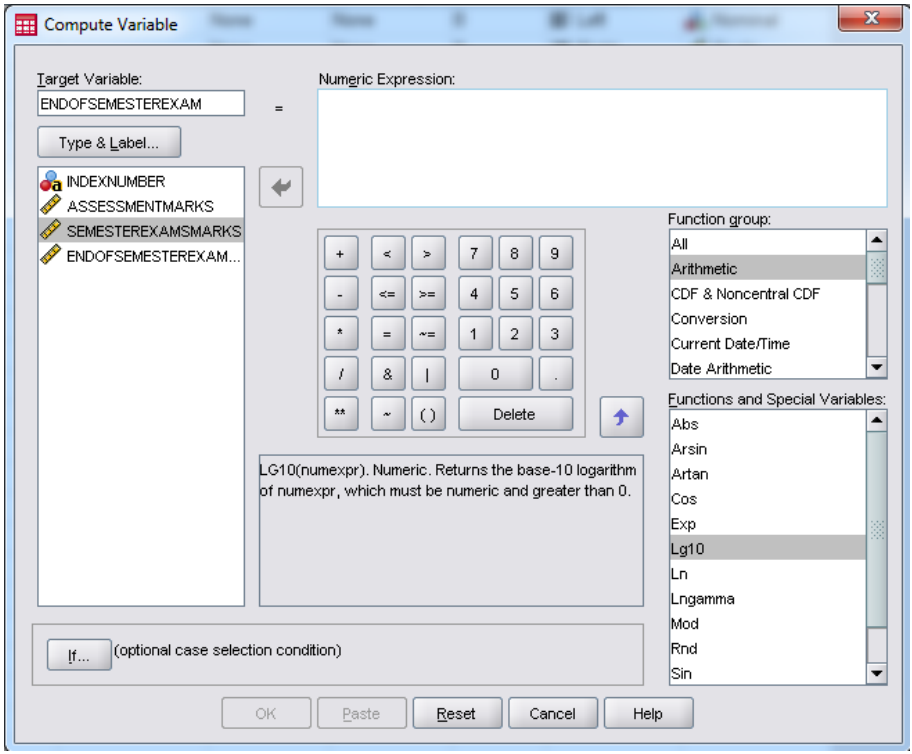


Figure 9.8.2 *Outcome of clicking the Arithmetic under the 'function group dialogue box'.*

Once this has been done the researcher can then move the cursor and click on the 'Ok' button to transform the target variable data to the indicated new label that have been assigned to it. The transformed data with the new label can be seen to be added to the other variables already in the data view window as shown in Figure 9.8.4.

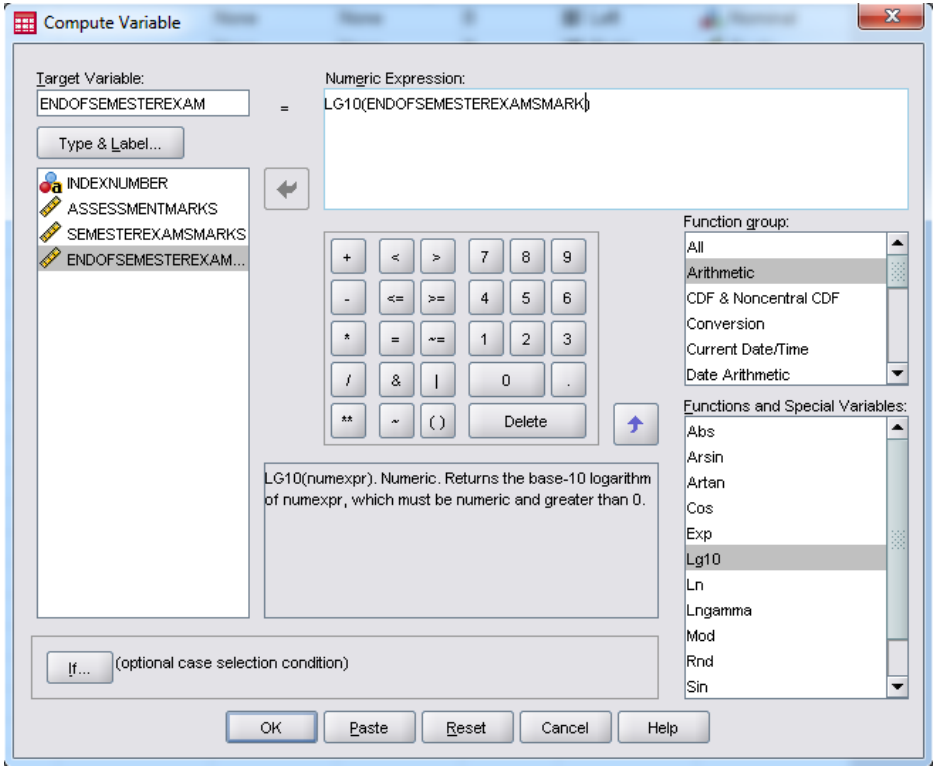


Figure 9.8.3 *Outcome of clicking on the function '$\log_{10}$' and moving the target variable into the numeric expression dialogue box.*

Now that the data on one dependent variable has been transformed, there is a need to test if the transformed variable would pass the variables normality tests and requirement as done previously. Thus needs to go through the process of testing the transformed data.

	INDEXNUMBER	ASSESSMENTMARKS	SEMESTEREXAMSMARKS	ENDOFSEMESTEREXAMSMARK	ENDOFSEMESTEREXAM	var
1	1120048	17.00	15.00	45.00	1.65	
2	1120049	18.00	16.00	48.00	1.68	
3	1120051	15.00	16.00	35.00	1.54	
4	1120052	15.00	17.00	32.00	1.51	
5	1120053	17.00	15.00	44.00	1.64	
6	1120055	15.00	17.00	42.00	1.62	
7	1120056	15.00	18.00	26.00	1.41	
8	1120057	16.00	17.00	38.00	1.58	
9	1120058	15.00	17.00	38.00	1.58	
10	1120061	15.00	17.00	48.00	1.68	
11	1120063	19.00	18.00	48.00	1.68	
12	1120066	16.00	19.00	23.00	1.36	
13	1120067	15.00	15.00	16.00	1.20	
14	1120069	16.00	18.00	30.00	1.48	
15	1120070	15.00	17.00	22.00	1.34	
16	1120073	15.00	17.00	46.00	1.66	
17	1120076	16.00	18.00	48.00	1.68	
18	1120077	10.00	13.00	15.00	1.18	
19	1120078	16.00	16.00	38.00	1.58	
20	1120080	12.00	14.00	17.00	1.23	
21	1120082	15.00	18.00	49.00	1.69	
22	1120083	16.00	17.00	33.00	1.52	
23	1120084	15.00	17.00	34.00	1.53	

Data View | Variable View

Figure 9.8.4 *Transformed data shown in the Data view window.*

Testing to check if the transformed data passed the normality tests and requirements:

Click on the analyse button, move the cursor to descriptive and then to explore. Click on the explore button to open the explore window. Place the cursor on the label of the transformed data and then use the enabled arrow to move it into the dependent list dialogue box. Continue by clicking on the ‘plots’ button in the window and then check the histogram and the normality tests boxes. Proceed by clicking the ‘Continue’ button to return to the explore window. In the explore window, check ‘both’ in the display dialogue box and

then click on ‘Ok’ for the results or the outputs of the normality tests. Now let us compare the results and check whether the transformation has succeeded in converting the data of the transformed variable to become normally distributed.

Check the values for the mean, median and standard deviation of the Data

Table 9.3.1 *Statistics.*

Mean		**1.5458**
Median		1.5798
Mode		1.68

Since the output generated from the normality shown in the table above reveals that the mean (1.55≈2), median (1.58≈2) and mode (1.68≈2) for the data are not equal. This shows that is not normally distributed.

Compare the value of Skewness and Kurtosis for the Data

Table 9.3.2 *Descriptives.*

			Statistic	**Std. Error**
ENDOFSEMESTEREXAMS2	Mean		1.5458	.02687
	95% Confidence Interval for Mean	Lower Bound	1.4911	
		Upper Bound	1.6006	
	5% Trimmed Mean		1.5526	
	Median		1.5798	
	Variance		.024	
	Std. Deviation		.15436	
	Minimum		1.18	
	Maximum		1.82	
	Range		.64	
	Interquartile Range		.20	
	Skewness		-.857	.409
	Kurtosis		.356	.798

For data to be considered as normally distributed, the z-scores for skewness and kurtosis must range between -1.96 to 1.96.

$$Z - scores\ for\ Skewness = \frac{-0.857}{0.409} = -2.09$$

$$Z - scores\ for\ Kurtosis = \frac{0.356}{0.798} = 0.45$$

The skewness falls outside the range whiles the kurtosis Z-score value values fall within the range. Since skewness fell within the range and the kurtosis outside it, it means the data is not normally distributed. They should all fall within the range before they can be said to be normally distributed.

The Normality test values

Table 9.3.3 *Tests of Normality.*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ENDOFSEMESTEREXAMS2	.130	33	.167	.925	33	.026
a. Lilliefors Significance Correction						

For data to pass the Kolmogorov-Smirnov test, since the p-value or sigma is 0.167 and greater than $\alpha = 5\%$ (0.05), then the data is not normally distributed. For Sharpiro-Wilk test of normality, since the p-value (sigma) = 0.026 is less than the level of significance $(\alpha) = 0.05$, the data is not normally distributed.

Examine the trend of Normal Q-Q Plot for the Data

For a normally distributed data, all or most of data points must lie exactly on or almost close to the trend line shown in a Normal Q-Q Plot. From the Normal Q-Q Plot for the data tested as shown above, most of the points do not fall on or close to the trend line and hence the data cannot be said to be normally distributed.

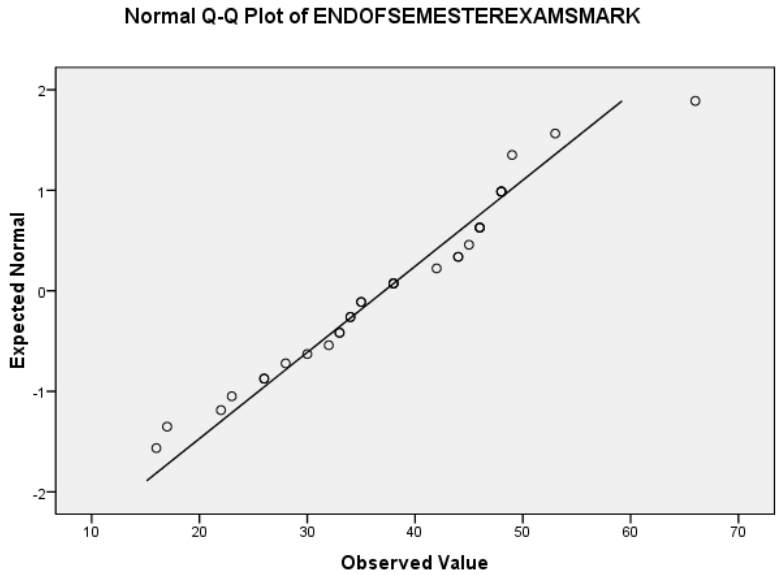


Figure 9.8.5 *Normal Q-Q Plot for the Data.*

Examine the shape of the Histogram plot for the Data

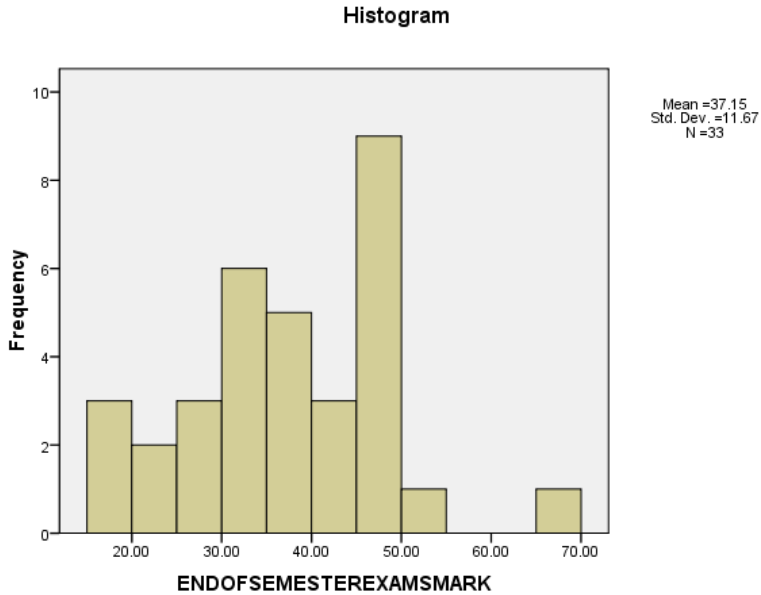


Figure 9.8.6 *Histogram for the Data.*

When a data is normally distributed, a drawn histogram for such data must look and have a shape of a normal curve with the data histogram formed being bilaterally symmetrical. However, the histogram for the data being tested as shown above does not show such characteristics and therefore cannot be said to be normally distributed.

Examine the Box –Plot for the Data

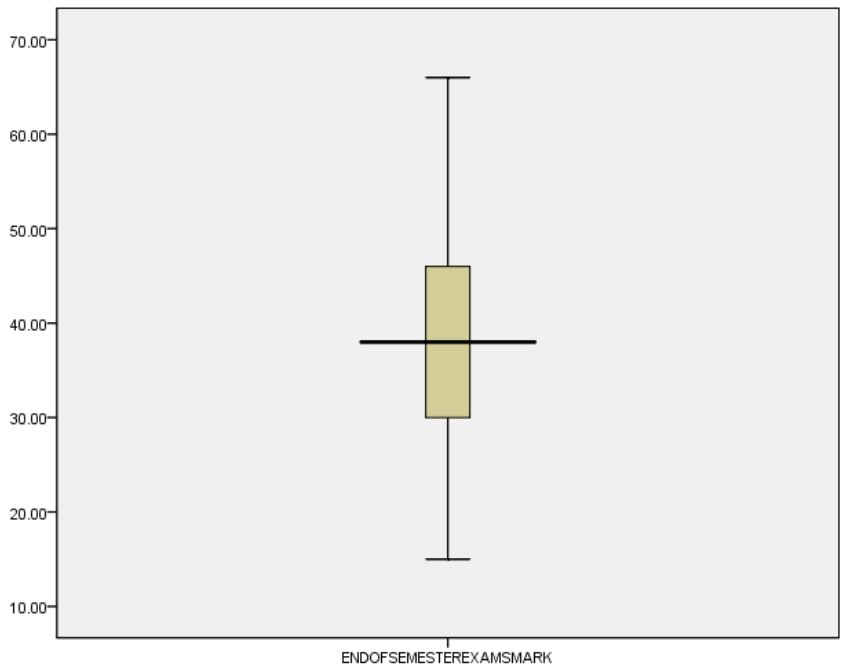


Figure 9.8.7 *Box-Plot for the Data.*

The Box-Plot just like the histogram must have two symmetrical halves for a data to be considered normally distributed. Since the histogram for the data does not show this characteristic, the data is not normally distributed.

If all transformations applied do not succeed, then the researcher can proceed to select an appropriate non parametric test to analyse the data.

Bibliography

[1] Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* 26:211-252.

[2] Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165.

[3] Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

[4] Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

[5] John, J. A., Draper, N. R. (1980). An alternative family transformations. *Applied Statistics* 29:190-197.

[6] Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

[7] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[8] Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics* 28:602-632.

[9] Swinscow, T. D., Campbell, M. J. (2003). Statistics at square one. 10th ed. New Delhi: Viva Books Private limited.

[10] Whittaker, J., Whitehead, J., Somers, M. (2005). The neglogtransformation and quantile regression for the analysis of a large credit scoring database. *Applied Statistics* 54:863-878.

[11] Yeo, I., Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87:954-959.