BOX-COX TRANSFORMATION APPROACH FOR DATA NORMALIZATION: A STUDY OF NEW PRODUCT DEVELOPMENT IN MANUFACTURING SECTOR OF PAKISTAN

Fozia Malik
Ph.D Scholar, Quaid-i-Azam School of Management Sciences
Quaid-i-Azam University Islamabad

AjmalWaheed Khan, Ph.D
Professor, Quaid-i-Azam School of Management Sciences
Quaid-i-Azam University Islamabad

Muhammad Tahir Ali Shah
AllamaIqbal Open University Islamabad

ABSTRACT

The aim of this paper is the application of Box-Cox transformation approach for data normalization. It is mostly noticed that in social science research discipline the data is not normally distributed which can cause various problems for researchers. These problems are related to decisions which statistical tools should apply in case of non-normality of data. A data set using two independent variables; (i) internal resources, (ii) external resources, one mediating variable which is new product development process and one dependent variable namely new product success from manufacturing sector of Pakistan is utilized to analyze normality of data through the Shapiro-Wilk statistics. When it was analyzed that data is not normal than box-cox transformation approach was employed. It was noticed that applying after box-cox transformation data was normal which can be utilized for further statistical analysis. Therefore, this paper contributes in suggesting statistical technique, for example, Box-Cox Transformation approach (Box & Cox, 1964) can be used for normalizing data. The research scholars can gain insight from this research regarding the procedure of Box-Cox Transformation approach.

Keywords: *Box-Cox Transformation approach, Data Normalization, Shapiro-Wilk Test,New Product Development, Manufacturing Sector*

Jel Classification: D13, E01, E02, E05

1. Fozia Malik : fmalik1980@gmail.com
2. AjmalWaheed Khan, Ph.D : awkhan2@yahoo.com
3. Muhammad Tahir Ali Shah : Tahirshah49@gmail.com

ILMA-JBS is published by the Ilma University – Formerly IBT
Main Ibrahim Hydri Road, Korangi Creek, Karachi-75190, Pakistan

1. INTRODUCTION

One of the major problems faced in social and behavioral science research is normality of data. It is noticed that data sets are seldom normal which resulted in inappropriate solutions and problematic conclusions while carrying out different statistical tests. If data is not normal it can lead towards inefficient estimates or sample covariance and there are outliers in a sample which can cause biased results. These biased as well as inefficient parameter estimates and incorrect test statistics can create problems of inaccurate model evaluation. Therefore, there is a need of normal data to get accurate results. Data normalization is related to the normal curve which means a bell shaped curve as presented in Figure 1.

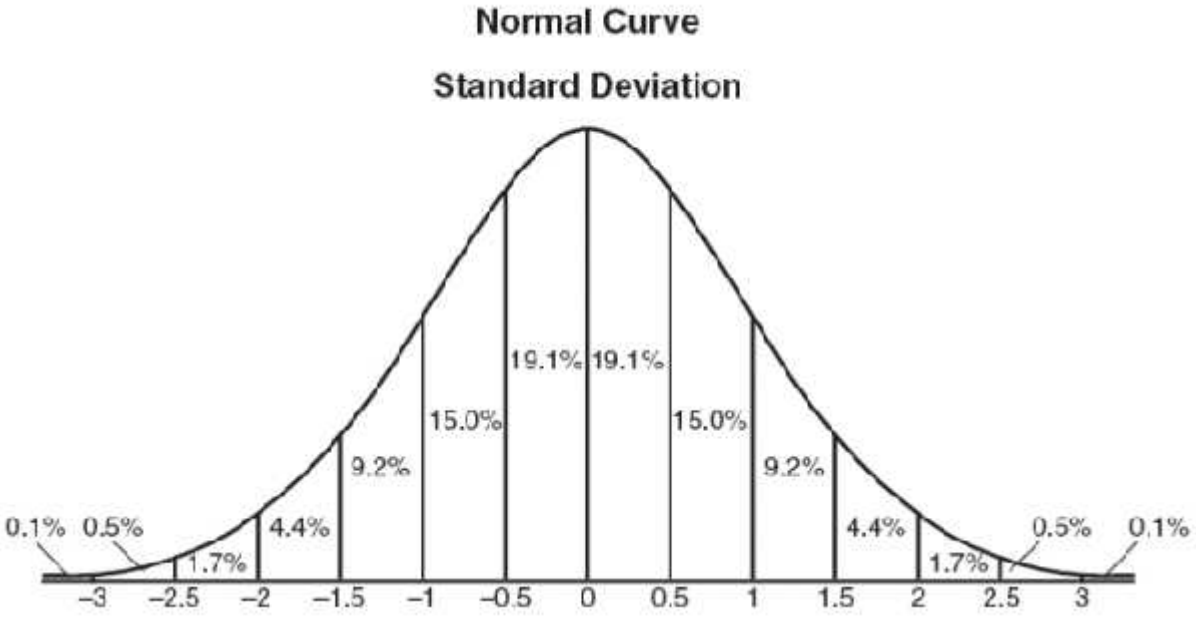


Figure 1 Normal Distribution

As discussed earlier that most of the data in social science research is not normally distrusted and due to it most curves are not perfect normal curves. These curves are skewed because it indicates the distribution of scores to one extreme end as majority of scores lie to one location either to positive side or negative side (Jackson, 2006). The positive and negative ends do not indicate that the skew is good or a bad but it only indicates that if more scores are towards positive direction then it is positively skewed and vice versa.

The Figure 2 below is a presenting a normal curve, compared to negatively and positively skewed curves.

 ISSN 2414-8393

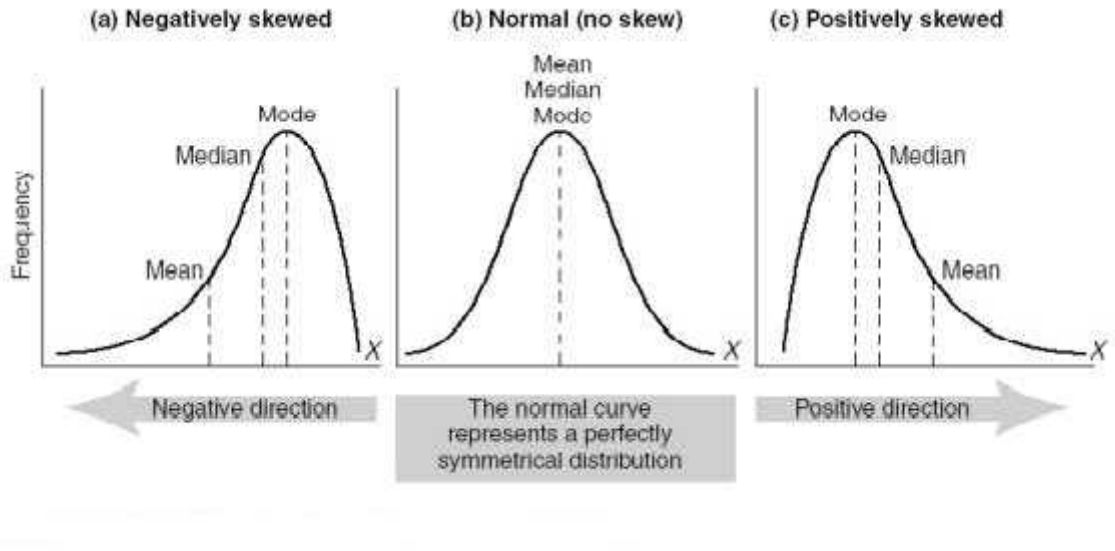


Figure 2(a) Negatively Skewed, (b) Normal Distribution, (c) Positively Skewed

However, the problem for social science researchers arise when these curves are badly skewed to any of above mentioned directions and data is not normal. These normality issues can resulted into inefficient and inaccurate results. Therefore, there is a need to check the data normality and when it becomes clear that data is not normal then it should assess what statistical techniques should apply for the normality of data. The famous tests for assessing normality of data include histogram and box plot, normal quintile plot also called normal probability plot, goodness of fit tests like Shapiro-wilk test, Kolmogorov-smirnov test and Anderson-darling test etc. Most of the researchers apply many transformation techniques but few of them report the data transformation or cleaning in their research (Osborne, 2008). However, utilizing such transformation can provide the benefits of meeting assumptions of various statistical analyses, reducing the chances of committing either a Type 1 or Type II error improving generalizability of the results, improving effect sizes etc. This is fortunate for social science researchers to have such statistical data cleaning or transformation techniques like Box-Cox technique for improving the results of analysis (Zimmerman, 1994-1995-1998). Keeping in mind these benefits in this research box-cox transformation technique is applied for data normalization. The study variables include internal resources, external resources as independent variables, new product development process as a mediating variable and new product success as dependent variable. Internal and external resources act as success drivers for any organization and utilizing these resources can help in creating proficient new product development process which can lead towards new product success.

2. LITERATURE REVIEW

Normality of data refers to the bell shape curve with normal distribution of mean, mode or median scores. Normal distribution of this curve can help in getting reliable, efficient and accurate results. As suggested by various researchers that normal distribution of data can be resulted in efficient parameter estimates. But the real problem with social science research is that data is seldom normally distributed. If it is not normally distributed than the statistical tools applied for statistical analysis can be questionable. For example, if data is normally distrusted researchers can use t-test and ANOVA tests for comparison of mean based on demographic characteristics and study variables. If data is not normally distributed than other statistical methods, for example, for comparing means based on demographic characteristics and study variables non-parametric tests such as Mann-Whitney U test and

 ISSN 2414-8393

Kruskul-Wallis can be applied as compare to parametric tests which include t-test and ANOVA testas suggested by Norusis (2008); Pallant (2013); and Field (2013). However, it is also assumed that parametric tests are more powerful as compare to non-parametric tests. Keeping in mind such situation, researchers can also apply different data transformations techniques. These techniques include (i) linear transformations; (ii) log transformation; (iii)tukey's ladder powers and (iv) box-cox transformation which is most advance technique for data transformation.

2.1 Linear Transformation

To normalize the data linear transformation technique can be used through this a transformation can be done by multiplying it with a constant and after that adding a second constant. For example, the equation for transformation is that if Y is the transformed value of X, then Y = aX + b. However, linear transformation technique normally does not change statistics like Pearson's r.

2.2 Log Transformation

This includes various log transformation techniques include and are useful to reduce skew such as (i) logarithms, it is assumed that growth rates are often exponential and in this case log transforms can be applied to normalize them. If in case that variances increase with the mean than most particularly log transforms can be appropriate technique; (ii) reciprocal, in that situation when log transforms are failed to normalize the data, researchers can apply reciprocal (1/x) transformation, this is frequently applied for enzyme reaction rate data; (iii) square root, researchers can use square root transform in that case where data is in counts e.g., blood cells on a haemocytometer or woodlice in a garden, applying a square root transform can convert data with a Poisson distribution to a normal distribution; and (iv) arcsine, this transformation is also called angular transformation and is particularly suitable for percent ages and proportions which are not normally distributed.

The following guidelines (see Table 1) should be used when transforming dataas suggested by Tabachnick and Fidell (2007) and Howell (2007).

Table 1
Situations and Guidelines for Data Transformation

If your data distribution is;	**Than use this Transformation;**
Moderately positive skewness	Square-Root • NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) • NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) • NEWX = LG10(X + C)
Moderately negative skewness	Square-Root • NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) • NEWX = LG10(K – X)

Source: Tabachnick and Fidell (2007) and Howell (2007). Where; C = a constant added to each score so that the smallest score is 1. K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

 ISSN 2414-8393

2.3 Tukey's Ladder Power Transformation

Another technique for data transformation is Tukey's ladders power introduced by Tukey (1977). This data transformation technique describes that by using a power transformation an orderly way of re-expressing variables can get. For example, if a transformation for x of the type x^{λ}, resulted in an effectively linear probability plot, then one should consider changing measurement scale for the rest of the statistical analysis. There is not any constraint considered on the values of λ because if choosing λ = 1 then it leaves the data unchanged. It is also noticed that negative values of λ are also reasonable. It is suggested by Tukey (1977) it is more convenient to simply define the transformation when λ = 0 to be the logarithmic function rather than the constant 1.

In this technique a typical ladder is sometimes included. This technique describes that goes up the ladder (positive λ) to remove left skewness and goes down the ladder (negative λ) to remove right skewness. The Table 2 describes the Tukey's ladder power transformation in which it is explained that for making sense of transformation special care must be taken when x takes on negative values. To avoid these considerations, researchers must limit themselves to variables where x>0. For some dependent variables according to the number of errors, before applying the transformation it is considered to add 1 to x. Also the transformedvariable xλ is reversed when the transformation parameter λ is negative. And in order to preserve the order of the variable after transformation, researchers choose to redefine the Tukey transformation to be -(xλ) if λ < 0.

Table 2
Tukey's Ladder Powers Transformation

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{1}{X^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	Log x	$\sqrt{x}$	x	X^2

Source:Tukey (1977)

Table 3 is the reproduction of table 2 and is representing the modified Tukey's ladder of transformation when λ < 0.

Table 3
Modified Tukey's Ladder of Transformations

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{-1}{X^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	Log x	$\sqrt{x}$	x	X^2

Source:Tukey (1977)

2.4 Box-Cox Transformation

This is the most advanced transformation technique introduced by two statisticians named as George Box and Sir David Cox and known as Box-Cox transformation (Box& Cox, 1964). They developed a procedure to identify an appropriate exponent (Lambda=l) for transforming data into a normal shape. In this data transformation technique the value of lambda defines

 ISSN 2414-8393

the power on which all data is raised and Box-Cox power transformation searches from lambda -5 to +5 until the best value is found (see Table 4).

Table 4

Common Box-Cox Transformations

L	Y'
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(Sqrt(Y))$
0	log(Y)
0.5	$Y^{0.5} = Sqrt(Y)$
1	$Y^1 = Y$
2	Y^2

Source: Box and Cox (1964). Where, 'Y' is the transformation of the of original data*Y*
Note that for Lambda = 0, the transformation is NOT Y^0 (because this would be 1 for every
value) but instead the logarithm of *Y*

The major benefit of Box-Cox (Box & Cox, 1964; Sakia, 1992) is that it provides a family of transformations rather than one particular transformation as discussed in other traditional transformation techniques. These families of transformations which include various traditional transformations are;

- λ = 1.00 (No Transformation Needed; Produces Results Identical to Original Data)
- λ = 0.50 (Square Root Transformation)
- λ = 0.33 (Cube Root Transformation)
- λ = 0.25 (Fourth Root Transformation)
- λ = 0.00 (Natural Log Transformation)
- λ = -0.50 (Reciprocal Square Root Transformation)
- λ = -1.00 (Reciprocal or Inverse Transformation and So Forth)

Box-Cox transformation technique include different families of transformations as discussed above, therefore, it can optimally normalize the required variable of study and to get the best option it also eliminates the need to randomly try different transformations (Osborne, 2010).Another usefulness of Box-Cox transformation technique is that it eliminates skewness and other distributional features that can complicate the analysis. The basic purpose of this technique is to find a simple way of transformation that leads to normality of data. The normal data lead to a straight line on the q-q plot of normality. This can maximize the correlation coefficients if the scatter diagram is linear. So, the Box-Cox transformation can normalize the data and correcting normality, linearity, and homoscedasticity.

3. METHODOLOGY

This research employs two types of probability sampling techniques for selection of manufacturing firms which include; (i) a stratified random sampling plan and (ii) simple random sampling technique for selection of large scale manufacturing firms of Pakistan.
Firstly a stratified sampling technique is used for this research which is also recommended by Akgün, Keskin, and Byrne (2010-2012) and sample firms from the directory of ministry of industries & production, Pakistan Stock Exchange and Securities & Exchange Commission of Pakistan (SECP) and industrial development board were selected on the basis of following eligibility criteria;

- Strata of local and multinational firms is identified

 ISSN 2414-8393

- In last three years (2010 to 2013) new products are launched in the market
- Firms which are involved in export of products
- Large scale manufacturing firms were identified, small and mediums firms were eliminated

According to the stratified random sampling underpinning the research; a sampling frame of 500 companies were collected from directory of ministry of industries & production, Pakistan Stock Exchange, SECP and Industrial Development Board. On identified firms simple random sampling technique was applied. From the random numbers table of Walpole (1990), the firms were selected based on simple random sampling procedure. The local and multinational firms were arranged according to years of their establishment and number of employees. Total 50 firms were selected out of 500 firms which constituted 10 percent of the total identified population. Since the sampled population was a three digits number, a group of three digits was read from left to right, starting from first row and first column of the random number's table of Walpole (1990). The number that was less than and equal to 500 was selected and repeated numbers were also skipped. Amongst the selected 50 companies, in each company 9 questionnaires were distributed based on purposive non-probability sampling technique. The reason for choosing purposive sampling technique was that there is no such published record found to determine the number of managers involved in new product development decisions. It is also suggested by researchers (Godambe, 1982) that to produce a powerful way of sampling; both random and purposive sampling can be combined. Therefore, purposive sampling can provide reliable and robust data. Hence, based on purposive sampling technique from 50 selected companies 9 questionnaires at their head offices were distributed in person.Keeping in view this, 450 questionnaires were distributed among 50 selected large scale manufacturing companies operating in Pakistan, 380 questionnaires were received back and 52 questionnaires were not completely filled by respondents, so the sample size for this research was 328. The response rate was 72% which is quite reasonable (see Table 5).

Table 5:Summary of Response Rate on Questionnaire and Variable Measurement

Questions	**Respondents**	**Population Size**	**Sample Size**	**No. of Responses**	**% of Responses**
101 Questions on Likert Five-Point Scale	Manufacturing Firms	500	50	50	100%
	Managers	Not Available	450 (9 from each company)	328	72%

4. RESULTS AND DISCUSSION

It is important to measure normality of the data before performing any statistical test for identifying that data is normally distributed or not. In this research Shapiro-Wilk test of normality is utilized.This test was carried out on SPSS. The procedure for Shapiro-wilk test is given below;

1. On the top menu Select Analyze> Descriptive Statistics > Explore
2. In the dialogue box, move the variables to be tested to the Dependent List box.
3. Click Plots and select Normality plots with tests. Click Continue.
4. Click OK to run the tests.

 ISSN 2414-8393

As discussed earlier that normality of data plays a significant role in different statistical analysis but in social science research it is very rare that data is normal. In many cases, for example, in carrying out structural equation modeling (SEM) the most important assumption is the normality of data which needs to be fulfilled. For testing normality the most authentic technique suggested by researchers is the Shapiro-wilk test. This test is used in this study and before performing transformation technique it is noticed that data is not normal as presented in Table 6 below because in this as shown in this Table that before Box-Cox transformation the p value of all study variables is significant and this is showing that data is not normal in this situation. For fulfilling normality assumption for various applying various statistical tests researchers (for example, Box & Cox, 1964; Yuan &Bentler, 1998; Yuan, Chan, &Bentler, 2000) have suggested Box-Cox transformation technique which is considered most advanced and reliable data transformation technique for normalizing the data.

The procedure for Box-Cox transformation is given but before applying this technique; we need to have mean and standard deviation values of our study variables. On SPSS Box-Cox transformation can be applied in two steps as given below;

Step1:
Transform >Rank Cases (Enter Variable e.g., IR) > Click on Rank Type > Check on Fractional Rank > ok

Step 2:

1. Transform >Compute variable
2. Function Group > click Inverse DF
3. Function and Special Variables >Click ldf Normal
4. In Target variable > e.g., normalIR
5. In numeric expression (fractional rank variable, mean value of IR, standard deviation of IR)

It is noticed that before Box-Cox transformation data was not normal as all values of variables are less than 0.05 significance level, however, after Box-Cox transformation the p-values are above 0.05 which shows the normal distribution of data as depicted in Table 5 and also see Figure 3 to 6.

Table 6
Shapiro-Wilk Test of Normality

Variables	Before Box-Cox Transformation			After Box-Cox Transformation		
	Statistic	df	Sig.	Statistc	df	Sig.
Internal Resource	.961	328	.000	.998	328	.992
External Resource	.969	328	.000	.995	328	.425
New Product Development Process	.930	328	.000	.998	328	.990
New Product Success	.957	328	.000	.997	328	.852

The normality plots as presented in Figure 3 is representing the first independent variable which is IR (internal resource) and the normal data leads to a straight line on the q-q plot of normality. The second independent variable of study was ER (external resource) as shown in Figure 4 that all data is on the straight line of q-q plot of normality. The Figure 5 is showing the mediating variable of the study i.e. NPDP (new product development process), data is normal in this case also after applying Box-Cox transformation technique. In Figure 5

 ISSN 2414-8393

of current research, normality plot of the dependent variable NPS (New Product Success) is showing a normal data after applying Box-Cox transformation technique.

5. NORMALITY PLOTS

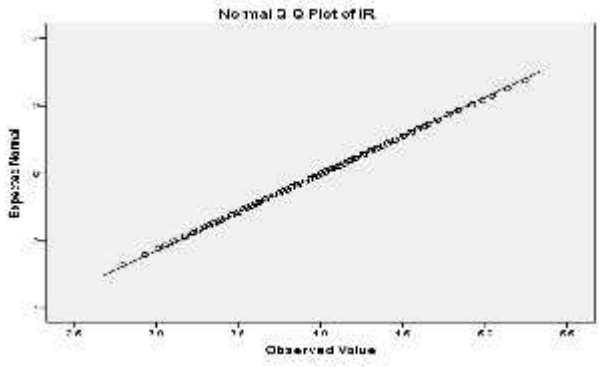


Figure 3
Normal Plot of Inter

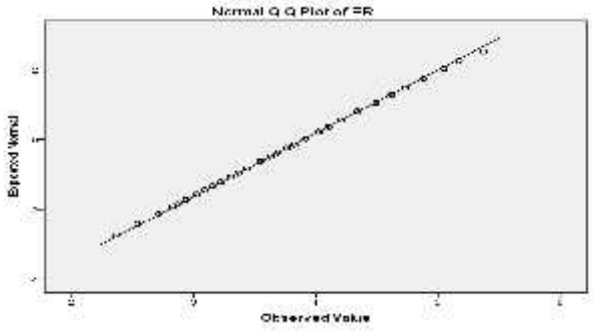
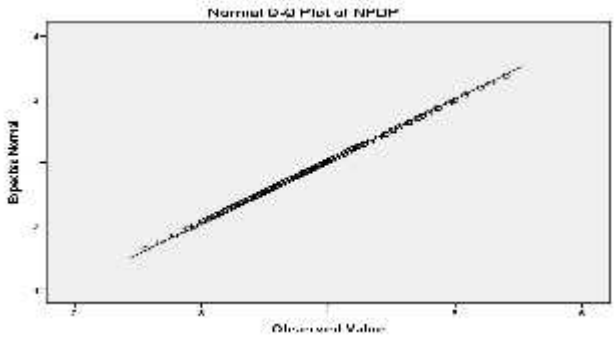


Figure 4
Normal Plot of External Resource Scale



final Resource Scale

Figure 5
Normal Plot of New Product Development Process Scale

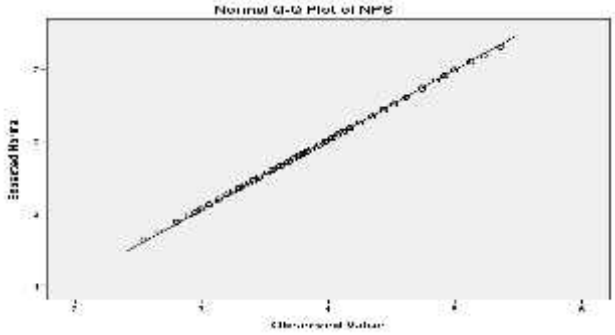


Figure 6
Normal Plot of New Product Success Scale

6. CONCLUSION AND FUTURE IMPLICATIONS

The basic aim of this research was to apply Box-Cox transformation technique to normalize the data. It is rarely noticed in social science research that data is normal. While applying various statistical technique e.g., comparison of mean, correlation coefficients, and especially in structural equation modeling, there is a need to fulfill normality assumptions in applying correct test statistics and to get efficient parameter estimates. Therefore, for normalizing a data a Box-Cox approach is utilized in this research. The results of current research showed that before applying Box-Cox technique data was not normal which is assessed through Shapiro-wilk test. However, it is noticed in this study that after applying Box-Cox transformation technique, data became normal. Hence, it is concluded that for normalizing data Box-Cox approach can be applied. This research is providing new and vital avenue for researchers that data can be normalized by using Box-Cox transformation technique. In this research the procedures for testing normality and applying Box-Cox transformation on SPSS are also given which can be helpful for researchers. With the

 ISSN 2414-8393

perspective of future research directions, the normalized variables can be tested for further statistical tests e.g., in carrying out regression analysis and even structural equation modeling.

REFERENCES

Akgün, A. E., Keskin, H., & Byrne, J. C. (2010). Procedural Justice Climate in New Product Development Teams: Antecedents and Consequences. *Journal of Product Innovation Management*, 27(7): 1096-1111.

Akgün, E.A., Keskin, H., & Byrne, J. (2012). The Role of Organizational Emotional Memory onDeclarative and Procedural Memory and Firm Innovativeness, *Journal of Product Innovation Management*, 29(3), 432–451.

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

Field, A. (2013). Discovering Statistics Using SPSS. 4th ed. London: Sage.

Godambe, V.P. (1982). Estimation in Survey Sampling: Robustness and Optimality. *Journal of theAmerican Statistical Association,* 77,393-403.

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

Jackson, S. (2006). *Research Methods and Statistics.A Critical Thinking Approach.* Second edition. Belmont, CS: Thomson.

Norusis, M. (2008).*SPSS 16.0 Advanced Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Osborne, J. W. (2010).Improving your data transformations: Applying the Box-Cox Transformation.*Practical Assessment, Research, andEvaluation,* 15 (12), 1-9.

Pallant, J. (2013). SPSS Survival Manual. 6th ed. Buckingham: Open University Press. Sakia, R. M. (1992). The Box-Cox Transformation Technique: A review. *The statistician,* 41, 169-178.

Tabachnick, B. G., &Fidell, L. S. (2007).*Using Multivariate Statistics* (5th ed.). Boston:Allyn and Bacon.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Walpole, R.E. (1990). *Introduction to Statistics*, 3rd Edition, Macmillan Publishing Co. Inc. NewYork.

Yuan, K.-H., &Bentler, P. M. (1998). Normal Theory Based Test Statistics in Structural Equation Modeling. *British Journal of Mathematical and Statistical Psychology*, 51, 289 –309.

Yuan, K-H., Chan, W., & Bentler., P. M. (2000). Robust Transformation with Applications To Structural Equation Modeling. *British Journal of Mathematical and Statistical Psychology,* 53, 31-50.

Zimmerman, D. W. (1994). A note on the Influence of Outliers on Parametric and Nonparametric Tests. *Journal ofGeneral Psychology*, 121(4), 391-401.

Zimmerman, D. W. (1995). Increasing the Power of Nonparametric Tests by Detecting and Down Weighting Outliers.*Journal of Experimental Education,* 64(1), 71-78.

Zimmerman, D. W. (1998). Invalidation of Parametric and Nonparamteric Statistical Tests by Concurrent Violation of Two Assumptions.*Journal of Experimental Education,* 67(1), 55-68.

 ISSN 2414-8393