Journal of the Nigerian Association of Mathematical Physics
***Volume* 22 (November, 2012), pp 241 – 248**
© J. of NAMP

Assessing the Validity of Normality Assumption Using Probability Plots and Letter Value Diagnostics

***Osemeke Reuben F. and Ehiwario J. C**

Department of Mathematics
College Of Education Agbor, Delta State, Nigeria

Abstract

The P-P plots, Q-Q plots as well as letter value diagnostics offer an effective and qualitative tool in examining a situation where the normality distributional assumption is assessed and validated. The normal probability plot is a graphical tool for comparing a sample data(x_1, $x_{2,\ldots,}x_n$) set with the normal distribution, based on a subjective visual examination of the data. The observed data(x_i i= 1,2,…,n) are first ordered $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq, \ldots, \leq x_{(n)}$ .Two types of probability plots have been suggested for normality validation. The plot of observed data against the cumulative probability ($P_i = \frac{i}{n+1}$), i= 1… n) for the P-P plots and the plot of observed data against the inverse cumulative probability ($m_i = \Phi^{-}(Pi)$) for the Q-Q plots and this is accompanied with relative tail thickness, symmetrically distributed and 45^0 (y=x) linearity curve of 95% confidence limits. In addition, equality of data for the mid-summary values was used to show a normal validation. Non-normality of the data was corrected through trial by error transformation. Square-root transformations were closer to normality

1.0 Introduction

The Q-Q and P-P plots compare a sample of sample data on the vertical axis to statistical population on the horizontal axis. The data points follow a strongly linear pattern suggesting that the sample data are normally distributed around the mean of zero and standard deviation of 1 or follow a 45% (y=x) linearity pattern from (0,0) to (1,1).The linearity pattern suggest that the data are normally distributed. If two distributions being compared are similar, the points in the Q-Q plots and P-P plots will approximately lie on the line y = x, but not necessarily on the line y=x. A Q-Q plot is used to compare the shapes of the distribution, providing a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions. A Q-Q plot is generally a more powerful approach to doing this than the common techniques of P-P plots and comparing histogram of the two samples, but requires a more statistical skill to interpret. This can provide an assessment of goodness of fit test that is graphical, rather than reducing to numerical summary

Another measure for assessing normality assumption is the letter value diagnostics which include the median, the hinges, the eights and the sixteen's. If the mid-summaries are approximately equal, then the values of the hinges, eights, and so on are nearly symmetric about the median. If the mid-summaries become progressively larger, the data is skewed towards the high side. If they decreased steadily, the data is skewed towards the low side. All this is an indication, that the data are not normally distributed.

To improve on the curve, a transformation is done. Transformation of real and simulated data's arises when the data does not satisfy the validity assumption of the normality term. However, picking a transformation is often a matter of trial by error. Different transformations are tried until one is found for which the data seem reasonable [1].

*Corresponding author: ***Osemeke Reuben F.,*** E-mail: reubitosmek@yahoo.com, Tel. +234 8065917356

2.0 Plotting Positions

The choice of quantiles from a theoretical distribution has occasioned much discussion. A natural choice, given a sample of size n is $\frac{\mathrm{k}}{\mathrm{n}}$ for k = 1… n as these are the quantiles that the sampling distribution realizes. Unfortunately, the last of theses, $\frac{\mathrm{k}}{\mathrm{n}}$, corresponds $\frac{k-0.5}{n}$ or instead space the points evenly in the uniform distribution using $\frac{k}{n+1}$. This last one was suggested early on in [2] and recently it has been argued to be the definitive position in [3].The claimed unique status of the estimates was rebutted in [4] .Many other choices have been suggested formal and heuristic [5] which use the following estimates for the uniform order statistic medians

$$\boxed{\mathrm{Mi} = \left\{\begin{array}{ll} 1-m(n) & \mathrm{i} = 1 \\ \frac{1-0.375}{n+0.365} & \mathrm{i} \ = 2, 3\ldots \mathrm{n}\text{-}1 \\ (.5)^{/\mathrm{n}} & \mathrm{i} = \mathrm{n} \end{array}\right\}}$$

The linearity of the plots suggest that the data are normally distributed.

Several different formula has been used and proposed as symmetrical plotting positions and such formulas are of the form $\mathrm{P_i} = \frac{i-c}{n-2c+1}$ for some value of c in the range from 0 to 1, which gives a range between $\frac{i}{n+1}$ and $\frac{i-1}{2/n}$. Others are $\mathrm{P_i} = \frac{k-0.5}{n}$, $\mathrm{P_i} = \frac{i-0.3175}{n+0.365}$, estimated as $\frac{k-0.5}{n}$ and $\frac{(0.5)}{n}$ as cumulative probability and inverse cumulative probability as $\Phi^{-1}(\mathrm{Pi}) = \Phi^{-1}\frac{i-c}{n-2c+1}$ for $0 \leq c \leq 1$. Others are $\Phi^{-1}(\mathrm{Pi}) = \Phi^{-1(}\frac{k-0.5}{n}$), $\Phi^{-}(\mathrm{Pi}) = \Phi^{-(}\frac{i-0.3175}{n+0.365}$), estimated as $\Phi^{-1}($ $\frac{k-0.5}{n}$) and $\Phi^{-1}($ $\frac{0.5}{n}$)

The criterion for selecting the above plotting positions have been discussed [6 - 9]. Recently, empirical investigations of the different plotting positions and power of test of fit based on correlation coefficient derived from Q-Q plots have been performed by [10] , examining how different plotting positions affects estimates of various features on the underlying distribution obtained from regression lines fitted to the corresponding Q-Q plots.

The plotting position of $\mathrm{Pi} = \frac{i}{n+1}$ suggested by [2] generally has superior properties and preferable for plotting normal curve.

P-P plot will not give the same relative emphasis to all regions of the hypothesized distribution as the corresponding Q-Q plot. The values of $\mathrm{m_i} = \Phi^{-}(\mathrm{Pi})$ on the horizontal axis of a Q-Q plot are more tightly bunched together in the modal regions of the underlying distribution than in the tail region whereas the values are usually spaced on the horizontal axis of the P-P plot.

3.0 Obtaining Cumulative Frequency: ($\mathrm{Pi} = \frac{i}{n+1}$) and Inverse Cumulative Frequency ($\mathrm{M_i} = \Phi^{-}(\mathrm{Pi})$)

This is characterized by mean of 0 and standard deviation of 1. Owing to its symmetry, the median from a standard normal distribution must also be 0.Therefore, in dealing with standard normal distribution, the quantile values below the median will be negative and the quantile values above the median will be positive. However, the question we must still answer is". How can we obtain the quantile values from this distribution? The process, by which we accomplished this task, is known as inverse normal scores transformations, and given a data containing observations (n) from a standardized normal distribution. Let $\mathrm{P_1}$ represent the first and smallest ordered or quantile value and let the symbol $\mathrm{P_n}$ represent the largest value. Because of symmetry, the standard normal quantiles $\mathrm{P_1}$ and $\mathrm{P_n}$ will have the same numerical value except for the sign $\mathrm{P_1}$ will be negative and $\mathrm{P_n}$ will be positive. Suppose we wish to obtain the first, and 19th standard normal ordered values corresponding to a sample of 19th observations as documented in Table 1. The first standard normal ordered value $\mathrm{P_1}$, is that value below the proportion $\frac{1}{n+1} = \frac{1}{20} = 0.05$.

See Table E.2 (b) of the cumulative standardized normal distribution [11] to locate area up to 0.05 so that from the body of our accompanying Table E.2 (b) would fall halfway between

-1.65 and -1.64. Because the standard normal values are usually reported with two decimal places, the value -1.65 is chosen. In addition, this is also applicable to the last standard normal ordered value $\mathrm{P_{19}}$.For Table E.2 (b) (see [11])

Table 1: An ordered array of midterm test scores obtained from19 students of a course in introductory finance & corresponding standard normal ordered values**.**

The aim of Table 1 is to use the observed data set to show that the P-P and Q-Q plots follows a linearity pattern without the influence of an outlier which is an indication of the validation of the normality assumption

I	Observed Data(Xi)	Cumulative Frequency ($P_i = \frac{i}{n+1}$)	Inverse Cumulative Frequency ($M_{i)} = \Phi^{-}(Pi)$)
1	48	0.05	-1.65
2	52	0.1	-1.28
3	55	0.15	-1.04
4	57	0.2	-0.84
5	58	0.25	-0.67
6	60	0.3	-0.52
7	61	0.35	-0.39
8	62	0.4	-0.25
9	64	0.45	-0.13
10	65	0.5	0.00
11	66	0.55	0.13
12	68	0.6	0.25
13	69	0.65	0.39
14	70	0.7	0.52
15	72	0.75	0.67
16	73	0.8	0.84
17	75	0.85	1.04
18	78	0.9	1.28
19	82	0.95	1.65

Source: [11]

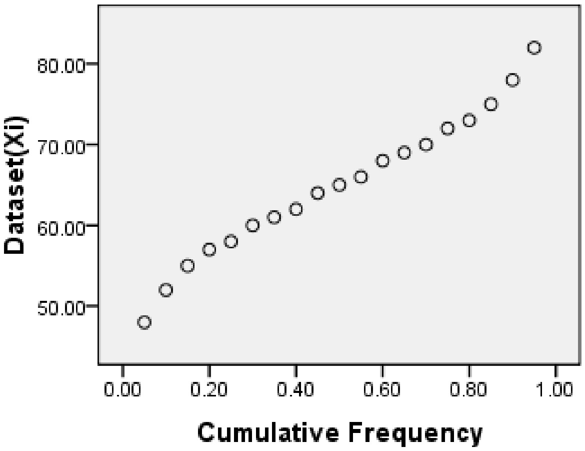

Fig1: P-P Plot Normality Validation

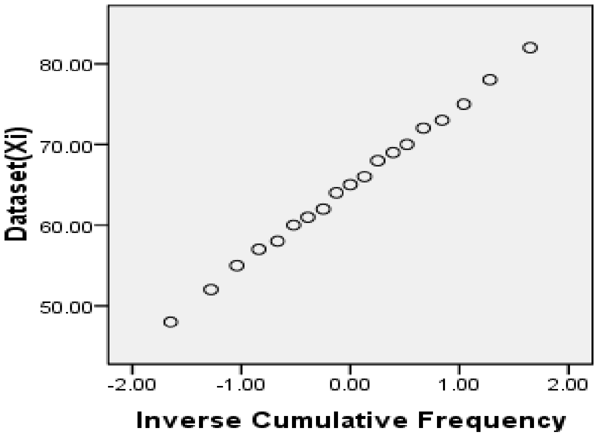

Fig2: Q-Q Plot Normality Validation

4.0 Letter Value Diagnostics (Median, Hinges and Other Summary Value)

Letter values are similar to percentiles of the data and are defined by their depth. The Median, the hinges, the eights, and the sixteenths are the start of the sequence of the letter values. They are defined as follows

Depth of the median d(M) = $\frac{n+1}{2}$. Depth of hinges d(H) = $\frac{([d(M)]+1)}{2}$, Depth of eights: d(E) = $\frac{([d(H)]+1}{2}$, Depth of sixteenths d(D) = $\frac{([d(E)]+1}{2}$

To find the letter values, first order the data set. The lower hinge is the observation at a distance d(H) from smallest observations, the upper hinge is the observation at a distance d(H) from the largest observations. Similarly, the lower and upper eights are the observations at a depth d(E) and so on. The midpoint for a given depth is the average of the upper and lower letter values at the depth. The spread is (upper –lower)

Normality of this letter values are ensured when the mid-summary values for the data batch are approximately equal.

5.0 Assessing Normality Using Letter Value Diagnostics

Table 2: Locating & Calculating the Letter Values for the Midterm Test Scores of a course in Introductory Finance

Depth of letter value	Depth	Dataset(xi)	Letter values
	1	48	Extreme = 48
D(D) = 3+1/2 = 2	2	52	D = 52
D(E) = 5+1/2 = 3	3	55	E = 55
	4	57	
D(H) = 10+1/2 = 5.5	5	58	H = 59
	6	60	
	7	61	
	8	62	
	9	64	
D(M) = 19+1/2 = 10	10	65	M = 65
	9	66	
	8	68	
	7	69	
	6	70	H = 71
	5	72	
	4	73	
	3	75	E= 75
	2	78	D = 78
	1	82	Extreme = 82

Source: [1]

Table 3: Letter Value Display

	Depth	Lower	Upper	Mid	Spread
N = 19					
M	19.0	65	65	65	
H	5	59	71	65	12
E	3	55	75	65	20
D	2	52	78	65	26
	1	48	82	65	34

From Table 3, we observed that the Mid-point values are the same, hence a validation of normal distributional assumption

6.0 Re-expression For Symmetric

Here we present data batch that are not normal, which shows an indication that transformation is needed to improve on the data

Table 4: Thirty Consecutive Values of March Precipitation at Minneapolis/St. Paul

0.77	1.74	0.81	1.20	1.95	1.20
0.47	1.43	3.37	2.20	3.00	3.09
1.51	2.10	0.52	1.62	1.31	0.32
0.59	0.81	2.81	1.87	1.18	1.35
4.75	2.48	0.96	1.89	0.90	2.05

Source: [1]

Table 5: The Output of Letter Value Display for the March Precipitation in Minneapolis/St Paul

	Depth	Lower	Upper	Mid	Spread
N =	30				
M	15.5	1.47	1.47	1.47	
H	8	0.90	2.10	1.50	1.20
E	4.5	0.68	2.905	1.79	2.225
D	2.5	0.495	3.23	1.86	2.735
C	1.5	0.395	4.06	2.23	3.665
	1	0.32	4.75	2.535	4.43

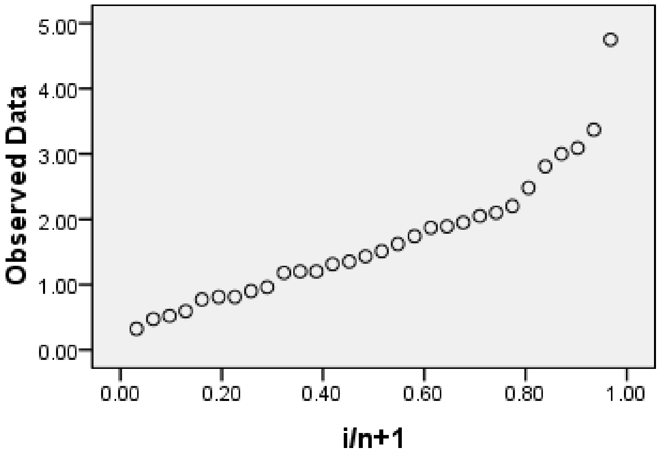

FIG 3: Observed Data versus Cumulative Frequency ($\frac{i}{n+1}$)

From Table 5, we noticed an upward trend of the mid-summaries which indicates skewness to the right. The P-P plots for the observed shows little skew. To move towards symmetry, we should try re-expression using square root and logarithm

Table 6. Square-Root Transformation

	Depth	Lower	Upper	Mid	Spread
N =	30				
M	15.5	1.212	1.212	1.212	
H	8	0.949	1.449	1.199	0.500
E	4.5	0.822	1.704	1.263	0.881
D	2.5	0.704	1.797	1.250	1.093
C	1.5	0.626	2.008	1.317	1.382
	1	0.566	2.179	1.372	1.614

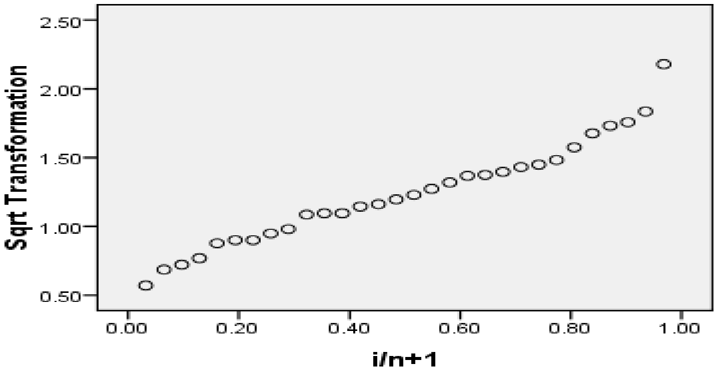

Fig 4: Square Root Transformation of Observed Data versus Cumulative Frequency ($\frac{i}{n+1}$)

Table 7: Logarithms Transformation

	Depth	Lower	Upper	Mid	Spread
N =	30				
M	15.5	0.167	0.167	0.167	
H	8	-0.046	0.322	0.138	0.368
E	4.5	-0.171	0.463	0.146	0.631
D	2.5	-0.306	0.509	0.101	0.815
C	1.5	-0.411	0.602	0.095	1.014
	1	-0.495	0.677	0.091	1.172

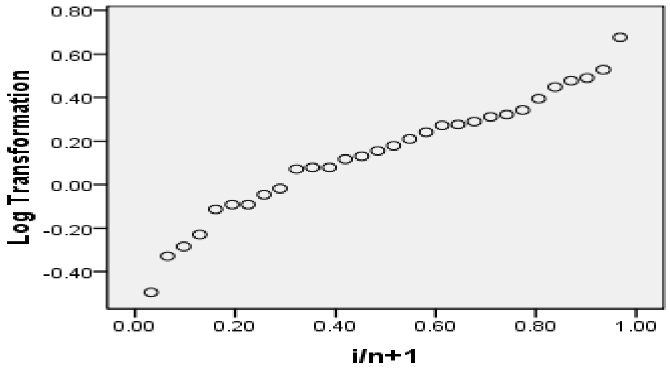

Fig 5: Logarithm Transformation of Observed Data versus Cumulative Frequency ($\frac{i}{n+1}$)

Table 8: Mid-summaries for several expressions

Tag	Raw	Root	Log
M	1.47	1.212	.1672
H	1.50	1.199	.1382
E	1.79	1.263	.1458
D	1.86	1.250	.1014
C	2.23	1.316	.0954
I	2.535	1.372	.0909

As we look for trends down each column of the mid-summaries of Table 8, if we need to choose among re-expression, we might select the square root for its simplicity. The log transformation has a miniature bend at the bottom of the graph, so there is little skewness

7.0 Findings/Conclusion

It has been shown that Probability plots and the letter value diagnostics provides a simple and effective qualitative tool for both assessing the fit of a proposed probability model and identifying viable alternatives .In addition, sample data were used to observed the behavioral pattern of probability plots for both P-P and Q-Q plot as well as letter value diagnostics.

Although the P-P and Q-Q plots shows substantial linearity but further analysis shows that the linearity pattern of Q-Q plot is much stronger. Stronger in the sense that the value of $m_i = \Phi^{-}(Pi)$ on the horizontal axis of a Q-Q plot are more tightly bunched together in the modal regions of the underlying than in the tail region whereas the P-P plot are usually spaced on the horizontal axis.

Above all, I recommend that researches especially those in statistical analysis should use Q-Q plots to asses' normality assumption than the P-P plots when the need arises because it has superior properties and have tendency of getting a valid and unbiased results

Reference

[1]. Tuckey W.J., (1981). Applications, Basics, and Computing of Exploratory Data Analysis. Abt Associates Inc pp 48-49.

[2]. Weibull W., (1939). The Phenomenon of Rapture in Solids, Ingeni Vetenskaps Akademien Handlingar, 153, 17

[3]. Makkonen L., (January 2008). Bringing Closure to the Plotting Position Controversy, Communications in Statistics-Theory and Methods 37(3):460-467.

[4]. Cook and Nicholas J., (January 2001). Comments on Plotting Positions in Extreme Value Analysis, .J. App. Meteor. Climate, 50(1):256-266

[5]. Filliben J.J., (February 1975). The Probability Plot Correlation Coefficient Test For Normality. Technometrics (American Society for Quality) 117(1):111-117

[6]. Kimball, B. F. (1960). On the Choice of Plotting Positions on Probability Paper. Journal of the American Statistical Association, 55, 546-560

[7]. Barnett, V., (1975). Probability Plotting Methods and Order Statistics, Applied Statistics, 24, 95-108.

[8]. Looney and Wiegund (1985). Uses of the correlation Coefficient with Normal Probability Plot, the American Statisticians, 34, 297-303

[9]. Harter, H. L., (1984). Another look at Plotting Positions, Communication in Statistics, Part A- Theory and Methods. 13 1613-1633.

[10]. Looney, S. W. and T.R.Gulledge, (1984). Regression Tests of fit and Probability Plotting Positions. Journals of Statistical Computation and Simulation, 20, 115-127.

[11]. David. Mark, B. and David, S. (1998) Statistics for Managers, 2rd, Prentice Hall International