Effects of Sample Size and Dispersion on Quantile-Based Plots for Detecting Normality

Isaac O. Ajao; Oluwafemi S. Obafemi; Folasade A. Bolarinwa

Department of Mathematics and Statistics,
The Federal Polytechnic, Ado-Ekiti, Nigeria.
e-mail: isaacoluwaseyiajao@gmail.com

Abstract **— Many instances in statistical modelling may require that the data are tested for normality before proceeding to further statistical analysis. The classical tests for the assessment of normality among others are Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Shapiro-Wilk test, Shapiro-Francia test, Anderson-Darling test, Cramer-von Mises test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino-Pearson omnibus test, and the Jarque-Bera test. The visual methods commonly used are the histogram, boxplot, pp-plot, Q-Q plot, and the stem-and-leaf plot. This paper seeks to find out the effect of sample and dispersion on quantile based plots for detecting normality in Monte Carlo simulated and the transformed data. It was observed that as the sample size increases the data approaches normality, while it suffers departure as standard deviation increases. It is therefore recommended that the visual methods, especially the Q-Q-plot be used for detecting normality only when the sample size is low and the standard deviation is high.**

Keywords - Normality, qq-plot, sample size, dispersion, classical tests.

.

I. INTRODUCTION

The statistical methods are based on various assumptions that uphold the methods. One of them is the normality assumption. It is often required to check the normality in many data analyses, although normality is implicitly or conveniently assumed in reality. If the assumption is violated, interpretations and inferences based on the models are not reliable, if not valid. There are two ways of checking normality.

The graphical methods visualize differences between the empirical distribution and the theoretical distribution like a normal distribution. The numerical methods conduct statistical tests on the null hypothesis that the variable is normally distributed. The graphical methods visualize the distribution using plots. They are grouped into descriptive and theoretical. The former method is based on empirical data, whereas the latter considers both empirical and theoretical distributions.

II. DESCRIPTIVE PLOTS AND THEORETICAL PLOTS

The frequently used descriptive plots are the stem-and-leaf-plot, (skeletal) boxplot, dot plot, and histogram. When N is small, a stem-and-leaf plot or dot plot is useful to summarize data; the histogram is more appropriate for large N samples. A stem-and-leaf plot assumes continuous variables, while a dot plot works for categorical variables.

A box plot presents the 25 percentile, 50 percentile (median), 75 percentile, and mean in a box. If a variable is normally distributed, its 25 and 75 percentile become symmetry, and its median and mean are located at the same point exactly in the middle. The P-P plot and Q-Q plot are more commonly used to check normality than the descriptive plots.

The probability-probability plot (P-P plot or percent plot) compares the empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function (e.g., the standard normal distribution function). Similarly, the quantile-quantile plot (Q-Q plot) compares ordered values of a variable with quantiles of a specific theoretical distribution (i.e., the normal distribution). If two distributions match, the points on the plot will form a linear pattern passing through the origin with a unit slope. So, the P-P plot and the Q-Q plot are used to see how well a theoretical distribution models the empirical data. Although visually appealing, these graphical methods do not provide objective criteria to determine the normality of variables. Interpretations are matter of judgments. Therefore this paper focuses on the necessity of numerical methods to determine normality rather than the graphical methods. It is also to compare numerical results with judgment.

2.1 Theoretical Statistics

The numerical methods of testing normality include the Kolmogorov-Smirnov (K-S) (Smirnov, 1948) D test (Lilliefors test, Lilliefors, 1967), Shapiro-Wilk' test, Anderson-Darling test, and Cramer-von Mises test (SAS Institute 1995, von Mises, 1928). The K-S D test and Shapiro-Wilk' W test are commonly used. The K-S, Anderson-Darling (Anderson and Darling, 1954), and Cramer-von Misers (Cramer, 1928) tests are based on the empirical distribution function (EDF), which is defined as a set of N independent observations x1, x2, …xn with a common distribution function F(x).

Table 1: Numerical tests of normality

Test	Statistic	Sample Size (N)	Distn
Jarque-Bera (S-K) test	χ^2	-	χ^2 (2)
Shapiro-Wilk	W	$7 \le N \le 2000$	-
Shapiro-Francia	W	$5 \le N \le 5000$	-
Kolmogorov-Smirnov	D	> 2000	EDF
Cramer-vol Mises	W^2	> 2000	EDF
Anderson-Darling	A^2	> 2000	EDF

III. Methodology

The Shapiro-Wilk statistic (1965) is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance (Royston, 1982). The statistic is positive and less than or equal to one; being close to one indicate normality. The W statistic requires that the sample size needs to greater than or equal to seven and less than or equal to 2,000 (Royston, 1992).

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (1)$$

where $x_{(i)}$ is the ith order statistic i.e. the ith-smallest number in the sample; $\bar{x} = (x_1 + \ldots + x_n)/n$ is the sample mean; the constant a_i are given by (Shapiro-Wilk, 1965):

$$(a_1, a_2,) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}} \quad (2)$$

where

$m = (m_1, \ldots, m_n)^T$ and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

The Shapiro-Francia test is an approximate test that modified the Shapro-Wilk test. The statistic was developed by Shapiro and Francia (1972) and Royston (1983). Let x_i be the ith ordered value from our size-n sample, also $m_{i:n}$ be the mean of the ith order statistic when making n independent draws from a normal distribution. The Pearson correlation coefficient between x and m is then given as:

$$W' = \frac{\operatorname{cov}(x,m)}{\sigma_x \sigma_m} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(m_i - \bar{m})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} \times \sqrt{\left[\sum_{i=1}^{n}(m_i - \bar{m})^2\right]}} \quad (3)$$

Under the null hypothesis that the data is drawn from a normal distribution, this correlation will be strong, so W' cluster just under 1, with the peak becoming narrower and closer to 1 as n increases. If the data deviate strongly from a normal distribution, W' will be smaller (Shapiro and Francia, 1972).

IV Data Analysis

Monte Carlo simulation Setup

To measure the effect of sample size on quantile based plots for detecting normality, we simulated random numbers following the normal distribution with $\mu = 50,\ \sigma = 2$ for various sample sizes (n = 20, 50, 100, 250, 500, 1000) and sets of transformed variables (cubic, square, square root, log, 1/square root, inverse, 1/square, and 1/cubic). Standard deviations of 3 and 5 were introduced into the simulation in order to test for the effect of dispersion. All analysis was done using MATLAB, R2017a and STATA 12 SE.

Table 1: Normality tests on transformed data when n = 20 and n = 50

Sample size	Transformation	Formula	chi2	Pr(chi2)	Swilk	Pr(Swilk)	Sfrancia	Pr(francia)
20	cubic	x^3	0.6200	0.7340	0.9789	0.9192	0.9874	0.9764
	square	x^2	0.5400	0.7650	0.9747	0.8496	0.9827	0.9209
	identity	x	0.5100	0.7750	0.9773	0.8949	0.9856	0.9591
	square root	sqrt(x)	0.5100	0.7740	0.9793	0.9250	0.9879	0.9802
	log	log(x)	0.5200	0.7690	0.9795	0.9273	0.9882	0.9818
	1/square root	1/sqrt(x)	0.5400	0.7620	0.9794	0.9261	0.9881	0.9816
	inverse	1/x	0.5700	0.7530	0.9791	0.9214	0.9878	0.9795
	1/square	1/(x^2)	0.6300	0.7280	0.9776	0.9002	0.9864	0.9677
	1/cubic	1/(x^3)	0.7200	0.6970	0.9753	0.8597	0.9840	0.9398
50	cubic	x^3	0.5500	0.7580	0.9599	0.0885	0.9729	0.2588
	square	x^2	1.1600	0.5610	0.9696	0.2244	0.9690	0.1829
	identity	x	1.9000	0.3870	0.9654	0.1508	0.9636	0.1139
	square root	sqrt(x)	2.3100	0.3150	0.9567	0.0647	0.9605	0.0862
	log	log(x)	2.7500	0.2530	0.9531	0.0460	0.9570	0.0636
	1/square root	1/sqrt(x)	3.2100	0.2010	0.9493	0.0319	0.9531	0.0459
	inverse	1/x	3.7000	0.1570	0.9451	0.0217	0.9490	0.0325
	1/square	1/(x^2)	4.5200	0.1040	0.9361	0.0095	0.9400	0.0156
	1/cubic	1/(x^3)	5.3400	0.0690	0.9261	0.0039	0.9299	0.0072

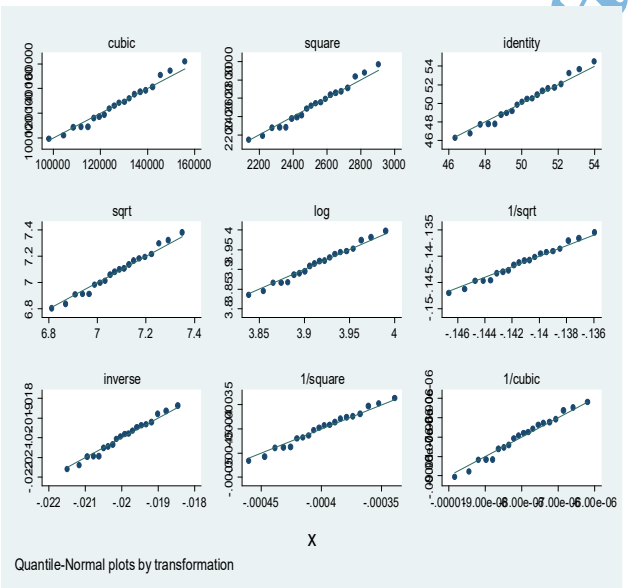

Fig. 1: Quantile plots of transformed data when n=20

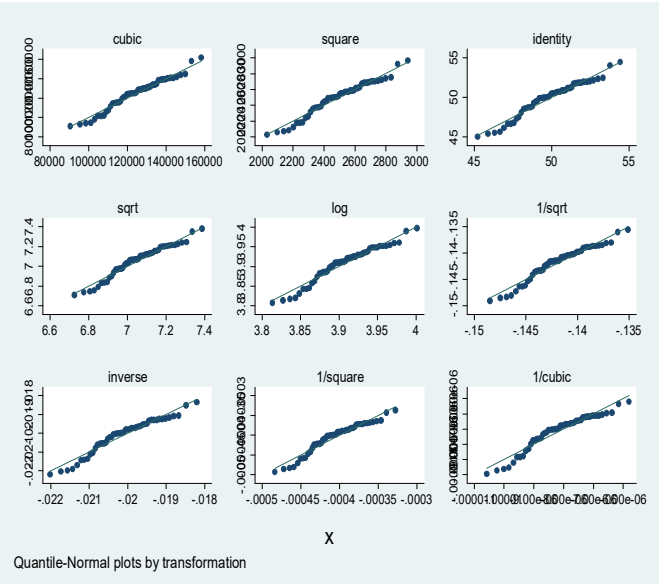

Fig. 2: Quantile plots of transformed data when n=50

Table 2: Normality tests on transformed data when n = 100 and n = 250

Sample size	Transformation	Formula	chi2	Pr(chi2)	Swilk	Pr(Swilk)	Sfrancia	Pr(francia)
100	cubic	x^3	1.5700	0.4560	0.9907	0.9171	0.9905	0.8401
	square	x^2	0.6200	0.7330	0.9929	0.7284	0.9927	0.6134
	identity	x	0.3400	0.8430	0.9935	0.8832	0.9933	0.7921
	square root	sqrt(x)	0.4500	0.7990	0.9932	0.9019	0.9930	0.8199
	log	log(x)	0.7100	0.7000	0.9925	0.8610	0.9924	0.7681
	1/square root	1/sqrt(x)	1.1400	0.5670	0.9915	0.7862	0.9913	0.6821
	inverse	1/x	1.7100	0.4250	0.9901	0.6732	0.9899	0.5662
	1/square	1/(x^2)	3.3100	0.1910	0.9861	0.3821	0.9858	0.3080
	1/cubic	1/(x^3)	5.2200	0.0740	0.9806	0.1504	0.9802	0.1234
250	cubic	x^3	0.7000	0.7060	0.9954	0.6552	0.9957	0.6421
	square	x^2	0.0100	0.9940	0.9961	0.7833	0.9966	0.7923
	identity	x	0.3700	0.8300	0.9955	0.6801	0.9961	0.7058
	square root	sqrt(x)	0.9000	0.6370	0.9947	0.5419	0.9954	0.5737
	log	log(x)	1.6500	0.4390	0.9937	0.3722	0.9943	0.4056
	1/square root	1/sqrt(x)	2.6000	0.2730	0.9923	0.2163	0.9930	0.2457
	inverse	1/x	3.7500	0.1540	0.9906	0.1066	0.9913	0.1283
	1/square	1/(x^2)	6.3700	0.0410	0.9863	0.0172	0.9870	0.0245
	1/cubic	1/(x^3)	9.2500	0.0100	0.9809	0.0019	0.9816	0.0034

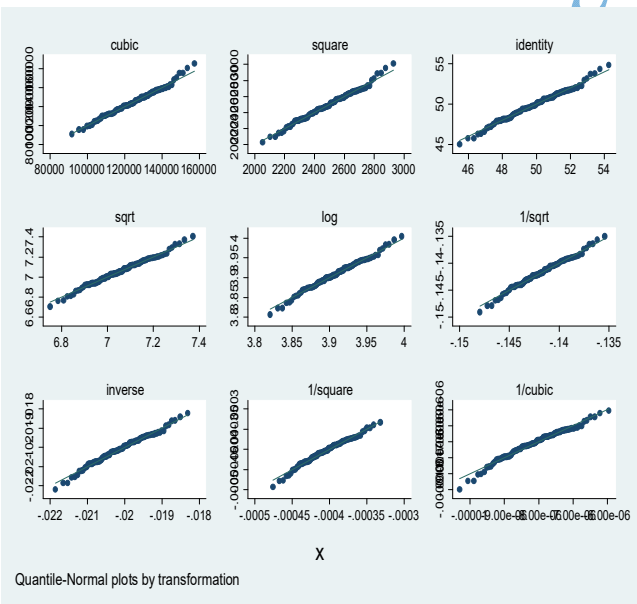

Fig. 3: Quantile plots of transformed data when n=100

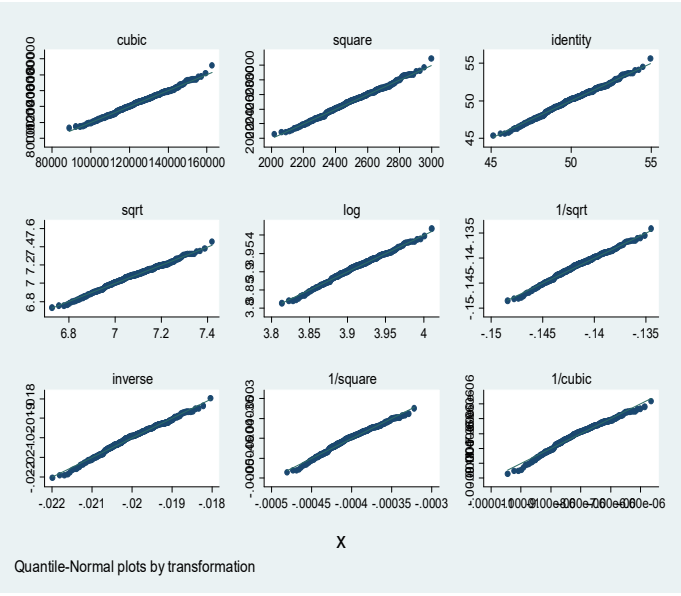

Fig. 4: Quantile plots of transformed data when n=250

Table 3: Normality tests on transformed data when n = 500 and n = 1000

Sample size	Transformation	Formula	chi2	Pr(chi2)	Swilk	Pr(Swilk)	Sfrancia	Pr(francia)
500	cubic	x^3	14.0500	0.0010	0.9896	0.0013	0.9891	0.0014
	square	x^2	7.4200	0.0240	0.9942	0.0555	0.9938	0.0395
	identity	x	2.5300	0.2820	0.9971	0.5331	0.9968	0.3624
	square root	sqrt(x)	1.1400	0.5660	0.9979	0.8037	0.9975	0.6014
	log	log(x)	0.5300	0.7660	0.9983	0.9013	0.9979	0.7189
	1/square root	1/sqrt(x)	0.7400	0.6910	0.9982	0.8752	0.9978	0.6781
	inverse	1/x	1.7900	0.4090	0.9976	0.6988	0.9972	0.4854
	1/square	1/(x^2)	6.5000	0.0390	0.9952	0.1192	0.9947	0.0749
	1/cubic	1/(x^3)	13.4500	0.0010	0.9909	0.0035	0.9903	0.0030
1000	cubic	x^3	28.6300	0.0000	0.9899	0.0000	0.9898	0.0000
	square	x^2	14.5300	0.0010	0.9944	0.0010	0.9944	0.0013
	identity	x	5.1400	0.0760	0.9973	0.0887	0.9973	0.0831
	square root	sqrt(x)	2.2300	0.3270	0.9980	0.2922	0.9980	0.2635
	log	log(x)	0.9200	0.6330	0.9983	0.4573	0.9984	0.4141
	1/square root	1/sqrt(x)	1.1300	0.5690	0.9982	0.3956	0.9983	0.3583
	inverse	1/x	2.8000	0.2470	0.9977	0.1785	0.9977	0.1642
	1/square	1/(x^2)	9.8400	0.0070	0.9954	0.0040	0.9954	0.0047
	1/cubic	1/(x^3)	20.1700	0.0000	0.9914	0.0000	0.9914	0.0000

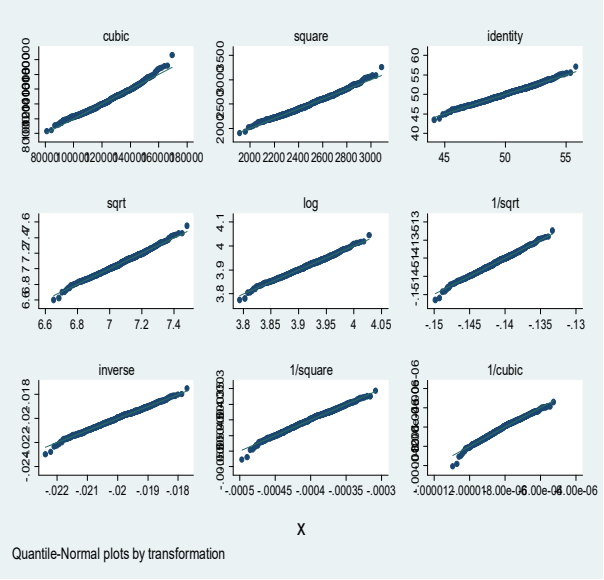

Fig. 5: Quantile plots of transformed data when n=500

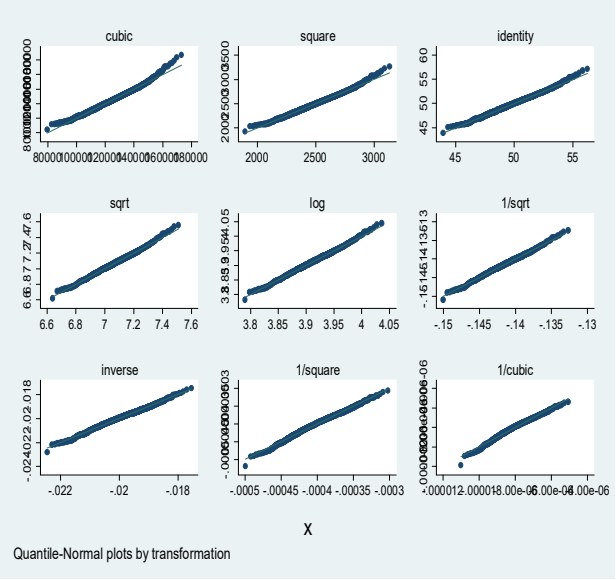

Fig. 6: Quantile plots of transformed data when n = 1000

V. Visual and Classical Results Compared

Making valid conclusions on whether data is normally distributed or not is mostly needed by researchers across the globe, and this can be done using the plots and the classical tests carried out in this paper. Considering the plot obtained on the squared transformed data when n = 1000, visually, one may quickly say the data is normally distributed, but its corresponding p-values using the classical tests Chi-square, Shapiro-Wilk, and Shapiro-Francia are 0.0010, 0.0010, and 0.0012 which indicate rejection of normality assumption. Therefore the quantile plots are most appropriate with few samples.

VI. Discussion of Results

It is discovered that normality increases as the sample sizes increase on the quantile plots. The transformed data is used to compare the normality with the introduction of Chi-square, Shapiro-Wilk and Shapiro-Francia test statistics. It is also discovered that dispersion has an effect on the normality of the transformed data, the departure from normality increases as the measure of dispersion increases.

Comparison studies have concluded that order statistic correlation tests such as Shapiro–Francia and Shapiro–Wilk are among the most powerful of the established statistical tests for normality (Razali and Wah, 2011). One might assume that the covariance-adjusted weighting of different order statistics used by the Shapiro–Wilk test should make it slightly better, but in practice, the Shapiro–Wilk and Shapiro–Francia variants are about equally good. In fact, the Shapiro–Francia variant actually exhibits more power to distinguish some alternative hypotheses (Ahmad and Khan, 2015).

VII. Conclusion and Recommendation

Using the results obtained so far, it can be concluded that sample sizes and measures of dispersion have a significant effect on the detection of normality in a set of data using the quantile plots. Normality increases as the sample sizes increase but decrease as the standard deviation increases. It is therefore recommended that the quantile plot be employed only when the sample size is small. However, classical tests always give objective and precise results.

References

[1] Ahmad F. and Khan R. A. (2015): "A power comparison of various normality tests", Pakistan Journal of Statistics and Operation Research II

[2] Anderson, T. W. and Darling, D. A. (1954) "A Test of Goodness of Fit", Journal of the American Statistical Association, 49, 765-769.

[3] Crammer, H. (1928): "On the composition of elementary errors", Scandinavian Actuarial Journal.

[4] Lilliefors, H. (1967): "On the Kolmogorov-Smirnov test for normality with mean and variance unknown",Journal of the American Statistical Association, 62, 399-402.

[5] Razali N. M. and Wah Y. B. (2011): "Power Comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling Tests", Journal of Statistical Modeling and Analytics 2

[6] Royston, P. (1982): An extension of Shapiro and Wilks's W test for normality to large samples. Applied Statistics 31: 115–124.

[7] Royston, P. (1983): A simple method for evaluating the Shapiro–Francia W' test of non-normality. Statistician 32: 297–300.

[8] Royston, P. (1991a): sg3.2: Shapiro–Wilk and Shapiro–Francia tests. Stata Technical Bulletin 3: 19. Reprinted in Stata Technical Bulletin Reprints, vol. 1, p. 105. College Station, TX: Stata Press.

[9] Royston, P. (1991b): Estimating departure from normality. Statistics in Medicine 10: 1283–1293.

[10] Royston, P. (1992): Approximating the Shapiro–Wilk W-test for non-normality. Statistics and Computing 2: 117–119.

[11] Royston, P. (1993a): A pocket-calculator algorithm for the Shapiro–Francia test for non-normality: An application to medicine. Statistics in Medicine 12: 181–184.

[12] Royston, P. (1993b): A toolkit for testing for non-normality in complete and censored samples. Statistician 42: 37–43.

[13] Shapiro, S. S., and Francia, R. S. (1972): An approximate analysis of variance test for normality. Journal of the American Statistical Association 67: 215–216.

[14] Shapiro, S. S., and Wilk, M. B. (1965): An analysis of variance test for normality (complete samples). Biometrika 52: 591–611.