Communications for Statistical Applications and Methods
2014, Vol. 21, No. 4, 309–316

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.4.309
Print ISSN 2287-7843 / Online ISSN 2383-4757

Goodness-of-Fit Test for the Normality based on the Generalized Lorenz Curve

Youngseuk Cho[a], Kyeongjun Lee[1,a]

[a]Department of Statistics, Pusan National University, Korea

Abstract

Testing normality is very important because the most common assumption is normality in statistical analysis. We propose a new plot and test statistic to goodness-of-fit test for normality based on the generalized Lorenz curve. We compare the new plot with the Q-Q plot. We also compare the new test statistic with the Kolmogorov-Smirnov (KS), Cramer-von Mises (CVM), Anderson-Darling (AD), Shapiro-Francia (SF), and Shapiro-Wilks (W) test statistic in terms of the power of the test through by Monte Carlo method. As a result, new plot is clearly classified normality and non-normality than Q-Q plot; in addition, the new test statistic is more powerful than the other test statistics for asymmetrical distribution. We check the proposed test statistic and plot using Hodgkin's disease data.

Keywords: Generalized Lorenz curve, goodness-of-fit, Lorenz curve, normality test, power.

1. Introduction

Testing normality is very important because the most common assumption is normality in statistical analysis. So normality was researched continuously by many scholars. Estimation of data distribution is used to histogram, Q-Q plot and P-P plot which uses a graph. In addition to using the graphic method, typical methods using the test statistic are the Kolmogorov-Smirnov test and Shapiro-Wilk test.

This study proposes a generalized Lorenz curve (GLC) test statistic and a graphical method, called the g-plot, for the normality test based on the generalized Lorenz curve proposed by Shorrocks (1983).

We first introduce a Lorenz curve and generalized Lorenz curve. Lorenz curve introduced by Lorenz (1905) provides the means to access the disparity in an income distribution optically and compare income disparity between two distributions. The theorem of Atkinson (1970) was the first result to give terms under which such a Lorenz inequality comparison has normative significance. In case of an increasing and strict concave utility function, the Atkinson theorem says that a distribution with the dominating Lorenz curves that do not cross is preferred. But Atkinson's theorem has the limitation that both the distributions have same means. So Shorrocks (1983) extended Atkinson's theorem to the case of unequal means. He proposed the generalized Lorenz curve as a tool to simultaneously account for differences in income mean and income disparity. The generalized Lorenz curve is a standard Lorenz curve scaled up by the mean. This scaling reveals a dominance in the relationship not apparent for an independent checkup of means and Lorenz curves. The height of generalized Lorenz curve indicate the degree of incomes; however, the curvature of the Lorenz curve shows the degree of income inequality (Arora and Jain, 2006).

[1] Corresponding author: Department of Statistics, Pusan National University, Busan 609-390, Korea.
E-mail: xellos74@pusan.ac.kr

Published online 31 July 2014 / journal homepage: http://csam.or.kr
© 2014 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

Gastwirth (1971) presented an alternative definition of the Lorenz curve in terms of the inverse of to discrete variables as well as to continuous ones.

Let F denotes the cumulative distribution function of income distribution. The income is assumed to be non-negative. For F a given non-negative fraction p, let

$$F^{-1}(p) = \inf_{y}\{y|F(y) \geq p\}, \quad 0 \leq p \leq 1, \tag{1.1}$$

denotes the inverse distribution function corresponding to F.

It shall be assumed that F is continuous function with finite support.

The Lorenz curves corresponding to the distributions with cdf F is defined as

$$L(p) = \frac{1}{\mu}\int_0^{F^{-1}(p)} xdF(x), \tag{1.2}$$

where μ denotes the respective means of the distribution with cdf F (Gastwirth, 1971). The generalized Lorenz curve corresponding to is the Lorenz curve scaled up by the mean (Shorrocks, 1983) and is given by

$$\mathrm{GL}(p) = \mu L(p) = \int_0^{F^{-1}(p)} xdF(x), \quad 0 \leq p \leq 1. \tag{1.3}$$

Arora and Jain (2006) consider a normality using the two generalized Lorenz curves over a specified interval of interest.

In Section 2, we propose a normalized sample generalized Lorenz curve (NSGLC) that uses transformed generalized Lorenz curve (TGLC) that convert generalized Lorenz curve. We also propose a new test statistic and plot which uses NSGLC and its empirical critical values are presented. In Section 3, we compare the new plot with Q-Q plot by using Hodgkin's disease data. We also compare the new test statistic with Shapiro-Wilk test. In Section 4, Monte Carlo simulation is conducted to investigate the performance of the proposed test under several alternative distributions.

2. New Plot and Test Statistic

In Section 2, we propose a new plot and test statistic. Before proposing new plot, we propose the transformed generalized Lorenz curve (TGLC) and normalized sample generalized Lorenz curve (NSGLC).

Let $X_1, X_2, \ldots, X_n$ be a random variables with order statistics $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$. Generalized Lorenz curve assumed that X is a non-negative incomes. However, all distributions does not have non-negative support. In order to solve this, all values of the order statistic were subtracted by the value of the first order statistic, and then each result was added all together. The above result is multiplied by $(1 - p_j)$ since a generalized Lorenz curve cannot show the characteristics of the skewed distribution. Then TGLC is obtained as

$$\mathrm{TGLC}(p_j) = \frac{1}{L_X(n)}\left[\sum_{i=1}^{j} L_X(i) + (1 - p_j)\sum_{i=1}^{n} L_X(i)\right], \tag{2.1}$$

where $L_X(i) = X_{i:n} - X_{1:n}$, $p_j = j/n$ and $j = 1, 2, \ldots, n$. Using the percentile points of standard normal distribution, uniform distribution on $(0, 1)$, exponential distribution, beta distribution with two

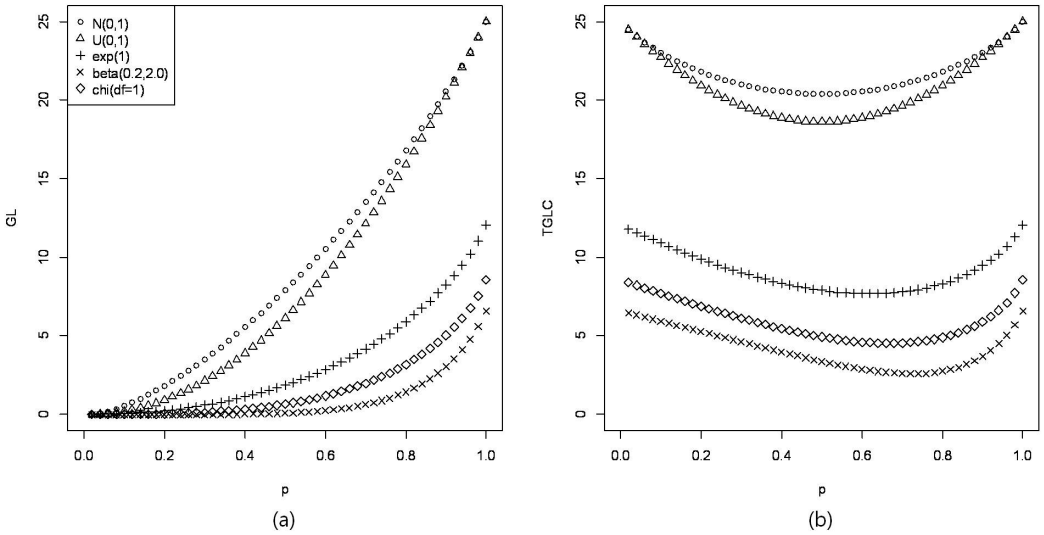

Figure 1: *(a) Generalized Lorenz curves and (b) Transformed generalized Lorenz curves*

parameters $(0.2, 2.0)$, and chi-squared distribution with 1 degrees of freedom, the results are given in Figure 1.

Let F denote the cdf of normal distribution, an NSGLC is obtained as

$$\text{NSGLC}(p_j) = \frac{\text{TGLC}_F(p_j)}{\text{TGLC}(p_j)}, \tag{2.2}$$

where $\text{TGLC}_F(p_j) = \{\sum_{i=1}^{j} T_F(i) + (1-p)\sum_{i=1}^{n} T_F(i)\}/T_F(n)$ and $T_F(i) = F^{-1}\{i/(n+1)\} - F^{-1}\{1/(n+1)\}$.

The NSGLC has the following result.

Lemma 1. *TGLC and $TGLC_F$ are a location and scale invariant statistic.*

Proof: Let X be a random variable with a location parameter μ and scale parameter σ. Let $Z = (X - \mu)/\sigma$, then $X = \mu + \sigma Z$. The distribution of Z does not depend on μ and σ.

$L_X(i)$ of TGLC is

$$\begin{aligned} L_X(i) &= (\sigma Z_{i:n} + \mu) - (\sigma Z_{1:n} + \mu) \\ &= \sigma L_Z(i). \end{aligned}$$

TGLC of NSGLC is

$$\frac{\sum_{i=1}^{j} L_X(i) + (1-p)\sum_{i=1}^{n} L_X(i)}{L_X(n)} = \frac{\sum_{i=1}^{j} L_Z(i) + (1-p)\sum_{i=1}^{n} L_Z(i)}{L_Z(n)}.$$

Let $F^{-1}(i/n+1)$ has a location parameter μ and scale parameter σ. If $G^{-1}(i/n+1) = [F^{-1}(i/n+1) - \mu]/\sigma$, then $F^{-1}(i/n+1) = \mu + \sigma G^{-1}(i/n+1)$. The distribution of $G^{-1}(i/n+1)$ does not depend on μ and σ.

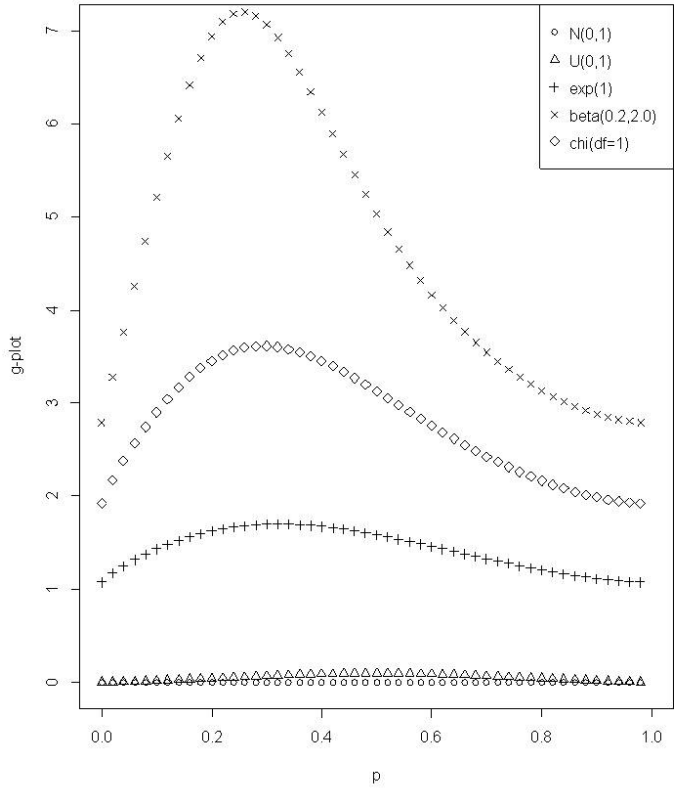

Figure 2: *g-plot of various distributions*

$T_F(i)$ of TGLC_F is

$$\begin{aligned} T_F(i) &= \left[\sigma G^{-1}\left(\frac{i}{n+1}\right)+\mu\right] - \left[\sigma G^{-1}\left(\frac{1}{n+1}\right)+\mu\right] \\ &= \sigma T_G(i). \end{aligned}$$

TGLC_F of NSGLC_F is

$$\frac{\sum_{i=1}^{j} T_F(i) + (1-p)\sum_{i=1}^{n} T_F(i)}{T_F(n)} = \frac{\sum_{i=1}^{j} T_G(i) + (1-p)\sum_{i=1}^{n} T_G(i)}{T_G(n)}.$$

Hence the required result is proved. □

Theorem 1. *NSGLC is a location and scale invariant statistic.*

Proof: Theorem 1 is straightforward by Lemma 1. □

Now, we propose new plot using NSGLC,

$$g\text{-plot}(p_j) = \left|1 - \mathrm{NSGLC}(p_j)\right|. \tag{2.3}$$

Table 1: The critical values for normality test

n	10	20	30	40	50	60	70	80	90	100
Critical value	.1888	.1222	.1002	.0872	.0785	.0733	.0691	.0657	.0630	.0606

Table 2: Number of T_4 cells/mm^3 in blood samples

Hodgkin's disease data	171, 257, 288, 295, 396, 397, 431, 435, 554, 568, 795 902, 958, 1004, 1104, 1212, 1283, 1378, 1621, 2415
Mean	823.2
SD	566.4

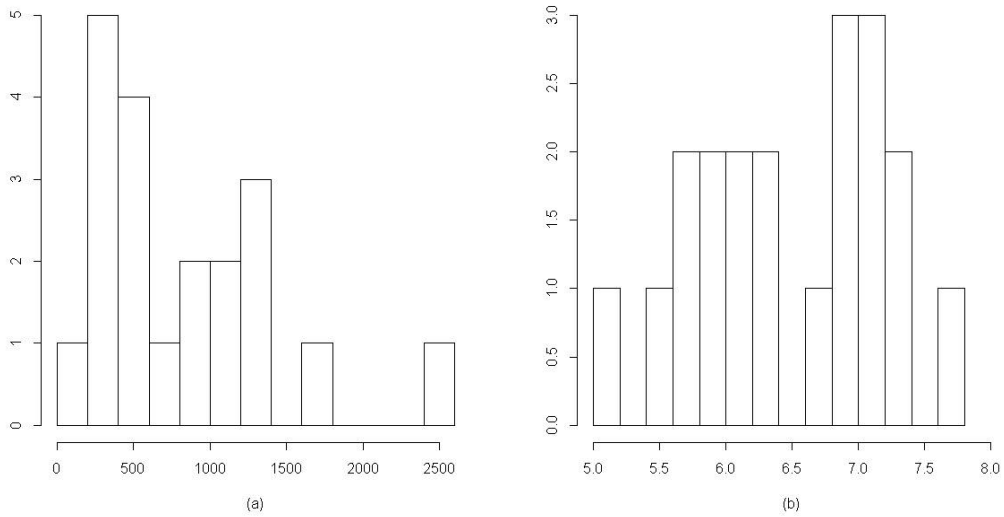

Figure 3: *Histograms of (a) Hodgkin's and (b)* ln*(Hodgkin's)*

If data is normally distributed, the NSGLC is 1 and g-plot will converge with the x-axis. Therefore, we are going to test if the data will follow the normal distribution by using the degree of how much the g-plot is apart from the x-axis.

To confirm the shape of g-plot of various distributions, we generate 50 data from $X_{i:n} = F^{-1}(i/((n+1))$ at standard normal distribution, uniform distribution on $(0, 1)$, exponential distribution, beta distribution with two parameters $(0.2, 2.0)$, and chi-squared distribution with 1 degrees of freedom. We draw the g-plot. Figure 2 shows the results of g-plot for five distributions.

Now, we propose new test statistic using NSGLC.

$$\text{GLC} = \max\left[\left|1 - \text{NSGLC}(p_j)\right|\right]. \tag{2.4}$$

If the data accurately follows a normal distribution, we expect the GLC test statistics to be small and consequently the GLC test statistics to be small. Hence, we may reject the normality if the GLC test statistics exceed the corresponding upper tail null critical values. The critical values are not available explicitly and so the percentage points need to be determined through Monte Carlo simulations since GLC test statistics have a drawback due to the difficulty in the distribution theory. Table 1 shows the 5% critical values for $n = 10$ to 100 by 10.

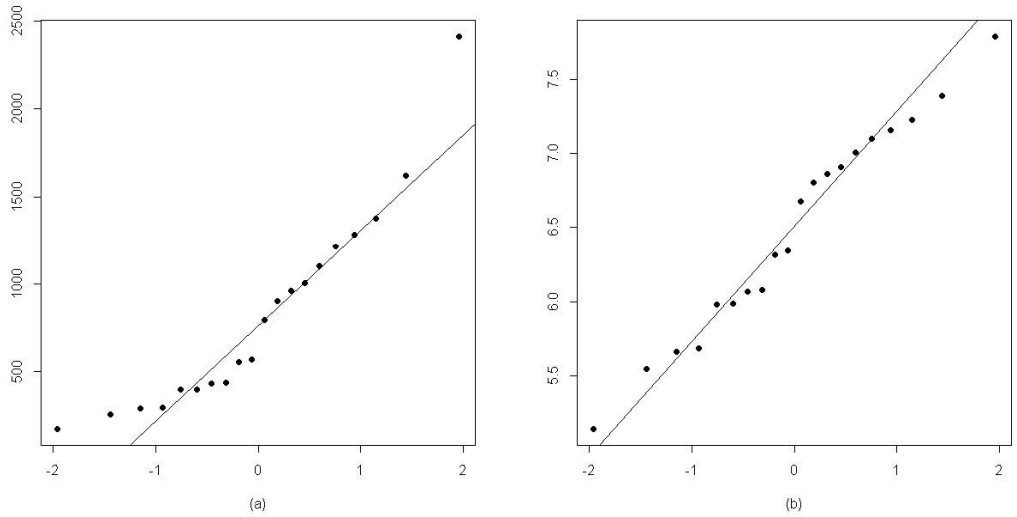

Figure 4: *Q-Q plots of (a) Hodgkin's and (b) ln(Hodgkin's)*

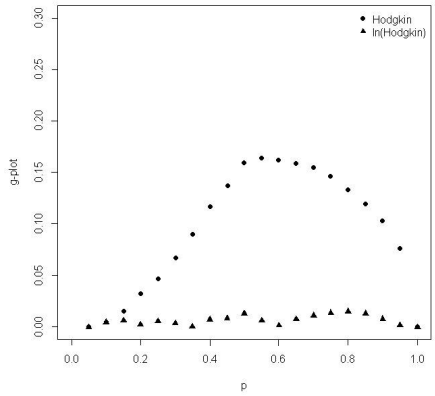

Figure 5: *g-plot of (a) Hodgkin's and (b) ln(Hodgkin's)*

3. Illustrative Example

As an illustrative example, we take Hodgkin's disease data (Alterman, 1992). It is numbers of T_4 cells per mm^3 in blood samples from 20 patients in improvement from Hodgkin's disease.

From Figure 3, distribution of the logarithm of the T cell data is more symmetric than row data.

A Q-Q plot of this data is appeared in Figure 4(a), p-value of W test is .029. Therefore, the W test conclude that the data is not normally distributed at significance 5%. A Q-Q plot of natural logarithmic data is appeared in Figure 4(b), p-value of W test is .812. Therefore, the W test conclude that the natural logarithmic data is normally distributed.

Figure 5 provides the g-plots of raw data and natural logarithmic data. From Figure 5, the g-plot of natural logarithmic data is closer to zero than g-plot of raw data. Thus, the g-plot conclude that the natural logarithmic data is normally distributed.

It is also able to confirm as the GLC test. From Table 3, the GLC of raw data is .1644 and the

Table 3: Comparison of W and GLC in Hodgkin's disease data

	W	p-value	GLC	critical value
raw data	.891	.029	.1644	.1222
natural logarithmic data	.973	.812	.0153	

Table 4: Estimated power of the GLC test the alternative distributions at the significance level 5%

Distribution	n	KS	CVM	AD	SF	W	GLC
uniform(0, 1)	10	.0656	.0720	.0803	.0510	.0843	.0867
	20	.0933	.1433	.1696	.0853	.2015	.2145
	50	.2559	.4398	.5742	.4752	.7468	.6504
$t(9)$	10	.0666	.0698	.0749	.0844	.0742	.0501
	20	.0788	.0904	.0978	.1312	.1073	.0813
	50	.0911	.1147	.1314	.2137	.1706	.1410
Cauchy	10	.5810	.6169	.6177	.6398	.5956	.4119
	20	.8451	.8796	.8809	.8926	.8684	.6860
	50	.9936	.9952	.9978	.9983	.9971	.9363
exponential	10	.2994	.3798	.4095	.4360	.4483	.5845
	20	.5769	.7222	.7729	.7980	.8344	.9317
	50	.9638	.9914	.9973	.9987	.9991	1.0000
Weibull(0.4, 0.1)	10	.8601	.9313	.9427	.9424	.9540	.9856
	20	.9963	.9990	.9997	.9999	.9999	1.0000
	50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
beta(0.2, 2.0)	10	.8227	.9077	.9241	.9261	.9411	.9854
	20	.9933	.9992	.9998	.9998	.9999	1.0000
	50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
gamma(1, 3)	10	.2976	.3821	.4111	.4314	.4408	.5800
	20	.5832	.7254	.7761	.8008	.8366	.9341
	50	.9610	.9887	.9973	.9987	.9993	1.0000
chi-squared(1)	10	.5423	.6711	.7054	.7139	.7405	.8693
	20	.8854	.9521	.9679	.9734	.9842	.9982
	50	.9998	.9999	1.0000	1.0000	1.0000	1.0000

GLC of natural logarithmic data is .0153. Therefore, the GLC test conclude that the raw data is not normally distributed and the natural logarithmic data is normally distributed.

4. Simulation Study

A Monte Carlo simulation study was conducted to determine the power under different alternatives and to assess the power of the GLC test statistics. As an alternative distribution, the eight alternative distributions were considered in the simulation study; uniform distribution on $(0, 1)$, Student's t distribution with 9 degrees of freedom, Cauchy distribution with two parameters $(1, 3)$, exponential distribution, Weibull distribution with shape parameter 0.4 and scale parameter 0.1, beta distribution with two parameters $(0.2, 2.0)$, gamma distribution with two parameter $(1, 3)$ and chi-squared distribution with 1 degrees of freedom. The GLC test statistics was compared to five other tests; Kolmogorov-Smirnov (KS), Cramer-von Mises (CVM), Anderson-Darling (AD), Shapiro-Francia (SF) and Shapiro-Wilks (W). To accomplish the power comparison, 10,000 samples of size $n = 10, 20, 50$ were generated from each of the alternative distributions, and the power of each test was estimated by counting a number of samples falling into the corresponding critical region given in Table 1. Table 4 presents the results of the empirical power of the test at the significance level 5%. The results report that the performance of GLC was generally good for all sample sizes and alternative distributions. For uniform distribution, GLC was found to be better than other test statistics except W. For t distribution with 9

degrees of freedom, GLC was found to be better than KS, CVM and AD, except $n = 10$. For the exponential, Weibull, beta, gamma and chi-squared distribution, GLC was outperformed the all different test statistics; consequently, the GLC test is a very effective test for asymmetrical distribution.

References

Alterman, D. G. (1992). *Practical Statistics for Medical Research*, Chapman and Hall, London.

Arora, S. and Jain, K. (2006). Testing for generalized Lorenz dominance, *Statistical Methods and Applications*, **15**, 75–88.

Atkinson, A. B. (1970). On the measurement of inequality, *Journal of Economic Theory*, **2**, 244–263.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve, *Econometrica*, **39**, 1037–1039.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth, *Publications of the American Statistical Association*, **9**, 209–219.

Shorrocks, A. F. (1983). Ranking income distributions, *Economica*, **50**, 3–17.

Received March 26, 2014; Revised July 3, 2014; Accepted July 16, 2014