Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 653~660

Invariance Properties for Statistics Based on the Sample Lorenz Curve

Suk-Bok Kang[1)] · Young-Suk Cho[2)]

Abstract

In this paper, we prove that the transformed sample Lorenz curve, normalized sample Lorenz curve, and the test statistics for testing of normality based on the normalized sample Lorenz curve and the modified Lorenz curve which were introduced by Kang and Cho (2001a, 2002) are location and scale invariant statistics.

1. Introduction

Consider a sample of n individuals, and let x_i be the income of individual i, $i=1,2,\ldots,n$, such that $x_1 \le x_2 \le \ldots \le x_n$. The sample Lorenz curve is the polygon joining the points $(h/n, L_h/L_n)$, $h=0,1,\ldots,n$, where $L_0=0$ and $L_h=\sum_{i=1}^{h} x_i$ is the total income of the poorest h individuals. Hence the Lorenz curve $q=L(y)$ has as its abscissa the cumulative proportion of income receivers, arrayed by increasing size of their incomes, and as its ordinate the corresponding proportion of income received. Its general representation is given by

$$L(y)=\int_0^y x\,dF(x)/E(Y) \tag{1.1}$$

where Y is a nonnegative income variable for which the mathematical expectation $\mu=E(Y)$ exists, and $p=F(y)$ is the cumulative distribution function (cdf) of the population of income receivers.

Gastwirth (1971) has given the following definition of the Lorenz curve $L(p)$;

---

1) Professor, Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea.
E-mail: sbkang@yu.ac.kr

2) Full-time Lecture, School of Free Major, Miryang National University, Kyongnam, 627-702, Korea.
E-mail : choys@mnu.ac.kr

$$L(p)=\int_0^p F^{-1}(x)dx/E(Y) \tag{1.2}$$

where

$$F^{-1}(p)=\inf_x\{x : F(x)\geq p\}. \tag{1.3}$$

If X represents annual income, $L(p)$ is the proportion of total income that accrues to individuals having the $100p\%$ lowest incomes. When all members of the population receive the same income, the Lorenz curve is the equidistribution or identity function $L(p)=p$. As the distribution becomes more unequal, the Lorenz curve bends downward and to the right within the unit square.

The Lorenz curve is extensively used in the study of inequality income distribution. It is easy to see that

(a) $L(0)=0$

(b) $L(1)=1$

(c) $L(p)\leq p$

(d) $L(p)$ is convex function.

The Lorenz curve proved to be a powerful tool for the analysis of a variety of scientific problems: e.g., (a) to measure the income inequality within a population of income receivers. (b) as a criterion to perform a partial ordering of social welfare states, (c) to assess the progressiveness of a tax system, (d) to extend the concept of the Lorenz curve to functions of income or other variables, (e) to study the stochastic properties of the sample Lorenz curve, and (f) to derive goodness-of-fit tests for exponential distribution functions.

Moothathu (1985) derived the maximum likelihood estimators (MLEs) of the Lorenz curve and the Gini index of a Pareto distribution, their exact and asymptotic distributions, and moments. Moothathu (1990) also obtained the uniformly minimum variance unbiased estimator (UMVUE) and a strongly consistent asymptotically normal unbiased estimator of the Lorenz curve, the Gini index and Theil entropy index of a Pareto distribution. Kang and Cho (1999a) proposed the several estimators of the Lorenz curve in the Pareto distribution.

Assume that $X_1, X_2, \ldots, X_n$ are positive random variables with order statistics $X_{(1)}<\cdots<X_{(n)}$. Let $r=[np]$ denote the greatest integer less than or equal to np. Then the sample Lorenz curve (Gail and Gastwirth (1978)) is defined by

$$L_n(p)=\frac{\sum_{i=1}^{r=[np]} X_{(i)}}{\sum_{i=1}^{n} X_{(i)}}. \tag{1.4}$$

Cho et al. (1999) proposed the transformed Lorenz curve that can be used in the study of symmetric distribution. The transformed Lorenz curve and the transformed sample Lorenz curve were defined by

$$TLC(p) \equiv L(p) - p + 1 \tag{1.5}$$

and

$$TSLC(p) \equiv \frac{\sum_{j=1}^{i}(X_{j:n} - X_{1:n})}{\sum_{j=1}^{n}(X_{j:n} - X_{1:n})} - p + 1, \quad p = i/n, \ i = 1, 2, \ldots, n. \tag{1.6}$$

Gail and Gastwirth (1978) studied the scale free goodness-of-fit test for the exponential distribution based on the Lorenz curve. Kang and Cho (1999b) studied the scale and location free goodness-of-fit test for the normal distribution based on the transformed sample Lorenz curve.

Kang and Cho (2000a) also studied the scale free goodness-of-fit test for the expnential distribution based on the transformed sample Lorenz curve. Kang and Cho (2000b) proposed usually powerful and easily computing test statistic for uniformity which does not depend on the unknown interval (a, b) by the transformed sample Lorenz curve.

In this paper, we prove that the transformed sample Lorenz curve, normalized sample Lorenz curve, the test statistics for testing of normality based on the normalized sample Lorenz curve which was introduced by Kang and Cho (2001a), and test statistic to test for normality based on the modified Lorenz curve are location and scale invariant statistics.

2. Invariant test statistics

Consider some distributions with the location parameter θ_1 and the scale parameter θ_2 ($\theta_2 > 0$). Let $f(x; \theta_1, \theta_2)$ and $F(x; \theta_1, \theta_2)$ be the probability density function (pdf) and cumulative distribution function of the underling distribution, respectively.

Let $X_1, \ldots, X_n$ be a random sample from the distribution with pdf $f(x; \theta_1, \theta_2)$, and let $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics.

Let

$$Y_i = \frac{X_i - \theta_1}{\theta_2}, \quad i = 1, \ldots, n \tag{2.1}$$

then the distribution of Y_i does not depend on the location parameter θ_1 and the scale parameter θ_2.

To test $H_0 : X \sim F(x; \theta_1, \theta_2)$, Kang and Cho (2001a) proposed normalized sample Lorenz curve ($NSLC$) as follows;

$$NSLC(p) = \frac{TSLC(p)}{TSLC_F(p)}, \quad p = i/n, \ i = 1, 2, \ldots, n \tag{2.2}$$

where

$$TSLC_F(p)=\frac{\sum_{j=1}^{i}[F^{-1}(j/(n+1))-F^{-1}(1/(n+1))]}{\sum_{j=1}^{n}[F^{-1}(j/(n+1))-F^{-1}(1/(n+1))]}-p+1. \tag{2.3}$$

Kang and Cho (2001b) proposed the test statistics for testing of normality that is very important test in statistical analysis based on the normalized sample Lorenz curve which was introduced by Kang and Cho (2001a) as follows;

$$TS_1=\frac{1}{n}\sum_{i=1}^{n}|1-NSLC(i/n)| \tag{2.4}$$

$$TS_2=\max_p NSLC(p)-\min_p NSLC(p). \tag{2.5}$$

Kang and Cho (2002) also proposed a new plot and test statistic to test for normality based on the modified Lorenz curve as follows;

$$Plot(p)=|NSLC_1(p)|+|NSLC_2(1-p)|,\quad p=i/n,\quad i=1,2,\ldots,n \tag{2.6}$$

where

$$NSLC_1(p) = 1-\frac{TSLC(p)}{TSLC_F(p)}, \tag{2.7}$$

$$NSLC_2(p) = 1-\frac{TSLC_1(p)}{TSLC_{F_1}(p)}, \tag{2.8}$$

$$TSLC_1(p)=\frac{\sum_{j=1}^{i}(X_{n:n}-X_{n-j+1:n})}{\sum_{j=1}^{n}(X_{n:n}-X_{n-j+1:n})}-p+1, \tag{2.9}$$

$$TSLC_{F_1}(p)=\frac{\sum_{j=1}^{i}[F^{-1}(n/(n+1))-F^{-1}((n-j+1)/(n+1))]}{\sum_{j=1}^{n}[F^{-1}(n/(n+1))-F^{-1}((n-j+1)/(n+1))]}-p+1 \tag{2.10}$$

and

$$TS=\max_p Plot(p). \tag{2.11}$$

From the equations (1.6) to (2.11), we have the following results:

Theorem 2.1. The transformed sample Lorenz curve $TSLC(p)$ is location and scale invariant statistic.

Proof. Let the transformed sample Lorenz curve for the random sample $Y_1,\ldots,Y_n$ be denoted by $TSLC_Y(p)$. Then

$$
\begin{aligned}
TSLC_Y(p) &= \frac{\sum_{j=1}^{i}(Y_{j:n}-Y_{1:n})}{\sum_{j=1}^{n}(Y_{j:n}-Y_{1:n})}-p+1 \\
&= \frac{\sum_{j=1}^{i}\left(\frac{Y_{j:n}-\theta_1}{\theta_2}-\frac{Y_{1:n}-\theta_1}{\theta_2}\right)}{\sum_{j=1}^{n}\left(\frac{Y_{j:n}-\theta_1}{\theta_2}-\frac{Y_{1:n}-\theta_1}{\theta_2}\right)}-p+1 \\
&= \frac{\sum_{j=1}^{i}\left(\frac{X_{j:n}-X_{1:n}}{\theta_2}\right)}{\sum_{j=1}^{n}\left(\frac{X_{j:n}-X_{1:n}}{\theta_2}\right)}-p+1 \\
&= \frac{\sum_{j=1}^{i}(X_{j:n}-X_{1:n})}{\sum_{j=1}^{n}(X_{j:n}-X_{1:n})}-p+1 \\
&= TSLC_X(p)
\end{aligned}
$$

which completes the proof. ◆

Theorem 2.2. The normalized sample Lorenz curve $NSLC(p)$ is location and scale invariant statistic.

Proof. Since the cdf of Y is

$$
\begin{aligned}
G(t) &= P[Y \le t] \\
&= P\left[\frac{X-\theta_1}{\theta_2} \le t\right] \\
&= P[X \le \theta_1+\theta_2 t] \\
&= F(\theta_1+\theta_2 t;\ \theta_1, \theta_2),
\end{aligned}
$$

we have the following equation;

$$\theta_1+\theta_2 G^{-1}(p) = F^{-1}(p), \quad 0 \le p \le 1$$

Therefore,

$$
\begin{aligned}
TSLC_G(p) &= \frac{\sum_{j=1}^{i}[G^{-1}(j/(n+1))-G^{-1}(1/(n+1))]}{\sum_{j=1}^{n}[G^{-1}(j/(n+1))-G^{-1}(1/(n+1))]}-p+1 \\
&= \frac{\sum_{j=1}^{i}[\theta_2 G^{-1}(j/(n+1))-\theta_2 G^{-1}(1/(n+1))]}{\sum_{j=1}^{n}[\theta_2 G^{-1}(j/(n+1))-\theta_2 G^{-1}(1/(n+1))]}-p+1
\end{aligned}
$$

$$= \frac{\sum_{j=1}^{i} [F^{-1}(j/(n+1)) - F^{-1}(1/(n+1))]}{\sum_{j=1}^{n} [F^{-1}(j/(n+1)) - F^{-1}(1/(n+1))]} - p + 1$$

$$= TSLC_F(p)$$ ◆

Corollary 2.1. The test statistics TS_1 and TS_2 are location and scale invariant statistics.

Theorem 2.3. The test statistics $NSLC_1(p)$, $NSLC_2(p)$, and TS are location and scale invariant statistics.

Proof. For the random sample $Y_1, \ldots, Y_n$, the $TSLC_2(p)$ is

$$TSLC_{1Y}(p) = \frac{\sum_{j=1}^{i} (Y_{n:n} - Y_{n-j+1:n})}{\sum_{j=1}^{n} (Y_{n:n} - Y_{n-j+1:n})} - p + 1$$

$$= \frac{\sum_{j=1}^{i} \left(\frac{X_{n:n} - \theta_1}{\theta_2} - \frac{X_{n-j+1:n} - \theta_1}{\theta_2} \right)}{\sum_{j=1}^{n} \left(\frac{X_{n:n} - \theta_1}{\theta_2} - \frac{X_{n-j+1:n} - \theta_1}{\theta_2} \right)} - p + 1$$

$$= \frac{\sum_{j=1}^{i} (X_{n:n} - X_{n-j+1:n})}{\sum_{j=1}^{n} (X_{n:n} - X_{n-j+1:n})} - p + 1$$

$$= TSLC_{1X}(p)$$

and

$$TSLC_{G_1}(p) = \frac{\sum_{j=1}^{i} [G^{-1}(n/(n+1)) - G^{-1}((n-j+1)/(n+1))]}{\sum_{j=1}^{n} [G^{-1}(n/(n+1)) - G^{-1}((n-j+1)/(n+1))]} - p + 1$$

$$= \frac{\sum_{j=1}^{i} [\theta_2 G^{-1}(n/(n+1)) - \theta_2 G^{-1}((n-j+1)/(n+1))]}{\sum_{j=1}^{n} [\theta_2 G^{-1}(n/(n+1)) - \theta_2 G^{-1}((n-j+1)/(n+1))]} - p + 1$$

$$= \frac{\sum_{j=1}^{i} [F^{-1}(n/(n+1)) - F^{-1}((n-j+1)/(n+1))]}{\sum_{j=1}^{n} [F^{-1}(n/(n+1)) - F^{-1}((n-j+1)/(n+1))]} - p + 1$$

$$= TSLC_{F_1}(p)$$

which completes the proof. ◆

Now, we perform three tests based on the test statistics TS_1, TS_2, and TS to illustrate the usefulness of the location and scale invariant test.

The following data are the numbers of T_4 cells per mm^3 in blood samples from 20 patients in remission from Hodgkin's disease (see Alterman(1992), p126).

171	257	288	295	396	397	431	435	554	568
795	902	958	1004	1104	1212	1283	1378	1621	2415

To take a simple example, we use these Hodgkin's disease data. These data were chosen because these data have been used previously in several literatures as example for testing the normality, and the p-values for the normality test of the Shapiro-Wilk for the Hodgkin's disease data and the logarithm of the Hodgkin's disease data are known as 0.031 and 0.772, respectively.

For the Hodgkin's disease data, we obtain the values of the three location and scale invariant test statistics to test normality as follows;

$$TS_1 = 0.094368,\ TS_2 = 0.164418,\ TS = 0.256582$$

For the logarithm of the Hodgkin's disease data, we obtain the values of the three location and scale invariant test statistics to test normality as follows;

$$TS_1 = 0.006637,\ TS_2 = 0.028466,\ TS = 0.051677$$

Kang and Cho (2001b, 2002) obtained the critical values of test statistics TS_1, TS_2, and TS for several sample sizes and significance levels. For $n = 20$, the cirtical values of TS_1, TS_2, and TS are given by

TS_1			TS_2			TS
$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
0.05801313	0.6882460	0.09048427	0.10392230	0.12304260	0.16650810	0.1931850

So, the normality is rejected at the significance level 0.05 for the Hodgkin's disease data, but not rejected for the logarithm of the Hodgkin's disease data. These results are identical with the results of Shapiro-Wilk test.

References

1. Altman, D. G. (1992). *Practical Statistics for Medical Research*. Chapman & Hall, New York.
2. Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999). A study on distribution based on the transformed Lorenz curve. *The Korean Journal of Applied*

Statistics, Vol. 12, 153-163.
3. Gail, M. H. and Gastwirth, J. L. (1978). A scale-free goodness-of-fit test for the exponential distribution based on Lorenz curve. *Journal of American Statistical Association*, Vol. 73, 787-793.
4. Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, Vol. 39, 1037-1038.
5. Kang, S. B. and Cho, Y. S. (1999a). Estimation of the Lorenz curve of the Pareto distribution. *The Korean Communications in Statistics,* Vol. 6(1), 285-292.
6. Kang, S. B. and Cho, Y. S. (1999b). Test of normality based on the transformed Lorenz curve. *The Korean Communications in Statistics,* Vol. 6, 901-908.
7. Kang, S. B. and Cho, Y. S. (2000a). Goodness-of-fit test for the exponential distribution based on the transformed sample Lorenz curve. *The Korean Communications in Statistics,* Vol. 7, 277-283.
8. Kang, S. B. and Cho, Y. S. (2000b). Test for the uniform distribution based on the transformed sample Lorenz curve, *Far East Journal of Theoretical Statistics*, Vol. 4, 285-295.
9. Kang, S. B. and Cho, Y. S. (2001a). A study on distribution based on the normalized sample Lorenz curve. *The Korean Communications in Statistics,* Vol. 8, 185-192.
10. Kang, S. B. and Cho, Y. S. (2001b). Test of normality based on the normalized sample Lorenz curve. *The Korean Communications in Statistics,* Vol. 8, 851-858.
11. Kang, S. B. and Cho, Y. S. (2002). More powerful test for normality based on the normalized sample Lorenz curve. *The Korean Journal of Applied Statistics,* Vol. 15, 415-421.
12. Moothathu, T. S. K. (1985) Sampling distribution of Lorenz curve and Gini index of the Pareto distribution. *Sankhyā* , Vol. 47(B), 247-278.
13. Moothathu, T. S. K. (1990). The best estimator of Lorenz curve, Gini index and Theil entropy index of Pareto distribution. *Sankhyā* , Vol. 52(B), 115-127.

[received date : May. 2003, accepted date : Jul. 2003]