The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 901-908

Test of Normality Based on the Transformed Lorenz Curve

Suk-Bok Kang[1] Young-Suk Cho[2]

Abstract

Using the Transformed Lorenz curve which is introduced by Cho et al.(1999), we propose the test statistic for testing of normality that is very important test in statistical analysis and compare the proposed test statistic with the Shapiro and Wilk's W test statistic in terms of the power of test through by Monte Carlo method.

1. 서론

실험이나 조사에서 얻은 데이터의 통계적 분포에 관한 추정은 통계학에 있어서 매우 중요하게 다루는 주제 가운데 하나이다. 특히 분포의 형태에 관한 추정은 히스토그램이나 Q-Q 플롯과 같은 그래프를 이용하여 접근하기도 하는데, 이들 연구는 Jackson et al. (1989), Endrenyi와 Patel (1991), Holmgren (1995), Lee et al. (1998), 그리고 Cho et al. (1999)등에 의해 연구되었다. 그래프를 이용한 방법 외에도 검정통계량을 이용한 대표적인 방법으로는 Kolmogorov-Smirnov 검정, Shapiro와 Wilk (1965)의 W 검정 통계량, Shapiro와 Francia (1972)의 W' 검정 통계량 등에 의해서 계속 연구되어 왔으며, 그리고 Looney (1995)는 다중정규성 (multivariate normality)에 대한 연구를 하였다.

한편, 소득불평등이 시간의 흐름에 따라 어떻게 변동했는가를 추적하거나 혹은 나라와 나라 사이의 소득분배를 비교할 때 무언가 측정할 척도가 있어야 하는데, 이런 목적으로 사용되는 도구를 불평등지표(지수)라 한다. 경제학분야에서 가장 대표적인 불평등지표는 Gini ratio (Gini coefficient)이고, 이를 구하기 위해서 Lorenz curve를 이용한다. 특정, 소득분포들의 Lorenz curve를 구하기 위해서 연구가 계속 진행되어 왔으며, Moothathu (1985a, b)는 척도모수와 위치모수를 가지는 지수분포와 위치모수와 형상모수를 가지는 파레토분포에서 Lorenz curve와 Gini index의 최우추정량에 대한 확률분포를 구하였고, 또한 Moothathu (1990)는 파레토분포에서 Lorenz curve, Gini index, 그리고 Theil Entropy Index에 대한 균일최소분산비편향추정량과 강일치 점근정규 비편향추정량도 제시하였다. 최근에는 Castillo et al. (1998)등이 주어진 데이터를 이용하여 Lorenz curve 함수와 그 모수를 추정하는 새로운 방법을 제시하였다.

본 논문의 2절에서는 일반적으로 정규성검정에 사용하는 검정통계량을 소개하고, 변환된 Lorenz curve를 이용한 새로운 검정통계량을 제시한다. 3절에서는 몬테칼로 모의실험을 통해 제시한 검정통계량과 기존의 검정통계량을 검정력 측면에서 비교한다.

1) Professor, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea
2) Adjunct Assistant Professor, Department of Statistics Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea

2. 정규성 검정에 대한 검정통계량

Shapiro와 Wilk (1965)는 소표본에서 정규성 검정을 위한 다음과 같은 검정통계량 W를 제시하였다. 표준정규분포 $N(0,1)$에서의 확률표본 $X_1, X_2, \cdots, X_n$의 순서통계량을 $X_{1:n}, X_{2:n}, \cdots, X_{n:n}$이라 하고, $\boldsymbol{m}' = (m_1, m_2, \cdots, m_n)$을 이 순서통계량의 기대값 벡터이고, $\boldsymbol{V} = (v_{ij})$를 $n \times n$ 공분산행렬이라 하자. 즉, i번째 순서통계량의 기대값은 $E(X_{i:n}) = m_i$ 이고, 공분산은 $\mathrm{Cov}(X_{i:n}, X_{j:n}) = v_{ij}$ $(i, j = 1, 2, \cdots, n)$이다.

$Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$을 n개 표본의 순서통계량이라 하고, $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$이 미지의 모평균 μ와 모분산 σ^2인 정규분포 $N(\mu, \sigma^2)$에서 추출한 표본인지를 검정하고자 한다. 만일 $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$이 정규분포 $N(\mu, \sigma^2)$에서 추출한 표본이라고 가정하면, $Y_{i:n} = \mu + \sigma X_{i:n}$, $i = 1, 2, \cdots, n$으로 표현할 수 있다. 정규성검정을 위한 검정통계량을 W라 하면,

$$W = \frac{(\sum_{i=1}^{n} a_i Y_{i:n})^2}{\sum_{i=1}^{n} (Y_{i:n} - \overline{Y})^2}$$

이다. 여기서

$$\boldsymbol{a}' = (a_1, a_2, \cdots, a_n) = \frac{\boldsymbol{m}' \boldsymbol{V}^{-1}}{(\boldsymbol{m}' \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m})^{1/2}}$$

이다.

행렬 $\boldsymbol{V}$의 계산은 Sarhan과 Greenberg (1956)가 단지 표본 $n = 20$까지만 계산하였고, 그 후 Shapiro와 Wilk (1965)는 표본 $n = 50$까지 계산하였다.

대표본에서 Gupta (1952)는 $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$를 독립으로 생각할 수 있고, 회귀직선의 기울기의 추정에서는 $\boldsymbol{V}^{-1}$를 단위행렬 $\boldsymbol{I}$로 대신 사용할 수 있다고 주장하였다. 이 결과를 이용하여 Shapiro와 Francia (1972)는 정규성 검정을 위한 다음과 같은 검정통계량 W'을 제시하였다.

$$W' = \frac{\left(\sum_{i=1}^{n} b_i Y_{i:n}\right)^2}{\sum_{i=1}^{n} (Y_{i:n} - \overline{Y})^2}$$

여기서

$$\boldsymbol{b}' = (b_1, b_2, \cdots, b_n) = \frac{\boldsymbol{m}'}{(\boldsymbol{m}' \boldsymbol{m})^{1/2}}$$

이다.

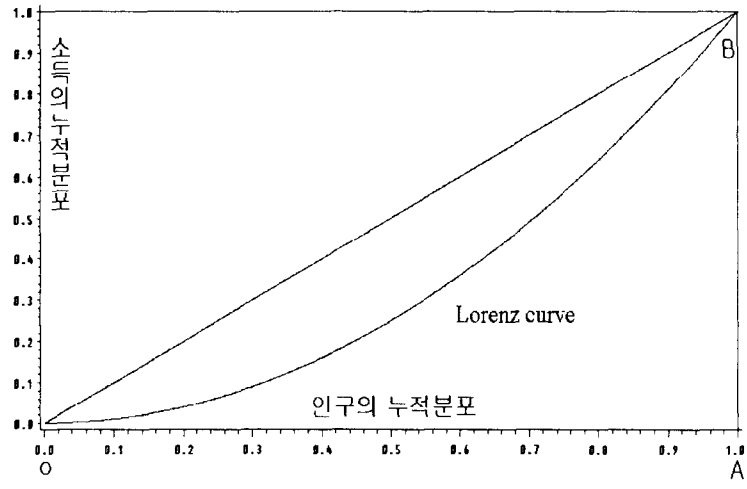

그림 2.1 : Lorenz curve

Lorenz curve는 사람들을 소득 크기대로 순서를 정한 뒤, 낮은 소득을 가진 사람부터 시작해서 수평축에 총인구에 대한 인구의 누적비를, 수직축에서는 총소득에 대한 그들의 소득 누적비를 그린 하나의 곡선이다. 그림 2.1에서 Lorenz curve는 원점 O에서 시작하여 가장 가난한 사람이 먼저 나오고 마지막에 최고의 부자가 나타남으로써 점 A에서 곡선이 끝난다. 이 곡선이 대각선 OB에 가까울수록 소득분배는 평등하고 대각선에서 멀리 떨어져 있을수록 불평등하다. 예를 들면, 현실적으로는 불가능하나 만일 모든 사람의 소득이 같다면, Lorenz curve는 대각선 OB에 일치하며, 이를 완전평등선이라 하고, 반대로 국민소득 전부를 한 사람이 가진다면 직각선 OAB가 되는데, 이것을 완전불평등선이라 한다.

이 Lorenz curve를 수학적으로 표시하면

$$L(y)=\int_0^y x\, dF(x)/E(Y) \tag{2.1}$$

이고, 여기서 Y는 기대값 $E(Y)$가 존재하는 음이 아닌 소득변수이고, $p=F(y)$는 전체 소득수입자의 누적분포함수(cdf)이다. 이와 같이 정의된 변수를 이용하여, $F(y)$를 수평축에 표시하고 $L(y)$를 수직축에 표시하여 Lorenz curve를 그릴 수 있다. $F^{-1}(p)=\inf_x\{x : F(x)\geq p\}$로 정의하면, Lorenz curve는 다음과 같이 나타날 수 있다.

$$L(p)=\int_0^p F^{-1}(x)dx/E(Y) \tag{2.2}$$

이 Lorenz curve를 이용하여 그래프적인 측면에서 특정분포의 좌우 치우침을 보다 잘 파악하기 위하여 Cho et al. (1999)는 그림 2.2와 같은 형태의 변환된 Lorenz curve ($TL(p)$)를 $TL(p)=1+L(p)-p$로 계산하고, 몇 가지 연속확률변수의 $TL(p)$를 제시하였다. 변환된 Lorenz curve의 시점과 종점이 1이므로 본 논문에서는 시점과 종점을 0으로 하고자 $TL(p)=L(p)-p$로 계산하였다.

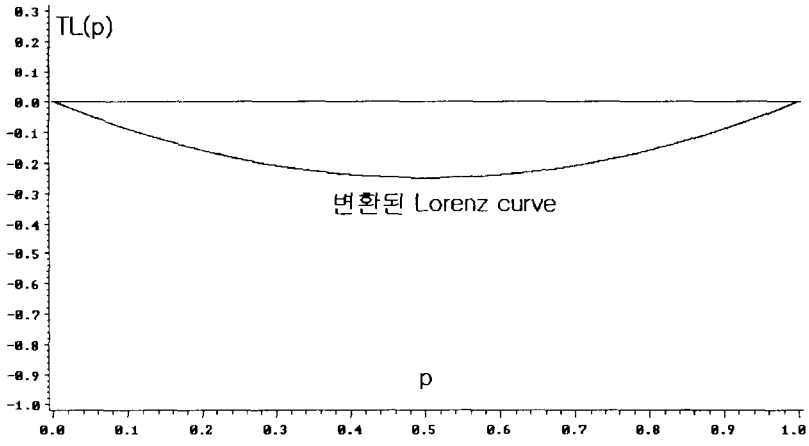

그림 2.2 : 변환된 Lorenz curve

일반적으로 확률변수의 누적분포함수가 특성함수로 표시되는 경우에 변환된 Lorenz curve를 정확히 계산 할 수 없으므로 다른 방법으로 추정해야 한다. 양의 값을 갖는 확률분포이거나 분포를 알 수 없는 경우에 다음과 같이 변환된 표본 Lorenz curve

$$TL(p)=\frac{\sum_{j=1}^{i}X_{j:n}}{\sum_{j=1}^{n}X_{j:n}}-p,\quad p=i/n,\quad i=1,2,\cdots,n \tag{2.3}$$

를 이용하고, 만일 정규분포와 같이 음수를 갖는 경우에는 표본을 양수로 변수변환하여 다음과 같이 변환된 표본 Lorenz curve를 제시하였다.

$$TL(p)=\frac{\sum_{j=1}^{i}(X_{j:n}-X_{1:n})}{\sum_{j=1}^{n}X_{j:n}-nX_{1:n}}-p,\quad 0\le p\le 1 \tag{2.4}$$

이 변환된 표본 Lorenz curve의 몇 가지 성질을 살펴보면,

(1) 변환된 표본 Lorenz curve는 수평축 아래의 볼록한 곡선 (convex curve)이 되고,

(2) $TL(0)=0$이고,

(3) $TL(1)=0$이다.

그리고 만일 데이터가 정규분포를 따른다면, $p=0.5$에서 변환된 Lorenz curve $TL(p)$의 값이 최소값을 가지며 이 점을 중심으로 좌우대칭이 됨을 보였다. 이러한 성질을 이용하여 정규성 검정을 위한 새로운 검정통계량을 제시하면, 수평축 점 $p=0.5$에서 변환된 표본 Lorenz curve의 높이를 계산한

$$TL(0.5)=\frac{\sum_{j=1}^{[n/2]}(X_{j:n}-X_{1:n})}{\sum_{j=1}^{n}X_{j:n}-nX_{1:n}}-0.5 \tag{2.5}$$

을 검정통계량으로 제시한다. 이 검정통계량을 이용하여 정규성 검정을 하면 그림 2.3과 같은 특정한 베타분포와 파레토분포를 제외하고는 표 4.1과 비슷한 결과를 얻을 수 있었다.

그러나 그림 2.3에서와 같이 좌우치우침이 있는 것 중에서 특히, 베타분포 BETA(0.4, 0.1)의 $TL(0.5)$, 파레토분포 PAR(3.0,1)의 $TL(0.5)=0.5-0.5^{2/3}$와 표준정규분포 $N(0,1)$의 $TL(0.5)$가 일치하는 특별한 경우에는 문제점을 가지고 있었다.

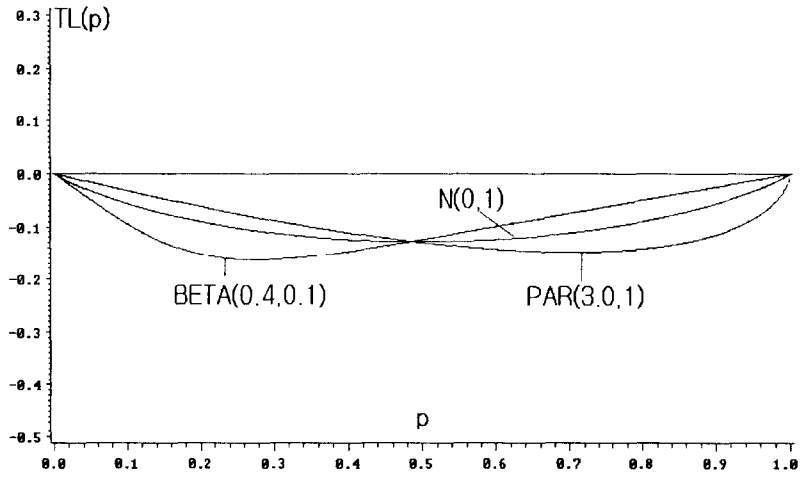

그림 2.3 : 특정분포의 변환된 Lorenz curve

이와 같은 문제점을 해결하기 위해 좌우대칭에 관한 검정을 고려한 $p=0.25$와 $p=0.75$에서 변환된 표본 Lorenz curve의 값을 계산하여 두 점에서의 차이인 $TL(0.25)-TL(0.75)$를 새로운 좌우대칭에 관한 통계량으로 제시하여 $TL(0.5)$와 동시에 검정하는 방법을 생각해 보았다.

3. 모의실험을 통한 검정력 비교

일반적으로 지금까지 통계패키지에서 사용하는 Shapiro와 Wilk의 W 검정통계량의 계산을 위하여 IMSL의 부프로그램 SPWLK를 이용한다. 이 SPWLK 부프로그램은 Royston (1982)의 몬테칼로 모의실험을 통하여 변수변환한 확률변수

$$Y=(1-W)^{\lambda}$$

(λ는 표본 수 n의 함수)가 근사적으로 정규분포를 따른다는 연구결과를 이용하여 근사적인 기각역 그리고 W에서 a_i를 계산하여 표본의 수가 3에서 2000까지에서 사용 가능한 부프로그램이다.

표 4.1의 기각역은 새로 제시한 검정통계량을 이용하여 주어진 데이터가 정규분포를 따른다는 가설에 대한 유의수준 α에서의 기각역이며, 이를 구하기 위하여 Parametric bootstrap의

bootstrap percentile를 이용하여 기각역을 구하였다. 이 방법을 간단히 소개하면, 먼저 표준정규분포 $N(0,1)$를 따르는 n개의 난수를 IMSL 부프로그램 RNNU로부터 발생하여 $TL(0.5)$와 $TL(0.25)-TL(0.75)$를 계산한다.

이 작업을 $B(=2,000)$번 반복하여 B개의 $TL(0.5)$와 $TL(0.25)-TL(0.75)$를 구한 후 이들을 작은 값부터 크기 순으로 나열하여 주어진 유의수준 α에 대하여 $B*\alpha_1/2$번째를 하한 $B*(1-\alpha_1/2)$번째를 상한으로 구했다. n개의 데이터에서 $TL(0.5)$와 $TL(0.25)-TL(0.75)$를 계산하여 위에서 구한 기각역에 포함되면 이 데이터는 정규분포를 따르지 않는다는 결정을 한다. 여기서 $\alpha_1=1-\sqrt{(1-\alpha)}$이다.

또한, 표 4.1에는 각각의 표본크기 n 과 각 문제에 대해 유의수준 0.05에서 반복횟수 1,000번의 몬테칼로 모의실험을 통하여 정규분포를 따른다는 귀무가설을 기각하는 경우의 수를 계산해 검정의 기각역을 구했다. 모의실험을 위한 분포로 좌우대칭인 t 분포와 균일분포를 이용하고, 좌우대칭이 아닌 분포 중 정규분포의 $TL(0.5)$ 값과 일치하는 베타분포와 파레토분포를 선택하였다. 그리고 지수분포와 와이블분포에서도 비교하였다.

4. 결론

경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve를 변환하여 정규성 검정을 위한 검정통계량 $TL(0.5)$와 $TL(0.25)-TL(0.75)$를 동시에 이용하는 경우와 W 검정통계량과 검정력 면에서 비교하면 표본이 작은 경우 ($n=20$) 좌우대칭인 균일분포와 t 분포에서는 조금 차이가 있다. 그러나 표본이 크면 자유도 5인 t 분포에 대해서는 우수함을 알 수 있다.

특히, 좌우 치우침이 있는 지수, 와이블, 파레토분포에 대한 $TL(0.5)$과 $TL(0.25)-TL(0.75)$를 동시에 이용한 검정통계량이 a_i를 복잡하게 계산해야 하는 W 통계량보다 계산이 간편할 뿐 아니라 검정력 측면에서도 더 우수함을 알 수 있었다.

표 4.1 :각 분포에서 $TL(0.5)$, $TL(0.25)-TL(0.75)$, W 검정통계량의 검정력 비교

n	분포	$TL(0.5)$와 $TL(0.25)-TL(0.75)$	W	기각역
20	UNIF(0,1)	.173	.195	
	t (df=5)	.123	.172	
	EXP(0,1)	.878	.834	$(-0.3205789, -0.1265670)^C$
	WIB(0.5,1.0)	1.00	.999	$(-0.0507388, 0.0773118)^C$
	BETA(0.4, 0.1)	.999	1.00	
	PAR(3, 1)	.970	.953	
100	UNIF(0,1)	.974	1.00	
	t (df=5)	.366	.333	
	EXP(0,1)	1.00	1.00	$(-0.2168230, -0.1112605)^C$
	WIB(0.5,1.0)	1.00	1.00	$(-0.0182834, 0.0216919)^C$
	BETA(0.4, 0.1)	1.00	1.00	
	PAR(3, 1)	1.00	1.00	
200	UNIF(0,1)	1.00	1.00	
	t (df=5)	.489	.417	
	EXP(0,1)	1.00	1.00	$(-0.1915111, -0.1044300)^C$
	WIB(0.5,1.0)	1.00	1.00	$(-0.0130968, 0.0145979)^C$
	BETA(0.4, 0.1)	1.00	1.00	
	PAR(3, 1)	1.00	1.00	

* 기각역에서 위는 $TL(0.5)$의 기각역이고 아래는 $TL(0.25)-TL(0.75)$의 기각역이다.

참고문헌

[1] Castillo, E., Hadi, A. S., and Sarabia, J. M. (1998). A method for estimating lorenz curves, *Communications in Statistics- Theory and Methods*, Vol. 27(8), 2037-2063.

[2] Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999), 변환된 Lorenz curve를 이용한 분포 연구, <응용통계연구>, 제12권 1호, 153-163.

[3] Endrenyi, L. and Patel, M. (1991). A new, sensitive graphical method for detecting deviations from the normal distribution of drug responses: the NTV plot, *British Journal Clinical Pharmacology*, Vol. 32, 159-166.

[4] Gupta, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample, *Biometrika*, Vol, 39, 260-273.

[5] Holmgren, E. B. (1995). The P-P plot as a method for comparing treatment effects, *Journal of American Statistical Association*, Vol. 90, 360-365.

[6] Jackson, P. R., Tucker, G. T. and Woods, H. F. (1989). Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism

histograms and probit plots, *British Journal Clinical Pharmacology*, Vol. 28, 647-653.

[7] Lee, J.-Y., Woo, J. S., and Chio, D. W. (1998). Using a normal test variable (NTV) for clinical reseach, <응용통계연구>, 제11권 1호, 1-12.

[8] Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality, *The American Statistician*, Vol. 49, 64-70.

[9] Moothathu, T. S. K. (1985a). Distribution of maximum likelihood estimators of Lorenz curve and Gini index of the exponential distribution, *Annals of Institute of Statistical Mathematics*, Vol. 37, 437-497.

[10] Moothathu, T. S. K. (1985b). Sampling distribution of Lorenz curve and Gini index of the Pareto distribution, *Sankhya*, Series B, Vol. 47, 247-278.

[11] Moothathu, T. S. K. (1990). The best estimator of Lorenz curve, Gini index and Theil entropy index of the Pareto distribution, *Sankhya*, Series B, Vol. 52, 115-127.

[12] Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples, *Applied Statistics*, Vol. 31, 115-124.

[13] Sarhan, A. E. and Greenberg, B. G. (1956). Estimation of location and scale parameters by order statistics from singly doubly censored samples part I, *Annals of Mathematical Statistics*, Vol. 27, 427-451.

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

[15] Shapiro, S. S. and Francia, R. S. (1972). An approximation analysis of variance test for normality, *Journal of American Statistical Association*, Vol. 67, 215-216.