응용통계연구 제 15권 2호, 2002년, pp. 415-421

NORMALIZED SAMPLE LORENZ CURVE를 이용한 검정력이 높은 정규성 검정 *

강 석 복 [1] 조 영 석 [2]

요 약

통계적분석에서 가장 대표적인 가정이 정규성 가정이므로 데이터의 정규성 검정은 매우 중요하다. 이 논문에서는 정규성 검정을 위해 경제학에서 소득분배의 불균형에 관한 척도로 널리 이용되는 Lorenz curve를 변형한 새로운 플롯과 검정통계량을 제시한다. 그리고 제한한 검정을 W검정 (Shapiro and Wilk (1965)), Lorenz curve를 이용한 TL검정 (Kang and Cho (1999))과 몬테칼로 방법을 이용하여 검정력을 비교한다. 제안된 검정이 특별한 대립분포의 경우를 제외하고는 대부분 검정력이 높았다.

주요용어: 검정력, 정규성검정, 표본 로렌쯔 곡선

1. 서론

데이터의 통계적 분석에서 우선 주어진 데이터가 어떤 통계적 분포를 따르는지에 관한 검정은 통계학에 있어서 매우 중요한 과제 중의 하나이다. 그래프를 이용하여 데이터의 형태에 관한 추정은 히스토그램이나 Q-Q 플롯, P-P플롯과 같은 방법이 널리 이용되어 왔으며 경제학분야에서 소득불평등의 척도로 사용되는 Lorenz curve를 이용하는 연구로 Cho et al. (1999)는 transformed Lorenz curve를 제시하였고, Kang과 Cho (2001)는 normalized sample Lorenz curve ($NSLC$)를 이용한 방법을 연구하였다.

그래프를 이용한 방법 외에도 검정통계량을 이용한 대표적인 방법으로는 Kolmogorov-Smirnov 검정, Shapiro와 Wilk (1965)의 W 검정통계량, Shapiro와 Francia (1972)의 W'검정 등이 있다. Kang과 Cho (1999)는 경제학분야에서 소득불평등의 척도로 사용되는 Lorenz curve로부터 새로운 transformed Lorenz curve를 제시하고 이를 이용하여 정규성을 검정하는 새로운 TL 검정통계량을 제시하였다. 본 논문의 2절에서는 데이터의 정규성 검정을 위해 Kang과 Cho (2001)가 제시한 normalized sample Lorenz curve를 이용한 새로운 플롯을 소개하고 이 연구를 바탕으로 정규성 검정을 위한 새로운 플롯과 검정통계량을 제시한다. 3절에서는 데이터의 정규성 검정을 위해 새로 제시한 검정 통계량과 일반적으로 가장 많이 사용하는 W검정 통계량과 Kang과 Cho (1999)의 TL 검정통계량을 몬테칼로 모의실험을 통해 검정력 측면에서 비교한다.

* 본 연구는 한국과학재단 목적기초연구(R05-2001-000-00077-0) 지원으로 수행되었음

1) (712-749) 경상북도 경산시 대동 214-1, 영남대학교 통계학과 교수
E-mail: sbkang@yu.ac.kr
2) (627-702) 경상남도 밀양시 내이동 1025-1, 밀양대학교 산업경제학과 전임강사
E-mail: choys@mnu.ac.kr

2. 정규성 검정에 대한 검정통계량과 플롯

주어진 n개 데이터 $X_1, X_2, \cdots, X_n$ 의 순서통계량을 $X_{1:n}, X_{2:n}, \cdots X_{n:n}$이라 하고, $X_1, X_2, \cdots, X_n$ 이 미지의 모평균 μ와 모분산 σ^2인 정규분포 $N(\mu, \sigma^2)$에서 추출한 표본인지를 검정하기 위해 Kang과 Cho (2001)는 그래프를 이용하여 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$에 대한 위치모수와 척도모수에 불변인 $NSLC$를 다음과 같이 제시하였다.

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)},\ p = i/n,\ i = 1, 2, \cdots, n$$

여기서

$$\begin{aligned} TSL(p) &= \frac{\sum_{j=1}^{i}(X_{j:n} - X_{1:n})}{\sum_{j=1}^{n}(X_{j:n} - X_{1:n})} - p + 1, \\ TSL_F(p) &= \frac{\sum_{j=1}^{i}\left(\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1))\right)}{\sum_{j=1}^{n}\left(\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1))\right)} - p + 1, \\ \Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \end{aligned}$$

이다. 이 곡선을 (x, y) 좌표 평면상에 $(1-p,\ 1-NSLC(p))$를 표시하는 정규성 검정을 위한 새로운 플롯으로 제시하였으며, 주어진 데이터가 정확히 정규분포를 따른다면 X축과 거의 일치하고, 좌우대칭인 분포에서는 $NSLC$가 $p = 0.5$에 대칭이고, 왼쪽으로 치우치는 분포에서는 곡선이 왼쪽으로 치우치는 경향을 알 수 있었다. 그러나 베타분포 Beta$(0.4, 0.1)$와 같이 오른쪽으로 치우치는 분포에 대해서는 X축을 교차하였으며, 베타분포 Beta$(0.1, 0.4)$와 Beta$(0.4, 0.1)$는 실제 서로 대칭이지만 $NSLC$들은 비대칭으로 나타남으로 이를 보완하고 오른쪽으로 치우치는 분포인 경우 그래프도 오른쪽으로 치우치는 그래프를 그리기 위해서 다음과 같은 새로운 플롯을 제시한다.

우선 데이터의 정규성을 생각한다면, 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$에 대한 위치모수와 척도모수에 불변인 새로운 플롯은 다음과 같다.

$$Plot(p) = |NSLC_1(p)| + |NSLC_2(1-p)|,\ p = i/n,\ i = 1, 2, \cdots, n$$

여기서

$$\begin{aligned} NSLC_1(p) &= 1 - \frac{TSL(p)}{TSL_F(p)}, \\ NSLC_2(p) &= 1 - \frac{TSL_2(p)}{TSL_F2(p)}, \end{aligned}$$

$$\begin{aligned} TSL_2(p) &= \frac{\sum_{j=1}^{i}(X_{n:n} - X_{n-j+1:n})}{\sum_{j=1}^{n}(X_{n:n} - X_{n-j+1:n})} - p + 1, \\ TSL_F2(p) &= \frac{\sum_{j=1}^{i}\left(\Phi^{-1}(n/(n+1)) - \Phi^{-1}((n-j+1)/(n+1))\right)}{\sum_{j=1}^{n}\left(\Phi^{-1}(n/(n+1)) - \Phi^{-1}((n-j+1)/(n+1))\right)} - p + 1 \end{aligned}$$

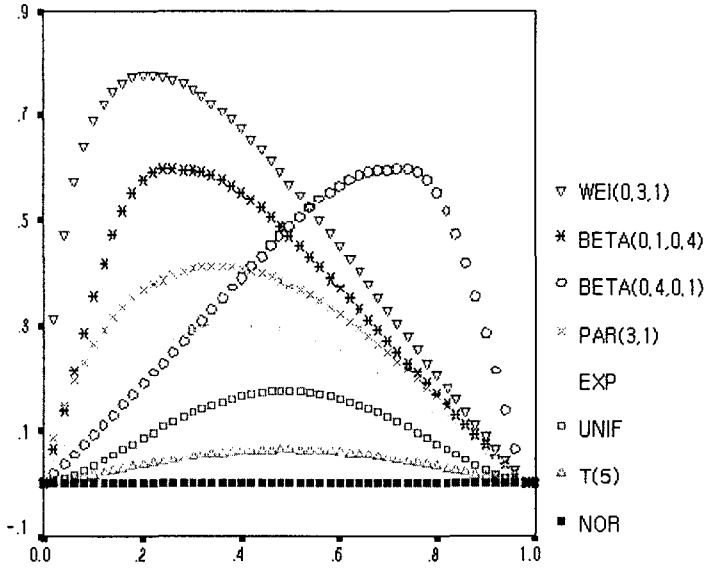

그림 2.1: 특정분포들의 플롯

이다.

이 곡선을 (x,y)좌표 평면상에 $(1-p,\ Plot(p))$를 표시하는 정규성 검정을 위한 새로운 플롯으로 제시한다. 만일 주어진 데이터가 정확히 정규분포를 따른다면 $E[\Phi(X_{i:n})] = i/(n+1)$이므로 $TSL(p)$는 $TSL_F(p)$와 $TSL_2(p)$는 $TSL_{F2}(p)$와 거의 일치 할 것이므로 플롯은 $y=0$인 X축에 가까울 것으로 기대된다. 따라서 $y=0$으로부터 이 직선이 떨어진 정도로 정규성을 판단하는 그래프를 제시한다. 우선 좌우대칭인 표준정규분포 $(N(0,1))$, t분포 $(t(5))$, 균일분포 $(U(0,1))$와 비대칭인 표준지수분포 (Exp(1)), 와이블분포 (Wei(0.3,1)), 파레토분포 (Par(3,1)), 베타분포 (Beta(0.1,0.4), Beta(0.4,0.1))에서 각각 50개의 데이터를 $X_{i:n} = F^{-1}(i/(n+1))$에 의해 생성하여 새로 제시한 플롯을 그림 2.1에 제시했다. 이 플롯으로부터 주어진 데이터가 정확히 정규분포를 따른다면 Kang과 Cho(2001)의 결과와 마찬가지로 X축에 일치하고, 좌우대칭인 분포에서는 $NSLC$가 $p=0.5$에 대칭이고, 왼쪽으로 치우치는 분포에서는 곡선이 왼쪽으로 치우친다. 그리고 오른쪽으로 치우치는 분포에 대해서도 새로운 플롯은 오른쪽으로 치우치며 서로 대칭인 베타분포 Beta(0.1,0.4)와 Beta(0.4,0.1)의 새로운 플롯은 $p=0.5$에서 서로 대칭으로 나타난다. 따라서 Kang과 Cho(2001)가 제시한 그래프의 문제점을 보완하였으므로 그래프적인 측면에서 정규성 검정을 위한 새로운 플롯으로 매우 타당하다고 생각한다.

다음은 검정통계량에 의한 정규성 검정을 하기 위해 새로운 플롯을 기초로 하여 기존의 검정통계량보다 검정력을 높일 수 있는 새로운 검정통계량으로 각 p에 대해 플롯을 계산하

여 그 중에서 가장 큰 값

$$TS = \max_p(Plot(p))$$

를 제시한다. 각 검정통계량들은 모두 양수이며 주어진 자료가 정확히 정규분포를 따른다면 검정통계량은 0이고, 그 값이 커지면 비정규분포라고 판단한다. 주어진 유의수준 α에서 새로 제시한 검정통계량들을 이용하여 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$을 검정하기 위한 기각역을 정확히 계산하기가 불가능하기 때문에 이를 계산하기 위해 parametric bootstrap방법 중 가장 간단한 bootstrap percentile 방법을 이용하여 기각역을 구하였다. 이 방법을 간단히 소개하면, 먼저 표준정규분포 $N(0,1)$를 따르는 n개의 난수를 IMSL 부프로그램 RNNOR로부터 발생하여 TS를 계산한다. 이 작업을 $B(=10,000)$번 반복하여 B개의 TS를 구한 후 이들을 작은 값부터 크기 순으로 나열하여 주어진 유의수준 α에 대하여 $B*(1-\alpha)$번째를 기각역으로 제시한다. 주어진 유의수준에서 정규성을 검정하고자 하는 새로운 n개의 데이터에 대하여 새로 제시한 검정통계량을 계산하여 그 값 이상이면 이 데이터는 정규분포를 따르지 않는다는 결정을 한다.

3. 모의실험을 통한 검정력 비교

각 분포와 표본크기 n에 대해 IMSL 부프로그램으로부터 난수를 발생하여 유의수준 0.05에서 반복횟수 10,000번의 몬테칼로 모의실험을 통하여 정규분포를 따른다는 귀무가설을 기각하는 경우의 수를 계산해 Shapiro와 Wilk의 W 검정통계량과 Kang과 Cho (1999)의 TL 검정통계량의 검정력을 구했다. Kang과 Cho의 TL 검정통계량은 변환된 표본 Lorenz curve의 값을 계산하여 두 점에서의 차이인 $TL(0.25) - TL(0.75)$를 좌우대칭에 관한 검정통계량으로 제시하여 $TL(0.5)$와 동시에 검정하는 방법이며 변환된 표본 Lorenz curve는 다음과 같다.

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

귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$의 검정을 위한 모의실험의 분포로 좌우대칭인 t분포와 균일분포를 이용하고, 좌우대칭이 아닌 분포 중 베타분포, 파레토분포, 지수분포, 그리고 와이블분포에서 비교하여 그 결과를 표 3.1에 제시하였다. 표 3.1로부터 각 경우에 대하여 검정력을 비교하면, 좌우대칭인 균일분포에서는 새로 제시한 TS 검정통계량이 다른 검정통계량보다 검정력이 조금 높으며, t분포에서는 표본이 $n = 10$, 20인 경우에 W 검정통계량이 조금 높으나 , $n = 50$에서는 새로 제시한 TS 검정통계량의 검정력이 조금 높다. 특히, 비대칭인 표준지수분포, 와이블분포, 파레토분포, 베타분포 Beta(0.1, 0.4), 베타분포 Beta(0.4, 0.1)에서는 새로 제시한 TS 검정통계량이 다른 검정통계량보다 일반적으로 더 우수함을 알 수 있다. 경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve를 변환하여 정규성 검정을 위해 제시한 검정통계량 $TL(0.5)$와 $TL(0.25) - TL(0.75)$를 동시에 이용하는 경우와 일반적으로 사용하는 W검정통계량과 새로 제시한 TS 검정통계량에 대한 각 경우에 검정력을 비교해 본 결과 두 단계를 거쳐 판단해야하는 TL 검정통계량과 복

잡하게 계산해야 하는 W통계량보다 계산이 간편할 뿐 아니라 여러 경우에 검정력 측면에서도 우수하여 정규성검정에 매우 유용함을 알 수 있었다.

표 3.1: 각 분포에서 W , TL, 그리고 TS 검정통계량의 검정력 비교(유의수준 5%)

n	본포	W	$TL(0.5)$	TS	TS의 기각역
	U(0,1)	.0726	.0710	.0910	
	t (df=5)	.1187	.0851	.0871	
	Exp(0,1)	.4415	.4410	.5015	
10	Wei(0.3,1.0)	.9849	.9946	.9997	0.2869169 이상
	Beta(0.1, 0.4)	.9170	.9486	.9986	
	Beta(0.4, 0.1)	.9696	.8490	.9839	
	Par(3, 1)	.6566	.6521	.7063	
	U(0,1)	.1986	.1410	.2991	
	t (df=5)	.1762	.1212	.1302	
	Exp(0,1)	.8315	.8548	.8856	
20	Wei(0.3,1.0)	1.000	1.000	1.000	0.1931850 이상
	Beta(0.1, 0.4)	.9998	.9999	1.000	
	Beta(0.4, 0.1)	.9999	.9977	1.000	
	Par(3, 1)	.9545	.9629	.9736	
	U(0,1)	.8868	.5447	.8879	
	t (df=5)	.2747	.2203	.3163	
	Exp(0,1)	.9998	1.000	.9999	
50	Wei(0.3,1.0)	1.000	1.000	1.000	0.1286721 이상
	Beta(0.1, 0.4)	.9998	.9999	1.000	
	Beta(0.4, 0.1)	.9999	.9977	1.000	
	Par(3, 1)	1.000	1.000	1.000	
	U(0,1)	.9999	.9689	.9996	
	t (df=5)	.3282	.3312	.5696	
	Exp(0,1)	1.000	1.000	1.000	
100	Wei(0.3,1.0)	1.000	1.000	1.000	0.0967448 이상
	Beta(0.1, 0.4)	1.000	1.000	1.000	
	Beta(0.4, 0.1)	1.000	1.000	1.000	
	Par(3, 1)	1.000	1.000	1.000	

참고문헌

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

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

[3] Kang, S. B. and Cho, Y. S. (1999). Test of normality based on the transformed Lorenz curve. *The Korean Communications in Statistics,* Vol. 6, 901-908.

[4] Kang, S. B. and Cho, Y. S. (2001). Test of normality based on the normalized sample Lorenz curve. *The Korean Communications in Statistics,* Vol. 8, 851-858.

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

[6] 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.

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

[2002년 3월 접수, 2002년 7월 채택]

More Powerful Test for Normality Based on the Normalized Sample Lorenz Curve *

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

ABSTRACT

Because most common assumption is normality in statistical analysis, testing normality is very important. We propose a new plot and test statistic to test for normality based on the modified Lorenz curve that is proved to be a powerful tool to measure the income inequality within a population of income receivers. We also compare the proposed test statistics with the W test (Shapiro and Wilk (1965)), TL test (Kang and Cho (1999)) in terms of the power of test through by Monte Carlo method. The proposed test is more usually powerful than the other tests except some case.

Keywords: Power; Sample Lorenz curve; Test of normality.

* This work was supported by grant No. R05-2001-000-00077-0 from the Basic Research Program of the Korea Science & Engineering Foundation

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

2) Senior Lecturer, Department of Applied Economics, Miryang National University.
E-mail: choys@mnu.ac.kr