

Test of Normality Based on the Normalized Sample Lorenz Curve¹⁾

Suk-Bok Kang²⁾ Young-Suk Cho³⁾

Abstract

Using the normalized sample Lorenz curve which is introduced by Kang and Cho (2001), we propose the test statistics for testing of normality that is very important test in statistical analysis and compare the proposed test with the other tests in terms of the power of test through by Monte Carlo method. The proposed test is more power than the other tests except some cases

Keywords : Bootstrap percentile, Lorenz curve, Normality, Normalized sample Lorenz curve, Power,

1. 서론

일반적으로 자료의 통계적 분석에서 데이터의 정규성에 관한 검정은 매우 중요한 가정이며 매우 일반화된 가정이라고 할 수 있다. 따라서 지금껏 수많은 학자들이 이 연구에 심혈을 기울여 왔다. 특히 분포의 형태에 관한 추정으로 히스토그램이나 Q-Q 플롯, P-P플롯 등과 같은 그래프를 이용한 연구 또한 활발하였지만, 판단이 다소 주관적인 단점을 보완하기 위해 검정통계량을 이용한 검정에 관해 많은 연구가 역시 진행되어 왔다. 검정통계량을 이용한 대표적인 방법으로는 Kolmogorov-Smirnov 검정, Shapiro와 Wilk (1965)의 검정, Shapiro와 Francia (1972)의 검정 등이 있다. 특히, Kang과 Cho (1999)는 경제학분야에서 소득불균형의 척도에 사용되는 Lorenz curve를 이용하여 변환된 Lorenz curve ($TL(p)$)를 소개하고 $TL(0.25) - TL(0.75)$, $TL(0.5)$ 을 동시에 이용하여 정규성을 검정하는 검정방법을 연구하였다.

본 논문의 2절에서는 일반적으로 정규성 검정에 이용되는 Shapiro와 Wilk (1965)의 W 검정통계량과 Kang과 Cho (1999)의 검정통계량을 소개하고, Kang과 Cho (2001)가 그래프를 통하여 특정분포의 적합도검정을 위해 제시한 normalized sample Lorenz curve를 이용한 새로운 검정통계량을 제시한다. 3절에서는 몬테칼로 모의실험을 통해 제시한 검정통계량들과 기존의 검정통계량들

1) 본 연구는 한국과학재단 목적기초연구(2001-1-10400-014-2)지원으로 수행되었음.

2) Professor, Department of Statistics and Institute of Natural Sciences, Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea
E-mail : sbkang@yu.ac.kr

3) Adjunct Assistant Professor, Department of Statistics Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea
E-mail : choys@yu.ac.kr

을 검정력 측면에서 비교한다.

2. 정규성 검정에 대한 검정통계량

주어진 n 개 데이터 X_1, X_2, \dots, X_n 순서통계량을 $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ 라 하고, X_1, X_2, \dots, X_n 이 미지의 모평균 μ 와 모분산 σ^2 인 정규분포 $N(\mu, \sigma^2)$ 에서 추출한 표본인지를 검정하고자 한다. Cho et al. (1999)는 경제학분야에서 소득불균형의 척도에 이용되는 Lorenz curve를 이용하여, 그래프적인 측면에서 특정분포의 좌우 치우침을 보다 잘 파악하기 위하여 다음과 같이 변환된 표본 Lorenz curve를 제시하였다.

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

그리고 Kang와 Cho (1999)는 만일 데이터가 정규분포를 따른다면, $p=0.5$ 에서 변환된 Lorenz curve $TL(p)$ 의 값이 최소이며 이 점을 중심으로 좌우대칭임을 보였다. 이러한 성질을 이용하여 정규성 검정을 위해 수평축 점 $p=0.5$ 에서 변환된 표본 Lorenz curve의 높이를 계산한

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

을 검정통계량으로 제시하였다. 그러나 좌우치우침이 있는 것 중에서 특히, 베타분포 BETA (0.4, 1.1) $TL(0.5)$ 파레토분포 PAR(3.0, 1) $TL(0.5) = 0.5 - 0.9^{2/3}$ 표준정규분포 $N(0, 1)$ 의 $TL(0.5)$ 가 일치하는 특별한 경우에는 문제점을 가지고 있어, 이와 같은 문제점을 해결하기 위해 좌우대칭에 관한 검정을 고려한 $p=0.25$ 과 $p=0.75$ 에서 변환된 표본 Lorenz curve의 값을 계산하여 두 점에서의 차이인 $TL(0.25) - TL(0.75)$ 새로운 좌우대칭에 관한 통계량으로 제시하여 $TL(0.5)$ 와 동시에 검정하는 방법을 제안하였다. 이를 TL 검정통계량이라 하자. Cho et al. (1999)가 제안한 정규성 검정은 귀무가설 $H_0: X \sim N(\mu, \sigma^2)$ 에 대하여 변환된 Lorenz curve를 그린 다음 검정하고자 하는 데이터의 변환된 Lorenz curve를 그려 비교해야 하는 번거로움이 있었다. 이를 보완하여 Kang과 Cho (2001)는 그래프를 이용하여 귀무가설 $H_0: X \sim F(x)$ 를 검정하기 위해 위치모수와 척도모수에 불변인 다음과 같은 새로운 Normalized Sample Lorenz Curve (NSLC)를 제시했다.

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

여기서

$$\begin{aligned} TSL(p) &= \frac{\sum_{j=1}^i (X_{j:n} - X_{1:n})}{\sum_{j=1}^n (X_{j:n} - X_{1:n})} - p + 1 \\ TSL_F(p) &= \frac{\sum_{j=1}^i (F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))}{\sum_{j=1}^n (F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))} - p + 1 \end{aligned}$$

이다. 이 곡선을 (x, y) 좌표 평면상에 $(1-p, 1-NSLC(p))$ 표시하는 새로운 플롯으로 제안했다. 데이터가 귀무가설 $H_0: X \sim F(x)$ 를 따른다면, 이 NSLC에서의 기대되는 직선은 $y=0$ 이다. 따라서 이 직선으로부터 떨어진 정도로 귀무가설 $H_0: X \sim F(x)$ 를 판단하는 그래프적인 방법을 제안하였다.

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

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

여기서

$$TSL_F(p) = \frac{\sum_{j=1}^i (\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1)))}{\sum_{j=1}^n (\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1)))} - p + 1$$

이다.

우리는 이 곡선을 이용하여 검정통계량에 의한 정규성 검정을 하기 위해 우선 표준정규(NOR), t(STT(3)), 균일(UNIF), 지수(EXP(0, 1)), 와이블(WEI(0.3, 1)), 파레토(PAR(3, 1)), 베타분포(BETA(0.1, 0.4), BETA(0.4, 0.1))에서 각각 50개의 데이터를 생성하여 그런 NSLC들을 그림 2.1에 나타내었다. 이 그림으로부터 주어진 데이터가 정확히 정규분포를 따른다면 x 축에 일치하고, 좌우대칭인 분포에서는 NSCL가 $p=0.5$ 관해 대칭이고, 왼쪽으로 치우치는 분포에서는 곡선이 왼쪽으로 치우치고, 오른쪽으로 치우치는 분포에 대해서는 x 축을 교차함으로 정규성 검정에서 검정력을 높일 수 있는 새로운 검정통계량들을 다음과 같이 제안할 수 있다.

$$\begin{aligned} TS_1 &= \sum_{i=1}^n |NSLC(i/n)| \\ TS_2 &= \text{MAX}_p(NSLC(p)) - \text{MIN}_p(NSLC(p)) \end{aligned}$$

각 검정통계량들은 모두 양수이며 주어진 자료가 정확히 정규분포를 따른다면 검정통계량의 값은 0이고, 그 값이 커지면 비정규분포라고 판단한다. 주어진 유의수준 α 에서 새로 제시한 검정통계량들을 이용하여 귀무가설 $H_0: X \sim N(\mu, \sigma^2)$ 을 검정하기 위한 기각역은 정확히 계산하기가 어렵기 때문에 Parametric bootstrap 방법 중 가장 간단한 bootstrap percentile 방법을 이용하여 기각역을 구하였다. 이 방법은 먼저 표준정규분포 $N(0,1)$ 를 따르는 n 개의 난수를 IMSL 부프로그램 RNNUR로부터 생성하여 TS_1 과 TS_2 를 각각 계산한다. 이 작업을 $B = 10,000$ 번 반복하여 B 개의 TS_1 과 TS_2 를 각각 구한 후 이들을 작은 값부터 크기 순으로 나열하여 주어진 유의수준 α 에 대하여 $B*(1-\alpha)$ 번째를 각각 기각역으로 표 2.1에 제시했다. 주어진 유의수준에서 정규성을 검정하고자 하는 새로운 n 개의 데이터에 대하여 새로 제시한 검정통계량을 계산하여 표 2.1의 값 이상이면 이 데이터는 정규분포를 따르지 않는다는 결정을 한다.

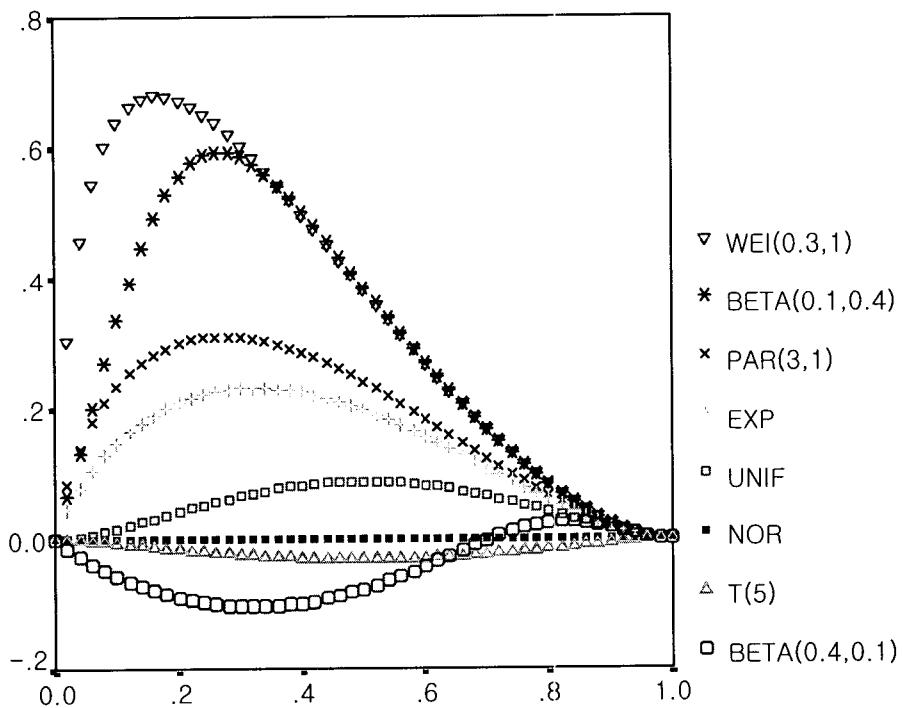


그림 2.1 특정분포의 NSLC

표 2.1 TS_1 와 TS_2 의 기각역

n	TS_1			TS_2		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
10	.07684759	.09017450	.12132010	.15200690	.18274790	.25311030
12	.07051306	.08435573	.11276400	.13701370	.16338190	.23035360
14	.06678420	.07970364	.10338800	.12571040	.14790820	.20882970
16	.06279560	.07394108	.09843467	.11607400	.13663300	.19040200
18	.06080971	.07190912	.09572175	.11053500	.13058730	.18076190
20	.05801313	.06882460	.09048427	.10392230	.12304260	.16650810
22	.05653959	.06673748	.08673760	.10060270	.11688630	.15702470
24	.05396953	.06366716	.08405161	.09520203	.11187290	.15062250
26	.05264060	.06204909	.07990111	.09207064	.10689770	.14123310
28	.05156487	.06006413	.07871835	.08941793	.10359560	.13483850
30	.05009012	.05884550	.07643305	.08524120	.10033630	.13383850
32	.04835622	.05693956	.07266425	.08276516	.09617245	.12198200
34	.04848027	.05688310	.07221615	.08235490	.09481555	.12211040
36	.04682245	.05457366	.06949200	.07961100	.09179437	.11843210
38	.04593222	.05379175	.06893744	.07694757	.08918035	.11481220
40	.04488901	.05286945	.06811346	.07553077	.08717859	.11385550
42	.04431644	.05213464	.06639955	.07388461	.08614385	.11191200
44	.04309854	.05060637	.06539948	.07163000	.08333129	.10616640
46	.04276692	.04985164	.06426141	.07090628	.08249491	.10356720
48	.04196423	.04927063	.06324614	.06925207	.08037424	.10171800
50	.04148623	.04880487	.06240071	.06858373	.07935107	.10128860

3. 모의실험을 통한 검정력 비교

일반적으로 정규성검정은 Shapiro와 Wilk (1965)가 지금까지 발표된 여러 검정통계량들과 W 검정통계량을 비교하여 대부분의 경우 W 검정통계량의 검정력이 높음을 보였고 여러 통계패키지에서 널리 사용하고 있기 때문에 W 검정통계량과 본 논문에서 새로 제시한 검정통계량의 검정력을 비교하고자 한다. Shapiro와 Wilk의 W 검정통계량의 계산을 위하여 IMSL의 부프로그램 SPWLK를 이용한다. SPWLK 부프로그램은 Royston (1982)의 몬테칼로 모의실험을 통하여 변수변환한 확률변수

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

(λ 는 표본 수 n 의 함수)가 근사적으로 정규분포를 따른다는 연구결과를 이용하여 표본의 수가 3에서 2000까지에서 근사적인 기각역 그리고 W 에서 a_i 의 계산을 가능하게 하는 부프로그램이다. 표 3.1에는 각 분포와 표본크기 n 대해 IMSL 부프로그램으로부터 난수를 발생하여 유의수준 0.05에서 반복횟수 10,000번의 몬테칼로 모의실험을 통하여 정규분포를 따른다는 귀무가설을 기각하는 경우의 수를 계산해 각 검정통계량의 검정력을 구했다. 모의실험을 위한 분포로 좌우대칭인 t 분포, 균일분포, normal mixture(1)($N(0,1)$ 와 $N(2,1)$), 그리고 normal mixture(2)($N(0,1)$ 와 $N(4,1)$)를 이용하고, 좌우대칭이 아닌 분포 중 χ^2 분포, 베타분포, 파레토분포, 지수분포, 와이블분포, 그리고 평균과 분산이 존재하지 않는 코시분포에서 비교하였다.

표 3.1로부터 각 경우에 대하여 검정력을 비교해보면, 좌우대칭인 균일분포에서는 새로 제시한 TS_2 에 의한 검정이 다른 검정보다 검정력이 조금 높으며, t 분포에서는 W 에 의한 검정이 검정력이 조금 높으며, 그리고 새로 제안한 검정통계량 중에는 TS_2 에 의한 검정보다는 TS_1 에 의한 검정이 검정력이 조금 높다. 특히, 우측으로 치우침이 있는 베타분포 BETA(0.4, 0.1)와 평균과 분산이 존재하지 않는 코시분포에서는 W 검정통계량이 더 우수하나, 치우침이 있는 χ^2 , log normal, 지수, 와이블, 파레토, 베타분포 BETA(0.1, 0.4)에서는 새로 제시한 검정통계량 TS_2 가 다른 검정통계량보다 더 우수하다. 정규 혼합분포인 경우 $n=1$ 에서는 검정통계량 TS_2 가 조금 더 우수하고, $n=2$ 에서 Normal mixture(1)인 경우는 TS_2 , Normal mixture(2)인 경우는 W 검정통계량이 다른 검정통계량보다 조금 더 우수함을 알 수 있었다.

경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve를 변환하여 정규성 검정을 위해 제안한 $TL(0.5)$ 와 $TL(0.25) - TL(0.75)$ 동시에 이용한 TL 검정통계량과 널리 이용되는 W 검정통계량은 각각 두 번을 판단해야하는 번거로움과 a_i 를 복잡하게 계산해야 하는 불편함이 있다. 새로 제안한 TS_2 검정통계량은 간편할 뿐 아니라 여러 경우에 검정력 측면에서도 우수함을 알 수 있었다. 그러나 좌우 대칭인 t 분포, 좌측으로 치우침이 있는 BETA(0.4, 0.1)그리고 코시분포의 경우 검정력이 떨어지고 있으므로 이에 대한 연구가 필요하다고 생각한다.

표 3.1 :각 분포에서 W , TL , TS_1 , 그리고 TS_2 검정통계량의 검정력 비교

n	분포	W	TL	TS_1	TS_2
10	UNIF(0,1)	.0726	.0710	.0792	.0985
	t (df=5)	.1187	.0851	.0943	.0737
	EXP(0,1)	.4415	.4410	.5976	.6080
	WIB(0.3,1.0)	.9849	.9946	.9983	.9978
	BETA(0.1, 0.4)	.9170	.9486	.9586	.9709
	BETA(0.4, 0.1)	.9696	.8490	.4547	.5196
	PAR(3, 1)	.6566	.6521	.7722	.7840
	$\chi^2(2)$.4415	.4410	.5976	.6080
	$\chi^2(4)$.2377	.1921	.3027	.3364
	$\chi^2(10)$.1223	.0810	.1465	.1643
	Log normal	.6066	.5716	.8991	.9497
	Cauchy	.6031	.4722	.4431	.4560
Normal mixture(1)	Normal mixture(1)	.0292	0.306	.0391	.0459
	Normal mixture(2)	.0915	.0892	.0479	.1278
20	UNIF(0,1)	.1986	.1410	.1437	.2083
	t (df=5)	.1762	.1212	.1271	.1219
	EXP(0,1)	.8315	.8548	.9268	.9323
	WIB(0.3,1.0)	1.000	1.000	1.000	1.000
	BETA(0.1, 0.4)	.9998	.9999	.9998	1.000
	BETA(0.4, 0.1)	.9999	.9977	.5032	.9196
	PAR(3, 1)	.9545	.9629	.9855	.9856
	$\chi^2(2)$.8314	.8548	.9262	.9309
	$\chi^2(4)$.5263	.4779	.5750	.6204
	$\chi^2(10)$.2412	.1714	.2183	.2585
	Log normal	.9350	.9335	.9621	.9666
	Cauchy	.8597	.7015	.6222	.7679
Normal mixture(1)	Normal mixture(1)	.0411	.0385	.0497	.0620
	Normal mixture(2)	.3567	.1890	.1396	.3185

참고문헌

- [1] Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999), 변환된 Lorenz curve를 이용한 분포 연구, <응용통계연구>, 제12권 1호, 153-163.
- [2] Kang, S. B. and Cho, Y. S. (1999). Test of Normality Based on the Transformed Lorenz Curve. *The Korean Communications in Statistics*, Vol. 6(3), 901-908.
- [3] Kang, S. B. and Cho, Y. S. (2001). A study on Distribution Based on the Normalized Sample Lorenz Curve. *The Korean Communications in Statistics*, Vol. 8(1), 185-192.

- [4] Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples, *Applied Statistics*, Vol. 31, 115–124.
- [5] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, Vol. 52, 591–611.
- [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.