응용통계연구 제 **20**권 *1*호, *2007*년, *pp. 79–89*

쿨백-라이블러 판별정보에 기반을 둔 정규성 검정의 개선*

최병진[1)]

요 약

Arizono와 Ohta(1989)에 의해 소개된 정규성 검정은 쿨백-라이블러 판별정보를 이용하고 있으며, 검정통계량의 유도에 기반이 되는 판별정보의 추정량을 얻기 위해 Vasicek(1976)의 표본엔트로피와 분산의 최대가능도 추정량을 사용했다. 그런데 두 추정량은 편향성을 가지게 되므로 보다 정확한 판별정보의 추정을 위해 비편향 추정량을 사용하는 것이 바람직하다. 본 논문에서는 편향을 수정한 엔트로피 추정량과 분산의 균일최소분산비편향 추정량을 사용하여 판별정보의 추정량을 구하고 이로부터 유도되는 검정통계량을 사용하는 개선된 정규성 검정을 제시한다. 제안한 검정의 특성을 규명하고 검정력 비교를 위해서 모의실험을 수행한다.

주요용어: 적합도, 정규성, 엔트로피, 쿨백-라이블러 판별정보, 검정력.

1. 서론

많은 연구자들은 현실세계의 다양하고 복잡한 현상을 설명하기 위해 분포를 가정한 모형을 설정하고 실험이나 조사에 의해 수집된 자료의 분석을 통해 중요한 결론을 도출하게 된다. 확률분포들 중에서 특히 정규분포는 수학적인 처리의 용이성과 유도된 결과의 우수성 등 좋은 특성들을 제공하기 때문에 통계학을 포함한 다양한 응용분야에서 중요한 역할을 담당하고 있을 뿐만 아니라 개발된 대부분의 통계적 분석 방법들은 자료의 정규성 가정을 전제로 하고 있다. 그런데 정규분포가 분석 모형으로 여러 장점을 가진다 하더라도 연구 중의 근원현상에 대한 적절한 모형인지를 자료의 적합도 평가를 통해 타당성을 검토해 보는 것은 중요하다. 이런 이유로 정규성 검정에 관한 많은 연구가 있어 왔고 지금까지 개발된 검정들은 카이제곱 형태의 검정, 경험적 분포함수에 기반한 검정, 적률에 관련된 검정, 회귀와 상관에 의한 검정과 기타 검정 등의 범주로 분류를 할 수가 있다. 이들 검정들에 대한 세부적인 논의는 D'Agostino와 Stephens(1986)의 9장을 참고하기 바란다.

한편 정보이론에서 불확실성의 측도로 사용되는 Shannon(1948)의 엔트로피를 적합도 검정에 이용하려는 시도가 있어 왔다. Vasicek(1976)은 엔트로피의 추정량으로 표본엔트로피를 제안하고 이를 이용하여 정규분포의 최대엔트로피 특성에 기초한 검정을 개발하였다. Arizono와 Ohta(1989)는 엔트로피의 확장된 개념으로 쿨백-라이블러(Kullback-Leibler) 판별정보를 이용한 검정을 연구한 바가 있다. 그런데 Vasicek(1976)과 Arizono-Ohta(1989)의

* 본 연구는 2005학년도 경기대학교 학술연구비(신진연구과제) 지원에 의하여 수행되었음.

1) (443-760) 경기도 수원시 영통구 이의동 산94-6, 경기대학교 경상대학 경제학부 응용정보통계학전공, 조교수
E-mail: bjchoi92@kyonggi.ac.kr

검정은 정규성에 대한 복합가설에서는 동일한 것으로 알려져 있고 이들 검정은 다른 분포에 대한 적합도 검정의 개발에도 많은 영향을 주었다. 이에 관해서는 Dudewicz와 van der Meulen(1981), Chandra 등(1982), Gokhale(1983), Ebrahimi 등(1992), 김종태와 이우동(1998), Kim 등(1999)을 참고하기 바란다.

본 논문에서는 판별정보에 기반을 둔 정규성에 대한 Arizono-Ohta 검정을 개선시키고자 한다. Arizono와 Ohta(1989)는 검정통계량의 도출을 위해 Vasicek(1976)의 표본엔트로피와 분산의 최대가능도추정량을 사용하고 있다. 그런데 이들 추정량들은 작은 표본에서는 편향을 가지게 되므로 검정력에 어느 정도 영향을 주게 될 것으로 짐작된다. 따라서 편향을 수정한 엔트로피 추정량과 분산의 균일최소분산비편향 추정량을 사용한 새로운 검정통계량을 제시하고 기존의 검정통계량과 검정력 관점에서 성능을 비교하고자 한다. 본 논문의 구성은 다음과 같다. 2절에서는 쿨백-라이블러 정보에 기반한 개선된 검정통계량을 제시하고 그 특성을 조사한다. 3절에서는 모의실험에 의해 추정된 표본크기에 따른 검정통계량의 기각값을 제시한다. 또한 표에서 제공되지 않는 표본크기에 대한 기각값을 근사적으로 얻기 위한 계산공식을 제시한다. 4절에서는 여러 대립분포 하에서 검정통계량들의 성능을 비교하기 위해 모의실험을 수행한다. 5절에서는 제안한 검정의 수행 과정을 실제자료를 통해 설명하고 6절에서는 결론을 맺는다.

2. 정규성에 대한 Arizono-Ohta 검정의 개선

Arizono와 Ohta(1989)의 정규성 검정은 밀도함수로 각각 $f(x)$와 $g(x)$를 가지는 두 분포 $F(x)$와 $G(x)$에 대해 Kullback과 Leibler(1951)가 정의한 다음의 판별정보(이하 KL 판별정보)

$$I(g:f)=\int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{f(x)}\,dx \tag{2.1}$$

에 기반을 두고 있다. KL 판별정보는 두 분포간의 불일치도를 나타내는 상대 엔트로피로 그 값은 항상 0보다 크거나 같게 되며 두 분포가 동일할 때 0이 되는 특성을 가진다.

$X_1,\ldots,X_n$을 임의의 확률밀도함수 $g(x)$를 가지는 분포함수 $G(x)$로부터 추출된 크기 n의 확률표본이라고 하고 $F\left(x;\mu,\sigma^2\right)$를 평균 μ와 분산 σ^2인 정규확률밀도함수 $f\left(x;\mu,\sigma^2\right)$을 가지는 정규분포함수라고 하자. 주어진 표본에 대한 영가설 $H_0: X_1,\ldots,X_n \sim F\left(x;\mu,\sigma^2\right)$ 검정을 위한 KL 판별정보는

$$I(g:f)=\int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{f\left(x;\mu,\sigma^2\right)}\,dx \tag{2.2}$$

가 되며 영가설이 참이라면 KL 판별정보는 아주 작은 값을 가지게 되고 대립가설이 참이면 큰 값을 가지게 되므로 정규성에 대한 검정기준으로 사용할 수가 있다. 그런데 판별정보 (2.2)에서 적분을 하려면 $g(x)$의 형태가 완전하게 규정되어야만 하지만 대부분의 경우에는 알 수 없기 때문에 추정을 해야 한다. 일반적으로 KL 판별정보의 직접적인 추정은 어렵지만 식 (2.2)는

$$I(g:f)=-H(g)+\frac{1}{2}\log(2\pi\sigma)+\frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^2 g(x)\,dx \tag{2.3}$$

로 쓸 수 있으므로 식 (2.3)에 있는 우측 항들의 개별적인 추정에 의해 $I(g:f)$를 추정해 낼 수가 있다. 여기서 $H(g)$는 대립분포 $G(x)$의 엔트로피로 다음과 같이

$$H(g) = -\int_{-\infty}^{\infty} g(x) \log g(x)\ dx \tag{2.4}$$

로 정의된다. 평균 μ와 분산 σ^2이 알려져 있는 단순 영가설 하에서 $I(g:f)$의 추정량은 $H(g)$에 대한 표본엔트로피와 $\int_{-\infty}^{\infty} \{(x-\mu)/\sigma\}^2 g(x)\,dx$에 대한 일치추정량을 사용하면

$$I_{m,n} = -H_{m,n} + \frac{1}{2}\log(2\pi\sigma) + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \tag{2.5}$$

로 얻을 수가 있다. 여기서 $H_{m,n}$은 표본엔트로피로 다음의 형태

$$H_{m,n} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}\left[X_{(i+m)} - X_{(i-m)}\right]\right\} \tag{2.6}$$

를 가지게 된다. $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$는 확률표본에 대한 순서통계량이 되며 $j < 1$이면 $X_{(j)} = X_{(1)}$, $j > n$이면 $X_{(j)} = X_{(n)}$가 되고 윈도크기 m은 $n/2$보다 작은 양의 정수다. 한편, 평균 μ와 분산 σ^2이 알려져 있지 않은 복합 영가설 하에서 $I(g:f)$는 각각의 최대가능도추정량 $\bar{X}$와 $S_n^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2/n$을 식 (2.5)에 대입하여

$$I_{m,n} = -H_{m,n} + \frac{1}{2}\log(2\pi S_n) + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_n}\right)^2 \tag{2.7}$$

로 추정이 된다. 식 (2.5)와 (2.7)의 단조변환을 통해 유도된

$$KLI_{m,n}^{S} = \frac{\exp(H_{m,n})}{\sigma\exp\left\{\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{(2n\sigma^2)}\right\}}, \tag{2.8}$$

$$KLI_{m,n}^{C} = \frac{\exp(H_{m,n})}{S_n\exp\left\{\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(2nS_n^2)}\right\}} = \frac{\exp(H_{m,n})}{S_n\sqrt{e}} \tag{2.9}$$

을 단순 영가설과 복합 영가설에 대한 검정통계량으로 사용하게 된다.

그런데 검정통계량을 얻기 위한 KL 판별정보의 추정에서 Arizono와 Ohta(1989)가 사용한 표본엔트로피 $H_{m,n}$과 분산의 최대가능도추정량 S_n^2은 일치성은 가지지만 비편향성을

보장하지는 못한다. 따라서 보다 정확하게 판별정보를 추정하기 위해선 비편향성을 가지는 추정량의 사용이 바람직하다고 할 수 있으므로 편향이 수정된 다음의 엔트로피 추정량

$$H^*_{m,n} = H_{m,n} - \log\frac{n}{2m} - \left(1 - \frac{2m}{n}\right)\psi(2m) + \psi(n+1) - \frac{2}{n}\sum_{k=1}^{m}\psi(k+m-1) \quad (2.10)$$

을 사용하고자 한다. 여기서 ψ는 디감마함수(digamma function)로 $\psi(t) = \Gamma(t)'/\Gamma(t)$로 정의된다. 이와 더불어 분산의 추정량으로는 $S^2_{n-1} = nS^2_n/(n-1)$을 사용한다. 표본엔트로피 $H_{m,n}$의 일치성 규명을 위해 Vasicek(1976)에 의해 언급이 된 바 있는 $H^*_{m,n}$는 Wieczorkowski와 Grzegorzewski(1999)가 다양한 모의실험을 통해서 $H_{m,n}$보다 제곱근 평균제곱오차와 편향이 일관되게 더 작게 나타나고 있음을 보고한 바가 있음에도 불구하고 적합도 검정의 영역에서는 거의 사용된 적이 없는 것 같다.

엔트로피와 분산의 추정량으로 $H^*_{m,n}$와 S^2_{n-1}을 사용한 식 (2.3)의 KL 판별정보 추정량은 단순 영가설의 경우는

$$I^S_{m,n,c} = -H^*_{m,n} + \frac{1}{2}\log(2\pi\sigma) + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2, \quad (2.11)$$

복합 영가설의 경우는

$$I^C_{m,n,c} = -H^*_{m,n} + \frac{1}{2}\log(2\pi S_{n-1}) + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_{n-1}}\right)^2 \quad (2.12)$$

로 구해지게 된다. Arizono와 Ohta(1989)의 방식에 따라 식 (2.11)과 (2.12)로부터 검정통계량을 유도해 보면

$$KLI^S_{m,n,c} = \frac{\exp\left(H^*_{m,n}\right)}{\sigma\exp\left\{\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{(2n\sigma^2)}\right\}}, \quad (2.13)$$

$$KLI^C_{m,n,c} = \frac{\exp\left(H^*_{m,n}\right)}{S_{n-1}\exp\left\{\dfrac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{(2nS^2_{n-1})}\right\}} = \frac{\exp\left(H^*_{m,n}\right)}{S_{n-1}\exp\left(\dfrac{n-1}{2n}\right)} \quad (2.14)$$

가 되고 이것을 개선된 KL 판별정보에 기반한 정규성 검정(이하 개선된 KL 검정)의 통계량으로 사용하고자 한다. 이들 통계량을 사용하는 개선된 KL 검정은 다음과 같은 특성들을 가지게 된다.

- 영가설 하에서 $n \to \infty$, $m \to \infty$, $m/n \to 0$이면 $I^S_{m,n,c}$, $I^C_{m,n,c} \to 0$이기 때문에 $KLI^S_{m,n,c}$, $KLI^C_{m,n,c} \to (2\pi)^{1/2}$가 되고 대립가설하에서는 $KLI^S_{m,n,c}, KLI^C_{m,n,c} < (2\pi)^{1/2}$가 되므로 $KLI^S_{m,n,c}$ 또는 $KLI^C_{m,n,c}$를 사용하는 검정은 일치성을 가진다.

- 정규분포의 엔트로피는 분산에만 의존하기 때문에 Vasicek(1976)의 검정은 평균이 알려진 경우에는 적용이 불가능하지만, KL 판별정보는 평균과 분산에 의존하게 되므로 이에 기초한 검정은 정규성에 대한 단순 또는 복합가설에도 적용이 가능하다.

- n이 커지면 $H^*_{m,n} \approx H_{m,n}$, $S_{n-1} \approx S_n$, $e^{(n-1)/2n} \approx e^{1/2}$가 되기 때문에 $KLI^S_{m,n,c} \approx KLI^S_{m,n}$, $KLI^C_{m,n,c} \approx KLI^C_{m,n}$임을 알 수가 있다. 따라서 검정통계량 (2.13) 또는 (2.14)를 사용하는 개선된 KL 검정은 Arizono와 Ohta(1989)의 검정과 근사적으로 같게 된다.

3. 검정통계량의 기각값

2절에서 제시한 개선된 KL 검정은 검정통계량이 주어진 유의수준 α에서의 기각값 $K(m,n,\alpha)$보다 작으면 영가설을 기각하게 된다. 그런데 $K(m,n,\alpha)$를 결정하려면 영가설 하에서 검정통계량의 표본분포를 알아야 하지만 어려운 일이기 때문에 모의실험을 통해서 기각값을 추정했다.

표본크기 $n \leq 100$에 대해 표준정규분포로부터 50000개의 표본들을 생성한 다음 주어진 n에 대해 모든 윈도크기 $m < n/2$에서 검정통계량 $KLI^C_{m,n,c}$를 계산했고, 이들 값으로부터 $100\alpha\%$가 되는 백분율들을 구했다. $K(m,n,\alpha)$는 Ebrahimi 등(1992)의 기준에 따라, 구한 백분율값들 가운데서 가장 큰 것을 선택했다. 다음의 표 3.1은 유의수준 5%에서 표본크기에 따른 검정통계량의 기각값과 이에 대응하는 윈도크기를 나타낸 것이다.

일반적으로 엔트로피나 KL 판별정보에 기반을 둔 적합도 검정에서 최적 윈도크기의 선택은 어려운 문제라 할 수 있지만 대체적으로 가장 작은 기각값을 산출하는 윈도크기의 사

표 3.1: 유의수준 5%에서 검정통계량 $KLI^C_{m,n,c}$의 기각값

n	m	기각값	n	m	기각값	n	m	기각값
5	2	2.0274	14	6	2.1340	26	4	2.2344
6	2	1.9547	15	7	2.1407	28	5	2.2518
7	3	2.0755	16	7	2.1455	30	5	2.2630
8	3	2.0562	17	8	2.1498	35	5	2.2952
9	4	2.1091	18	7	2.1575	40	6	2.3196
10	4	2.0987	19	4	2.1660	45	7	2.3360
11	5	2.1233	20	4	2.1748	50	8	2.3540
12	5	2.1212	22	4	2.1964			
13	6	2.1334	24	4	2.2143			

표 3.2: 표본크기에 따른 윈도크기

n	m	n	m	n	m	n	m
$21-26$	4	$55-60$	9	72	14	$85-88$	22
$27-37$	5	$61-64$	10	73	15	$89-91$	24
$38-41$	6	$65-66$	11	$74-78$	16	$92-93$	26
$42-48$	7	$67-69$	12	$79-80$	17	$94-95$	29
$49-54$	8	$70-71$	13	$81-84$	20	$96-100$	30

용은 검정력이 제일 낮게 되는 반면, 가장 큰 기각값을 제공하는 윈도크기는 가장 높은 검정력을 주게 된다(Ebrahimi 등(1992)을 참고). 표 3.2는 표본크기 $n \leq 100$에 대해 모든 윈도크기를 사용하여 여러 대립분포 하에서 $KLI^C_{m,n,c}$의 검정력을 모의실험을 통해 얻은 다음, 검정력이 가장 높게 나타나는 윈도크기를 얻은 것으로 표본크기가 증가하면 이에 대응되는 윈도크기도 증가하는 형태를 볼 수 있다.

그림 3.1은 각 표본크기 $n = 5(1)100$에 대해 표 3.2에 주어진 윈도크기를 사용하여 $\alpha = 5\%$에서 얻은 $KLI^C_{m,n,c}$의 기각값을 그린 것으로 표본크기가 커짐에 따라 $K(m,n,0.05)$는 $(2\pi)^{1/2} = 2.5066...$에 접근함을 볼 수 있다. 따라서 표본크기가 100이상일 때에는 $KLI^C_{m,n,c}$이 충분히 $(2\pi)^{1/2}$에 가깝지 않으면 영가설을 기각하게 된다.

표본크기에 대한 기각값이 표 3.1에서 제공되지 않는 경우에는 표에서 제시된 값들을 보간하여 근사적인 기각값을 구해야 한다. 이를 위해 고려한 계산공식은

$$K(m,n,0.05) = \beta_0 + \beta_1\sqrt{n} + \beta_2/\sqrt{n} + \beta_3/n + \beta_4/n^2 \tag{3.1}$$

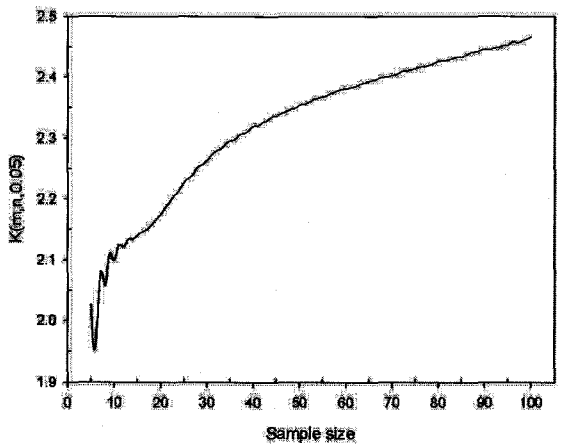

그림 3.1: 표본크기($n \leq 100$)에 대한 검정통계량 $KLI^C_{m,n,c}$의 기각값 플롯

표 3.3: 근사 기각값의 계산을 위한 추정된 계수

계수	β_0	β_1	β_2	β_3	β_4
추정값	0.5739	0.09129	14.58006	−49.74706	167.97566

이고 계수추정을 위해 $21 \leq n \leq 100$에 대한 기각값들을 이용하여 회귀분석을 수행했다. 적합된 회귀선의 결정계수는 99.98%로 나왔고 추정된 계수들은 표 3.3과 같다. 근사공식 (3.1)을 이용하면 손쉽게 기각값을 계산할 수가 있다. 예를 들면 $n = 60$일 때 기각값은 $K(9, 60, 0.05) = 0.5739 + 0.09129\sqrt{60} + 14.58006/\sqrt{60} - 49.74706/60 + 167.97566/60^2 \approx 2.3808$로 얻게 된다.

4. 개선된 KL 검정의 검정력

$KLI^C_{m,n,c}$ 검정의 성능을 모의실험을 통해 비교해 보기 위해 $KLI^C_{m,n}$ 검정 및 EDF 검정(Cramér-von Mises W^2 검정, Watson U^2 검정, Anderson-Darling A^2 검정)과 Shapiro-Wilk W 검정을 선택했다. 대립분포로는 (1) 표준지수분포($Exponential(1)$), (2) $\alpha = 1$, $\beta = 2$인 와이블분포($Weibull(1,2)$), (3) $\alpha = 1.5$, $\beta = 2$인 감마분포($Gamma(1.5,2)$), (4) $\mu = 1$, $\sigma^2 = 0.8$인 로그정규분포($LN(1,0.8)$), (5) 자유도 3인 카이제곱분포($\chi^2(3)$), (6) 자유도 1인 스튜던트의 t 분포($t(1)$), (7) $\alpha = 3$, $\beta = 1$인 베타분포($Beta(3,1)$), (8) $\mu = 1$, $\lambda = 1$인 역가우스분포($IG(1,1)$)를 고려했다.

각각의 대립분포로부터 생성시킨 크기 $n = 10, 20, 30$인 50000개의 표본들을 기초로 유의수준 5%에서 검정들의 검정력을 각각의 기각영역에 속하는 표본들의 상대빈도로 계산했다. $KLI^C_{m,n}$의 기각값과 윈도크기는 Arizono와 Ohta(1989)에 제시된 표에 있는 값을 사용했다.

표 4.1은 대립가설 하에서 모의실험을 통해 추정된 각 검정의 검정력을 보여주고 있으며, 제시된 결과를 통해 다음과 같은 사항들을 알 수가 있다.

- $KLI^C_{m,n,c}$이 $KLI^C_{m,n}$에 비해 일관되게 좋은 성능을 보이고 있다. 특히 표본크기가 작을 때($n = 10, 20$) 모든 대립분포에서 $KLI^C_{m,n,c}$ 검정력이 $KLI^C_{m,n}$보다 훨씬 더 높게 관측이 되고 있다. 이것은 $KLI^C_{m,n,c}$이 보다 정확하게 추정된 판별정보를 사용한 효과로 추측이 된다.

- 표본크기가 커질 경우에는 전반적으로 $KLI^C_{m,n,c}$와 $KLI^C_{m,n}$의 검정력은 높아지는 경향을 보이지만 여전히 $KLI^C_{m,n,c}$ 검정이 우수함을 일 수 있다. 그러나 그 검정력의 차이는 많이 줄어든 것을 볼 수 있다.

- $KLI^C_{m,n,c}$ 검정은 $t(1)$을 제외한 모든 대립분포에서 표본크기에 상관없이 W^2, U^2 와 A^2 검정들보다 일관되게 좋은 성능을 보인다. 특히 표본크기가 작은 경우($n = 10$), 검정력 차이는 더 크게 됨을 알 수가 있다.

표 4.1: 여러 대립분포에 대해서 추정된 W^2, U^2, A^2, W, $KLI^C_{m,n}$, $KLI^C_{m,n,c}$의 검정력

대립분포	n	W^2	U^2	A^2	W	$KLI^C_{m,n}$	$KLI^C_{m,n,c}$
	10	0.384	0.371	0.415	0.445	0.178	0.472
$Exponential(1)$	20	0.728	0.691	0.779	0.837	0.663	0.853
	30	0.900	0.870	0.934	0.968	0.912	0.974
	10	0.383	0.371	0.414	0.446	0.175	0.473
$Weibull(1,2)$	20	0.726	0.692	0.779	0.840	0.665	0.855
	30	0.898	0.869	0.934	0.969	0.912	0.974
	10	0.264	0.252	0.286	0.309	0.078	0.312
$Gamma(1.5,2)$	20	0.540	0.499	0.594	0.663	0.380	0.638
	30	0.732	0.679	0.793	0.869	0.662	0.853
	10	0.421	0.408	0.448	0.471	0.182	0.469
$LN(1,0.8)$	20	0.757	0.724	0.794	0.835	0.615	0.805
	30	0.913	0.888	0.937	0.961	0.865	0.948
	10	0.261	0.251	0.284	0.309	0.079	0.311
$\chi^2(3)$	20	0.537	0.495	0.590	0.659	0.373	0.632
	30	0.736	0.684	0.796	0.870	0.664	0.852
	10	0.612	0.609	0.613	0.601	0.307	0.423
$t(1)$	20	0.880	0.880	0.882	0.865	0.648	0.684
	30	0.965	0.966	0.966	0.954	0.834	0.853
	10	0.169	0.167	0.186	0.200	0.058	0.240
$Beta(3,1)$	20	0.366	0.340	0.415	0.485	0.296	0.571
	30	0.544	0.500	0.626	0.743	0.578	0.813
	10	0.502	0.487	0.531	0.558	0.250	0.570
$IG(1,1)$	20	0.843	0.817	0.875	0.908	0.758	0.900
	30	0.959	0.946	0.974	0.986	0.946	0.984

- W 검정은 W^2, U^2 와 A^2 검정들보다 더 좋은 성능을 발휘하고 있지만, $KLI^C_{m,n,c}$ 검정과 비교해보면 지수, 와이블, 감마, 베타분포들에서는 검정력이 낮게, $t(1)$ 분포에서는 높게, 그리고 그 밖의 분포들에서는 비슷하게 나타나고 있다.
- $KLI^C_{m,n,c}$의 검정력은 대립분포 $t(1)$에서 W^2, U^2, A^2 와 W들의 검정력보다 낮게 나타나고 있다. 따라서 $KLI^C_{m,n,c}$는 정규분포보다 긴 꼬리를 가지는 분포에서 좋은 성능을 보이지 않음을 알 수 있다.

표 4.2는 표 3.1의 기각값을 사용하는 $KLI^C_{m,n,c}$ 검정이 유의수준을 제대로 유지하는 지를 알아보기 위해 $N(1,1)$, $N(3,9)$와 $N(5,25)$로부터 생성한 50000개의 표본을 기초로 추정한 결과를 보여주고 있다. 유의수준은 5%로 설정했으며 표본크기와 분포에 걸쳐서 유의수준이 잘 통제되고 있음을 알 수 있다.

표 4.2: 추정된 $KLI^C_{m,n,c}$ 검정의 제 1종 오류($\alpha = 5\%$)

표본크기	$N(1,1)$	$N(3,9)$	$N(5,25)$
10	0.051	0.049	0.051
20	0.048	0.050	0.049
30	0.049	0.048	0.050

5. 예제

개선된 $KLI^C_{m,n,c}$ 검정을 수행하는 절차는 복잡하지가 않으며 그 과정을 요약하면 다음과 같다.

1. 주어진 표본크기에 대한 윈도크기를 표 3.1 또는 표 3.2에서 찾는다.
2. 검정통계량 $KLI^C_{m,n,c}$를 계산한다.
3. 표본크기에 해당하는 기각값을 표 3.1에서 찾는다. 표에 없는 경우에는 (3.1)을 이용하여 근사적인 기각값을 구한다.
4. 계산된 통계량의 값이 기각값보다 작으면 가설을 기각한다.

실제자료에 대해 $KLI^C_{m,n,c}$에 의한 검정을 수행해 보기로 한다. 사용된 자료는 보잉 720 항공기에 있는 냉난방기의 연속적인 고장사이에 운전한 시간간격을 기록한 것으로 자료값들은 다음과 같다: 14, 27, 32, 34, 54, 57, 59, 61, 66, 67, 102, 134, 152, 209, 230. Seshadri(1999)에 의하면 이 자료는 역가우스분포를 따르는 것으로 알려져 있다. 검정에 앞서, 기초통계량으로 왜도와 첨도를 구해보면 각각 1.2236과 3.5994로 정규분포의 이론적인 값과 다소 차이가 있음을 알 수 있으므로 아마도 정규분포를 따르지 않을 것으로 판단이 된다. 표 3.1에서 $n = 16$에 대한 윈도크기를 보면 $m = 7$이고 이것을 사용하여 검정통계량을 계산해 보면 $KLI^C_{m,n,c} \approx 1.9554$가 된다. 기각값은 표 3.1에 제시된 바와 같이 2.1455이고 계산된 값이 이보다 작으므로 유의수준 5%에서 영가설을 기각하게 된다.

6. 결론

본 논문에서는 Arizono와 Ohta(1989)의 정규성 검정을 개선한 검정을 제시하고 그 특성을 조사했다. 개선된 KL 검정은 엔트로피와 분산에 대한 비편향 추정량을 사용하여 얻은 KL 판별정보 추정량에 바탕을 두고 있고, 어떠한 대립분포에 대해서도 점근적으로 일치성을 가지게 된다. 모의실험을 통한 검정력 비교에서 개선된 KL 검정이 기존의 다른 검정들보다 일관되게 좋은 검정력을 제공하는 것으로 나타났고, 특히 표본의 크기가 작을 때 좋은 성능을 발휘하는 것으로 관측됐다. 따라서 응용에서 개선된 KL 검정이 검정력 이득의 측면에서 기존 검정들의 대안으로 사용될 수 있을거라 기대된다.

참고문헌

김종태, 이우동 (1998). 쿨백-레이블러 정보함수에 기초한 와이블분포와 극단값 분포에 대한 적합도 검정, <응용통계연구>, 11, 351–362.

Arizono, I. and Ohta, H. (1989). A test for normality based on Kullback-Leibler information, *The American Statistician*, **43**, 20–22.

Chandra, M., De Wet, T. and Singpurwalla, N. D. (1982). On the sample redundancy and a test for exponentiality, *Communications in Statistics-Theory and Methods*, **11**, 429–438.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques,* Marcel Dekker, New York.

Dudewicz, E. J. and van der Meulen, E. C. (1981). Entropy-based tests of uniformity, *Journal of the American Statistical Association*, **76**, 967–974.

Ebrahimi, N., Habibullah, M. and Soofi, E. S. (1992). Testing exponentiality based on Kullback-Leibler information, *Journal of the Royal Statistical Society*, Ser. B, **54**, 739–748.

Gokhale, D. V. (1983). On entropy-based goodness-of-fit tests, *Computational Statistics & Data Analysis*, **1**, 157–165.

Kim, J. T., Lee, W. D., Ko, J. H., Yoon, Y. H. and Kang, S. G. (1999). Goodness of fit test for normality based on Kullback-Leibler information, *The Korean Communications in Statistics*, **6**, 909–917.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics*, **22**, 79–86.

Seshadri, V. (1999). *The Inverse Gaussian Distribution: Statistical Theory and Applications,* Springer, New York.

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell System Technical Journal*, **27**, 349–423, 623–656.

Vasicek, O. (1976). A test for normality based on sample entropy, *Journal of the Royal Statistical Society*, Ser. B, **38**, 54–59.

Wieczorkowski, R. and Grzegorzewski, P.(1999). Entropy estimators-improvements and comparisons, *Communications in Statistics-Simulation and Computation*, **28**, 541–567.

[2006년 6월 접수, 2006년 8월 채택]

Improving a Test for Normality Based on Kullback-Leibler Discrimination Information*

Byungjin Choi[1)]

ABSTRACT

A test for normality introduced by Arizono and Ohta(1989) is based on Kullback-Leibler discrimination information. The test statistic is derived from the discrimination information estimated using sample entropy of Vasicek(1976) and the maximum likelihood estimator of the variance. However, these estimators are biased and so it is reasonable to make use of unbiased estimators to accurately estimate the discrimination information. In this paper, Arizono-Ohta test for normality is improved. The derived test statistic is based on the bias-corrected entropy estimator and the uniformly minimum variance unbiased estimator of the variance. The properties of the improved KL test are investigated and Monte Carlo simulation is performed for power comparison.

Keywords: Normality, entropy, Kullback-Leibler discrimination information, power.

* This work was supported by Kyonggi University Research Grant.

1) Assistant Professor, Department of Applied Information Statistics, Kyonggi University, Suwon, Gyeonggi-Do 443-760, Korea
E-mail: bjchoi92@kyonggi.ac.kr