Journal of the Korean Data & Information Science Society
2013, **24**(4), 847–856

http://dx.doi.org/10.7465/jkdi.2013.24.4.847
한국데이터정보과학회지

모수적 엔트로피 추정량과 비모수적 엔트로피 추정량에 기초한 정규분포에 대한 적합도 검정

최병진[1]

[1]경기대학교 응용정보통계학과

접수 2013년 6월 20일, 수정 2013년 7월 11일, 게재확정 2013년 7월 16일

요 약

본 논문에서는 모수적과 비모수적 엔트로피 추정량들에 기초한 정규분포에 대한 적합도 검정을 다룬다. 정규분포의 엔트로피에 대한 모수적 추정량으로 사용할 최소분산비편향추정량을 유도한다. 이 추정량과 대립가설 하에서의 자료생성분포에 대한 비모수적 엔트로피 추정량으로 표본엔트로피와 이것의 변형된 추정량들을 이용하여 검정통계량들을 구축했고 이 검정통계량들을 사용하는 새로운 엔트로피 기반 적합도 검정들을 제시한다. 제안한 검정들의 기각값들을 모의실험을 통해 추정해서 표의 형태로 제시한다. 성능의 조사를 위해 수행한 모의실험에서 제안한 검정들이 기존의 Vasicek (1976) 검정보다는 더 좋은 검정력을 가지는 것으로 나타난다. 응용에서 새로운 검정들이 정규성 검정을 위한 경쟁적인 도구로 시용될 수 있을 것으로 기대된다.

주요용어: 검정력, 엔트로피, 엔트로피 추정량, 적합도, 정규분포.

1. 서론

응용에서 연구자들은 여러 변수들을 고려하여 측정한 자료를 분석한 결과를 바탕으로 관심의 대상이 되는 현상들의 인과관계나 특성을 파악하게 된다. 통계적 추론 또는 모형화를 위한 목적으로 사용할 통계적 방법은 자료에 대한 분포적 가정을 요구한다. 이용가능한 대부분의 통계적 방법들은 자료가 평균이 μ이고 분산이 σ^2인 정규분포 $N(\mu, \sigma^2)$에서 추출되었다는 가정하에서 출발한다. 이런 정규성 가정의 배경은 정규분포가 다른 확률분포에 비해 수리적으로 다루기 쉽고 정규성 가정하에서 유도되는 분포와 관련된 결과의 우수성 등 여러 장점들을 가지고 있는 것에 기인한다. 물론 정규분포가 이와 같은 여러 장점들을 가지고 있는 매력적인 확률모형이기는 하지만 자료분석에 앞서서 주어진 자료가 정규분포를 따르는지를 확인해 보는 것은 매우 중요하다. 그 이유는 정규성 가정하에 개발된 통계적 방법들을 비정규적인 자료에 적용할 경우에는 얻어진 분석 결과는 타당하지 못하게 되며 이로부터 잘못된 추론과 해석을 할 가능성이 커지기 때문이다.

자료의 정규성 여부를 알아보는 방법 중의 하나는 적합도 검정을 수행해 보는 것이다. 크기 n의 확률표본 $X_1, X_2, \ldots, X_n$이 확률밀도함수 $g(x)$를 가지는 임의의 분포에서 추출되었다고 하면, 정규분포에 대한 적합도 검정은 영가설과 대립가설이 각각 $H_0 : g(x) = f(x)$와 $H_1 : g(x) \neq f(x)$로 설정이 되는 분포적 가설에 대한 검정이다. 여기서 $f(x)$는 평균과 분산이 각각 μ와 σ^2인 정규분포의 확률밀도함수이다. Pearson이 1900년에 정규성에 대한 카이제곱 적합도 검정을 고안한 이후에 정규성에 대한 검정 문

[1] (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 응용정보통계학과, 부교수.
E-mail: bjchoi92@kyonggi.ac.kr

제는 상당한 관심을 받아 왔고 꽤 많은 수의 검정들을 문헌에서 쉽게 찾아 볼 수가 있다. 지금까지 개발된 검정들에 관해서는 D'Agostino와 Stephens (1986)의 9장, Mardia (1980)와 그 속에 있는 참고문헌들을 보기 바란다.

Shannon (1948)은 정보이론분야에서 불확실성의 측도로 널리 사용되는 엔트로피를 처음 소개했고 주어진 분산을 가지는 모든 확률분포들 중에서 정규분포의 엔트로피가 최대가 됨을 밝혔다. 그 이후에 추론통계학분야에서 엔트로피를 추정하고 분포적 가설 검정에 활용하기 위한 많은 연구가 있어 왔다. Vasicek (1976)은 표본의 순서통계량으로부터 정의되는 표본엔트로피를 엔트로피 추정량으로 제안했고 표본엔트로피는 엔트로피에 대한 일치성을 가짐을 보였다. Correa (1995)는 표본엔트로피를 변형한 새로운 형태의 엔트로피 추정량을 제시했다. 엔트로피에 기초한 적합도 검정의 구축을 처음 시도한 학자는 Vasicek (1976)으로 Shannon (1948)이 발견한 정규분포의 엔트로피 특성짓기를 이용한 정규성 검정을 제시했다. 이 검정은 다른 분포들에 대한 적합도 검정의 개발에 많은 영향을 주었다 (Dudewicz와 van der Meulen, 1981; Grzegorzewski와 Wieczorkowski, 1999; Choi와 Kim, 2006, Lee 등, 2013). 자료생성분포와 정규분포의 엔트로피 불일치에 기초를 두고 있는 Vasicek의 검정은 자료생성분포에 대한 비모수적 엔트로피 추정량으로 표본엔트로피를 사용하고 정규분포의 엔트로피에 대한 모수적 추정량으로 최대가능도추정량을 이용해서 엔트로피 불일치의 추정량을 얻고 이것을 변형한 형태를 검정통계량으로 사용하고 있다. 그러나, 검정통계량에서 모수적 엔트로피의 추정량으로 사용한 최대가능도추정량은 점근적 일치성과 효율성 등을 가지고 있지만 비편향성과 최소분산은 보장해주질 못하기 때문에 최대가능도추정량보다 더 좋은 추정량의 사용이 바람직할 것으로 판단된다.

본 논문에서는 정규분포의 엔트로피에 대한 모수적 추정량으로 균일최소분산비편향추정량을 유도한다. 자료생성분포의 엔트로피에 대한 비모수적 추정량은 표본엔트로피, Correa의 엔트로피 추정량과 함께 Wieczorkowski와 Grzegorzewski (1999)에서 표본엔트로피보다 평균제곱오차와 편향의 관점에서 더 좋은 성능을 가지는 것으로 보고된 편향이 수정된 엔트로피 추정량을 사용한다. 비모수적 엔트로피 추정량들과 모수적 엔트로피 추정량을 이용하여 검정통계량들을 도출하고 이 검정통계량들을 이용하는 정규분포에 대한 엔트로피 기반 적합도 검정을 제안한다. 2절에서는 정규분포의 엔트로피에 대한 모수적 추정량을 사용할 균일최소분산비편향추정량을 유도한다. 정규분포의 적합을 알아보기 위해 사용할 검정통계량들을 기존에 개발된 비모수적 엔트로피 추정량들과 유도한 모수적 추정량을 활용하여 구축한다. 3절에서는 표본크기와 윈도도크기에 따른 검정통계량들의 기각값을 모의실험을 통해 추정하여 표의 형태로 제시한다. 4절에서는 제안한 검정들의 검정력 비교를 위해서 모의실험을 수행한다. 5절에서는 결론을 맺는다.

2. 엔트로피 기반 적합도 검정통계량

정보이론분야에서 불확실성의 측도로 많이 사용되는 엔트로피는 Shannon (1948)에 의하면 확률밀도함수 $f(x)$를 가지는 확률변수 X에 대해서

$$H(f) = -\int_{-\infty}^{\infty} f(x)\ \log f(x)\ dx \tag{2.1}$$

로 정의된다. Shannon (1948)은 엔트로피에 관해서 다음과 같이 정규분포를 특성지울 수가 있음을 보였다: 주어진 분산 σ^2을 가지는 모든 확률분포들 중에서 정규분포는 식 (2.1)의 엔트로피를 최대화시키고 그 값은 $H(f) = \left\{ \log\left(2\pi e\sigma^2\right) \right\}/2$가 된다. 정규분포에 대한 엔트로피 특성짓기는 적합도 검정을 개발하기 위한 자연스러운 방법을 제공해 줄 수가 있다.

크기 n의 확률표본 $X_1, X_2, \ldots, X_n$이 확률밀도함수 $g(x)$를 가지는 임의의 분포에서 추출되었다고 하면, 정규분포의 적합도 검정을 위한 영가설과 대립가설은 $H_0 : g(x) = f(x)$ 대 $H_1 : g(x) \neq f(x)$로 설

정이 된다. 여기서, $f(x)$는 평균이 μ이고 분산이 σ^2인 정규분포의 확률밀도함수이다. 확률밀도함수로 각각 $f(x)$와 $g(x)$를 가지는 두 분포간의 엔트로피 불일치 정도를 재는 측도로 엔트로피 차이

$$D(g:f) = H(g) - H(f) \tag{2.2}$$

를 고려하자. 영가설하에서는 두 분포는 일치하게 되어(즉, $f(x) = g(x)$) $D(g:f) = 0$이 됨을 알 수 있다. 정규분포의 엔트로피가 주어진 분산을 가지는 모든 확률분포들 중에서 최대가 되는 점에 주목하면 대립가설하에서의 엔트로피 $H(g)$는 정규분포의 엔트로피 $H(f)$보다 작게 되므로 $D(g:f) < 0$이 된다. 그러므로, $D(g:f)$가 음의 방향으로 큰 값을 가지면 주어진 표본은 정규분포가 아닌 다른 분포에서 추출되었음을 나타낸다. 검정통계량의 유도를 위해서 $D(g:f)$를 단조변환한

$$T = \sqrt{2\pi e}\exp\left\{(D(g:f)\right\} = \sqrt{2\pi e}\exp\left\{H(g) - H(f)\right\} \tag{2.3}$$

를 고려한다. T는 0과 $\sqrt{2\pi e}$ 사이의 값을 가지게 되고 작은 값을 가지게 되면 영가설을 기각하고 대립가설을 받아들이게 된다.

식 (2.3)을 검정통계량으로 사용하기 위해서는 $H(f)$와 $H(g)$를 표본으로부터 추정을 해야 한다. $H(f)$는 영가설하에서의 엔트로피, 즉 정규분포의 엔트로피로 Vasicek (1976)은 $H(f)$의 σ^2을 $S_n^2 = \sum_{i=1}^n \left(X_i - \bar{X}\right)^2/n$으로 대체한 최대가능도추정량 $\hat{H}_{ml} = \left\{\log\left(2\pi e S_n^2\right)\right\}/2$을 사용했다. 하지만, $\hat{H}_{ml}$은 $H(f)$에 대한 편향성을 가지고 있을 뿐만 아니라 최소분산 또한 보장해 주지는 않는다. 보다 정확한 $H(f)$의 추정을 위해서는 비편향성과 함께 최소분산을 가지는 추정량의 사용이 바람직하다. 이런 목적으로 $H(f)$에 대한 균일최소분산비편향추정량을 유도하여 이용하기로 한다.

영가설이 사실이라면 표본은 정규분포에서 추출된 것이고 정규분포는 지수족의 일원이다. 통계량 $\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$은 σ^2에 대해서 충분하면서도 완비적이다. $W = \sum_{i=1}^n (X_i - \bar{X})^2$으로 정의하고 통계량 $\varphi(W)$를 W의 함수라고 하면 $H(f)$에 대한 균일최소분산비편향추정량 $\hat{H}_{umvu}$는 $E[\varphi(W)] = H(f)$가 되는 $\varphi(W)$를 찾으면 된다. $\varphi(W)$를 결정하기 위해 W를 σ^2으로 나눈 $V = W/\sigma^2$를 고려한다. V는 자유도가 $(n-1)$인 카이제곱분포를 따르게 되고 $V^* = \log V$의 기대값은 V^*의 적률생성함수

$$\phi_{V^*}(t) = E\left(e^{tV^*}\right) = E\left(V^t\right) = \frac{2^t\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{2.4}$$

를 이용하여 다음과 같이 구할 수 있다. 식 (2.4)의 $\phi_{V^*}(t)$를 t에 대해서 1차 미분을 하게 되면

$$\phi'_{V^*}(t) = \frac{2^t\left\{\Gamma'\left(\frac{n-1}{2} + t\right) + \Gamma\left(\frac{n-1}{2} + t\right)\log 2\right\}}{\Gamma\left(\frac{n-1}{2}\right)} \tag{2.5}$$

로 얻게 된다. 식 (2.5)에서 $t = 0$을 대입하면 $E\left(V^*\right) = \phi'_{V^*}(0) = \Gamma'\left\{\left(n-1\right)/2\right\}/\Gamma\left\{\left(n-1\right)/2\right\} + \log 2 = \psi\left\{\left(n-1\right))/2\right\} + \log 2$ 가 된다. 여기서 $\psi(k)$은 디감마함수로 $\psi(k) = \Gamma'(k)/\Gamma(k)$로 정의된다. $E\left(V^*\right)$로부터 얻은 $E(\log W) = \log\sigma^2 + \psi\{(n-1)/2\} + \log 2$를 이용하면

$$E\left(\frac{\log W}{2}\right) = \frac{1}{2}\log\sigma^2 + \frac{1}{2}\left[\psi\left(\frac{n-1}{2}\right) + \log 2\right] \tag{2.6}$$

가 됨을 알 수 있다. 식 (2.6)으로부터 $E\left[\varphi\left(W\right)\right] = \log\sigma^2/2 + \log\left(2\pi e\right)/2$가 되는 $\varphi\left(W\right)$는

$$\varphi\left(W\right) = \frac{1}{2}\log W - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) - \frac{1}{2}\log 2 + \frac{1}{2}\log\left(2\pi e\right) \tag{2.7}$$

로 얻게 된다. 따라서, 레만-쉐페 정리에 의해 식 (2.7)의 $\varphi(W)$는 $H(f)$에 대한 균일최소분산비편향추정량 $\hat{H}_{umvu}$이 된다. $\hat{H}_{umvu}$는 또한 $S_{n-1}^2 = W/(n-1)$인 관계에 의해

$$\hat{H}_{umvu} = \frac{1}{2}\log S_{n-1}^2 - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) + \frac{1}{2}\log\frac{n-1}{2} + \frac{1}{2}\log(2\pi e) \tag{2.8}$$

와 같다.

자료생성분포의 엔트로피 $H(g)$에 대한 추정문제는 Vasicek (1976), Györfi와 van der Meulen (1987), Dudewicz와 van der Meulen (1987), van Es (1992), Ebrahimi 등 (1994), Correa (1995) 등 여러 학자들에 의해 연구가 되었고 제안된 추정량들 중에서 Vasicek (1976)이 제안한 표본엔트로피는 가장 간단하고 용도가 넓은 추정량으로 분포에 대한 적합도 검정에서 많이 사용되고 있다. 표본엔트로피는

$$H_{m,n} = \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{n}{2m}\left\{X_{(i+m)} - X_{(i-m)}\right\}\right] \tag{2.9}$$

로 정의되고 $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$은 X_i들의 순서통계량으로 $i > n$이면 $X_{(i)} = X_{(n)}$, $i < 1$이면 $X_{(i)} = X_{(1)}$이고 m은 $n/2$보다 작은 양의 정수값을 갖는 윈도크기이다. Vasicek (1976)은 또한 표본엔트로피의 편향을 수정한 엔트로피 추정량으로

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

를 제시했다. $H_{m,n}^{C}$는 Wieczorkowski와 Grzegorzewski (1999)의 모의실험을 통한 엔트로피 추정량의 비교에서 표본엔트로피보다 평균제곱오차와 편향의 측면에서 더 좋은 성능을 보이는 것으로 나타나고 있지만 엔트로피을 기반으로 하는 적합도 검정에서는 거의 사용된 적이 없다. Correa (1995)에 의해 개발된 엔트로피 추정량은 표본엔트로피의 변형 형태로

$$C_{m,n} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\sum_{j=i-m}^{i+m}(j-i)\left\{X_{(j)} - \bar{X}_{(i)}\right\}}{n\sum_{j=i-m}^{i+m}\left\{X_{(j)} - \bar{X}_{(i)}\right\}^2} \tag{2.11}$$

가 된다. 여기서, $i > n$이면 $X_{(i)} = X_{(n)}$, $i < 1$이면 $X_{(i)} = X_{(1)}$, m은 $n/2$보다 작은 양의 정수값을 갖는 윈도크기이며 $\bar{X}_{(i)} = \sum_{j=i-m}^{i+m} X_{(j)}/(2m+1)$이다.

식 (2.3)에 제시된 T의 추정량을 얻기 위해서 자료생성분포의 엔트로피 $H(g)$를 식 (2.9)-(2.11)의 비모수적 엔트로피 추정량들로 각각 대체하고 정규분포의 엔트로피 $H(f)$에 대해서는 식 (2.8)의 모수적 엔트로피 추정량 $\hat{H}_{umvu}$를 대입한다. 그러면 T의 추정량은 각각

$$TH_{m,n} = \left\{\frac{2}{(n-1)S_{n-1}^2}\right\}^{\frac{1}{2}}\exp\left\{\frac{1}{2}\psi\left(\frac{n-1}{2}\right)\right\}\exp(H_{m,n}), \tag{2.12}$$

$$THC_{m,n} = \left\{\frac{2}{(n-1)S_{n-1}^2}\right\}^{\frac{1}{2}}\exp\left\{\frac{1}{2}\psi\left(\frac{n-1}{2}\right)\right\}\exp\left(H_{m,n}^{C}\right), \tag{2.13}$$

$$TC_{m,n} = \left\{\frac{2}{(n-1)S_{n-1}^2}\right\}^{\frac{1}{2}}\exp\left\{\frac{1}{2}\psi\left(\frac{n-1}{2}\right)\right\}\exp(C_{m,n}) \tag{2.14}$$

가 되고 이 추정량들을 주어진 표본이 정규분포을 따른다는 영가설을 검정하기 위한 검정통계량으로 사용하고자 한다.

식 (2.12)-(2.14)를 검정통계량으로 사용하는 제안된 엔트로피 기반 검정은 Vasicek의 검정과 마찬가지로 모든 대립가설에 대해서 일치성을 가지게 됨을 다음과 같이 보일 수가 있다. 여기서 일치성은 $m \to \infty$, $n \to \infty$이면서 $m/n \to 0$일 때 검정력이 1로 수렴하게 됨을 의미한다. $n \to \infty$이면 $\sqrt{2/(n-1)}\exp\left[\psi\{(n-1)/2\}\right]/S_{n-1} \to 1/\sigma$가 된다. 자료생성분포의 엔트로피에 대한 추정량 $H_{m,n}$, $H^C_{m,n}$, $C_{m,n}$들은 $m \to \infty$, $n \to \infty$, $m/n \to 0$이면 영가설하에서 정규분포의 엔트로피 $H(f)$로 수렴하게 되어 $\exp(H_{m,n})$, $\exp\left(H^C_{m,n}\right)$, $\exp(C_{m,n})$들은 $\sigma\sqrt{2\pi e}$에 수렴한다. 따라서, $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$들은 영가설하에서 $m \to \infty$, $n \to \infty$이면서 $m/n \to 0$이면 $\sqrt{2\pi e}$로 가게 된다. 한편 대립가설하에서는 $\exp(H_{m,n}), \exp\left(H^C_{m,n}\right), \exp(C_{m,n}) \to \exp\{H(g)\}$이므로 $TH_{m,n}, THC_{m,n}, TC_{m,n} \to \exp\{H(g)\}/\sigma = \exp\{H(g) - \log\sigma^2/2\} = \sqrt{2\pi e}\exp\{H(g) - H(f)\}$가 된다. 정규분포의 확률밀도함수는 식 (2.1)의 엔트로피를 최대로 하므로 $\sqrt{2\pi e}\exp\{H(g) - H(f)\} < \sqrt{2\pi e}$가 된다. 따라서, $m \to \infty$, $n \to \infty$, $m/n \to 0$이면 $TH_{m,n}$, $THC_{m,n}$, $TC_{m,n}$의 검정력은 1로 수렴하게 된다.

3. 검정통계량들의 기각값

자료로부터 계산한 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$은 영가설 $H_0 : g(x) = f(x)$이 사실이면 $\sqrt{2\pi e}$에 가까운 값을 가지게 될 것이고 대립가설 $H_1 : g(x) \neq f(x)$가 맞다면 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$의 값은 $\sqrt{2\pi e}$보다 작게 될 것이다. 영가설의 채택여부를 결정하려면 영가설하에서 검정통계량 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$의 표본분포를 이론적으로 유도해서 유의수준 α에서의 기각값을 각각 구해야 한다. 검정통계량들의 표본분포는 Cressie (1976), Dudewicz와 van der Meulen (1981), van Es (1992) 등에 의한 표본엔트로피와 유사 표본엔트로피들을 포함하는 통계량들의 점근적 분포성질들을 이용하면 구할 수 있을 것처럼 보인다. 하지만 제안된 이들 연구결과의 적용은 극히 제한적이기 때문에 검정통계량들의 표본분포의 해석적인 유도는 아주 어려운 문제이다. 그러므로, 모의실험을 이용하여 각 검정통계량의 100α 백분위수를 추정하고 이것을 기각값으로 사용하기로 한다.

Table 3.1 Estimated critical values of $TH_{m,n}$ for the significance level 5%

	m									
n	1	2	3	4	5	6	7	8	9	10
5	0.917	1.315								
6	1.094	1.464								
7	1.242	1.582	1.663							
8	1.357	1.701	1.782							
9	1.476	1.815	1.883	1.866						
10	1.575	1.926	1.974	1.970						
12	1.736	2.106	2.145	2.130	2.085					
14	1.858	2.246	2.300	2.273	2.230	2.169				
16	1.970	2.369	2.430	2.404	2.358	2.298	2.231			
18	2.053	2.467	2.543	2.523	2.475	2.417	2.354	2.286		
20	2.131	2.554	2.636	2.624	2.580	2.522	2.461	2.395	2.329	
25	2.272	2.714	2.817	2.826	2.793	2.744	2.685	2.627	2.568	2.507
30	2.377	2.830	2.943	2.965	2.948	2.913	2.864	2.812	2.758	2.703
35	2.450	2.915	3.043	3.075	3.069	3.042	3.003	2.958	2.909	2.859
40	2.508	2.982	3.116	3.158	3.163	3.144	3.115	3.077	3.035	2.990
45	2.561	3.038	3.178	3.226	3.236	3.226	3.203	3.172	3.137	3.098
50	2.598	3.080	3.224	3.278	3.292	3.287	3.271	3.247	3.216	3.184
60	2.666	3.150	3.304	3.364	3.387	3.390	3.381	3.364	3.344	3.319
70	2.709	3.200	3.358	3.425	3.455	3.463	3.462	3.452	3.437	3.419
80	2.747	3.244	3.403	3.474	3.506	3.519	3.521	3.515	3.505	3.493
100	2.801	3.300	3.464	3.540	3.579	3.599	3.608	3.608	3.605	3.598

Table 3.2 Estimated critical values of $THC_{m,n}$ for the significance level 5%

	m									
n	1	2	3	4	5	6	7	8	9	10
5	1.974	2.631								
6	2.169	2.667								
7	2.322	2.696	2.925							
8	2.427	2.758	2.954							
9	2.552	2.830	2.979	3.079						
10	2.634	2.903	3.005	3.103						
12	2.800	3.036	3.092	3.152	3.199					
14	2.909	3.131	3.185	3.210	3.242	3.263				
16	3.018	3.221	3.267	3.278	3.293	3.304	3.311			
18	3.090	3.288	3.339	3.346	3.349	3.353	3.356	3.355		
20	3.160	3.349	3.395	3.406	3.402	3.399	3.398	3.395	3.391	
25	3.290	3.462	3.511	3.526	3.523	3.515	3.502	3.494	3.487	3.479
30	3.384	3.541	3.585	3.602	3.604	3.602	3.591	3.581	3.571	3.561
35	3.448	3.599	3.649	3.668	3.676	3.674	3.668	3.659	3.649	3.639
40	3.503	3.648	3.691	3.711	3.721	3.723	3.721	3.716	3.710	3.703
45	3.547	3.685	3.731	3.754	3.766	3.771	3.772	3.769	3.765	3.758
50	3.581	3.715	3.761	3.781	3.796	3.804	3.805	3.807	3.805	3.802
60	3.642	3.762	3.807	3.829	3.843	3.853	3.857	3.860	3.861	3.861
70	3.678	3.795	3.838	3.862	3.878	3.887	3.896	3.900	3.904	3.906
80	3.714	3.828	3.866	3.890	3.904	3.914	3.922	3.927	3.932	3.937
100	3.764	3.866	3.903	3.925	3.940	3.952	3.961	3.967	3.974	3.979

주어진 유의수준 α에 해당하는 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$의 기각값을 결정하기 위해서 표본크기가 $n = 5, 6, \ldots, 100$인 50000개의 표본들을 표준정규분포로부터 각각 독립적으로 생성했다. 그리고 추출된 이들 표본을 이용해서 $m < n/2$인 모든 윈도크기에 대해서 각 검정통계량을 계산했다. 그런 다음 각각의 표본크기와 윈도크기에 대해서 계산한 각 검정통계량의 값들을 바탕으로 경험적 분포를 추정하고 이것을 이용하여 기각값으로 사용할 100α 백분율을 구했다.

Table 3.1은 표본크기와 윈도크기에 따른 유의수준 5%에 대한 $TH_{m,n}$의 추정된 기각값들이다. 또한 Table 3.2와 Table 3.3은 유의수준 5%에 대한 $THC_{m,n}$과 $TC_{m,n}$의 추정된 기각값들을 보여 주고 있다. 표들에서 보는 바와 같이 기각값들은 주어진 윈도크기에 대해서 표본크기가 커짐에 따라 증가하는 형태로 나타나고 있음을 볼 수 있다.

Table 3.3 Estimated critical values of $TC_{m,n}$ for the significance level 5%

	m									
n	1	2	3	4	5	6	7	8	9	10
5	1.116	1.743								
6	1.304	1.870								
7	1.467	1.967	2.142							
8	1.593	2.072	2.231							
9	1.722	2.181	2.302	2.343						
10	1.818	2.284	2.371	2.414						
12	2.001	2.477	2.524	2.529	2.527					
14	2.129	2.624	2.675	2.652	2.628	2.603				
16	2.248	2.750	2.803	2.772	2.731	2.695	2.659			
18	2.339	2.854	2.917	2.885	2.838	2.790	2.747	2.706		
20	2.417	2.938	3.008	2.984	2.932	2.880	2.833	2.789	2.746	
25	2.568	3.106	3.187	3.177	3.135	3.085	3.033	2.983	2.937	2.892
30	2.679	3.221	3.310	3.307	3.278	3.237	3.191	3.144	3.098	3.053
35	2.754	3.310	3.407	3.415	3.394	3.363	3.323	3.280	3.239	3.198
40	2.817	3.376	3.477	3.492	3.479	3.454	3.424	3.389	3.351	3.313
45	2.871	3.430	3.539	3.560	3.551	3.531	3.506	3.476	3.445	3.412
50	2.914	3.476	3.586	3.609	3.604	3.589	3.567	3.545	3.518	3.489
60	2.980	3.546	3.659	3.686	3.685	3.674	3.659	3.641	3.622	3.600
70	3.025	3.595	3.714	3.742	3.747	3.740	3.729	3.716	3.701	3.685
80	3.068	3.637	3.756	3.788	3.794	3.789	3.780	3.769	3.758	3.745
100	3.121	3.693	3.814	3.848	3.858	3.859	3.854	3.847	3.839	3.830

Table 3.4 Estimated type-I errors of the proposed tests for the significance level 5%

		$N(1,1)$			$N(3,9)$			$N(5,25)$		
n	m	$TH_{m,n}$	$THC_{m,n}$	$TC_{m,n}$	$TH_{m,n}$	$THC_{m,n}$	$TC_{m,n}$	$TH_{m,n}$	$THC_{m,n}$	$TC_{m,n}$
10	2	0.049	0.048	0.047	0.050	0.049	0.049	0.051	0.050	0.050
20	3	0.049	0.048	0.049	0.048	0.047	0.048	0.050	0.050	0.050
30	3	0.050	0.050	0.050	0.051	0.050	0.050	0.051	0.050	0.052

각 검정통계량으로부터 모의실험을 통해 추정한 기각값이 주어진 유의수준, 즉 제 1종 오류 α를 잘 유지하는지를 평가해보기로 한다. 이를 위해서 표본크기 $n = 10, 20, 30$에 대해서 $N(1,1)$, $N(3,9)$와 $N(5,25)$로부터 각각 독립적으로 생성해서 얻은 10000개의 표본을 이용하여 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$들의 제 1종 오류를 추정해 보았다. 추정한 제 1종 오류가 주어진 값에 비해서 아주 작거나 크게 나온다면 기각값의 정확성에 문제가 있음을 나타내는 것이어서 검정의 결과를 신뢰할 수가 없게 된다. Table 3.4는 $\alpha = 0.05$로 했을 때 모의실험을 통해 추정한 제 1종 오류이다. 10000개의 표본에 기초하여 추정한 α의 표준오차를 계산해보면 $\sigma_\alpha = \sqrt{0.05(1-0.05)/10000} \approx 0.0022$가 되고 이것을 이용해서 구한 α에 대한 95% 신뢰구간은 대략 (0.0456,0.0544)가 된다. Table 3.4의 추정값들을 보면 0.047-0.52 사이에 나타나고 있으므로 Table 3.1-Table 3.3에 주어진 기각값을 사용하는 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$ 검정들은 제 1종 오류를 잘 통제함을 알 수 있다.

4. 검정력 비교

2절에서 제안한 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$ 검정들의 검정력을 모의실험을 통해서 비교해 보기로 한다. 모의실험에서 표본크기는 $n = 10, 20, 30$으로 했고 대립가설에서의 분포로는 (a) 균일분포 $U(0,1)$, (b) 평균이 θ인 지수분포 $E(\theta)$, (c) 자유도가 ν인 카이제곱분포 χ^2_ν, (d) 형상모수 θ와 척도모수 β를 가지는 감마분포 $G(\theta,\beta)$, (e) 형상모수 θ와 척도모수 β를 가지는 와이블분포 $W(\theta,\beta)$, (f) 모수 μ와 σ^2을 가지는 로그정규분포 $LN(\mu,\sigma^2)$, (g) 위치모수 μ와 척도모수 λ를 가지는 역가우스분포 $IG(\mu,\lambda)$ 등 7개의 분포들을 선택했다.

Table 4.1 Estimated powers for the proposed entropy-based tests for the significance level 5%

			Test Statistic			
Distribution	n	m	$TH_{m,n}$	$THC_{m,n}$	$TC_{m,n}$	$K_{m,n}$
$U(0,1)$	10	2	0.173	0.170	0.170	0.154
	20	3	0.424	0.421	0.430	0.408
	30	3	0.654	0.652	0.661	0.639
$E(2)$	10	2	0.430	0.427	0.418	0.403
	20	3	0.851	0.849	0.839	0.840
	30	3	0.966	0.966	0.964	0.964
χ^2_3	10	2	0.259	0.257	0.250	0.234
	20	3	0.618	0.615	0.608	0.604
	30	3	0.828	0.827	0.823	0.818
$G(2,1)$	10	2	0.190	0.187	0.181	0.169
	20	3	0.464	0.462	0.452	0.449
	30	3	0.655	0.653	0.646	0.641
$W(1.3,1)$	10	2	0.221	0.217	0.215	0.199
	20	3	0.529	0.527	0.521	0.514
	30	3	0.754	0.752	0.746	0.743
$LN(0,0.2)$	10	2	0.146	0.144	0.136	0.130
	20	3	0.337	0.335	0.325	0.324
	30	3	0.481	0.480	0.469	0.469
$IG(1,4)$	10	2	0.169	0.167	0.160	0.152
	20	3	0.393	0.391	0.379	0.380
	30	3	0.567	0.565	0.565	0.552

제안한 검정들의 추정된 검정력을 얻고자 각 표본크기에 대해서 10000개의 표본들을 선택한 대립가설의 분포 각각으로부터 독립적으로 생성했다. 검정통계량들을 계산하기 위해서는 윈도크기 m의 값이 정해져야만 한다. 최적의 윈도크기를 결정하는 이론적인 방법은 아직까지는 제안되어져 있지 않다. 문제 해결을 위한 대안으로 Vasicek (1976)은 검정력이 가장 크게 나오는 윈도크기를 선택할 것을 권장했다. 모의실험을 통해 얻은 검정력 연구의 결과를 토대로 Vasicek (1976)은 윈도크기에 대한 최적의 값으로 $n = 10$일 때는 $m = 2$, $n = 20, 30$일 때는 $m = 3$, $n = 50$일 때는 $m = 4$를 사용할 것을 추천했다. 이 추천에 따라 표본크기에 따른 이들 윈도크기의 값을 이용하여 검정통계량들을 계산했다. 각 검정통계량에 대한 추정된 검정력은 검정통계량별로 계산한 10000개의 값들 중에서 주어진 표본크기와 윈도크기에 해당하는 기각값보다 작게 나온 값들의 빈도를 세고 이것을 10000으로 나눈 값으로 구했다.

Table 4.1은 유의수준 5%에 대한 각 검정의 추정된 검정력이다. 표의 마지막 열은 제안한 검정들의 경쟁자로 선택한 Vasicek (1976)의 $K_{m,n}$ 검정에 대한 검정력이다. $K_{m,n}$ 검정은 Vasicek (1976)의 모의실험을 통한 검정력 비교에서 경험적 분포함수에 기초한 EDF 검정과 Shapiro-Wilk 등 기존에 개발된 표준적인 정규성 검정보다 더 좋은 검정력을 가지는 것으로 나타난다. 제시된 표의 결과에서 보는 바와 같이 모든 대립분포와 표본크기에서 $TH_{m,n}$은 $THC_{m,n}$, $TC_{m,n}$과 $K_{m,n}$에 비해서 가장 좋은 검정력을 가지는 것으로 나타난다. $THC_{m,n}$, $TC_{m,n}$과 $K_{m,n}$에 대한 $TH_{m,n}$의 검정력 차이는 표본크기가 작을 때에는 크게 되고 표본크기가 커짐에 따라 감소하는 경향을 보인다. $THC_{m,n}$는 검정통계량들 중에서 두번째로 좋은 성능을 보이고 $TH_{m,n}$과는 거의 비슷한 수준의 검정력을 가짐을 볼 수 있다. $TC_{m,n}$은 표본크기가 $n = 20, 30$일 때 균일분포 $U(0, 1)$에서 가장 높은 검정력을 보이게 되지만 이 경우를 제외하고는 $TH_{m,n}$과 $THC_{m,n}$에 비해서 낮은 검정력을 가지게 된다. $K_{m,n}$의 경우는 검정력이 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$보다는 낮게 관측이 되므로 전반적으로 다른 검정통계량들보다는 성능이 다소 떨어지는 것을 알 수 있다. 그러나, 표본크기가 커짐에 따라 $K_{m,n}$과 다른 검정들과의 검정력 차이는 어느 정도 줄어드는 것을 볼 수 있다.

5. 결론

본 논문에서는 모수적과 비모수적 엔트로피 추정량들에 기초한 정규분포에 대한 적합도 검정을 다루었다. 정규분포의 엔트로피에 대한 모수적 추정량으로 사용할 최소분산비편향추정량을 유도했다. 이 추정량과 대립가설 하에서의 자료생성분포에 대한 비모수적 엔트로피 추정량으로 표본엔트로피와 이것의 변형된 추정량들을 사용하여 구축한 검정통계량 $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$들을 제시했다. 이들 검정통계량을 이용하는 검정은 표본크기와 윈도크기가 커짐에 따라 모든 대립가설에 대해서 검정력이 1이 되는 일치성을 가짐을 보였다. 제안한 검정들에서 사용할 기각값은 영가설하에서의 검정통계량들의 표본분포로부터 결정이 된다. 그러나, 해석적인 방법에 의한 표본분포의 도출은 아직까지도 해결이 되지 못하고 있는 문제이다. 이런 이유로 주어진 유의수준에 대한 기각값은 모의실험을 통해 추정했고 결과의 일부분을 표의 형태로 제시했다. $TH_{m,n}$, $THC_{m,n}$과 $TC_{m,n}$ 검정들의 성능을 조사하기 위해서 모의실험을 수행했다. 결과에서 $TH_{m,n}$, $THC_{m,n}$, $TC_{m,n}$의 순으로 검정력이 높게 나타났고 $THC_{m,n}$는 $TH_{m,n}$와 거의 대등한 수준의 검정력을 보이는 것을 확인할 수 있었다. 기존의 개발되어 많이 활용되는 EDF 검정과 Shapiro-Wilk 검정보다 더 좋은 검정력을 보여주는 Vasicek (1976)의 $K_{m,n}$ 검정과의 비교에서 제안한 검정들은 더 나은 검정력을 보였다. 응용에서 새로운 검정들이 정규성 검정을 위한 경쟁적인 도구로 사용될 수 있을 것으로 기대된다.

References

Choi, B. and Kim, K. (2006). Testing goodness-of-fit for Laplace distribution based on maximum entropy. *Statistics*. **40**, 517–531.

Correa, J. C. (1995). A new estimator of entropy. *Communications in Statistics-Theory and Methods*, **24**, 2439-2449.

Cressie, N. (1976). On the logarithms of high-order spacings. *Biometrika*, **63**, 343–355.

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 test for uniformity. *Journal of the American Statistical Association*, **76**, 967–974.

Dudewicz, E. J. and van der Meulen, E. C. (1987). *New perspectives in theoretical and applied statistics*, Wiley, New York.

Ebrahimi, N., Pflughoeft, K. and Soofi, E. S. (1994). Two measures of sample entropy. *Statistics and Probability Letters*, **20**, 225–234.

Grzegorzewski. P. and Wieczorkowski, R. (1999). Entropy-based goodness-of-fit test for exponentiality. *Communications in Statistics-Theory and Methods*, **28**, 1183–1202.

Györfi, L. and van der Meulen, E. C. (1987). Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis*, **5**, 425-436.

Lee, S., Lee, J. and Noh, J. (2013). Maximum entropy test for infinite order autoregressive models. *Journal of the Korean Data & Information Science Society*, **24**, 637-642.

Mardia, K. V. (1980). Tests of univariate and multivariate normality. In *Handbook of Statistics*, Vol. 1., edited by P. K. Krishnaiah, North-Holland, Amsterdam, 279–320.

Shannon, C. E. (1948). A mathematical theory of communications. *Bell System Technical Journal*, **27**, 379–423, 623–656.

van Es, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, **19**, 61–72.

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

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

Journal of the Korean Data & Information Science Society
2013, **24**(4), 847–856

http://dx.doi.org/10.7465/jkdi.2013.24.4.847
한국데이터정보과학회지

Goodness-of-fit test for normal distribution based on parametric and nonparametric entropy estimators

Byungjin Choi[1]

[1]Department of Applied Information Statistics, Kyonggi University

Received 20 June 2013, revised 11 July 2013, accepted 16 July 2013

Abstract

In this paper, we deal with testing goodness-of-fit for normal distribution based on parametric and nonparametric entropy estimators. The minimum variance unbiased estimator for the entropy of the normal distribution is derived as a parametric entropy estimator to be used for the construction of a test statistic. For a nonparametric entropy estimator of a data-generating distribution under the alternative hypothesis sample entropy and its modifications are used. The critical values of the proposed tests are estimated by Monte Carlo simulations and presented in a tabular form. The performance of the proposed tests under some selected alternatives are investigated by means of simulations. The results report that the proposed tests have better power than the previous entropy-based test by Vasicek (1976). In applications, the new tests are expected to be used as a competitive tool for testing normality.

Keywords: Entropy, entropy estimator, goodness-of-fit, normal distribution, power.

[1] Associate professor, Department of Applied Information Statistics, Kyonggi University, Gyeonggi-Do 443-760, Korea. E-mail: bjchoi92@kyonggi.ac.kr