한국통계학회논문집
2011, 18권, 5호, 625–636

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.5.625

로그정규분포의 엔트로피에 대한 두 모수적 추정량의 비교

최병진[1,a]

[a]경기대학교 응용정보통계학과

요 약

본 논문에서는 로그정규분포의 엔트로피에 대한 모수적 추정량으로 최소분산비편향추정량과 최대가능도추정량을 제시하고 성질을 비교한다. 각 추정량의 분산을 유도해서 일치성을 밝히고 최대가능도추정량의 편향이 추정에 미치는 영향을 분석한다. 델타근사방법을 이용해서 얻은 추정량의 분포를 제시하고 적합도 평가를 통한 유도한 분포의 확증을 위해서 모의실험을 수행한다. 평균제곱오차에 의한 상대적 효율성에 대한 조사를 통해 두 추정량의 성능을 비교한다. 모의실험의 결과에서 최소분산비편향추정량은 최대가능도추정량보다 더 좋은 효율을 보이는 것으로 나타나며, 특히 표본크기와 분산이 동시에 작아짐에 따라 효율이 점점 높아지게 되어 월등히 나은 성능을 발휘함을 볼 수 있다.

주요용어: 로그정규분포, 엔트로피, 모수적 엔트로피추정량, 일치성, 정규근사, MSE 효율성.

1. 서론

많은 응용 분야에서 관심의 대상이 되는 변수를 측정한 자료에 나타나는 변동을 기술하고 분석하기 위해서 거의 대부분 정규분포를 가정한다. 주된 이유는 정규성하에서 개발된 다양한 통계적 방법을 분석 도구로 활용할 수 있을 뿐만 아니라 목적에 부합하는 새로운 방법의 유도가 다른 분포에 비해 용이하고 우수한 결과를 제공하기 때문이다. 분석을 위한 확률모형으로 정규분포의 사용이 정당화되려면 측정한 자료는 양의 값만 가지고 있지만, 이론적으로는 양과 음의 값 모두를 가질 수 있어야 하는 지지가 필요하다.

그러나, 무게, 높이와 밀도 등과 같이 근원적으로 양의 값만 가지는 변수를 측정한 자료를 분석해야 하는 경우가 흔히 발생하게 되고 적어도 이런 이론적 측면에서 정규분포 대신에 로그정규분포가 더 현실적인 확률모형이 될 수 있다. 게다가 척도모수를 충분히 작게 함으로써 로그정규분포를 정규분포와 매우 닮은 형태로 구축하는 것이 가능하게 되므로 비록 정규분포가 아주 적합하다는 생각을 하더라고 적절한 로그정규분포로 대체할 수가 있다.

로그정규분포는 서로 독립이고 양의 값만 가지는 n개의 확률변수 $X_1, X_2, \ldots, X_n$으로부터 얻은 $\log T_n = \sum_{i=1}^{n} \log X_i$의 표준화된 확률변수가 중심극한정리를 적용하면 표준정규분포를 하게 될 것이라는 사실로부터 $T_n = \prod_{i=1}^{n} X_i$의 극한분포로 유도가 될 수 있음을 Galton (1879)에 의해 처음으로 언급이 되었다. Kapteyn (1903)이 위에 기술한 같은 방침으로 로그정규분포의 발생을 다시 고찰하기 전까지는 별다른 관심을 받지 못했으며, Kapteyn과 van Uven (1916)이 모수를 추정하기 위한 표본분위수에 기초한 도식적인 방법을 제시한 이후 로그정규분포에 대한 많은 연구가 이루어져 왔다 (Nydell, 1919; Davies, 1929; Finney, 1941; Aitchison과 Brown, 1957; Koopmans 등, 1964; Wu, 1966; Nakamura, 1991). 로그정규분포는 지질학, 대기과학, 생태학, 미생물학, 사회과학과 경제학 등 다양한 분야에서 폭넓은

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

응용성을 가짐을 쉽게 찾아볼 수 있고 이들 응용은 경험적 관측에 바탕을 두고 있을 뿐만 아니라 어떤 경우에는 이론적인 논거에 의해 지지를 받을 수가 있다. 이와 관련해서는 Johnson 등 (1994), Crow와 Shimizu (1988)를 참고하기 바란다.

정보이론에서 확률변수와 관련된 불확실성의 측도로 사용하는 엔트로피는 확률변수 X가 확률밀도함수 $f(x)$를 가진다고 하면, 아래와 같이

$$H(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \tag{1.1}$$

로 정의되어진다. 엔트로피의 개념은 통신이론에서 Shannon (1948)에 의해 처음으로 소개된 이후 통계학을 포함한 여러 분야에서 이론 및 응용적인 측면에서 많은 관심을 받아왔으며, Shannon의 엔트로피를 확장하고 일반화하기 위한 많은 시도가 있어 왔다. 그 결과 개발된 다양한 형태의 엔트로피를 문헌에서 찾을 수가 있으며 (Havrda와 Charvat, 1967; Burbea와 Rao, 1982; Kapur과 Kesavan, 1992; Ullah, 1996), 특히 Kullback과 Leibler (1951)에 의해 두 분포간 불일치 정도를 나타내는 측도로 소개된 판별정보함수(discrimination information function)는 모형진단을 위한 다양한 정보지표의 개발에 중요한 역할을 담당하고 있다 (Soofi와 Retzer, 2002).

엔트로피의 추정은 확률밀도함수 $f(x)$가 미지인 경우에는 비모수적 방법을, 확률밀도함수 $f(x)$가 기지인 경우에는 모수적 방법을 이용하게 된다. 비모수적 엔트로피추정량은 커널방법에 의해 추정한 확률밀도함수 $f(x)$의 추정량을 이용하거나, 표본 순서통계량의 차이로 정의되는 m-spacing을 이용해서 얻게 된다. 반면에 모수적 엔트로피추정량은 일치추정량을 사용하게 되는데, 이론적으로 주어지는 엔트로피에 미지의 모수가 포함되어 있으면 이 모수에 대한 일치추정량으로 대체해서 얻게 된다.

본 논문에서는 로그정규분포의 엔트로피에 대한 두 가지 모수적 추정량을 소개한다. 2절에서는 최소분산비편향추정량과 최대가능도추정량을 제시하고 두 추정량의 비교를 위해 평균과 분산을 유도한다. 3절에서는 모의실험을 통해 최대가능도추정량의 편향성 정도를 알아보고 분포적 관점에서 두 추정량의 성질을 비교하기 위해 델타근사방법을 이용해서 각 추정량의 분포를 유도한다. 4절에서는 평균제곱오차에 의한 상대적 효율성의 조사를 통해 두 추정량의 성능을 분석한다. 마지막으로 5절에서는 결론을 맺는다.

2. 엔트로피에 대한 모수적 추정량

로그정규분포 $\mathrm{LN}(\mu,\sigma^2)$를 따르는 확률변수 X는 확률밀도함수로

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\log x-\mu)^2}{2\sigma^2}\right\}, \quad x>0,\ -\infty<\mu<\infty,\ \sigma>0 \tag{2.1}$$

를 가지게 된다. 확률변수 X의 엔트로피는

$$H(f) = \frac{1}{2}\log\sigma^2 + \mu + \frac{1}{2}\log(2\pi e) \tag{2.2}$$

로 주어지며, 두 개의 모수 μ와 σ^2에 의존하고 있음을 볼 수 있다.

$H(f)$의 추정을 위해서 $X_1, X_2, \ldots, X_n$이 $\mathrm{LN}(\mu,\sigma^2)$으로부터 추출된 크기 $n \geq 2$의 확률표본이라고 하자. $Z_i = \log X_i$, $i = 1,\ldots,n$으로 정의하면, $\mathrm{LN}(\mu,\sigma^2)$은 지수족이기 때문에 $(\overline{Z}, \sum_{i=1}^n (Z_i-\overline{Z})^2)$은 (μ,σ^2)에 대해 충분하면서 완비적인 결합통계량이 됨을 알 수 있다. $W = \sum_{i=1}^n (Z_i-\overline{Z})^2$으로 정의하고 통계량 $\varphi(\overline{Z},W)$는 $(\overline{Z},W)$의 함수라고 하면, $H(f)$에 대한 최소분산비편향추정량 H_{mvu}는 $E[\varphi(\overline{Z},W)] = H(f)$가 되는 $\varphi(\overline{Z},W)$를 찾으면 된다.

이제, $\varphi(\overline{Z}, W)$의 결정을 위해 W를 σ^2으로 나눈 통계량 $V = W/\sigma^2$를 고려한다. V에 로그함수를 적용하면 $\log V = \log W - \log \sigma^2$을 얻게 되고, 이것으로부터 $E(\log W) = E(\log V) + \log \sigma^2$이 됨을 알 수 있다. $\log V$의 평균은 $Z_1, Z_2, \ldots, Z_n$이 독립적으로 $N(\mu, \sigma^2)$을 따르기 때문에 V가 자유도 $n-1$인 카이제곱분포를 하게 되는 사실을 이용해서 구하면 된다.

$\log V$의 적률생성함수를 $M_{\log V}(t)$라 하면

$$M_{\log V}(t) = E\left(e^{t \log V}\right) = E\left(V^t\right) = 2^t \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{2.3}$$

로 유도되고, t에 대해 1차 미분한 적률생성함수는

$$M'_{\log V}(t) = \frac{2^t \Gamma'\left(\frac{n-1}{2} + t\right) + 2^t \log 2 \Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{2.4}$$

가 된다. $\log V$의 평균은 $M'_{\log V}(0)$이기 때문에 식 (2.4)에서 $t = 0$을 대입하면

$$M'_{\log V}(0) = E(\log V) = \frac{\Gamma'\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} + \log 2 = \psi\left(\frac{n-1}{2}\right) + \log 2 \tag{2.5}$$

를 얻게 되고, 여기서 $\psi(k)$는 $d \log \Gamma(k)/dk$로 정의되는 디감마함수이다. 따라서, $\log W$의 평균은 식 (2.5)의 결과를 이용하면

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

가 됨을 알 수 있다.

$\log W/2$와 $\overline{Z}$의 합의 평균은 식 (2.6)과 $E(\overline{Z}) = \mu$를 이용하면

$$E\left(\frac{1}{2}\log W + \overline{Z}\right) = \frac{1}{2}\log \sigma^2 + \mu + \frac{1}{2}\psi\left(\frac{n-1}{2}\right) + \frac{1}{2}\log 2 \tag{2.7}$$

가 되는 사실로부터, $E[\varphi(\overline{Z}, W)] = \log \sigma^2/2 + \mu + \log(2\pi e)/2$가 되는 $\varphi(\overline{Z}, W)$는

$$\begin{aligned} \varphi\left(\overline{Z}, W\right) &= \frac{1}{2}\log W + \overline{Z} - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) - \frac{1}{2}\log 2 + \frac{1}{2}\log(2\pi e) \\ &= \frac{1}{2}\log S^2_{n-1} + \overline{Z} - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) + \frac{1}{2}\log\frac{n-1}{2} + \frac{1}{2}\log(2\pi e) \end{aligned} \tag{2.8}$$

로 결정할 수가 있으며, 여기서 $S^2_{n-1} = W/(n-1)$이다. 따라서, 레만-쉐페 정리에 의해 $\varphi(\overline{Z}, W)$는 $H(f)$에 대한 유일한 최소분산비편향추정량이 된다.

H_{mvu}는 비편향추정량이기 때문에 평균은 당연히 $H(f)$가 되고 H_{mvu}의 분산 σ^2_{mvu}은 두 통계량 $\overline{Z}$와 $\log S^2_{n-1}/2$의 합의 분산이 된다. 그런데, $\overline{Z}$와 S^2_{n-1}이 독립이고 $\log S^2_{n-1}$의 분산은 $\log V$의 분산과 같기 때문에 σ^2_{mvu}은 $\log V/2$와 $\overline{Z}$의 분산의 합이 됨을 알 수 있다. $\phi_{\log V}(t) = \log M_{\log V}(t)$로 정의하면 $\log V$의 분산은 $\phi''_{\log V}(0)$으로 구할 수가 있으므로 식 (2.3)으로부터 얻은

$$\phi_{\log V}(t) = \log \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} + t \log 2 \tag{2.9}$$

를 t에 대해 1차 미분해 보면

$$\phi'_{\log V}(t) = \frac{\Gamma'\left(\frac{n-1}{2}+t\right)}{\Gamma\left(\frac{n-1}{2}+t\right)} + \log 2 = \psi\left(\frac{n-1}{2}+t\right) + \log 2 \tag{2.10}$$

가 된다. 이것을 t에 대해 다시 미분하면

$$\phi''_{\log V}(t) = \psi'\left(\frac{n-1}{2}+t\right) \tag{2.11}$$

가 되고 이 식에 $t = 0$을 대입하면 $\log V$의 분산은 $\psi'((n-1)/2)$가 됨을 알 수 있다. 이 결과와 $\overline{Z}$의 분산이 σ^2/n인 사실을 이용하면 H_{mvu}의 분산은

$$\sigma^2_{mvu} = \frac{1}{4}\psi'\left(\frac{n-1}{2}\right) + \frac{\sigma^2}{n} \tag{2.12}$$

로 얻게 된다.

H_{mvu}의 분산이 $H(f)$에 대한 비편향추정량 $\tilde{H}$의 분산의 최소가 되는 지는 $\tilde{H}$의 분산에 대한 크래머-라오 부등식

$$\sigma^2_{\tilde{H}} \geq \frac{1}{n}\nabla H(f)^t \left\{I\left(\mu,\sigma^2\right)\right\}^{-1} \nabla H(f) \tag{2.13}$$

을 구해보면 된다. 기울기 벡터 $\nabla H(f)$와 단위 피셔 정보행렬 $I(\mu,\sigma^2)$는 각각

$$\nabla H(f) = \left(\frac{\partial H(f)}{\partial \mu}, \frac{\partial H(f)}{\partial \sigma^2}\right)^t = \left(1, \frac{1}{2\sigma^2}\right)^t, \tag{2.14}$$

$$\begin{aligned} I\left(\mu,\sigma^2\right) &= \begin{pmatrix} E\left[-\frac{\partial^2 \log f(X_1)}{\partial \mu^2}\right] & E\left[-\frac{\partial^2 \log f(X_1)}{\partial \mu \partial \sigma^2}\right] \\ E\left[-\frac{\partial^2 \log f(X_1)}{\partial \sigma^2 \partial \mu}\right] & E\left[-\frac{\partial^2 \log f(X_1)}{\partial \sigma^2}\right] \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{pmatrix} \end{aligned} \tag{2.15}$$

로 계산되므로 $\tilde{H}$의 분산에 대한 크래머-라오 부등식은 $\sigma^2_{\tilde{H}} \geq (\sigma^2 + 1/2)/n$로 얻게 된다. 표 1은 표본크기 n에 대해서 $d_n = \sigma^2_{mvu} - (\sigma^2 + 1/2)/n = \psi'((n-1)/2)/4 - 1/(2n)$의 값을 계산해 본 것으로 H_{mvu}의 분산은 크래머-라오 하한을 약간 초과한다는 것을 알 수 있다. 그러나, d_n 값은 표본크기가 커짐에 따라 0으로 수렴하게 되어서 H_{mvu}의 분산은 점근적으로 크래머-라오 하한이 되는 것을 볼 수 있다.

표본크기가 무한히 커질 경우 H_{mvu}의 분산의 극한행태를 알아보고자 디감마함수에 대한 급수전개식을 적용하면 $\psi'((n-1)/2) = \sum_{k=0}^{\infty}\{2/(n+2k-1)\}^2$로 얻게 된다. 이것을 이용하면 $\sigma^2_{mvu} = \sum_{k=0}^{\infty}\{1/(n+2k-1)\}^2 + \sigma^2/n$로 표현이 된다. 주어진 k에 대해서 $\lim_{n\to\infty}\{1/(n+2k-1)\}^2 = 0$이 되므로 $\lim_{n\to\infty}\sum_{k=0}^{\infty}\{1/(n+2k-1)\}^2 = \sum_{k=0}^{\infty}\lim_{n\to\infty}\{1/(n+2k-1)\}^2 = 0$이 됨을 알 수 있다. 또한 $\lim_{n\to\infty}\sigma^2/n = 0$이므로 $\lim_{n\to\infty}\sigma^2_{mvu} = 0$이 된다. 따라서, H_{mvu}는 $H(f)$로 확률수렴하게 되므로 일치성을 가짐을 알 수 있다.

표 1: n에 대해서 계산된 d_n의 값

n	d_n	n	d_n	n	d_n
2	0.9837005	20	0.0027493	80	0.0001599
4	0.1087005	25	0.0017255	100	0.0001019
6	0.0392561	30	0.0011829	125	0.0000649
8	0.0200894	35	0.0008612	150	0.0000450
10	0.0121813	40	0.0006549	175	0.0000330
15	0.0050530	60	0.0002865	200	0.0000252

로그정규분포에 대한 모수적 엔트로피추정량으로 엔트로피에 포함된 μ와 σ^2을 $\hat{\mu} = \overline{Z}$와 $\hat{\sigma^2} = S_n^2 = \sum_{i=1}^{n}(Z_i - \overline{Z})^2/n$으로 대체한 추정량

$$\begin{aligned} H_{ml} &= \frac{1}{2}\log S_n^2 + \overline{Z} + \frac{1}{2}\log(2\pi e) \\ &= \frac{1}{2}\log S_{n-1}^2 + \overline{Z} + \frac{1}{2}\log\frac{n-1}{n} + \frac{1}{2}\log(2\pi e) \end{aligned} \tag{2.16}$$

을 고려하면, $\hat{\mu}$와 $\hat{\sigma^2}$은 μ와 σ^2의 최대가능도추정량이므로 불변성에 의해서 엔트로피에 대한 최대가능도추정량이 된다. H_{ml}은 H_{mvu}에 관해서

$$H_{ml} = H_{mvu} + \frac{1}{2}\psi\left(\frac{n-1}{2}\right) - \frac{1}{2}\log\frac{n}{2} \tag{2.17}$$

으로 표현될 수가 있고, 이것을 이용하면 평균과 분산은 각각

$$\mu_{ml} = H(f) + \frac{1}{2}\psi\left(\frac{n-1}{2}\right) - \frac{1}{2}\log\frac{n}{2}, \tag{2.18}$$

$$\sigma_{ml}^2 = \frac{1}{4}\psi'\left(\frac{n-1}{2}\right) + \frac{\sigma^2}{n} \tag{2.19}$$

가 됨을 알 수 있다. H_{ml}은 H_{mvu}과 같은 분산을 가지게 되며, 표본크기가 커짐에 따라 0으로 수렴하게 되어 일치성을 가지게 된다.

한편, H_{ml}의 편향은 $b_n = \psi((n-1)/2)/2 - \log(n/2)/2$이 되고 유한한 n에 대해서 $b_n \neq 0$이 되기 때문에 H_{ml}은 $H(f)$에 대해 편향적이 된다. 편향이 추정에 미치는 효과를 알아보고자 b_n의 디감마함수를 정적분의 형태로 표현해 보면

$$\frac{1}{2}\psi\left(\frac{n-1}{2}\right) = \frac{1}{2}\log\frac{n-1}{2} - \frac{1}{2(n-1)} - \int_0^\infty \frac{4t}{\left(4t^2 + (n-1)^2\right)\left(e^{2\pi t} - 1\right)}\,dt \tag{2.20}$$

가 된다. 식 (2.20)의 마지막 항은 $y = 2\pi t$로 치환하고 $y \geq 0$에 대해서 $e^y - 1 \geq y$를 이용하면

$$\begin{aligned} I = \int_0^\infty \frac{4t}{\left(4t^2 + (n-1)^2\right)\left(e^{2\pi t} - 1\right)}\,dt &= \int_0^\infty \frac{y}{\left(y^2 + (n-1)^2\pi^2\right)\left(e^y - 1\right)}\,dy \\ &\leq \int_0^\infty \frac{1}{y^2 + (n-1)^2\pi^2}\,dy \\ &\leq \frac{1}{n-1}\int_0^\infty \frac{n-1}{y^2 + (n-1)^2\pi^2}\,dy \end{aligned} \tag{2.21}$$

로 얻게 된다. 그런데, 식 (2.21)의 우변에서 적분함수는 코쉬분포 Cauchy$(0,(n-1)\pi^2)$의 확률밀도함수이므로 $\int_0^\infty (n-1)/(y^2+(n-1)^2\pi^2)\,dy = 1/2$이 되고, 이것으로부터 $0 \le I \le 1/(2(n-1))$가 유도된다. 이 결과와 $n \ge 2$에 대해 $1/(n-1) > 0$인 사실로부터,

$$\frac{1}{2}\psi\left(\frac{n-1}{2}\right) \le \frac{1}{2}\log\frac{n-1}{2} - \frac{1}{2(n-1)} < \frac{1}{2}\log\frac{n-1}{2} < \frac{1}{2}\log\frac{n}{2} \tag{2.22}$$

가 얻어지므로 $b_n < 0$이 됨을 알 수 있다. 따라서, $E(H_{ml}) < H(f)$가 되므로 H_{ml}은 $H(f)$에 대해 과소추정을 하게 된다.

표본크기를 무한히 크게 할 경우, b_n의 극한을 보기 위해 식 (2.20)과 I에 관한 부등식으로부터 얻게 되는

$$\frac{1}{2}\log\frac{n-1}{2} - \frac{1}{n-1} \le \frac{1}{2}\psi\left(\frac{n-1}{2}\right) \le \frac{1}{2}\log\frac{n-1}{2} - \frac{1}{2(n-1)} \tag{2.23}$$

를 이용해서 b_n의 대소관계를 표현해 보면

$$\frac{1}{2}\log\left(1-\frac{1}{n}\right) - \frac{1}{n-1} \le b_n \le \frac{1}{2}\log\left(1-\frac{1}{n}\right) - \frac{1}{2(n-1)} \tag{2.24}$$

가 된다. 따라서, $\lim_{n\to\infty}\{\log(1-1/n)/2 - 1/(n-1)\} = 0$, $\lim_{n\to\infty}\{\log(1-1/n)/2 - 1/(2(n-1))\} = 0$이므로 $\lim_{n\to\infty} b_n = 0$이 된다. 그러므로, H_{ml}은 $H(f)$에 수렴하게 되어 점근적 비편향성을 가지게 된다.

3. 추정량의 성질 비교

2절에서 본 바와 같이, 로그정규분포의 엔트로피 $H(f)$에 대해서 H_{mvu}는 비편향성을 가지는 반면에 H_{ml}은 편향성을 가지게 된다. H_{mvu}의 비편향성과 표본크기에 따른 H_{ml}의 편향성 정도를 확인하기 위해서 몬테칼로 모의실험을 수행해 보기로 한다. 몬테칼로 연구에서 선택한 분포는 LN$(0,0.5)$와 LN$(0,2)$이고 각 분포에 대한 이론적인 엔트로피는 $H(f) = 1.0724, 1.7655$가 된다. 크기 $n = 2,4,\ldots,100$인 표본을 두 분포에서 독립적으로 생성해서 H_{mvu}와 H_{ml}의 값을 계산하고 이런 과정을 1000번 반복을 해서 얻은 각 추정량의 계산값들의 평균을 $H(f)$에 대한 추정값으로 사용한다.

그림 1은 이론적인 엔트로피와 모의실험에 의해 표본크기별로 추정한 H_{mvu}와 H_{ml}의 값을 그린 것이다. 그림 1(a)는 LN$(0,0.5)$에 대한 것으로 모든 표본크기에서 H_{mvu}의 값은 $H(f)$의 값과 유사하게 나타나므로 비편향성을 가지는 것을 볼 수 있다. 그러나, H_{ml}의 값은 표본크기가 작아짐에 따라 $H(f)$의 값에 비해 현저히 작게 나타나고 있어서 편향성이 커지게 되고 이로 인해 $H(f)$를 상당한 정도로 과소추정을 하게 되는 현상을 볼 수 있다. 그러나, 표본크기가 커짐에 따라 H_{ml}의 값은 $H(f)$의 값에 근접하게 되므로 비편향성을 가지게 된다. 이런 현상은 LN$(0,2)$에 대한 그림 1(b)에서도 확인할 수가 있다.

다음은 분포적 관점에서 두 추정량의 성질을 비교해 보기로 한다. H_{mvu}와 H_{ml}의 분포행태는 두 통계량 $\log S_{n-1}^2/2$와 $\overline{Z}$의 분포행태에 의해 결정이 된다. 먼저, $\overline{Z}$의 분포를 구해보면 $N(\mu,\sigma^2/n)$가 되는 것을 알 수 있다. $\log S_{n-1}^2/2$의 분포는 V에 관한 표현식 $\log S_{n-1}^2/2 = \log V/2 + \log\sigma^2/2 - \log(n-1)/2$과 V가 자유도 $n-1$인 카이제곱분포를 따르게 되는 사실을 이용해서 구할 수가 있지만, 로그변환된 카이제곱 확률변수는 정규분포에 의해 근사가 잘 되는 것으로 알려져 있으므로 (Olshen, 1937), 델타근사방법을 적용하여 추정량의 분포를 구해보기로 한다.

통계량 T_n을

$$T_n = \frac{1}{2}\log S_{n-1}^2 - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) + \frac{1}{2}\log\frac{n-1}{2} \tag{3.1}$$

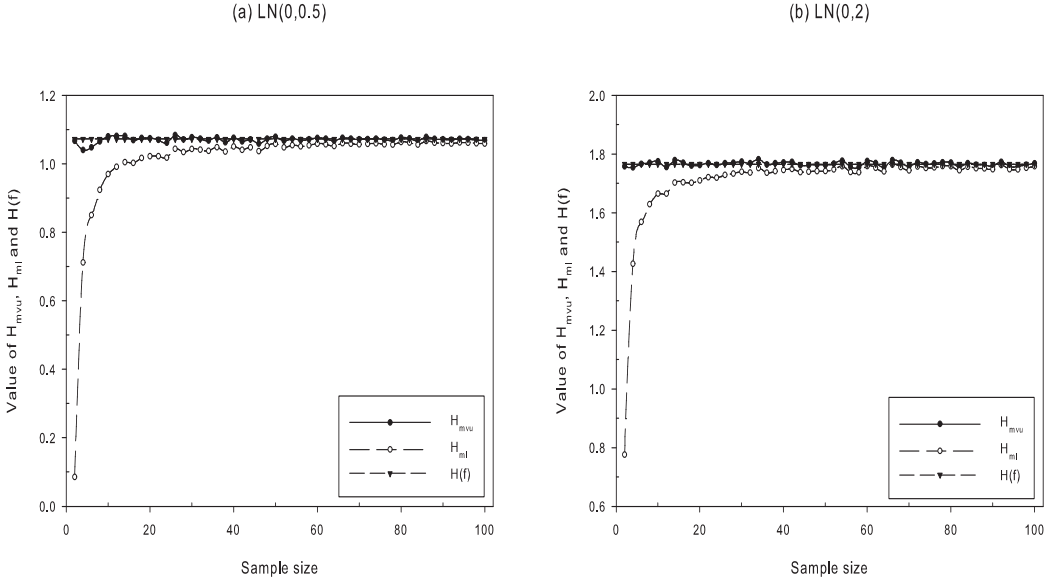

그림 1: *이론적인 엔트로피와 모의실험에 의한 표본크기별 H_{mvu}와 H_{ml}의 추정값*

으로 정의하면 $H_{mvu} = T_n + \overline{Z} + \log(2\pi e)/2$로 표현할 수 있다. $\log S_{n-1}^2/2$을 S_{n-1}^2의 평균에 대해서 테일러전개를 시키면

$$\frac{1}{2}\log S_{n-1}^2 \simeq \frac{1}{2}\log\sigma^2 + \frac{\left(S_{n-1}^2 - \sigma^2\right)}{2\sigma^2} \tag{3.2}$$

가 되고 이 결과를 이용해서 얻게 되는 T_n의 근사식은

$$T_n \simeq \frac{1}{2}\log\sigma^2 + \frac{\left(S_{n-1}^2 - \sigma^2\right)}{2\sigma^2} - \frac{1}{2}\psi\left(\frac{n-1}{2}\right) + \frac{1}{2}\log\frac{n-1}{2} \tag{3.3}$$

가 된다. 식 (3.3)에서 양변을 $2\sigma^2$을 곱한 다음, S_{n-1}^2의 분산의 제곱근으로 나누어서 정리하면

$$\sqrt{2(n-1)}\left(T_n - \frac{1}{2}\log\sigma^2\right) \simeq \frac{\left(S_{n-1}^2 - \sigma^2\right)}{\sigma^2\sqrt{\frac{2}{n-1}}} - \sqrt{\frac{n-1}{2}}\left\{\psi\left(\frac{n-1}{2}\right) - \log\frac{n-1}{2}\right\} = Z_{S_{n-1}^2} - a_{1,n} \tag{3.4}$$

가 된다. $n \to \infty$이면 $Z_{S_{n-1}^2}$은 중심극한정리에 의해서 $N(0,1)$로 수렴하게 된다. $a_{1,n}$은 식 (2.23)에 의해 $-2/(n-1) \le \psi((n-1)/2) - \log((n-1)/2) \le -1/(n-1)$가 되므로 0으로 수렴하게 된다. 그러므로, 좌변은 $N(0,1)$로 수렴하게 되고 T_n은 근사적으로 $N(\log\sigma^2/2, 1/\{2(n-1)\})$을 따른다고 할 수 있다. 이 결과를 이용해서 T_n과 $\overline{Z}$의 합의 분포를 구해보면 S_{n-1}^2과 $\overline{Z}$가 독립이기 때문에 근사적으로 $N(\log\sigma^2/2 + \mu, 1/\{2(n-1)\} + \sigma^2/n)$가 됨을 알 수 있다. 그러므로, H_{mvu}의 분포는 근사적으로 $N(H(f), 1/\{2(n-1)\} + \sigma^2/n)$가 된다.

H_{ml}의 분포 또한 같은 방식으로 유도할 수가 있다. $\log S_n^2/2 = \log S_{n-1}^2/2 + \log((n-1)/n)$로부터 식 (3.2)를 이용해서 얻게 되는 근사식은

$$\frac{1}{2}\log S_n^2 \simeq \frac{1}{2}\log\sigma^2 + \frac{\left(S_{n-1}^2 - \sigma^2\right)}{2\sigma^2} + \frac{1}{2}\log\frac{n-1}{n} \tag{3.5}$$

표 2: n에 대해서 계산된 $1/4\,\psi'(n-1)/2$, $1/\{2\,(n-1)\}$과 d_n^*

n	$\frac{1}{4}\psi'\left(\frac{n-1}{2}\right)$	$\frac{1}{2(n-1)}$	d_n^*	n	$\frac{1}{4}\psi'\left(\frac{n-1}{2}\right)$	$\frac{1}{2(n-1)}$	d_n^*
20	0.0277493	0.0263158	0.0014335	120	0.0042372	0.0042017	0.0000355
40	0.0131549	0.0128205	0.0003343	140	0.0036231	0.0035971	0.0000260
60	0.0086198	0.0084746	0.0001453	160	0.0031645	0.0031447	0.0000199
80	0.0064099	0.0063291	0.0000808	180	0.0028090	0.0027933	0.0000157
100	0.0051019	0.0050505	0.0000514	200	0.0025252	0.0025126	0.0000127

표 3: 정규성에 대한 검정 결과

검정	$N(1.419, 0.015)$				$N(1.968, 0.035)$			
	H_{mvu}	유의확률	H_{ml}	유의확률	H_{mvu}	유의확률	H_{ml}	유의확률
D 검정	0.0785	$p > 0.25$	0.0595	$p > 0.25$	0.0542	$p > 0.25$	0.0468	$p > 0.25$
W^2 검정	0.2051	$p > 0.25$	0.0505	$p > 0.25$	0.0638	$p > 0.25$	0.0327	$p > 0.25$
A^2 검정	1.1035	$p > 0.25$	0.4027	$p > 0.25$	0.5178	$p > 0.25$	0.2260	$p > 0.25$

이다. 식 (3.5)의 양변을 $2\sigma^2$을 곱한 다음, S_{n-1}^2의 분산의 제곱근으로 나누어서 정리해 보면

$$\sqrt{2\,(n-1)}\left(\frac{1}{2}\log S_n^2 - \frac{1}{2}\log\sigma^2\right) \simeq \frac{\left(S_{n-1}^2 - \sigma^2\right)}{\sigma^2\sqrt{\frac{2}{n-1}}} + \sqrt{\frac{n-1}{2}}\log\frac{n-1}{n} = Z_{S_{n-1}^2} + a_{2,n} \tag{3.6}$$

가 된다. $n \to \infty$이면 $Z_{S_{n-1}^2}$은 중심극한정리에 의해서 $N(0,1)$로, $a_{2,n}$은 0으로 수렴하게 되므로 좌변 또한 $N(0,1)$로 수렴함을 알 수 있다. 따라서, $\log S_n^2/2$는 근사적으로 $N(\log\sigma^2/2, 1/\{2\,(n-1)\})$를 따르게 되고 $\overline{Z}$ 역시 정규분포를 하게 되므로 H_{ml}의 근사분포는 $N(H\,(f)\,, 1/\{2\,(n-1)\} + \sigma^2/n)$가 된다.

H_{mvu}와 H_{ml}은 동일한 근사분포를 가지지만 분산은 2절에서 유도한 것과는 같지 않음을 볼 수 있다. 그런데, 트리감마함수 $\psi'\,((n-1)\,/2)$는

$$\psi'\left(\frac{n-1}{2}\right) \simeq \frac{2}{n-1} + \frac{1}{2}\left(\frac{2}{n-1}\right)^2 + \frac{1}{6}\left(\frac{2}{n-1}\right)^3 - \frac{1}{30}\left(\frac{2}{n-1}\right)^5 + \cdots \tag{3.7}$$

로 근사시킬 수가 있으므로 충분히 큰 n에 대해서 $\psi'((n-1)\,/2)/4 = 1/\{2\,(n-1)\,\} + o\,(1/n)$가 될 것으로 예상할 수 있다. 표 2는 n에 대해서 $\psi'((n-1)\,/2)/4$, $1/\{2\,(n-1)\}$과 두 값의 차이 $d_n^* = \psi'((n-1)\,/2)/4 - 1/\{2\,(n-1)\}$을 계산한 것으로 n이 증가함에 따라서 차이는 0으로 접근하는 것을 볼 수 있다. 따라서, $\psi'((n-1)\,/2)/4$는 $1/\{2\,(n-1)\}$에 의해 잘 근사가 되어지므로 σ_{mvu}^2과 σ_{ml}^2은 근사적으로 $1/\{2\,(n-1)\} + \sigma^2/n$와 같게 됨을 알 수 있다. 이 결과로부터 H_{mvu}와 H_{ml}는 평균이 $H\,(f)$이고 분산이 $\psi'((n-1)\,/2)/4 + \sigma^2/n$인 정규분포로 근사된다고 판단할 수 있다.

그림 2와 그림 3은 $n = 100$인 표본을 각각 $\mathrm{LN}(0,1)$과 $\mathrm{LN}(0,3)$에서 500번의 반복 추출을 통해 계산한 값으로부터 얻은 추정량의 분포를 보여주고 있다. 그림 2와 그림 3에서의 실선은 근사분포의 적합곡선을 나타낸 것으로 각각은 $N(1.419, 0.015)$와 $N(1.968, 0.035)$의 확률곡선이다. H_{mvu}의 분포(왼쪽 그림)와 H_{ml}의 분포(오른쪽 그림) 모두는 주어진 정규분포에 잘 적합이 됨을 볼 수 있다. 또한 표 3은 각 추정량에 대해서 계산한 500개의 값이 정규분포를 따르는지를 알아보고자 콜모고로프-스미르노프 검정(D 검정), 크래머-미제스 검정(W^2 검정)과 앤더슨-달링 검정(A^2 검정)을 수행해서 얻은 결과로 모든 검정통계량에 대한 유의확률은 0.25보다 크게 나왔음을 볼 수가 있다. 유의수준 5%에서 정규성에 대한 영가설을 기각할 수가 없기 때문에 두 추정량은 근사적으로 정규분포를 하게 되는 것을 확인할 수가 있다.

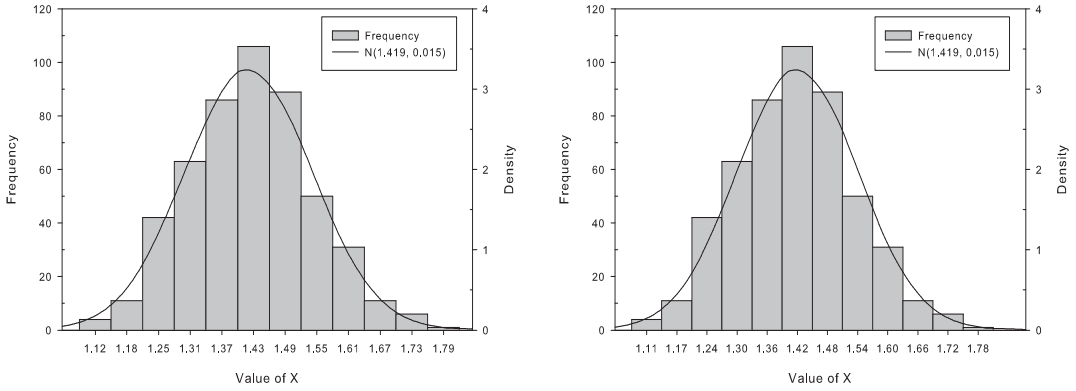

그림 2: $n = 100$일 때 $LN(0,1)$로부터 500번의 반복을 통해 얻은 추정량의 분포(왼쪽 그림은 H_{mvu}의 분포, 오른쪽 그림은 H_{ml}의 분포이고 실선은 $N(1.419, 0.015)$의 확률곡선임)

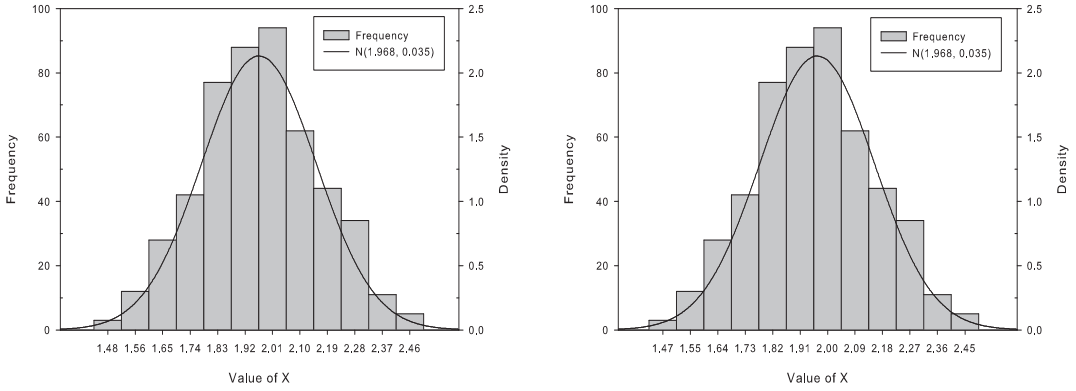

그림 3: $n = 100$일 때 $LN(0,3)$으로부터 500번의 반복을 통해 얻은 추정량의 분포(왼쪽 그림은 H_{mvu}의 분포, 오른쪽 그림은 H_{ml}의 분포이고 실선은 $N(1.968, 0.035)$의 확률곡선임)

4. 추정량의 성능 비교

두 추정량의 성능을 효율성의 평가를 통해 비교해 보기로 한다. 변동성의 관점에서 보면, 두 추정량의 분산은 동일하게 주어지므로 같은 효율을 가지게 된다. 그러나, H_{ml}이 편향적이므로 평균제곱오차(MSE)의 관점에서 비교하는 것이 타당하다. 두 추정량의 평균제곱오차를 각각 구해보면, $\mathrm{MSE}(H_{mvu}) = \sigma^2_{mvu}$과 $\mathrm{MSE}(H_{ml}) = \sigma^2_{ml} + \psi((n-1)/2)/2 - \log(n/2)/2$이 된다. 그러므로, H_{ml}에 대한 H_{mvu}의 상대적 효율성은

$$\mathrm{RE}(H_{ml}, H_{mvu}) = \frac{\mathrm{MSE}(H_{mvu})}{\mathrm{MSE}(H_{ml})} = \frac{\frac{1}{4}\psi'\left(\frac{n-1}{2}\right) + \frac{\sigma^2}{n}}{\frac{1}{4}\psi'\left(\frac{n-1}{2}\right) + \frac{\sigma^2}{n} + \frac{1}{4}\left\{\psi\left(\frac{n-1}{2}\right) - \log\frac{n}{2}\right\}^2} \tag{4.1}$$

로 주어지게 되며 n과 σ^2에 의존하게 된다. n과 σ^2의 변화에 따른 상대적 효율성을 살펴보고자 $\sigma^2 = 0.5, 2$로 하고 $n = 2, 4, \ldots, 100$에 대해서 식 (4.1)의 값을 계산해 보기로 한다. 이와 함께, $\mathrm{LN}(0, 0.5)$와 $\mathrm{LN}(0, 2)$에서 동일한 크기의 표본을 1000번 반복 생성해서 얻은 H_{mvu}와 H_{ml}의 평균제곱오차를 이용해서 $\mathrm{RE}(H_{ml}, H_{mvu})$를 추정해 본다.

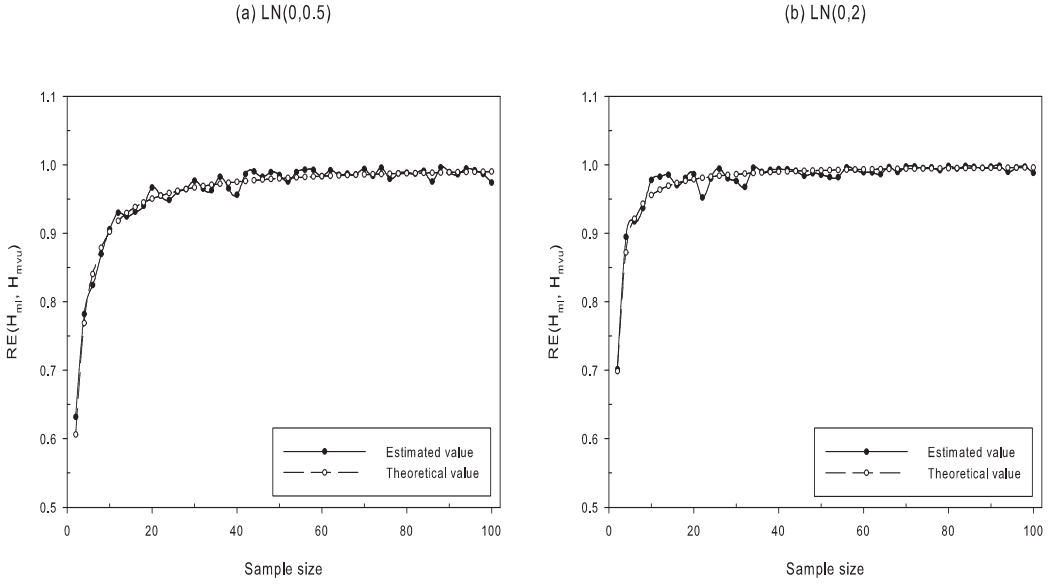

그림 4: *표본크기에 따른 $RE(H_{ml}, H_{mvu})$의 이론값과 모의실험에 의한 추정값*

그림 4는 n과 σ^2에 따른 $\mathrm{RE}(H_{ml}, H_{mvu})$의 이론값과 모의실험에 의한 추정값을 그려놓은 것이다. 그림에서 알 수 있듯이, σ^2이 주어진 경우에는 n이 작아질 때 H_{mvu}가 H_{ml}에 비해 효율이 더 좋게 나타난다. n이 주어진 경우에는 $\sigma^2 = 2$보다는 $\sigma^2 = 0.5$일 때 H_{ml}에 대한 H_{mvu}의 효율이 더 높게 된다. 결과를 종합해 보면, n과 σ^2이 동시에 작아짐에 따라 H_{mvu}는 H_{ml}에 비해 월등히 나은 성능을 발휘하게 되고 n이 커지면 σ^2에 상관없이 $\mathrm{RE}(H_{ml}, H_{mvu})$는 1로 접근하게 되어 두 추정량은 같은 효율을 보이게 된다.

5. 결론

본 논문에서는 로그정규분포의 엔트로피에 대한 모수적 추정량으로 최소분산비편향추정량과 최대가능도추정량을 제시했고 두 추정량의 성질을 변동성과 비편향성 관점에서 살펴 보았다. 추정량의 분포적 성질을 알아보기 위해 델타근사방법을 이용해서 각 추정량의 분포를 유도했고 적합도 평가를 통한 유도한 분포의 확증을 위해서 모의실험을 수행했다. 또한 평균제곱오차에 의한 상대적 효율성에 대한 조사를 통해 두 추정량의 성능을 평가해 보았다.

본 논문에서 도출된 결과를 요약하면 다음과 같다. 유도된 두 추정량의 분산은 같게 나왔으며 이를 통해 동일한 변동성을 가짐을 알 수 있었다. 최대가능도추정량은 최소분산비편향추정량과는 다르게 엔트로피에 대한 편향성을 가지게 되고 편향이 추정에 미치는 영향을 조사해 본 결과, 유한한 표본크기에 대해서 편향은 항상 음의 값을 가지게 되어 과소추정을 하게 함을 알 수 있었다. 그러나, 표본크기가 무한히 커짐에 따라 편향은 0으로 수렴하게 되어 점근적 비편향성을 가지는 것을 확인할 수 있었다. 델타근사방법에 의한 두 추정량의 분포는 동일한 정규분포를 가지는 것으로 나타났고 적합도 평가를 위해 수행한 모의실험의 결과를 통해 적합이 잘 되는 것을 확인할 수 있었다. 성능 비교를 위해 수행한 모의실험으로부터 얻은 평균제곱오차에 의한 상대적 효율성 결과에서 최소분산비편향추정량은 최대가능도추정량보다 더 좋은 효율을 보이는 것으로 나타났으며, 특히 표본크기와 분산이 동시에 작아짐에 따라 효율이 점점 높아지게 되어 월등히 나은 성능을 발휘함을 볼 수 있었다.

참고 문헌

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*, Cambridge University Press, Cambridge.

Burbea, J. and Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach, *Journal of Multivariate Analysis*, **12**, 576–579.

Crow, E. L. and Shimizu, K. (1988). *Lognormal Distributions: Theory and Applications*, Marcel Dekker, New York.

Davies, G. R. (1929). The analysis of frequency distributions, *Journal of the American Statistical Association*, **24**, 467–480.

Finney, D. J. (1941). On the distribution of a variate whose logarithmic is normally distributed, *Journal of the Royal Statistical Society, Series B*, **7**, 155–161.

Galton, F. (1879). The geometric mean in vital and social statistics, *Proceedings of the Royal Society of London*, **29**, 965–967.

Havrda, J. and Charvat, F. (1967). Quantification method in classification processes: Concept of structural α-entropy, *Kybernetika*, **3**, 30–35.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*, John Wiley & Sons, New York.

Kapteyn, J. C. (1903). *Skew Frequency Curves in Biology and Statistics*, Astronomical Laboratory Noordhoff, Groningen.

Kapteyn, J. C. and van Uven, M. J. (1916). *Skew Frequency Curves in Biology and Statistics*, Hotsema Brothers, Groningen.

Kapur, J. N. and Kesavan, H. K. (1992). *Entropy Optimization Principles with Applications*, Academic Press, San Diego.

Koopmans, L. H., Owen, D. B. and Rosenblatt, J. I. (1964). Confidence intervals for the coefficient of variation for the normal and lognormal distributions, *Biometrika*, **51**, 25–32.

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

Nakamura, T. (1991). Existence of maximum likelihood estimates for interval-censored data from some three-parameter models with a shift origin, *Journal of the Royal Statistical Society, Series B*, **53**, 211–220.

Nydell, S. (1919). The mean errors of the characteristics in logarithmic-normal distribution, *Skandinavisk Aktuarietidskrift*, **1**, 134–144.

Olshen, A. C. (1937). Transformations of the Pearson Type III distributions, *The Annals of Mathematical Statistics*, **8**, 176–200.

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

Soofi, E. S. and Retzer, J. J. (2002). Information indices: Unification and applications, *Journal of Econometrics*, **107**, 17–40.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications, *Journal of Statistical Planning and Inference*, **49**, 137–162.

Wu, C. Y. (1966). The types of limit distribution for some terms of variational series, *Scientia Sinica*, **15**, 745–762.

2011년 6월 접수; 2011년 8월 채택

Comparison of Two Parametric Estimators for the Entropy of the Lognormal Distribution

Byungjin Choi[1,a]

[a]Department of Applied Information Statistics, Kyonggi University

Abstract

This paper proposes two parametric entropy estimators, the minimum variance unbiased estimator and the maximum likelihood estimator, for the lognormal distribution for a comparison of the properties of the two estimators. The variances of both estimators are derived. The influence of the bias of the maximum likelihood estimator on estimation is analytically revealed. The distributions of the proposed estimators obtained by the delta approximation method are also presented. Performance comparisons are made with the two estimators. The following observations are made from the results. The MSE efficacy of the minimum variance unbiased estimator appears consistently high and increases rapidly as the sample size and variance, n and σ^2, become simultaneously small. To conclude, the minimum variance unbiased estimator outperforms the maximum likelihood estimator.

Keywords: Lognormal distribution, entropy, parametric entropy estimator, consistency, normal approximation, MSE efficacy.

[1] Associate Professor, Department of Applied Information Statistics, Kyonggi University, Iui-Dong, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do 443-760. E-mail: bjchoi92@kyonggi.ac.kr