한국통계학회논문집
2012, 19권, 6호, 837–847

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.837

Power Investigation of the Entropy-Based Test of Fit for Inverse Gaussian Distribution by the Information Discrimination Index

Byungjin Choi[1,a]

[a]Department of Applied Information Statistics, Kyonggi University

Abstract

Inverse Gaussian distribution is widely used in applications to analyze and model right-skewed data. To assess the appropriateness of the distribution prior to data analysis, Mudholkar and Tian (2002) proposed an entropy-based test of fit. The test is based on the entropy power fraction(EPF) index suggested by Gokhale (1983). The simulation results report that the power of the entropy-based test is superior compared to other goodness-of-fit tests; however, this observation is based on the small-scale simulation results on the standard exponential, Weibull $W(1, 2)$ and lognormal LN$(0.5, 1)$ distributions. A large-scale simulation should be performed against various alternative distributions to evaluate the power of the entropy-based test; however, the use of a theoretical method is more effective to investigate the powers. In this paper, utilizing the information discrimination(ID) index defined by Ehsan *et al.* (1995) as a mathematical tool, we scrutinize the power of the entropy-based test. The selected alternative distributions are the gamma, Weibull and lognormal distributions, which are widely used in data analysis as an alternative to inverse Gaussian distribution. The study results are provided and an illustrative example is analyzed.

Keywords: Inverse Gaussian distribution, entropy, entropy power fraction, information discrimination, test of fit, power.

1. 서론

오른쪽으로 긴 꼬리를 보이는 자료를 분석하기 위한 확률모형으로 폭넓게 사용되는 역가우스분포 IG(μ, λ)는

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0,\ \mu > 0,\ \lambda > 0 \tag{1.1}$$

을 밀도함수로 가지게 되며 μ와 λ는 각각 위치와 척도를 나타내는 모수들이다. 브라운운동에서 첫 통과시간의 분포로 Schrödinger (1915)와 Smoluchowsky (1915)에 의해 독립적으로 처음 제시된 역가우스분포는 Tweedie (1957a, 1957b)의 초기 연구 이후에 자료분석을 위한 확률모형으로 많은 관심을 받아왔다. 이론적 또는 응용적인 측면에서 역가우스분포의 유용성을 보여주는 다양한 사례들은 Chhikara와 Folks (1989)와 Seshadri (1999)에서 찾아볼 수 있다.

자료분석에 앞서서 사용할 확률모형으로 역가우스분포의 적합성을 확인하는 것은 매우 중요하다. 역가우스분포에 대한 적합도 검정법을 다루는 문헌은 정규분포나 지수분포에 비해서 비교적 많지 않지만 Edgeman 등 (1988)의 콜모고로프-스미르노프 검정, 경험적 분포함수에 기초를 두는 Edgeman

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

(1990)의 EDF 검정들과 Shannon (1948)의 엔트로피를 이용하고 있는 Mudholkar와 Tian (2002)의 엔트로피 기반 검정이 개발되어 있다. 특히 엔트로피 기반 검정은 모의실험에서 로그정규분포에 대한 검정력은 설정된 유의수준 5%를 조금 상회하는 정도의 아주 낮은 검정력을 가지는 것으로 나타나지만 대체적으로 기존의 적합도 검정들보다 좋은 성능을 보이는 것으로 Mudholkar와 Tian (2002)은 보고하고 있다. 그러나, 그들은 대립가설로 선택한 표준지수분포, 와이블분포 $W(1,2)$와 로그정규분포 LN(0.5, 1)에 대한 소규모의 모의실험에 의한 검정력만을 제시하고 있어 이 결과를 통해 검정력을 평가하기에는 제한적일 수 밖에 없다. 검정력에 대한 전반적인 조사를 하기 위해서는 여러 다양한 대립분포에 대한 대규모의 모의실험을 수행해서 얻은 결과가 필요하다. 하지만, 이런 경험적 방법을 이용하기 보다는 이론적인 도구을 통해 검정력을 규명하는 것이 더 효율적일 수가 있다.

엔트로피 기반 검정은 두 분포 f와 f^*에 대한 EPF(entropy power fraction) 지표인 $\mathrm{EPF}(f) = \exp\{H(f)\}/\exp\{H(f^*)\}$를 고려하여 제안된 Gokhale (1983)의 검정법으로 볼 수가 있고 $H(f)$와 $H(f^*)$는 각각 f와 f^*의 Shannon (1948)의 엔트로피이다. Ehsan 등 (1995)은 두 분포 f와 f^*간의 정보 불일치를 재는 측도로 알려져 있는 쿨백-라이블러 함수 $K(f : f^*) = \int f(x) \ln\{f(x)/f^*(x)\}\, dx$에 기초를 둔 ID 지표로 $\mathrm{ID}(f : f^*) = 1 - \exp\{-K(f : f^*)\}$를 제시했다. 주어진 제약하에서 f^*가 최대엔트로피를 가지는 분포라면, $\mathrm{ID}(f : f^*) = 1 - \mathrm{EPF}(f)$가 성립한다. Ehsan 등 (1995)이 언급한 응용적인 측면에서의 ID 지표의 유용성에도 불구하고 아직까지는 최대엔트로피에 기초를 둔 검정의 검정력 연구에서 많이 활용되고 있지는 않는 것 같다.

본 본문에서는 ID 지표 $\mathrm{ID}(f : f^*)$를 이론적 도구로 활용하여 엔트로피 기반 검정의 검정력을 조사하고자 한다. 대립가설의 분포로는 감마분포, 와이블분포와 로그정규분포를 고려한다. 이들 분포는 치우친 모습을 보이는 자료를 분석하고 모형화하기 위한 확률모형으로 역가우스분포의 대안으로 많이 사용된다. 본 논문의 구성은 다음과 같다. 2장에서는 Ehsan 등 (1995)이 분포들간의 정보 불일치 척도로 제시한 ID 지표를 소개한다. 3장에서는 감마분포, 와이블분포와 로그정규분포에 대해서 엔트로피 기반 검정의 검정력을 조사하기 위해 사용할 ID 지표를 유도하고 이 지표를 통해 얻은 분석결과를 제시한다. 4장에서는 쥐의 생존기간을 분석한 사례를 소개하고 끝으로 5장에서는 결론을 맺는다.

2. ID 지표

확률변수 X가 $f(x)$를 확률밀도함수로 가지고 $T_1(X), \ldots, T_s(X)$는 $f(x)$에 대해 적분가능한 X의 함수라 할 때, Jaynes (1957)에 따르면 s개의 제약 $E_f\{T_1(X)\} = \theta_1, \ldots, E_f\{T_s(X)\} = \theta_s$를 만족하는 분포들 $\mathcal{D} = \{f(x) : E_f\{T_j(X)\} = \theta_j,\ j = 1, \ldots, s\}$ 중에서 최대엔트로피를 가지는 분포(이하 최대엔트로피분포)가 존재하며 최대엔트로피분포의 확률밀도함수는 $f^*(x) = \exp\{-\lambda_0 - \sum_{j=1}^{s} \lambda_j T_j(x)\}$의 형태를 취한다. 여기서 $\lambda_0, \lambda_1, \ldots, \lambda_s$들은 s개의 제약으로부터 결정이 되는 라그랑지 승수들이다.

엔트로피에 기반을 두는 모형화에서 최대엔트로피분포의 엔트로피는 $\mathcal{D}$에 속한 분포들을 비교하기 위한 기준이 되고 두 분포 $f^*(x)$와 $f(x)$간의 정보 불일치는 엔트로피의 차이

$$\begin{aligned} \Delta_H &= H(f^*) - H(f) \\ &= -\int f^*(x) \ln f^*(x)\, dx + \int f(x) \ln f\,(x)\ dx \end{aligned} \tag{2.1}$$

로 측정한다. Δ_H를 기초로 정의된 몇몇 지표를 중에서 Gokhale (1983)의 EPF 지표는 분포적 가설에 대한 엔트로피 기반 적합도 검정의 개발에 주로 활용이 되었다. $\mathcal{D}$에 속한 분포의 EPF 지표는

$$\mathrm{EPF}(f) = \frac{\exp\{H(f)\}}{\exp\{H(f^*)\}} = \exp(-\Delta_H) \tag{2.2}$$

로 정의된다. EPF 지표는 정규화된 측도로 범위는 $0 < \text{EPF}(f) \leq 1$이 되고 $\text{EPF}(f) \approx 1$은 $f(x) \approx f^*(x)$를 나타낸다.

Ehsan 등 (1995)은 Δ_H 대신에 정보이론에서 두 분포들간의 정보 불일치를 재는 척도로 잘 알려져 있는 쿨백-라이블러 함수에 기초를 둔 새로운 형태인 ID 지표를 제시했고 모형선택을 위한 모수적 또는 비모수적 통계방법의 강건성과 검정력 연구를 계획하기 위한 도구로써 ID 지표의 유용성을 사례를 통해 보였다. $\mathcal{D}$에 속한 분포의 ID 지표는

$$\text{ID}(f : f^*) = 1 - \exp\{-K(f : f^*)\} \tag{2.3}$$

로 정의된다. 여기서 $K(f : f^*)$는 $f^*(x)$와 $f(x)$에 대한 쿨백-라이블러 함수로

$$K(f : f^*) = \int f(x) \ln \frac{f(x)}{f^*(x)}\, dx \tag{2.4}$$

가 된다. ID 지표 또한 EPF 지표와 마찬가지로 정규화된 측도로 $0 \leq \text{ID}(f : f^*) \leq 1$인 범위를 가지며 EPF 지표와는 반대로 $\text{ID}(f : f^*) \approx 0$은 $f(x)$는 근사적으로 $f^*(x)$와 같음을 나타낸다. Ehsan 등 (1995)에 따르면 ID 지표는 EPF 지표와는 $\text{ID}(f : f^*) = 1 - \text{EPF}(f)$인 관계를 가진다.

3. ID 지표를 통한 검정력 분석

응용에서 역가우스분포의 대안으로 주로 사용되는 확률모형은 감마분포, 와이블분포와 로그정규분포들이다. 이들 분포를 대립가설의 분포로 선택하여 Mudholkar와 Tian (2002)이 제안한 엔트로피 기반 검정의 검정력을 ID 지표를 통해서 규명하고자 한다.

3.1. 엔트로피 기반 검정

Mudholkar와 Tian (2002)은 영가설 $H_0 : X \sim \text{IG}(\mu, \lambda)$의 검정을 구축하기 위해서 다음의 결과들을 이용했다. 변환된 확률변수 $Y(= X^{-1/2}) > 0$가 $g(y)$를 확률밀도함수로 가지고 $T_1(Y) = Y^{-2}$과 $T_2(Y) = Y^2$이 $g(y)$에 대해 적분가능한 함수라고 하면, 제약 $E_g\{T_1(Y)\} = \theta_1 = 1/\nu$과 $E_g\{T_2(Y)\} = \theta_2 = \nu + \xi^2$을 만족하는 모든 분포들, $\mathcal{D} = \{g(y) : E_g(Y^{-2}) = 1/\nu, E_g(Y^2) = \nu + \xi^2\}$에서 최대엔트로피분포는 확률밀도함수

$$g^*\left(y; \nu, \xi^2\right) = \frac{2}{\sqrt{2\pi}\xi} \exp\left\{-\frac{(y^2 - \nu)^2}{2\xi^2 y^2}\right\}, \quad y > 0,\ \nu > 0,\ \xi^2 > 0 \tag{3.1}$$

을 가지는 제곱근 반비례(squared root-reciprocal) 역가우스분포 $\text{SRIG}(\nu, \xi^2)$가 된다. 또한 확률변수 Y가 $\text{SRIG}(\nu, \xi^2)$를 따르면 $X = 1/Y^2$은 μ와 λ가 각각 $1/\nu$와 $1/\xi^2$인 $\text{IG}(\mu, \lambda)$를 한다.

$\mathcal{D}$의 분포들은 제약 $E_g(Y^2) - 1/E_g(Y^{-2}) = \xi^2$을 만족하는 분포들 $\mathcal{D}^* = \{g(y) : E_g(Y^2) - 1/E_g(Y^{-2}) = \xi^2\}$이 된다. $\text{SRIG}(\nu, \xi^2)$의 엔트로피가 $H(g^*) = \ln(\pi e \xi^2/2)/2$임을 이용하면 $\mathcal{D}^*$에 속한 분포의 EPF 지표는

$$\text{EPF}(g) = \frac{2\exp\{H(g)\}}{\sqrt{2\pi e}\xi} \tag{3.2}$$

가 된다.

엔트로피 기반 검정은 $K(g) = \sqrt{2\pi e}\,\text{EPF}(g)$가 $\sqrt{2\pi e}$이면 영가설을 채택하고 $\sqrt{2\pi e}$보다 작으면 대립가설을 받아들이는 검정규칙을 사용한다. 검정통계량은 $K(g)$를 추정한

$$K_{m,n} = \frac{2\exp(H_{m,n})}{W} \tag{3.3}$$

을 이용한다. 식 (3.3)에서 $H_{m,n}$은 크기 n의 표본 $X_1, X_2, \ldots, X_n$으로부터 변환된 표본 $Y_1, Y_2, \ldots, Y_n$에 대한 표본엔트로피인

$$H_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{n}{2m}\left\{Y_{(i+m)} - Y_{(i-m)}\right\}\right] \tag{3.4}$$

이다. 여기서 $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$은 Y_i들의 순서통계량으로 $i > n$이면 $Y_{(i)} = Y_{(n)}$, $i < 1$이면 $Y_{(i)} = Y_{(1)}$이고 m은 $n/2$보다 작은 양의 정수값을 갖는 윈도크기이다. W는 ξ의 추정량으로 $W^2 = (\sum_{i=1}^{n} Y_i^2 - n^2/\sum_{i=1}^{n} Y_i^{-2})/(n-1)$의 양의 제곱근이다. 엔트로피 기반 검정은 유의수준 α에서 기각값 $K_{m,n}(\alpha)$에 대해 $K_{m,n} \le K_{m,n}(\alpha)$이면 영가설을 기각하게 된다. $K_{m,n}$에서 $H_{m,n}$은 $H(g)$에 대해 일치성을 가지고 (Vasicek, 1976) W는 ξ에 대한 일치추정량이기 때문에 Slutsky의 보조정리에 의해 검정통계량은 $n, m \to \infty$, $m/n \to 0$이면 $K(g)$로 확률수렴하게 된다.

3.2. ID 지표의 유도

확률변수 X가 감마분포 $G(\alpha,\beta)$를 따르면, 밀도함수는

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right), \quad \alpha > 0,\ \beta > 0,\ x > 0$$

로 주어진다. $X^{-1/2}$로 변환한 Y는

$$g(y) = \frac{2}{\Gamma(\alpha)\beta^{\alpha}} y^{-2\alpha-1}\exp\left(-\frac{1}{\beta y^2}\right), \quad \alpha > 0,\ \beta > 0,\ y > 0$$

를 밀도함수로 가지는 분포 G_1을 따른다. EPF 지표에 사용될 G_1의 엔트로피를 구해보면

$$\begin{aligned} H(g) &= -\int_0^{\infty} g(y)\ln g(y)\,dy \\ &= \ln\frac{\Gamma(\alpha)\beta^{\alpha}}{2} + \left(\alpha+\frac{1}{2}\right)\int_0^{\infty}\ln y^2 g(y)\,dy + \frac{1}{\beta}\int_0^{\infty}\frac{g(y)}{y^2}\,dy \\ &= \ln\frac{\Gamma(\alpha)\beta^{\alpha}}{2} + \left(\alpha+\frac{1}{2}\right)E_g\left(\ln Y^2\right) + \frac{1}{\beta}E_g\left(Y^{-2}\right) \end{aligned} \tag{3.5}$$

가 된다. 식 (3.5)에서 $E_g(\ln Y^2)$과 $E_g(Y^{-2})$은 각각 $E_g(\ln Y^2) = -E_f(\ln X) = -m'(0) = -\psi(\alpha) - \ln\beta$와 $E_g(Y^{-2}) = E_f(X) = \alpha\beta$로 구해진다. 여기서, $m(t)$는 $E_f(e^{t\ln X}) = E_f(X^t) = \beta^t\Gamma(\alpha+t)/\Gamma(\alpha)$로 주어지는 $\ln X$의 적률생성함수이고 $\psi(\alpha)$는 $\Gamma'(\alpha)/\Gamma(\alpha)$로 주어지는 디감마함수이다. 이렇게 구한 $E_g(\ln Y^2)$과 $E_g(Y^{-2})$를 식 (3.5)에 대입해서 정리하면, G_1의 엔트로피는

$$H(g) = \ln\frac{\Gamma(\alpha)}{2} - \left(\alpha+\frac{1}{2}\right)\psi(\alpha) - \frac{1}{2}\ln\beta + \alpha \tag{3.6}$$

가 된다.

G_1의 $\xi^2 = E_g(Y^2) - 1/E_g(Y^{-2})$은 $E_g(Y^2) = E_f(1/X)$과 $E_g(Y^{-2}) = E_f(X)$를 이용해서 다음과 같이 얻으면 된다. X의 기대값은 $E_f(X) = \alpha\beta$로 어렵지 않게 구해진다. 그러나, $1/X$의 기대값 $E_f(1/X)$은 구해야 할 정적분 $\{\Gamma(\alpha)\beta^{\alpha}\}^{-1}\int_0^{\infty} x^{\alpha-2}e^{-x/\beta}\,dx$가 폐쇄형으로 주어지지 않기 때문에 수식의 형태로 나타낼 수 없다. 그렇지만, $\alpha > 1$에 대해서는 $E_f(1/X) = \Gamma(\alpha-1)/\{\Gamma(\alpha)\beta\} = \{(\alpha-1)\beta\}^{-1}$로 표현이 된

다. 이 결과를 이용하게 되면 ξ^2에 대한 식으로 $\xi^2 = 1/\{\alpha(\alpha-1)\beta\}, \alpha > 1$을 얻는다. β는 ξ^2에 관하여 $\beta = 1/\{\alpha(\alpha-1)\xi^2\}$로 표현이 되고 이것을 식 (3.6)에 대입하면, G_1의 엔트로피는

$$H(g) = \ln \frac{\Gamma(\alpha)}{2} - \left(\alpha + \frac{1}{2}\right)\psi(\alpha) + \frac{1}{2}\ln\{\alpha(\alpha-1)\} + \ln\xi + \alpha \tag{3.7}$$

로 주어진다.

G_1의 EPF 지표는 식 (3.2)의 $H(g)$를 식 (3.7)로 대체하면

$$\mathrm{EPF}(g) = \sqrt{\frac{\alpha(\alpha-1)}{2\pi e}}\,\Gamma(\alpha)\exp\left\{\alpha - \left(\alpha + \frac{1}{2}\right)\psi(\alpha)\right\}, \quad \alpha > 1 \tag{3.8}$$

이 된다. 따라서, G_1의 ID 지표는

$$\begin{aligned}\mathrm{ID}(g : g^*) &= 1 - \mathrm{EPF}(g)\\ &= 1 - \sqrt{\frac{\alpha(\alpha-1)}{2\pi e}}\,\Gamma(\alpha)\exp\left\{\alpha - \left(\alpha + \frac{1}{2}\right)\psi(\alpha)\right\}, \quad \alpha > 1\end{aligned} \tag{3.9}$$

의 형태가 됨을 알 수 있다.

와이블분포 $W(\alpha,\beta)$를 따르는 확률변수 X는

$$f(x) = \frac{\alpha}{\beta}x^{\alpha-1}\exp\left(-\frac{x^\alpha}{\beta}\right), \quad \alpha > 0,\ \beta > 0,\ x > 0$$

을 밀도함수로 가진다. $X^{-1/2}$로 변환한 확률변수 Y는 밀도함수가

$$g(y) = \frac{2\alpha}{\beta}y^{-2\alpha-1}\exp\left(-\frac{1}{\beta y^{2\alpha}}\right)$$

로 주어지는 분포 G_2를 따르게 된다. EPF 지표에 이용할 G_2의 엔트로피를 구해보면

$$\begin{aligned}H(g) &= -\int_0^\infty g(y)\ln g(y)\,dy\\ &= \ln\frac{\beta}{2\alpha} + \left(\alpha + \frac{1}{2}\right)\int_0^\infty \ln y^2 g(y)\,dy + \frac{1}{\beta}\int_0^\infty y^{-2\alpha}g(y)\,dy\\ &= \ln\frac{\beta}{2\alpha} + \left(\alpha + \frac{1}{2}\right)E_g\left(\ln Y^2\right) + \frac{1}{\beta}E_g\left(Y^{-2\alpha}\right)\end{aligned} \tag{3.10}$$

이 된다. $E_g(\ln Y^2) = -E_f(\ln X)$는 $\ln X$의 적률생성함수 $m(t) = E_f(e^{t\ln X}) = E_f(X^t) = \beta^{t/\alpha}\Gamma(t/\alpha+1)$을 이용해서 구하면 $E_g(\ln Y^2) = -m'(0) = -\{\psi(1) + \ln\beta\}/\alpha$가 된다. $E_g(Y^{-2\alpha})$은 $E_f(X^\alpha)$와 같으므로 β가 된다. 이들 결과를 식 (3.10)에 대입해서 정리하면, G_2의 엔트로피는

$$H(g) = 1 - \ln(2\alpha) - \left(1 + \frac{1}{2\alpha}\right)\psi(1) - \frac{1}{2\alpha}\ln\beta \tag{3.11}$$

로 얻게 된다.

G_2의 ξ^2은 G_1의 경우와 동일한 방식으로 구할 수 있다. $E_f(X)$는 $\beta^{1/\alpha}\Gamma(1+1/\alpha)$로 쉽게 얻어진다. $E_f(1/X)$의 경우는 해석적인 방법을 통해 폐쇄형태로 주어지는 식으로 표현할 수가 없다. 그렇지

만, $\alpha > 1$에 대해서 $1/X$의 기대값은 $y = x^{\alpha}$로 치환해서 구해보면 $E_f(1/X) = \int_0^{\infty} y^{-1/\alpha}\, e^{-y/\beta}\, dy/\beta = \beta^{-1/\alpha}\Gamma(1-1/\alpha)$로 표현이 된다. 이 결과를 이용하면 $\xi^2 = \beta^{-1/\alpha}\{\Gamma(1-1/\alpha) - \Gamma^{-1}(1+1/\alpha)\}$, $\alpha > 1$이 된다. β는 ξ^2에 관해서 $\beta = \xi^{-2\alpha}\{\Gamma(1-1/\alpha) - 1/\Gamma(1+1/\alpha)\}$로 표현이 되고 이것을 식 (3.11)에 대입해서 얻은 G_2의 엔트로피는

$$H(g) = 1 - \ln(2\alpha) - \left(1 + \frac{1}{2\alpha}\right)\psi(1) - \frac{1}{2}\ln\left\{\Gamma\left(1 - \frac{1}{\alpha}\right) - \Gamma^{-1}\left(1 + \frac{1}{\alpha}\right)\right\} + \ln\xi + 1 \tag{3.12}$$

이 된다.

EPF 지표는 식 (3.2)의 $H(g)$를 식 (3.12)로 대체하여

$$\mathrm{EPF}(g) = \sqrt{\frac{e}{2\pi\alpha^2}}\left\{\Gamma\left(1 - \frac{1}{\alpha}\right) - \Gamma^{-1}\left(1 + \frac{1}{\alpha}\right)\right\}^{-\frac{1}{2}} \exp\left\{-\left(1 + \frac{1}{2\alpha}\right)\psi(1)\right\}, \quad \alpha > 1 \tag{3.13}$$

로 얻게 된다. 따라서, G_2의 ID 지표는

$$\begin{aligned}\mathrm{ID}(g : g^*) &= 1 - \mathrm{EPF}(g) \\ &= 1 - \sqrt{\frac{e}{2\pi\alpha^2}}\left\{\Gamma\left(1 - \frac{1}{\alpha}\right) - \Gamma^{-1}\left(1 + \frac{1}{\alpha}\right)\right\}^{-\frac{1}{2}} \exp\left\{-\left(1 + \frac{1}{2\alpha}\right)\psi(1)\right\}, \quad \alpha > 1\end{aligned} \tag{3.14}$$

로 주어진다.

로그정규분포 $\mathrm{LN}\left(\mu, \sigma^2\right)$을 따르는 확률변수 X의 밀도함수가

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

로 주어지면 $Y = X^{-1/2}$의 분포 G_3은

$$g(y) = \frac{2}{\sigma\sqrt{2\pi}y}\exp\left\{-\frac{\left(-\ln y^2 - \mu\right)^2}{2\sigma^2}\right\}, \quad \infty < \mu < \infty,\ \sigma^2 > 0,\ y > 0$$

을 밀도함수로 가진다. $E_g(\ln Y^2) = -E_f(\ln X) = -\mu$와 $E_g(-\ln Y^2 - \mu)^2 = E_f(\ln X - \mu)^2 = \sigma^2$을 이용해서 EPF 지표에 사용할 G_3의 엔트로피를 구해보면

$$\begin{aligned}H(g) &= -\int_0^{\infty} g(y)\ln g(y)\, dy \\ &= \ln\frac{\sigma\sqrt{2\pi}}{2} + \frac{1}{2}\int_0^{\infty}\ln y^2 g(y)\, dy + \frac{1}{2\sigma^2}\int_0^{\infty}\left(-\ln y^2 - \mu\right)^2 g(y)\, dy \\ &= \ln\frac{\sigma\sqrt{2\pi}}{2} + \frac{1}{2}E_g\left(\ln Y^2\right) + \frac{1}{2\sigma^2}E_g\left(-\ln Y^2 - \mu\right)^2 \\ &= \ln\frac{\sigma\sqrt{2\pi}}{2} - \frac{\mu}{2} + \frac{1}{2}\end{aligned} \tag{3.15}$$

가 된다.

G_3의 ξ^2은 G_1의 경우와 동일한 방식으로 다음과 같이 얻을 수 있다. $1/X$의 기대값 $E_f(1/X) = \int_0^{\infty} x^{-2}e^{-(\ln x-\mu)^2/(2\sigma^2)}\, dx/(\sigma\sqrt{2\pi})$는 $y = (\ln x - \mu)/\sigma$와 $z = y + \sigma$로 치환해서 구해보면 $E_f(1/X) =$

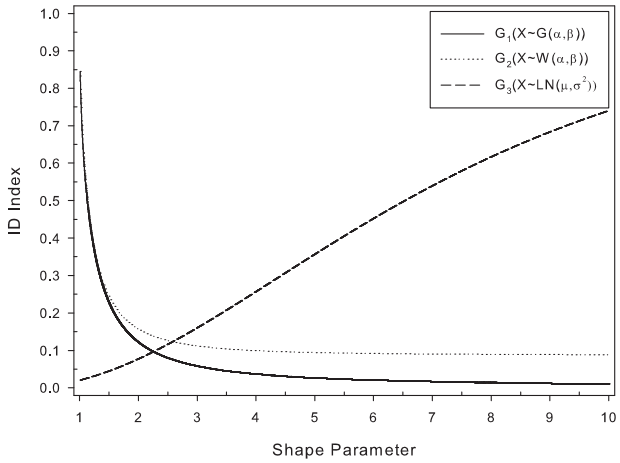

Figure 1: *ID indices of G_1, G_2 and G_3 distributions*

$e^{-\mu+\sigma^2/2}\int_{-\infty}^{\infty} e^{-(y+\sigma)^2/2}\,dy/\sqrt{2\pi} = e^{-\mu+\sigma^2/2}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz/\sqrt{2\pi} = e^{-\mu+\sigma^2/2}$가 된다. 이 결과와 $E_f(X) = e^{\mu+\sigma^2/2}$을 이용하면 $\xi^2 = e^{-\mu}(e^{\sigma^2/2} - e^{-\sigma^2/2})$을 얻게 된다. μ는 ξ^2에 대해서 $\mu = \ln(e^{\sigma^2/2} - e^{-\sigma^2/2}) - \ln\xi^2$으로 표현이 되고 이것을 식 (3.15)에 대입하여 얻게 되는 G_3의 엔트로피는

$$H(g) = \ln\frac{\sigma\sqrt{2\pi e}}{2} - \frac{1}{2}\ln\left(e^{\frac{\sigma^2}{2}} - e^{-\frac{\sigma^2}{2}}\right) - \frac{\ln\xi^2}{2} \tag{3.16}$$

이 된다.

EPF 지표는 식 (3.2)의 $H(g)$를 식 (3.16)으로 바꾸고 정리하면

$$\mathrm{EPF}(g) = \frac{\sigma e^{\frac{\sigma^2}{4}}}{\sqrt{e^{\sigma^2}-1}} \tag{3.17}$$

이 된다. 따라서, G_3의 ID 지표는

$$\begin{aligned}\mathrm{ID}(g:g^*) &= 1 - \mathrm{EPF}(g)\\ &= 1 - \frac{\sigma e^{\frac{\sigma^2}{4}}}{\sqrt{e^{\sigma^2}-1}}\end{aligned} \tag{3.18}$$

로 주어진다.

3.3. 분석결과

앞 절에서 유도한 ID 지표들은 형상모수(G_1과 G_2는 α, G_3은 σ^2)에만 의존하게 된다. 형상모수의 값이 변함에 따라 ID 지표의 값은 어떤 형태로 나타나는지를 일아보고자 G_1과 G_2에서는 1.01부터 10까지 0.01의 등간격으로 주어지는 899개의 α를, G_3에서는 0.01에서 10까지 등간격이 0.01인 999개의 σ^2을 선택해서 각 지표들을 계산했다.

Figure 1은 계산된 ID 지표들을 바탕으로 작성한 플롯이다. G_1과 G_2의 ID 지표는 α가 증가하면 감소하는 패턴을 보이고 있다. 이것으로부터 $K_{m,n}$은 작은 형상모수의 값을 가지는 감마분포와 와이블분

Table 1: Simulated powers of the 5% entropy-based test for $n = 30$ and $m = 3$

Distribution	Shape parameter						
	1.2	1.5	2.0	2.5	3.0	5.0	10.0
$G(\alpha, 1)$	0.546	0.421	0.283	0.209	0.163	0.101	0.073
$G(\alpha, 3)$	0.545	0.418	0.284	0.214	0.163	0.098	0.071
$W(\alpha, 1)$	0.556	0.474	0.391	0.355	0.342	0.309	0.305
$W(\alpha, 3)$	0.560	0.473	0.384	0.355	0.340	0.308	0.307
$\text{LN}(0, \sigma^2)$	0.083	0.118	0.186	0.263	0.359	0.670	0.963
$\text{LN}(1, \sigma^2)$	0.088	0.116	0.180	0.261	0.358	0.668	0.959

포에서는 높은 검정력을 보이지만 형상모수의 값이 커지는 경우엔 검정력이 낮아지는 것으로 예상된다. 그리고 G_2의 ID 지표는 α가 커짐에 따라서 G_1의 ID 지표보다 높게 나타나고 있어서 $K_{m,n}$은 감마분포보다는 와이블분포에서 더 높은 검정력을 보일 것으로 예측된다. G_3의 ID 지표는 $\sigma^2 \leq 1$에 대해서는 Figure 1에 나타내지 않았지만 σ^2이 작아지면 감소하고 커지면 증가하는 모습을 보인다. $K_{m,n}$의 검정력은 로그정규분포의 σ^2이 작아지면 낮아지고 반대로 커지면 높아지는 것을 예상할 수 있다. 또한 Figure 1에서 보듯이 형상모수가 대략 2.2보다 작은 값으로 주어지는 경우에는 G_3의 ID 지표가 가장 낮게 나타나는 반면 G_1과 G_2의 ID 지표는 아주 작은 차이를 보인다. 따라서, $K_{m,n}$은 로그정규분포에서 가장 낮은 검정력을 보이고 감마분포와 와이블분포에서는 거의 같은 검정력을 가질 것으로 판단된다. 형상모수가 2.2보다 큰 값으로 주어지는 경우에는 G_1의 ID 지표가 가장 작게 나타나는 반면에 G_3의 ID 지표는 가장 높게 관측된다. 그러므로, $K_{m,n}$의 검정력은 감마분포에서 가장 낮고 로그정규분포에서 가장 높게 나올 것으로 기대된다.

Table 1은 대립가설로 고려한 감마분포, 와이블분포와 로그정규분포에 대해서 $K_{m,n}$의 검정력을 반복이 10000인 모의실험을 통해 얻은 것이다. 표본크기와 윈도크기는 각각 $n = 30$과 $m = 3$으로 했고 유의수준은 5%로 설정했다. 주어진 표본크기와 윈도크기에 해당하는 기각값 $K_{3,30}(0.05)$는 Mudholkar와 Tian (2002)이 제시한 값을 사용했다. 추정된 검정력은 Figure 1에서 도출했던 것과 같은 결과를 보여줌을 확인할 수 있다.

대립가설의 분포들로부터 유도한 ID 지표값에 따라 $K_{m,n}$의 검정력이 어떻게 나타나는지를 알아보고자 반복인 50000인 모의실험을 수행했다. 표본크기와 윈도크기는 각각 $n = 30$과 $m = 3$으로 했고 유의수준은 5%로 설정했다. ID 지표값은 0.09에서 2.0까지 0.1씩 등간격으로 주었다. Figure 2는 감마분포 $G(\alpha, 1)$, 와이블분포 $W(\alpha, 1)$과 로그정규분포 $\text{LN}(0, \sigma^2)$에 대해서 추정한 $K_{m,n}$의 검정력을 보여준다. ID 지표값이 커지면 모든 대립분포에 대해서 $K_{m,n}$은 증가하는 검정력을 보인다. ID 지표값이 작아지면 와이블분포, 감마분포와 로그정규분포의 순으로 검정력이 높게 되고 ID 지표값이 커지게 되면 와이블분포, 로그정규분포와 감마분포의 순으로 검정력이 좋음을 알 수 있다. 2 이상의 큰 ID 지표값에 대해서는 그림에서 제시하지 않았지만 로그정규분포, 와이블분포와 감마분포의 순으로 검정력이 높게 나타난다.

4. 사례분석

분석에 사용한 rat 생존자료는 고수준 방사선에 노출된 20마리 수컷 쥐의 생존기간을 주 단위로 측정한 것으로 Lawless (1982)에 실려있다. 이 자료에 대한 적합모형으로 역가우스분포, 감마분포와 와이블분포를 고려하기로 한다. 먼저, 역가우스분포에 대한 적합을 알아보고자 엔트로피 기반 검정을 수행해 본다. 윈도크기를 Vasicek (1976)의 지침에 따라 $m = 3$으로 해서 검정통계량을 계산해보면 $K_{3,20} = 2.773$이 된다. 역가우스분포의 두 모수를 자료로부터 추정해 보면 $\hat{\mu} = 113.450$과 $\hat{\lambda} = 767.987$이 된다. 반복인 50000인 모수적 붓스트랩 방법을 사용하여 IG(113.450, 767.987)로부터 추정

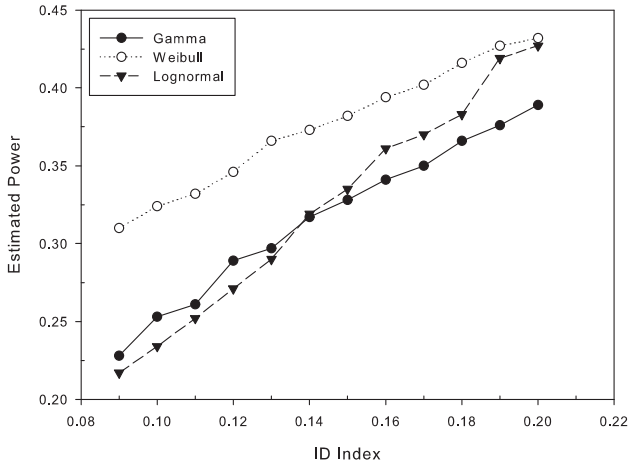

Figure 2: *Powers of the entropy-based test by values of ID index*

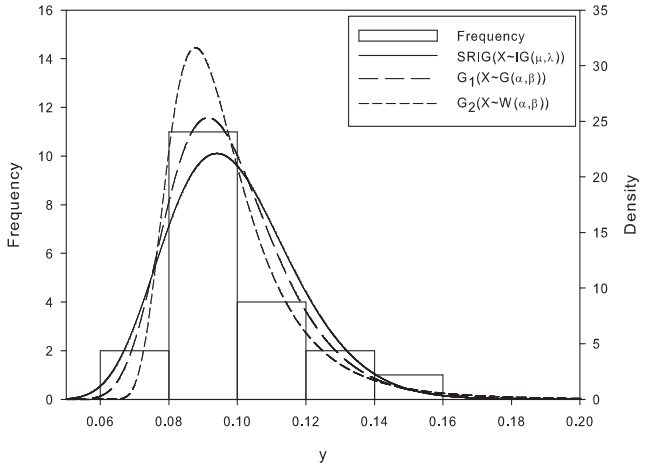

Figure 3: *Estimated density curves for rat survival data*

한 유의확률은 $p(K_{3,20} \leq 2.773) = 0.070$으로 유의수준 5%에서 유의하지 않다.

그런데, 감마분포와 와이블분포의 형상모수 α를 자료로부터 추정해 보면 각각 8.799와 3.799가 되고 이들 값을 식 (3.9)와 (3.14)에 대입하면 역가우스분포에 의한 ID 지표은 0.013과 0.101로 추정된다. 이것은 주어진 자료가 감마와 역가우스분포, 또한 와이블과 역가우스분포를 확실히 식별해 낼 수가 없음을 나타낸다. Figure 3은 변환한 자료를 기초로 추정된 확률밀도함수를 보여 주고 있다. 그림에서 볼 수 있듯이 밀도함수 모두가 유사한 모습으로 나타난다. 특히 감마분포와 역가우스분포로부터 유도된 밀도함수의 경우, 형상에서 아주 비슷한 형태를 보인다.

추가적으로 원자료에 대해서 감마분포와 와이블분포에 대한 적합도 검정을 실시해 보면 앤더슨-달링 검정 통계량 A^2은 각각 0.415와 0.285로 계산이 된다. 두 값에 대한 유의확률 모두 0.25보다 크게 나와서 유의수준 5%에서 유의적이지 않으므로 자료가 감마와 와이블분포 모두에 잘 적합이 됨을 알 수

있다. 이 결과는 엔트로피 기반 검정이 자료가 역가우스분포와 아주 유사한 형상을 보이는 다른 분포에서 추출된 경우에는 탐지를 잘 못해 주는 경향이 있음을 나타낸다. 따라서, 응용에서 역가우스분포를 확률모형으로 결정하기 위해 엔트로피 기반 검정을 사용한다면 제공되는 결과을 신중하게 받아들일 필요가 있다.

5. 결론

역가우스분포는 치우친 모습을 보이는 자료의 분석과 모형화를 위한 확률분포로 폭넓은 유용성을 가진다. Mudholkar와 Tian (2002)은 자료 분석에서 역가우스분포의 적합성을 알아보기 위한 방법으로 Gokhale (1983)의 EPF 지표에 기초한 엔트로피 기반 검정을 제안했다. 모의실험에서 로그정규분포에 대한 검정력은 설정된 유의수준 5%를 조금 상회하는 정도로 아주 낮게 나타나지만 대체적으로 기존의 다른 적합도 검정들보다 좋은 성능을 보여 주는 것으로 그들은 보고했다. 그러나, 대립가설로 선택한 표준지수분포, 와이블분포 $W(1,2)$와 로그정규분포 LN(0.5, 1)에 대한 소규모의 모의실험에 의한 검정력만을 제시하고 있어 이 결과를 통해 검정력을 평가하기에는 제한적일 수 밖에 없다. 검정력에 대한 전반적인 조사를 하기 위해서는 여러 다양한 대립분포에 대한 대규모의 모의실험을 수행해서 얻은 결과가 필요하다. 하지만, 이런 경험적 방법을 이용하기 보다는 이론적인 도구을 통해 검정력을 규명하는 것이 더 효율적일 수가 있다.

본 논문에서는 Ehsan 등 (1995)이 제시한 ID 지표를 이론적 도구로 이용하여 검정력 측면에서 엔트로피 기반 검정의 성능을 규명하고자 했다. ID 지표는 비교 대상이 되는 두 분포들 f와 f^*간의 정보 불일치를 재는 일반화된 지표이고 f^*가 최대엔트로피분포일 경우에는 EPF 지표는 ID 지표의 특수한 형태가 된다. Ehsan 등 (1995)은 ID 지표가 모형진단과 검정력 연구의 계획 등을 위한 유용한 도구가 될 수 있음을 언급했지만 아직까지는 많이 활용되고 있지는 않는 것 같다. 감마분포, 와이블분포와 로그정규분포는 치우친 모습의 자료를 분석하고 모형화를 위한 확률모형으로써 역가우스분포의 대안으로 많이 사용되므로 이들 분포를 대립가설의 분포로 고려했다. 각 분포로부터 유도한 ID 지표를 통해 다음의 결과를 얻을 수 있었다. 형상모수 값이 작아지면 엔트로피 기반 검정은 로그정규분포에 대해서는 가장 낮은 검정력을, 감마분포와 와이블분포에 대해서는 거의 같은 검정력을 보였다. 형상모수의 값이 커지면 로그정규분포에 대한 검정력은 가장 높게 나타났고 감마분포에 대한 검정력은 가장 낮게 나왔다. 또한 형상모수의 값에 상관없이 감마분포보다는 와이블분포에 대한 검정력이 더 높게 나타났다.

본 연구에서는 감마분포와 와이블분포의 형상모수가 $\alpha \leq 1$로 주어지는 경우에는 ξ^2이 폐쇄형으로 주어지지 않기 때문에 ID 지표의 유도를 통해 검정력을 조사할 수가 없었다. 이에 대해서는 향후의 연구과제로 남기고자 한다.

References

Chhikara, R. S. and Folks, J. L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, Marcel Dekker, New York.

Edgeman, R. L. (1990). Assessing the inverse Gaussian distribution assumption, *IEEE Transactions on Reliability*, **39**, 352–355.

Edgeman, R. L., Scott, R. C. and Pavur, R. J. (1988). A modified Kolmogorov-Smirnov test for the inverse density with unknown parameters, *Communications in Statistics-Simulation and Computation*, **17**, 1203–1212.

Ehsan, S., Ebrahimi, N. and Habibullah, M. (1995). Information distinguishability with application to analysis of failure data, *Journal of the American Statistical Association*, **90**, 657–668.

Gokhale, D. V. (1983). On entropy-based goodness-of-fit tests, *Computational Statistics and Data Analy-*

sis, **1**, 157–165.

Jaynes, E. T. (1957). Information theory and statistical mechanics, *Physicasl Review*, **106**, 620–630.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley, New York.

Mudholkar, G. S. and Tian, L. (2002). An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test, *Journal of Statistical Planning and Inference*, **102**, 211–221.

Schrödinger, E. (1915). Zur theorie der fall und steigversuche an teilchen mit Brownscher bewegung, *Physikalische Zeitschrift*, **16**, 289–295.

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

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

Smoluchowsky, M. V. (1915). Notiz über die berechning der Brownschen molkularbewegung bei des ehrenhaft-milikanchen versuchsanordnung, *Physikalische Zeitschrift*, **16**, 318–321.

Tweedie, M. K. (1957a). Statistical properties of inverse Gaussian distributions-I, *Annals of Mathematical Statistics*, **28**, 362–377.

Tweedie, M. K. (1957b). Statistical properties of inverse Gaussian distributions-II, *Annals of Mathematical Statistics*, **28**, 696–705.

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

2012년 8월 21일 접수; 2012년 10월 5일 수정; 2012년 11월 5일 채택