The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.4.611
26(4), 611–622

A Test of Fit for Inverse Gaussian Distribution Based on the Probability Integration Transformation

Byungjin Choi[a,1]

[a]Department of Applied Information Statistics, Kyonggi University

(Received April 11, 2013; Revised May 30, 2013; Accepted August 6, 2013)

Abstract

Mudholkar and Tian (2002) proposed an entropy-based test of fit for the inverse Gaussian distribution; however, the test can be applied to only the composite hypothesis of the inverse Gaussian distribution with an unknown location parameter. In this paper, we propose an entropy-based goodness-of-fit test for an inverse Gaussian distribution that can be applied to the composite hypothesis of the inverse Gaussian distribution as well as the simple hypothesis of the inverse Gaussian distribution with a specified location parameter. The proposed test is based on the probability integration transformation. The critical values of the test statistic estimated by simulations are presented in a tabular form. A simulation study is performed to compare the proposed test under some selected alternatives with Mudholkar and Tian (2002)'s test in terms of power. The results show that the proposed test has better power than the previous entropy-based test.

Keywords: Inverse Gaussian distribution, entropy, probability integration transformation, goodness-of-fit, power.

1. 서론

사회적 또는 물리적 현상의 설명을 위해 수행되는 과학적 연구의 한 단계는 수집된 자료에 대해 목적에 맞는 통계적 방법을 사용해서 분석한 결과를 얻는 것이다. 적용할 대부분의 통계적 방법은 자료에 대한 분포적인 가정을 요구하게 되고 주어진 자료가 가정을 충실히 따르지 않으면 도출된 결과를 신뢰할 수가 없게 된다. 이런 측면에서 자료분석의 첫 단계로써 분석에 사용할 통계적 방법의 적절성을 파악하는 것은 매우 중요하다고 할 수 있다. 자료에 대한 분포적인 가정을 점검하는 방법 중의 하나는 적합도 검정을 수행해보는 것이다. 크기 n의 확률표본 $X_1, X_2, \ldots, X_n$이 알려져 있지 않은 임의의 분포 F에서 추출되었다고 하면, 적합도 검정은 분포적 가설 $H_0 : X_1, X_2, \ldots, X_n \sim F \in \mathcal{F}_0$ 대 $H_1 : X_1, X_2, \ldots, X_n \not\sim F \in \mathcal{F}_0$에 대한 검정이다. 여기서, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$은 실값을 가지는 모수벡터이고 $\mathcal{F}_0 = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$은 모수 공간 Ω에서 정의되는 분포족이다. 적합도 검정에 관한 일반적인 방법론은 D'Agostino와 Stephens (1986)에 상세히 소개되어 있으므로 참고하기 바란다.

양의 값으로 측정된 자료가 오른쪽으로 긴 꼬리가 있는 형상을 보이는 경우에는 주로 비대칭 분포를 확률모형으로 사용하여 분석하게 된다. 적용가능한 확률모형들 중에서 역가우스분포는 정규분포에서와

[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

유사한 통계적 방법의 개발이 용이하기 때문에 다른 비대칭 분포들에 비해서 많은 유용성을 가진다. 역가우스분포로의 자료의 적합을 알아보기 위한 방법으로 Mudholkar와 Tian (2002)은 정보이론에서 불확실성의 측도로 사용되는 Shannon (1948)의 엔트로피에 기반을 둔 적합도 검정을 제안했다. 검정력 조사를 위해서 수행한 소규모의 모의실험 결과에서 Mudholkar와 Tian (2002)의 엔트로피 기반 적합도 검정은 Edgeman (1990)이 제안한 적합도 검정보다 더 좋은 검정력을 가지는 것으로 나타난다. 그러나, 엔트로피 기반 적합도 검정은 위치모수와 척도모수인 μ와 λ가 모두 알려져 있지 않거나 λ만 알려져 있는 역가우스분포의 적합을 알아보고자 하는 경우에만 사용이 가능하다. 응용에서 μ와 λ가 모두 알려져 있거나 μ만 알려져 있는 역가우스분포에 대한 적합을 시도해야 하는 경우가 발생할 수가 있고 이런 경우에 대해서도 사용할 수 있는 적합도 검정이 필요하다.

본 논문에서는 역가우스분포의 단순 또는 복합가설 모두를 검정할 수 있는 새로운 형태의 엔트로피 기반 적합도 검정을 제안하고 Mudholkar와 Tian (2002)의 검정과 검정력 측면에서 성능을 비교하고자 한다. 검정의 개발을 위해 다음의 분포적 결과를 고려한다. 역가우스분포 IG(μ, λ)에서 추출된 크기 n인 표본 $X_1, X_2, \ldots, X_n$으로부터 확률적분변환을 적용하여 얻은 변환된 표본 $U_1 = F_0(X_1), U_2 = F_0(X_2), \ldots, U_n = F_0(X_n)$은 균일분포 $U(0,1)$을 따르게 된다. 여기서, F_0는 역가우스분포의 분포함수이다. U_i들로부터 변환 $Y = -2\ln(U)$에 의해 얻게 되는 새로운 표본 $Y_1, Y_2, \ldots, Y_n$은 지수분포를 하게 된다. 그러므로 표본 $X_1, X_2, \ldots, X_n$에 대한 역가우스분포의 적합 문제는 변환된 표본 $Y_1, Y_2, \ldots, Y_n$에 대한 지수분포의 적합 문제가 된다. 평균이 μ인 모든 비음의 확률변수들 중에서 지수분포를 따르는 확률변수의 엔트로피는 최대가 되는 것으로 잘 알려져 있다. Gokhale (1983)은 엔트로피 특성짓기에 기초하여 적합도 검정의 구축을 위한 일반적인 방법론을 논의했고 이 방법론에 따라 지수분포의 엔트로피 특성짓기를 이용한 적합도 검정을 구축한다.

본 논문의 구성은 다음과 같다. 2장에서는 확률적분변환에 기초한 역가우스분포에 대한 엔트로피 기반 적합도 검정을 소개한다. 3장에서는 모의실험을 통해서 추정한 표본크기와 윈도크기에 따른 검정통계량의 1%, 5%와 10%에서의 기각값과 근사기각값을 얻기 위한 계산공식을 제시한다. 4장에서는 대립가설의 분포로 선택한 감마분포, 와이블분포, 로그정규분포와 표본크기 $10, 20, 30, 50$에 대해서 제안한 검정과 Mudholkar와 Tian (2002)의 검정을 검정력 측면에서의 성능을 비교하고자 모의실험을 수행한다. 5장에서는 실제 자료에 대해 제안한 검정을 수행해 본다. 6장에서는 결론을 내린다.

2. 검정방법

크기 n의 표본 $X_1, X_2, \ldots, X_n$이 분포함수 $F(x)$를 가지는 임의의 확률분포에서 추출되었다고 하자. 주어진 표본에 대한 역가우스분포 IG(μ, λ)의 적합 여부를 알아보기 위한 문제는

$$H_0 : X_1, X_2, \ldots, X_n \sim F_0(x) \quad vs. \quad H_1 : X_1, X_2, \ldots, X_n \not\sim F_0(x) \tag{2.1}$$

에 대한 가설검정의 문제가 된다. 여기서, IG(μ, λ)의 분포함수인 $F_0(x)$는

$$F_0(x) = \Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} - 1\right)\right] + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} + 1\right)\right] \tag{2.2}$$

로 주어지고 (Shuster, 1968), Φ는 표준정규분포의 분포함수이다. 표본 $X_1, X_2, \ldots, X_n$이 역가우스분포 IG(μ, λ)에서 추출되었다고 하고 이들 표본으로부터 확률적분변환 $U = F_0(X)$에 의해 얻은 표본을 $U_1, U_2, \ldots, U_n$이라 하자. U_i들은 식 (2.2)로부터 다음과 같은 방법으로 얻을 수가 있다. $\mu = \mu_0$와 $\lambda = \lambda_0$로 알려져 있으면 이들 값을 대입한다. $\mu = \mu_0$ 또는 $\lambda = \lambda_0$로 알려져 있다면 이 값과 알려져 있지 않은 모수의 추정량으로 μ의 경우는 $\bar{X}$를, λ의 경우는 $\hat{\lambda} = (n-1)/V$를 대입한다. 여기서,

$V = \sum_{i=1}^{n} \left(1/X_i - 1/\bar{X}\right)$이다. μ와 λ 모두가 알려져 있지 않다면 이들 모수를 각각의 추정량 $\bar{X}$와 $\hat{\lambda} = (n-1)/V$로 대체한다. 확률적분변환에 의해 얻은 $U_1, U_2, \ldots, U_n$은 균일분포 $U(0,1)$을 따르는 확률표본이 되고 U_i들로부터 변환 $Y = -2\ln(U)$을 이용하여 얻게 되는 새로운 표본 $Y_1, Y_2, \ldots, Y_n$은 평균이 2인 지수분포를 따르게 된다. 따라서, 식 (2.1)의 역가우스분포에 대한 적합도 검정 문제는

$$H_0 : Y_1, Y_2, \ldots, Y_n \sim F_0(y) \quad vs \quad H_1 : Y_1, Y_2, \ldots, Y_n \not\sim F_0(y) \tag{2.3}$$

로 주어지는 지수분포에 대한 적합도 검정 문제와 같게 된다. 여기서, $F_0(y) = 1 - \exp(-y/2)$로 평균이 2인 지수분포의 분포함수이다.

식 (2.3)을 검정하기 위한 엔트로피 기반 검정통계량의 유도를 위해 다음의 결과를 이용하기로 한다. 확률변수 Z가 $f_Z(z)$를 확률밀도함수로 가지고 $S_1(Z), \ldots, S_k(Z)$는 $f_Z(z)$에 대해 적분가능한 함수라고 하자. Jaynes (1957)의 최대엔트로피 추론원리에 의하면 주어진 k개의 제약들 $E\{S_1(Z)\} = \theta_1, \ldots, E\{S_k(Z)\} = \theta_k$ 하에서 최대엔트로피를 가지는 분포가 존재하게 되고 이 분포의 밀도함수는 $f_Z^*(z) = \exp\{-\delta_0 - \sum_{i=1}^{k} \delta_i S_i(Z)\}$로 주어진다. 여기서 δ_i들은 라그랑지 승수들로 k개의 제약으로부터 결정이 된다. 이제, $S_1(Z) = Z$와 $\theta_1 = \beta$로 두면 제약 $E(Z) = \beta$를 만족하는 분포들 $\mathcal{D} = \left\{f_Z(z) : E(Z) = \beta\right\}$ 중에서 최대엔트로피를 가지는 분포는 밀도함수가 $f_Z^*(z) = \exp(-z/\beta)/\beta$로 주어지는 지수분포가 되고 엔트로피는 $H(f_Z^*) = 1 + \ln\beta$가 된다.

확률변수 Z의 실현값으로 임의의 밀도함수 $f_Z(z)$를 가지는 분포에서 표본 $Z_1, Z_2, \ldots, Z_n$을 얻었다면 지수분포에 대한 적합을 알아보기 위한 가설은

$$H_0 : Z_1, Z_2, \ldots, Z_n \sim F_0(z) \quad vs \quad H_1 : Z_1, Z_2, \ldots, Z_n \not\sim F_0(z) \tag{2.4}$$

로 설정이 되고 $F_0(z) = 1 - \exp(-z/\beta)$로 주어지게 된다. 식 (2.4)의 가설을 검정하기 위한 엔트로피 기반 검정의 구축을 위해 Gokhale (1983)이 제시한 다음의 EPF(entropy power fraction) 지표

$$\begin{aligned} \mathrm{EPF}(f_Z|\beta) &= \frac{\exp\{H(f_Z)\}}{\exp\{H(f_Z^*)\}} \\ &= \frac{\exp\{H(f_Z)\}}{\beta e} \end{aligned} \tag{2.5}$$

로부터 얻게 되는 $T(f_Z) = e\mathrm{EPF}(f_Z|\beta) = \exp\{H(f_Z)\}/\beta$를 이용하기로 한다. 여기서, $H(f_Z)$는 확률변수 Z에 대한 Shannon (1948)의 엔트로피로

$$H(f_Z) = -\int_{-\infty}^{\infty} f_Z(z)\ln f_Z(z)\, dz \tag{2.6}$$

로 정의된다. 영가설 하에서는 $f_Z(z) = f_Z^*(z)$이므로 $T(f_Z) = e$가 된다. 그리고 대립가설 하에서는 $H(f_Z) < H(f_Z^*)$이기 때문에 $T(f_Z) < e$가 된다. 따라서, $T(f_Z)$가 작은 값을 가지게 되면 영가설을 기각하게 된다.

$T(f_Z)$를 검정통계량으로 이용하기 위해서는 표본으로부터 $T(f_Z)$를 추정해야만 한다. 이를 위해서는 $T(f_Z)$에 포함되어 있는 엔트로피 $H(f_Z)$와 β의 추정량이 필요하다. 비모수적인 방법에 의한 엔트로피의 추정은 여러 학자들에 의해 연구되었고 개발된 추정량들의 형태는 커널방법에 의한 확률밀도함수의 추정량 또는 표본 순서통계량의 차이로 정의되는 m-spacing에 기초하고 있다. 이에 관해서는 Vasicek (1976), Györfi와 van der Meulen (1987,1990), Dudewicz와 van der Meulen (1987), van Es (1992), Ebrahimi 등 (1994), Correa (1995) 등을 참고하기 바란다. 제안된 엔트로피 추정량들 중에서 Vasicek (1976)의 표본엔트로피는 가장 간단하고 응용이 넓은 추정량으로 분포적 가설을 검정하는데 있어서 폭

넓게 사용되고 있다. 이런 이유로 표본 $Z_1, Z_2, \ldots, Z_n$에 기초한 표본엔트로피 $H_{m,n}(Z)$를 $H(f_Z)$의 추정량으로 사용하기로 한다. 표본엔트로피 $H_{m,n}(Z)$는

$$H_{m,n}(Z) = \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{n}{2m}\left\{Z_{(i+m)} - Z_{(i-m)}\right\}\right] \tag{2.7}$$

로 정의되고 $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n)}$은 Z_i들의 순서통계량으로 $i > n$이면 $Z_{(i)} = Z_{(n)}$, $i < 1$이면 $Z_{(i)} = Z_{(1)}$이고 m은 $n/2$보다 작은 양의 정수값을 갖는 윈도크기이다. β의 경우는 $\beta = \beta_0$로 알려져 있으면 이 값를 사용하면 되고 알려져 있지 않다면 최대가능도추정량인 표본평균 $\bar{Z}$를 이용하면 된다. 따라서, 검정통계량의 형태는 평균이 $\beta = \beta_0$로 알려진 지수분포에 대한 적합도 검정의 경우에는

$$T_{m,n} = \frac{\exp\{H_{m,n}(Z)\}}{\beta_0} \tag{2.8}$$

이고 평균이 알려져 있지 않는 지수분포에 대한 적합도 검정의 경우에는

$$T_{m,n} = \frac{\exp\{H_{m,n}(Z)\}}{\bar{Z}} \tag{2.9}$$

가 된다.

역가우스분포의 적합에 관한 식 (2.3)의 가설 검정을 위한 검정통계량은 식 (2.8)에서 $H_{m,n}(Z)$를 $Y_1, Y_2, \ldots, Y_n$으로부터 얻은 표본엔트로피

$$H_{m,n}(Y) = \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{n}{2m}\left\{Y_{(i+m)} - Y_{(i-m)}\right\}\right] \tag{2.10}$$

로 대체하고 β_0의 값으로 2를 대입하면

$$T_{m,n} = \frac{\exp\{H_{m,n}(Y)\}}{2} \tag{2.11}$$

로 얻게 된다. 이 검정통계량을 사용하는 새로운 엔트로피 기반 적합도 검정을 제안하고자 한다. $T_{m,n}$ 검정은 유의수준 α에서 기각역으로 $\{T_{m,n} \le c_{m,n}(\alpha)\}$를 가지게 되고 $c_{m,n}(\alpha)$는 영가설 하에서 $T_{m,n}$의 분포로부터 결정되는 주어진 표본크기와 윈도크기 n과 m에 대응하는 기각값인 $100\alpha\%$가 되는 백분위수이다.

주어진 표본으로부터 식 (2.11)의 검정통계량 $T_{m,n}$을 계산하기 위해서는 표본엔트로피 $H_{m,n}(Y)$의 계산에 사용할 윈도크기가 정해져야 한다. 주어진 표본크기에 대한 최적의 윈도크기를 결정하기 위한 해석적 방법은 아직까지는 제시되어져 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.

3. 검정통계량의 기각값

주어진 유의수준 α에 대한 기각값의 결정을 위해서는 영가설 하에서 검정통계량의 표집분포를 알아야 한다. 표집분포는 표본엔트로피 또는 유사 표본엔트로피를 포함하는 통계량의 점근적 분포에 관한 Cressie (1976), Dudewicz와 van der Meulen (1981), Hall (1984, 1986), van Es (1992) 등의 연구결과를 적용하면 구할 수 있을 것처럼 보인다. 그러나, 제시된 결과의 이용은 매우 제한적일 수 밖에 없어서

Table 3.1. Critical values of $T_{m,n}$ at the significance level 1%

n	m									
	1	2	3	4	5	6	7	8	9	10
10	0.8985	1.2285	1.3743	1.3872						
11	0.9494	1.2748	1.4180	1.4596	1.4024					
12	1.0010	1.3333	1.4647	1.5142	1.4749					
13	1.0426	1.3750	1.4934	1.5526	1.5387	1.4755				
14	1.0856	1.4103	1.5257	1.5847	1.5901	1.5427				
15	1.1289	1.4650	1.5687	1.6213	1.6365	1.5989	1.5385			
16	1.1662	1.4940	1.5977	1.6530	1.6746	1.6515	1.5959			
18	1.2174	1.5551	1.6555	1.7059	1.7280	1.7264	1.6919	1.6400		
20	1.2749	1.6083	1.7127	1.7536	1.7726	1.7761	1.7588	1.7218	1.6748	
25	1.3801	1.7109	1.8183	1.8598	1.8751	1.8830	1.8803	1.8692	1.8458	1.8117
30	1.4523	1.7948	1.9015	1.9461	1.9624	1.9662	1.9647	1.9561	1.9434	1.9262
35	1.5077	1.8489	1.9579	2.0032	2.0211	2.0263	2.0275	2.0225	2.0137	2.0019
40	1.5548	1.8968	2.0091	2.0568	2.0761	2.0826	2.0834	2.0787	2.0710	2.0608
45	1.5927	1.9335	2.0464	2.0964	2.1180	2.1264	2.1291	2.1249	2.1188	2.1104
50	1.6227	1.9671	2.0835	2.1325	2.1551	2.1652	2.1662	2.1621	2.1576	2.1508
100	1.7801	2.1226	2.2422	2.2993	2.3301	2.3480	2.3585	2.3642	2.3651	2.3649
250	1.9041	2.2409	2.3606	2.4220	2.4571	2.4798	2.4951	2.5060	2.5136	2.5188
500	1.9588	2.2915	2.4093	2.4697	2.5062	2.5298	2.5465	2.5590	2.5682	2.5755

표집분포의 해석적 유도는 아주 어려운 문제이다. 그러므로, 모의실험을 이용하여 검정통계량의 분포함수를 추정해서 100α 백분위수를 결정하고 이것을 기각값으로 사용하기로 한다.

추정된 기각값을 얻기 위해서 역가우스분포 IG$(1,1)$로부터 표본크기 $n = 10, 11, \ldots, 100, 250, 500$인 10^5개의 표본들을 Michael 등 (1976)의 알고리즘을 사용하여 독립적으로 생성했다. 그런 다음, $m < n/2$인 모든 윈도크기에 대해 표본으로부터 계산한 검정통계량 $T_{m,n}$의 값들로부터 경험적 분포를 추정하여 얻었고 이 분포로부터 1%, 5%와 10%가 되는 백분위수를 구해서 기각값으로 정했다.

Table 3.1–Table 3.3은 주어진 표본크기에 대해서 유의수준 1%, 5%와 10%에서의 검정통계량 $T_{m,n}$의 추정된 기각값들이다. 고정된 윈도크기에 대해서 표본크기가 증가함에 따라 대응되는 기각값도 증가하는 현상을 보인다. 주어진 표본크기에 해당하는 기각값이 표에 제시되어 있지 않다면 인접한 표본크기에 대한 기각값들을 사용하여 보간한 값을 기각값에 대한 근사값으로 사용하면 된다. 근사기각값의 계산을 위해 고려한 선형보간식의 형태는

$$c_{m,n}(\alpha) = \beta_0 + \beta_1\sqrt{n} + \frac{\beta_2}{\sqrt{n}} + \frac{\beta_3}{n} + \beta_4\sqrt{m} + \frac{\beta_5}{\sqrt{m}} + \frac{\beta_6}{\sqrt{nm}} \tag{3.1}$$

와 같다. 계수들의 추정을 위해서 표본크기 $n = 20, 21, \ldots, 100$에 대해서 $m = 1, 2, \ldots, 10$인 윈도크기 각각으로부터 반복이 10^5인 모의실험을 통해 얻은 기각값들을 이용하여 회귀분석을 수행했다. Table 3.4는 유의수준 1%, 5%와 10%에 대한 근사기각값을 얻기 위한 회귀계수들을 추정한 결과로 적합된 모든 회귀직선의 결정계수는 99.6% 이상으로 높게 나타났다.

모의실험에 의해 결정된 $T_{m,n}$의 기각값이 주어진 유의수준을 잘 유지하는지를 알아볼 필요가 있다. 추정한 유의수준이 검정에서 주어진 값보다 아주 작거나 크게 나온다면 기각값의 정확성에 문제가 있음을 나타내는 것이어서 검정의 결과를 신뢰할 수가 없게 된다. 표본크기 $n = 10, 20, 30, 50$에 대해서 IG$(1, 0.7)$, IG$(1,2)$와 IG$(1,4)$로부터 각각 독립적으로 생성해서 얻은 10^4개의 표본을 이용하여 제안한

Table 3.2. Critical values of $T_{m,n}$ at the significance level 5%

n	m									
	1	2	3	4	5	6	7	8	9	10
10	1.1302	1.4472	1.5496	1.5624						
11	1.1800	1.4936	1.5904	1.6210	1.5700					
12	1.2285	1.5422	1.6320	1.6668	1.6357					
13	1.2660	1.5805	1.6694	1.7013	1.6884	1.6313				
14	1.3002	1.6135	1.6992	1.7340	1.7318	1.6905				
15	1.3403	1.6527	1.7384	1.7675	1.7698	1.7407	1.6862			
16	1.3642	1.6812	1.7650	1.7951	1.7996	1.7810	1.7374			
18	1.4187	1.7341	1.8226	1.8500	1.8539	1.8420	1.8133	1.7705		
20	1.4608	1.7791	1.8702	1.8981	1.9037	1.8951	1.8747	1.8438	1.8037	
25	1.5454	1.8636	1.9589	1.9917	1.9997	1.9949	1.9837	1.9668	1.9447	1.9180
30	1.6031	1.9296	2.0277	2.0630	2.0740	2.0713	2.0628	2.0512	2.0365	2.0183
35	1.6459	1.9739	2.0756	2.1147	2.1282	2.1304	2.1243	2.1144	2.1021	2.0888
40	1.6843	2.0129	2.1164	2.1579	2.1746	2.1787	2.1758	2.1674	2.1575	2.1460
45	1.7136	2.0438	2.1478	2.1928	2.2114	2.2176	2.2150	2.2094	2.2014	2.1913
50	1.7362	2.0689	2.1750	2.2207	2.2413	2.2487	2.2493	2.2451	2.2380	2.2293
100	1.8574	2.1900	2.3037	2.3572	2.3862	2.4025	2.4119	2.4169	2.4187	2.4185
250	1.9487	2.2791	2.3947	2.4526	2.4866	2.5084	2.5232	2.5336	2.5409	2.5461
500	1.9906	2.3168	2.4317	2.4898	2.5249	2.5480	2.5642	2.5762	2.5851	2.5920

Table 3.3. Critical values of $T_{m,n}$ at the significance level 10%

n	m									
	1	2	3	4	5	6	7	8	9	10
10	1.2637	1.5551	1.6353	1.6393						
11	1.3097	1.6009	1.6762	1.6921	1.6482					
12	1.3511	1.6470	1.7188	1.7352	1.7080					
13	1.3856	1.6846	1.7555	1.7721	1.7553	1.7065				
14	1.4173	1.7138	1.7874	1.8038	1.7948	1.7579				
15	1.4503	1.7498	1.8231	1.8382	1.8313	1.8019	1.7571			
16	1.4736	1.7756	1.8489	1.8657	1.8613	1.8392	1.8014			
18	1.5180	1.8219	1.9020	1.9199	1.9146	1.8978	1.8709	1.8350		
20	1.5574	1.8647	1.9460	1.9674	1.9641	1.9511	1.9301	1.9010	1.8678	
25	1.6310	1.9417	2.0277	2.0550	2.0580	2.0501	2.0347	2.0180	1.9968	1.9735
30	1.6809	1.9958	2.0886	2.1208	2.1280	2.1235	2.1134	2.0999	2.0846	2.0683
35	1.7181	2.0357	2.1327	2.1689	2.1803	2.1793	2.1717	2.1610	2.1479	2.1340
40	1.7499	2.0700	2.1687	2.2078	2.2222	2.2244	2.2205	2.2127	2.2022	2.1906
45	1.7755	2.0981	2.1978	2.2389	2.2565	2.2610	2.2587	2.2521	2.2435	2.2331
50	1.7960	2.1183	2.2204	2.2641	2.2829	2.2899	2.2892	2.2847	2.2778	2.2695
100	1.8971	2.2236	2.3334	2.3851	2.4131	2.4290	2.4380	2.4428	2.4447	2.4451
250	1.9737	2.2991	2.4119	2.4682	2.5012	2.5224	2.5369	2.5470	2.5543	2.5594
500	2.0074	2.3302	2.4431	2.5002	2.5344	2.5571	2.5729	2.5845	2.5934	2.6003

$T_{m,n}$ 검정과 Mudholkar와 Tian (2002)의 $K_{m,n}$ 검정의 제 1종 오류를 추정해 보았다. Table 3.5는 $\alpha = 0.05$로 주어졌을 때 두 검정들의 제 1종 오류를 모의실험을 통해 얻은 결과이다. 10^4개의 표본에 기초하여 추정한 제 1종 오류 α의 표준오차를 계산해보면 $\sigma_\alpha = \sqrt{0.05(1-0.05)/10000} \approx 0.0022$가

Table 3.4. Coefficients for approximately calculating critical values of $T_{m,n}$

	Regression Coefficient						
	β_0	β_1	β_2	β_3	β_4	β_5	β_6
$c_{m,n}(0.01)$	4.86087	−0.04061	−10.64239	9.89039	−0.21088	−1.71322	2.00697
$c_{m,n}(0.05)$	4.86830	−0.04118	−9.92253	8.61889	−0.21601	−1.75943	2.62818
$c_{m,n}(0.10)$	4.83510	−0.04034	−9.44170	7.74189	−0.21474	−1.76878	2.94676

Table 3.5. Estimated type I error rate of $T_{m,n}$ and $K_{m,n}$ tests

n	IG(1, 0.7)		IG(1, 2)		IG(1, 4)	
	$T_{m,n}$	$K_{m,n}$	$T_{m,n}$	$K_{m,n}$	$T_{m,n}$	$K_{m,n}$
10	0.048	0.048	0.050	0.050	0.050	0.049
20	0.051	0.050	0.048	0.050	0.048	0.053
30	0.053	0.047	0.048	0.053	0.048	0.051
50	0.050	0.047	0.051	0.051	0.051	0.050

된다. 이것을 사용해서 제 1종 오류 α에 대한 95% 신뢰구간을 구해보면 대략 (0.0456, 0.0544)가 된다. Table 3.5에 제시된 값들은 0.048에서부터 0.053에 걸쳐 나타나고 있으므로 $T_{m,n}$과 $K_{m,n}$ 검정들의 추정된 기각값은 제 1종 오류를 잘 통제함을 알 수 있다.

4. 검정력 분석

2절에서 제시한 $T_{m,n}$ 검정과 Mudholkar와 Tian (2002)이 제안한 엔트로피 기반 $K_{m,n}$ 검정을 검정력 측면에서의 성능을 비교해 보기 위해서 몬테-칼로 모의실험을 수행해 보기로 한다. 대립가설의 분포로는 감마분포 $G(\theta, \beta)$, 와이블분포 $W(\theta, \beta)$ 및 로그정규분포 $\mathrm{LN}(\mu, \sigma^2)$을 고려했다. 이들 분포들은 자료분석에서 역가우스분포의 대안으로 많이 활용되는 확률분포들이다. 표본크기는 $n = 10, 20, 30, 50$으로 했고 유의수준은 $\alpha = 0.05$로 설정했다.

각각의 표본크기에 대해서 10^4개의 표본들을 고려한 대립분포들에서 각각 독립적으로 생성하여 검정통계량 $T_{m,n}$을 계산했다. 검정통계량의 계산에 필요한 윈도크기는 Vasicek (1976)의 추천에 따라서 $n = 10$일 때는 $m = 2$, $n = 20, 30$일 때는 $m = 3$, $n = 50$일 때는 $m = 4$로 선택했다. 그런 다음, 얻은 $T_{m,n}$의 값들 중에서 Table 3.2에 제시된 기각값보다 작게 나온 빈도를 세고 이것을 10^4으로 나눈 값을 추정된 검정력으로 사용했다. Mudholkar와 Tian (2002)의 검정의 경우에도 $T_{m,n}$과 같은 방식으로 검정력을 계산했고 표본크기와 윈도크기에 해당하는 기각값은 Mudholkar와 Tian (2002)이 제시한 표에서 찾아서 사용했다.

Table 4.1은 대립가설의 분포가 감마분포 $G(\theta, \beta)$로 주어졌을 때 유의수준 5%에 대한 $T_{m,n}$과 $K_{m,n}$의 검정력을 보여주고 있다. 두 검정 모두 표본의 크기가 커짐에 따라 검정력이 증가하게 되고 주어진 표본크기에 대해서는 모수 θ와 β가 커짐에 따라 검정력이 감소하게 되는 경향을 가진다. 표본크기가 클 때보다 작을 때에 $T_{m,n}$의 검정력이 $K_{m,n}$의 검정력보다 더 높게 나타난다. 또한 θ와 β가 작아짐에 따라 $T_{m,n}$은 $K_{m,n}$에 비해서 더 좋은 성능을 가짐을 알 수 있다. 대체적으로 $T_{m,n}$이 $KL_{m,n}$보다는 더 좋은 검정력을 가지는 것으로 나타난다.

Table 4.2는 대립가설의 분포가 와이블분포 $W(\theta, \beta)$일 때 유의수준 5%에 대한 $T_{m,n}$과 $K_{m,n}$의 검정력을 비교한 것이다. $\theta = 2.0$인 와이블분포에서는 표본크기가 20, 30과 50에 대해서 $K_{m,n}$의 성능이 $T_{m,n}$보다 더 좋게 나타나고 있지만 전반적으로 감마분포 $G(\theta, \beta)$에 대한 결과와 유사한 현상을 관측할 수 있다.

Table 4.1. Estimated powers of $T_{m,n}$ and $K_{m,n}$ tests against the gamma distribution $G(\theta, \beta)$

θ	β	Test	n			
			10	20	30	50
0.5	0.5	$T_{m,n}$	0.617	0.905	0.980	0.999
		$K_{m,n}$	0.512	0.845	0.959	0.997
0.5	1.0	$T_{m,n}$	0.615	0.906	0.979	0.998
		$K_{m,n}$	0.508	0.845	0.956	0.996
0.5	2.0	$T_{m,n}$	0.608	0.908	0.980	0.999
		$K_{m,n}$	0.500	0.850	0.956	0.997
1.0	0.5	$T_{m,n}$	0.280	0.561	0.742	0.915
		$K_{m,n}$	0.209	0.472	0.661	0.863
1.0	1.0	$T_{m,n}$	0.292	0.555	0.739	0.915
		$K_{m,n}$	0.220	0.469	0.653	0.868
1.0	2.0	$T_{m,n}$	0.285	0.550	0.744	0.919
		$K_{m,n}$	0.215	0.462	0.659	0.867
2.0	0.5	$T_{m,n}$	0.111	0.210	0.320	0.488
		$K_{m,n}$	0.094	0.186	0.285	0.449
2.0	1.0	$T_{m,n}$	0.108	0.217	0.322	0.483
		$K_{m,n}$	0.091	0.192	0.287	0.442
2.0	2.0	$T_{m,n}$	0.112	0.209	0.316	0.483
		$K_{m,n}$	0.093	0.186	0.283	0.439

Table 4.2. Estimated powers of $T_{m,n}$ and $K_{m,n}$ tests against the Weibull distribution $W(\theta, \beta)$

θ	β	Test	n			
			10	20	30	50
0.5	0.5	$T_{m,n}$	0.632	0.920	0.985	0.999
		$K_{m,n}$	0.513	0.857	0.963	0.997
0.5	1.0	$T_{m,n}$	0.641	0.919	0.985	1.000
		$K_{m,n}$	0.526	0.858	0.961	0.998
0.5	2.0	$T_{m,n}$	0.648	0.921	0.987	0.999
		$K_{m,n}$	0.520	0.860	0.968	0.998
1.0	0.5	$T_{m,n}$	0.290	0.563	0.746	0.918
		$K_{m,n}$	0.215	0.475	0.661	0.866
1.0	1.0	$T_{m,n}$	0.284	0.554	0.739	0.910
		$K_{m,n}$	0.214	0.470	0.655	0.858
1.0	2.0	$T_{m,n}$	0.280	0.560	0.742	0.911
		$K_{m,n}$	0.210	0.473	0.651	0.860
2.0	0.5	$T_{m,n}$	0.128	0.257	0.382	0.586
		$K_{m,n}$	0.122	0.263	0.389	0.604
2.0	1.0	$T_{m,n}$	0.129	0.249	0.387	0.589
		$K_{m,n}$	0.125	0.258	0.393	0.605
2.0	2.0	$T_{m,n}$	0.124	0.257	0.379	0.583
		$K_{m,n}$	0.120	0.267	0.385	0.598

로그정규분포 LN(μ, σ^2)에 대해서 검정력을 비교한 결과는 Table 4.3에 제시되어 있다. 주어진 μ와 σ^2에 대해서 표본크기가 증가함에 따라 두 검정의 검정력은 증가하게 되고 표본크기가 주어졌을 때에는 μ와 σ^2이 커짐에 따라서 두 검정의 검정력은 증가하는 양상을 보이게 된다. 대체적으로 $T_{m,n}$은

Table 4.3. Estimated powers of $T_{m,n}$ and $K_{m,n}$ tests against the lognormal distribution $\mathrm{LN}(\mu, \sigma^2)$

μ	σ^2	Test	n			
			10	20	30	50
0.5	0.5	$T_{m,n}$	0.066	0.076	0.095	0.130
		$K_{m,n}$	0.052	0.058	0.066	0.088
0.5	1.0	$T_{m,n}$	0.104	0.197	0.286	0.456
		$K_{m,n}$	0.062	0.117	0.174	0.296
0.5	2.0	$T_{m,n}$	0.170	0.369	0.531	0.739
		$K_{m,n}$	0.093	0.221	0.348	0.550
1.0	0.5	$T_{m,n}$	0.060	0.077	0.095	0.125
		$K_{m,n}$	0.049	0.058	0.067	0.089
1.0	1.0	$T_{m,n}$	0.107	0.196	0.305	0.454
		$K_{m,n}$	0.064	0.118	0.189	0.295
1.0	2.0	$T_{m,n}$	0.179	0.371	0.530	0.744
		$K_{m,n}$	0.100	0.228	0.355	0.556
2.0	0.5	$T_{m,n}$	0.064	0.071	0.097	0.123
		$K_{m,n}$	0.053	0.053	0.071	0.086
2.0	1.0	$T_{m,n}$	0.112	0.198	0.300	0.455
		$K_{m,n}$	0.066	0.116	0.183	0.290
2.0	2.0	$T_{m,n}$	0.178	0.366	0.531	0.734
		$K_{m,n}$	0.101	0.229	0.352	0.546

$K_{m,n}$보다는 점정력이 더 높게 관측되고 있으며, 특히 표본크기와 μ, σ^2이 커짐에 따라서 $T_{m,n}$은 $K_{m,n}$에 비해서 더 좋은 성능을 발휘하는 것을 알 수 있다.

5. 예제

2절에서 제안한 $T_{m,n}$ 검정을 Proschan (1963)에 실려 있는 냉난방기 고장시간에 관한 자료에 적용해 보기로 한다. 이 자료는 보잉 720 항공기에 장착되어 있는 냉난방기가 고장이 나서 수리를 한 후에 다음 번의 고장이 날 때까지의 운전한 시간을 기록하여 얻은 것으로 다음과 같다: 102, 209, 14, 57, 54, 32, 67, 59, 134, 152, 27, 14, 230, 66, 61, 34. 고장시간들이 μ와 λ가 미지인 역가우스분포를 따르는지를 알아보기 위해 다음과 같이 검정을 수행한다.

1. 자료로부터 μ와 λ의 추정량을 계산하여 $\bar{X} = 82$, $\hat{\lambda} = 84.3199$로 얻는다.
2. 각 자료값에 대한 확률적분변환된 값은 식 (2.2)의 μ와 λ 대신에 계산한 $\bar{X}$와 $\hat{\lambda}$의 값을 대입하여 구하면 다음과 같다: $U_1 = 0.7493$, $U_2 = 0.9320$, $U_3 = 0.0368$, $U_4 = 0.5088$, $U_5 = 0.4842$, $U_6 = 0.2551$, $U_7 = 0.5810$, $U_8 = 0.5244$, $U_9 = 0.8359$, $U_{10} = 0.8687$, $U_{11} = 0.1915$, $U_{12} = 0.0368$, $U_{13} = 0.9458$, $U_{14} = 0.5744$, $U_{15} = 0.5394$, $U_{16} = 0.2796$.
3. 확률적분변환에 의해 얻은 U_i들로부터 $Y = -2\ln(U)$로 변환된 값들을 계산하여 다음과 같이 얻는다: $Y_1 = 0.5771$, $Y_2 = 0.1409$, $Y_3 = 6.6042$, $Y_4 = 1.3515$, $Y_5 = 1.4504$, $Y_6 = 2.7325$, $Y_7 = 1.0861$, $Y_8 = 1.2911$, $Y_9 = 0.3586$, $Y_{10} = 0.2815$, $Y_{11} = 3.3054$, $Y_{12} = 6.6042$, $Y_{13} = 0.1115$, $Y_{14} = 1.1090$, $Y_{15} = 1.2348$, $Y_{16} = 2.5491$.
4. 윈도크기를 $m = 2$로 선택하여 Y_i들로부터 계산한 표본엔트로피는 $H_{2,16}(Y) = 1.3766$이 되고 이것을 사용해서 구한 검정통계량의 값은 $T_{2,16} = 1.9808$이 된다.

5. 유의수준을 5%로 하고 표본크기 $n = 16$과 윈도크기 $m = 2$에 해당하는 기각값을 Table 3.2에서 찾아보면 $c_{2,16}(0.05) = 1.6812$가 된다.
6. 검정통계량의 값이 기각값보다 크기 때문에 영가설을 기각할 수 없다. 그러므로 냉난방기에 대한 고장시간들이 역가우스분포를 따르게 된다는 결과를 얻는다.

6. 결론

본 논문에서는 역가우스분포에 대한 적합을 알아보기 위한 방법으로 확률적분변환에 기초를 둔 엔트로피 기반 적합도 검정을 제안했다. 검정에서 사용하는 검정통계량의 기각값을 구하기 위해서는 검정통계량의 표집분포의 이론적인 유도가 필요하다. 그러나 표집분포의 해석적 유도는 매우 어렵기 때문에 모의실험을 이용하여 추정한 기각값을 표의 형태로 제시했다. 이와 함께 주어진 표본크기에 대한 기각값을 제시한 표에서 찾을 수 없는 경우에 근사적인 기각값을 계산하기 위한 공식을 제시했다. Mudholkar와 Tian (2002)의 엔트로피 기반 검정과의 검정력 측면에서의 성능을 비교하기 위해서 모의실험을 수행했다. 검정력을 비교한 결과에서 제안한 검정은 기존의 엔트로피 기반 검정보다 더 좋은 검정력을 가지는 것으로 나타났다.

Mudholkar와 Tian (2002)의 엔트로피 기반 검정은 위치모수와 척도모수가 모두 알려져 있지 않거나 척도모수만 알려져 있는 역가우스분포에 대한 적합을 알아보고자 하는 경우에만 사용이 가능하기 때문에 적합할 역가우스분포의 위치모수와 척도모수가 모두 알려져 있거나 위치모수만 알려져 있는 경우에는 적용할 수 없는 단점을 가지고 있다. 그러나, 제안한 검정은 위치모수와 척도모수가 모두 알려져 있거나 위치모수만 알려져 있는 역가우스분포의 적합에도 적용할 수 있는 장점이 있고 검정력 또한 Mudholkar와 Tian (2002)의 검정보다 더 우수하게 나타나고 있으므로 응용에서 기존의 엔트로피 기반 검정보다는 활용도가 더 높을 것으로 기대된다.

References

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.

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

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

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.

Györfi, L. and van der Meulen, E. C. (1990). An entropy estimate based on a kernel density estimation. In: Limits Theorems in Probability and Statistics, *Colloquia Mathematica Societatis János Bolyai*, **57**, 229–240.

Hall, P. (1984). Limit theorems for sums of general functions of m-spacings, *Mathematical Statistics and Data Analysis*, **1**, 517–532.

Hall, P. (1986). On powerful distributional tests on sample spacings, *Journal of Multivariate Analysis*, **19**, 201–255.

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

Michael, J. R., Schucany, W. R. and Hass, R. W. (1976). Generating random variables using transformation with multiple roots, *The American Statistician*, **30**, 88–90.

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.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate, *Technometrics*, **5**, 375–384.

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

Shuster, J. J. (1968). On the inverse Gaussian distribution function, *Journal of the American Statistical Association*, **63**, 1514–1516.

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, Series B*, **38**, 54–59.

확률적분변환에 기초한 역가우스분포에 대한 적합도 검정

최병진[a,1]

[a]경기대학교 응용정보통계학과

(2013년 4월 11일 접수, 2013년 5월 30일 수정, 2013년 8월 6일 채택)

요 약

Mudholkar와 Tian (2002)이 제시한 엔트로피 기반 검정은 위치모수와 척도모수가 모두 알려져 있지 않거나 척도모수만 알려져 있는 역가우스분포의 적합을 알아보고자 하는 경우에만 사용이 가능하다. 본 논문에서는 위치모수와 척도모수가 모두 알려져 있거나 위치모수만 알려져 있는 역가우스분포의 적합에도 적용할 수 있는 엔트로피 기반 적합도 검정을 소개한다. 이 검정은 확률적분변환에 기초를 두고 있다. 모의실험을 통해서 추정한 표본크기와 윈도크기에 따른 검정통계량의 기각값과 근사기각값을 얻기 위한 계산공식을 제시한다. 제안한 검정과 Mudholkar와 Tian (2002)의 검정을 검정력 측면에서의 성능을 비교하고자 모의실험을 수행한다. 모의실험 결과에서 제안한 검정은 기존의 엔트로피 기반 검정보다 더 좋은 검정력을 가지는 것으로 나타난다.

주요용어: 역가우스분포, 엔트로피, 확률적분변환, 적합도, 검정력.

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