

역가우스분포에 대한 변형된 엔트로피 기반 적합도 검정

최병진¹

¹경기대학교 응용정보통계학과

(2011년 2월 접수, 2011년 3월 채택)

요약

이 논문에서는 역가우스분포의 적합을 위한 변형된 엔트로피 기반 검정을 제시한다. 이 검정은 자료생성분포와 역가우스분포의 엔트로피 차이에 기초를 두고 있으며 검정통계량은 엔트로피 차이의 추정량을 사용한다. 엔트로피 차이의 추정량은 자료생성분포에 대한 엔트로피 추정량으로 Vasicek의 표본엔트로피와 역가우스분포에 대한 엔트로피 추정량으로 균일최소분산불편추정량을 사용하여 얻는다. 모의실험을 통해 얻은 표본크기와 윈도크기에 따른 검정통계량의 기각값들을 표의 형태로 제공한다. 제안한 검정의 검정력을 알아보기 위해 여러 대립분포와 표본크기에 대해서 모의실험을 수행하고 기존의 엔트로피 기반 검정과 비교한다.

주요어: 역가우스분포, 엔트로피, 엔트로피 특성짓기, 엔트로피 추정량, 엔트로피 기반 검정, 검정력.

1. 서론

통신이론에서 불확실성의 측도로써 Shannon (1948)에 의해 처음으로 소개된 엔트로피는 확률밀도함수 $f(x)$ 를 가지는 연속확률변수 X 에 대해서 아래와 같이

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

로 정의된다. Shannon (1948)은 제약조건으로 분산이 주어졌을 때 정규분포의 엔트로피가 최대가 된다는 발견을 통해 엔트로피에 의해서 확률분포를 특성지을 수가 있음을 언급했다. 그의 발견 이후 여러 학자들에 의해 통계학 분야에서 잘 알려진 많은 확률분포들에 대한 엔트로피 특성짓기가 이루어져 왔고 특정한 제약조건과 이에 대한 최대 엔트로피를 갖는 확률분포의 세부적인 목록에 대해서는 Kapur와 Kesavan (1992), Ahmed와 Gokhale (1989)를 보기 바란다.

분포에 대한 엔트로피 특성짓기 결과는 검정할 분포에 대한 확률밀도함수가 $f_0(x)$ 일 때, 다음의 가설 $H_0 : f(x) = f_0(x)$ 대 $H_1 : f(x) \neq f_0(x)$ 에 대한 적합도 검정을 개발하기 위한 자연스러운 방법을 제공해 줄 수가 있다. Vasicek (1976)은 Shannon (1948)의 발견을 기초로 정규성에 대한 엔트로피 기반 검정을 처음으로 구축했으며 Dudewicz와 Meulen (1981)은 균일분포에 대한 엔트로피 기반 검정을 개발한 바가 있다. Grzegorzewski와 Wiczorkowski (1999)는 지수분포에 대한 엔트로피 기반 검정을 제안했고 Choi와 Kim (2006)은 라플라스분포에 대한 엔트로피 기반 검정을 소개한 바가 있다.

두 개의 모수 μ 와 λ 를 갖는 역가우스분포 $IG(\mu, \lambda)$ 는 오른쪽 긴 꼬리를 보이는 비음의 자료를 분석하기 위한 확률모형으로 유연성을 제공하기 때문에 넓은 범위의 응용성을 가진다. 이와 관련한 구체적인

¹(443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 응용정보통계학과, 부교수.

E-mail: bjchoi92@kyonggi.ac.kr

내용은 Chhikara와 Folks (1989), Seshadri (1999) 등을 참고하기 바란다. 역가우스분포는 또한 이론적으로 정규분포와 매우 유사한 성질을 가지고 있어서 다른 분포에 비해서 통계적 추론을 위한 과정이 비교적 수월하기 때문에 정규성에 기반을 둔 통계적 방법 못지않게 다양한 분석도구가 개발되어져 왔고 이들을 이용할 수가 있다. 그렇지만 자료분석에서 사용하고자 하는 분석도구가 타당하기 위해서는 적합도 평가를 통해 자료분포가 역가우스분포를 따르는지를 알아보는 작업이 선행되어야만 한다. 이런 목적으로 다양한 형태의 적합도 검정들이 Edgeman 등 (1988,1992), Edgeman (1990), O'Reilly와 Rueda (1992) 등에 의해 제안된 바가 있다.

역가우스분포는 통상적인 방법으로는 엔트로피에 관해서 특성을 지을 수가 없지만 Mudholkar와 Tian (2002)은 변환된 역가우스분포를 통해 엔트로피 특성짓기가 가능함을 보이고 적합도 검정을 개발하기 위해 이 결과를 이용했다. 이 검정은 $\Delta(g, g_0) = \exp\{H(g)\} / \exp\{H(g_0)\}$ 로 주어지는 엔트로피 차이에 기초를 두고 있으며 $H(g)$ 와 $H(g_0)$ 는 각각 미지의 변환된 자료생성분포와 변환된 역가우스분포의 엔트로피이다. 검정통계량으로 사용할 $\Delta(g, g_0)$ 의 추정량은 $H(g)$ 에 대해서는 Vasicek (1976)의 표본 엔트로피를 사용하고 $H(g_0)$ 에 대해서는 모수 $\xi^2 = 1/\lambda$ 에만 의존하므로 이것의 일치 추정량으로 대체한 추정량을 사용하여 얻게 된다. 그런데 그들이 사용한 $H(g_0)$ 의 추정량은 일치성은 확보되지만 최소분산은 보장해주질 못하기 때문에 이보다 더 좋은 추정량의 사용이 바람직할 것으로 판단된다. 따라서 $H(g_0)$ 에 대한 추정량으로 균일최소분산불편추정량을 사용해서 얻은 엔트로피 차이의 추정량을 검정통계량으로 이용하는 검정이 Mudholkar와 Tian (2002)에 의해 제안된 검정보다는 더 좋은 성능을 보여줄 것이라고 기대된다.

이 논문의 구성은 다음과 같다. 2절에서는 변환된 역가우스분포의 엔트로피에 대한 균일최소분산불편추정량을 소개하고 이것을 이용하여 얻은 검정통계량을 역가우스분포의 적합도 평가를 위해 사용하고자 한다. 3절에서는 2절에서 얻은 검정통계량을 사용하는 변형된 엔트로피 기반 검정을 Mudholkar와 Tian (2002)이 제시한 검정과 검정력 측면에서 성능을 비교하기 위해 모의실험을 수행하고 그 결과를 제시한다. 4절에서는 제안한 검정의 수행방법을 소개하고 실제자료에 적용해 본다. 마지막으로 결론을 5절에서 요약한다.

2. 변형된 검정통계량

확률변수 X 가 아래의 밀도함수

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0, \mu > 0, \lambda > 0 \quad (2.1)$$

를 가지는 역가우스분포 $IG(\mu, \lambda)$ 를 따른다고 하면, 변환된 확률변수 $Y = 1/\sqrt{X}$ 의 밀도함수는

$$g(y; \nu, \xi^2) = \sqrt{\frac{2}{\pi \xi^2}} \exp\left\{-\frac{(y^2 - \nu)^2}{2\xi^2 y^2}\right\}, \quad y > 0, \nu > 0, \xi^2 > 0 \quad (2.2)$$

가 되며, 여기서 $\nu = 1/\mu$ 와 $\xi^2 = 1/\lambda$ 이다. 이 사실을 기초로 Mudholkar와 Tian (2002)은 제약조건 $E(Y^{-2}) = 1/\nu$ 와 $E(Y^2) = 1/\nu + \xi^2$ 을 만족하는 모든 비율의 연속확률변수들 중에서, 식 (2.2)의 밀도함수를 가지는 확률변수 Y 의 엔트로피는 $\log(\xi^2 \pi e/2) / 2$ 로 최대가 되는 성질을 통해서 역가우스분포를 엔트로피에 관해서 특성을 지을 수가 있음을 서술했고 이 결과에 기초하여 다음의 가설 $H_0 : g(y) = g_0(y) (= g(y; \nu, \xi^2))$ 대 $H_1 : g(y) \neq g_0(y)$ 에 대한 적합도 검정의 구축을 시도했다. 그들의 방법론을 간략히 소개하면 다음과 같다.

크기 n 의 표본 X_1, X_2, \dots, X_n 이 $IG(\mu, \lambda)$ 를 따른다고 하면 $Y_i = \sqrt{X_i}$, $i = 1, \dots, n$ 에 의해 변환된 표본들의 분포는 (2.2)의 밀도함수를 가지게 된다. $H(g)$, $H(g_0)$ 와 $\Delta(g, g_0)$ 의 추정량을 각각 $H_{m,n}$,

H_{MT} 와 $\Delta_{m,n}$ 으로 표기하면 엔트로피 차이의 추정량은 $\Delta_{m,n} = \exp(H_{m,n}) / \exp(H_{MT})$ 로 얻게 된다. $H_{m,n}$ 은 Y_i 들로부터 얻어지는 표본엔트로피로 아래와 같이

$$H_{m,n} = \frac{1}{n} \sum_{i=1}^n \log \left[\frac{n}{2m} \{Y_{(i+m)} - Y_{(i-m)}\} \right] \tag{2.3}$$

로 주어지게 되며 $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ 은 Y_i 들의 순서통계량으로 $i > n$ 이면 $Y_{(i)} = Y_{(n)}$, $i < 1$ 이면 $Y_{(i)} = Y_{(1)}$ 이고 m 은 $n/2$ 보다 작은 양의 정수값을 갖는 윈도크기이다. H_{MT} 는 $H(g_0) = \log(\xi^2 \pi e / 2) / 2$ 에서 ξ^2 을 $V = \sum_{i=1}^n Y_i^2 / (n-1) - n^2 / \{ (n-1) \sum_{i=1}^n Y_i^{-2} \}$ 로 대체한 추정량으로 $H_{MT} = \log(V \pi e / 2) / 2$ 가 되고 $H(g_0)$ 에 대해 일치성을 가진다. $\Delta_{m,n}$ 으로부터 얻게 되는

$$K_{m,n} = \frac{2 \exp(H_{m,n})}{\sqrt{V}} \tag{2.4}$$

을 설정된 영가설을 검정하기 위한 통계량으로 사용하고 계산된 검정통계량의 값이 크게 나오면 X_i 들은 $IG(\mu, \lambda)$ 를 따르는 것으로 판단하게 된다.

$K_{m,n}$ 에서 사용한 H_{MT} 는 일치성이란 좋은 성질을 가지고 있지만 이보다 더 좋은 추정량이 있으면 이것을 검정통계량에 사용하는 것이 바람직하다고 여겨지므로 H_{MT} 대신에 $H(g_0)$ 의 균일최소분산불편추정량을 이용하고자 한다. $H(g_0)$ 의 균일최소분산불편추정량은 Choi (2006)에 의해 유도되고 분포적 성질이 규명된 바가 있으며, 추정량은

$$H_{MV} = \frac{1}{2} \log \left(\frac{V_W \pi e}{4} \right) - \frac{1}{2} \psi \left(\frac{n-1}{2} \right) \tag{2.5}$$

로 주어진다. 여기서 $W_i = Y_i^2$, $i = 1, \dots, n$, $V_W = \sum_{i=1}^n (W_i - \bar{W}_H)$, $\bar{W}_H = n / \sum_{i=1}^n W_i^{-1}$ 이고 $\psi(\cdot)$ 는 디감마함수(digamma function)로 $\psi(n) = d \log \Gamma(n) / dn$ 으로 정의된다. $H_{m,n}$ 과 H_{MV} 를 사용해서 얻어지는 $\Delta(g, g_0)$ 의 추정량은

$$\Delta_{m,n} = \frac{2 \exp \left\{ \frac{1}{2} \psi \left(\frac{n-1}{2} \right) \right\} \exp(H_{m,n})}{\sqrt{V_W \pi e}} \tag{2.6}$$

가 되고 이것을 $\sqrt{\pi e}$ 로 나눈 아래의 추정량

$$K_{m,n}^* = \frac{2 \exp \left\{ \frac{1}{2} \psi \left(\frac{n-1}{2} \right) \right\} \exp(H_{m,n})}{\sqrt{V_W}} \tag{2.7}$$

을 검정통계량으로 이용하고자 한다.

$K_{m,n}^*$ 검정은 대표본에서 기존의 $K_{m,n}$ 검정과 동일하게 됨을 두 검정통계량에 대한 아래의 관계식

$$K_{m,n}^* = \frac{\exp \left\{ \frac{1}{2} \psi \left(\frac{n-1}{2} \right) \right\}}{\sqrt{n-1}} K_{m,n} \tag{2.8}$$

을 통해서 보일 수가 있다. 식 (2.8)에서 $a_n = \exp \left\{ \psi \left((n-1)/2 \right) / 2 \right\} / \sqrt{n-1}$ 라 하면 a_n 은 $b_n = \exp \left\{ \psi \left((n-1)/2 \right) / 2 - \log \left((n-1)/2 \right) / 2 \right\} / \sqrt{2}$ 과 같게 되므로 Gradsbteyn과 Pzybik (2000)의 결과에 따라 $\lim_{n \rightarrow \infty} \left\{ \psi \left((n-1)/2 \right) / 2 - \log \left((n-1)/2 \right) / 2 \right\} = 0$ 이 된다. 따라서 $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = 1/\sqrt{2}$ 이 되므로 $K_{m,n}^*$ 는 $K_{m,n}/\sqrt{2}$ 로 근사되어진다. 이 사실을 통해서 대표본에서 두 검정통계량은 동일한 검정결과를 제공함을 알 수 있다.

$K_{m,n}^*$ 검정은 다른 엔트로피 기반 검정과 마찬가지로 모든 대립가설에 대해서 일치성을 가지게 됨을 다음과 같이 보일 수가 있으며, 여기서 일치성은 $m \rightarrow \infty$, $n \rightarrow \infty$ 이면서 $m/n \rightarrow 0$ 일 때 검정력이

표 2.1. 모의실험에 의해 추정된 유의수준 5%에서의 기각값 $K_{m,n}^*(0.05)$

n	m						
	1	2	3	4	5	6	7
5	0.659	0.959					
6	0.786	1.060					
7	0.889	1.160	1.210				
8	0.976	1.243	1.304				
9	1.036	1.306	1.361	1.346			
10	1.122	1.392	1.432	1.431			
12	1.228	1.506	1.555	1.550	1.509		
14	1.321	1.611	1.668	1.652	1.621	1.568	
16	1.403	1.698	1.746	1.746	1.704	1.663	1.610
18	1.465	1.757	1.820	1.813	1.782	1.747	1.702
20	1.515	1.825	1.890	1.885	1.860	1.823	1.779
25	1.623	1.933	2.008	2.017	1.999	1.966	1.931
30	1.683	2.004	2.093	2.110	2.100	2.076	2.047
35	1.732	2.060	2.153	2.178	2.179	2.162	2.137
40	1.771	2.106	2.206	2.237	2.241	2.231	2.212
45	1.812	2.146	2.250	2.284	2.293	2.286	2.271
50	1.836	2.180	2.283	2.322	2.332	2.328	2.317

1로 확률수렴(이하 수렴)하게 됨을 의미한다. Vasicek (1976)의 연구결과에 의해 $m \rightarrow \infty, n \rightarrow \infty, m/n \rightarrow 0$ 이면 $H_{m,n}$ 은 영가설하에서 $H(g_0)$ 로 수렴하게 되므로 $\exp(H_{m,n})$ 는 $\xi\sqrt{\pi e/2}$ 로 수렴함을 알 수 있다. 또한 $2\exp\{\psi((n-1)/2)/2\}/\sqrt{V_W}$ 는 영가설하에서 $n \rightarrow \infty$ 일 때 $\sqrt{2}/\xi$ 로 수렴을 하므로 $K_{m,n}^*$ 는 $\sqrt{\pi e}$ 로 수렴하게 된다. 한편 대립가설하에서는 $K_{m,n}^* < \sqrt{\pi e}$ 이 되기 때문에 $m \rightarrow \infty, n \rightarrow \infty, m/n \rightarrow 0$ 이면 $K_{m,n}^*$ 의 검정력은 1로 수렴하게 된다.

자료로부터 계산된 $K_{m,n}^*$ 의 값은 영가설이 사실이면 $\sqrt{\pi e}$ 에 가깝게 될 것이고 대립가설이 사실이면 $\sqrt{\pi e}$ 보다 더 작게 될 것이므로 영가설의 기각여부를 판단하기 위해서는 영가설 하에서 $K_{m,n}^*$ 의 표본분포를 통해서 유의수준 α 에서의 기각값 $K_{m,n}^*(\alpha)$ 를 결정해야만 한다. 그러나 표본엔트로피와 유사 표본엔트로피들을 포함하는 통계량들의 점근적 분포성질들이 Cressie (1976), Dudewicz와 Meulen (1981), van Es (1992) 등에 의해 밝혀졌다고는 하지만 이들 연구결과와의 적용은 제한적이어서 $K_{m,n}^*$ 의 표본분포를 이론적으로 도출하는 것은 여전히 처리하기 어려운 문제이다. 따라서 영가설하에서 검정통계량의 표본분포로부터 100α 백분위수를 결정하기 위해 모의실험을 수행하고 추정된 값을 검정의 수행에 필요한 기각값 $K_{m,n}^*(\alpha)$ 로 사용하고자 한다. 표 2.1은 표본크기 n 과 윈도우 크기 m 에 대한 유의수준 5%에서의 기각값으로 Michael 등 (1976)의 알고리즘을 사용해서 IG(1, 1)로부터 생성된 자료를 변환해서 검정통계량을 계산하는 과정을 10000번 반복을 해서 얻은 것이다.

3. 검정력 분석

$K_{m,n}^*$ 검정의 성능을 검정력 관점에서 분석하기 위해서 모의실험을 수행한다. 표본크기 n 은 각각 10, 20, 30, 50으로 했으며 대립분포로는 표준지수분포($E(1)$), 균일분포($U(0, 1)$), 카이제곱분포(χ^2), 형상모수 θ 와 척도모수 β 를 갖는 와이블분포($W(\theta, \beta)$), 형상모수 θ 와 척도모수 β 를 갖는 감마분포($G(\theta, \beta)$), 모수 μ 와 σ^2 을 갖는 로그정규분포($LN(\mu, \sigma^2)$) 등을 선택했다.

검정력을 분석하기 위해 각 대립분포로부터 표본크기별로 자료를 생성하여 변환을 시킨 다음 두 검정

표 3.1. 유의수준 5%에 대한 $K_{m,n}^*$ 검정의 추정된 검정력

n	m	IG(1, 0.5)		IG(1, 2)		E(1)		U(0, 1)	
		$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$
10	2	4.79	4.37	5.31	4.82	22.24	21.33	50.81	49.48
20	3	5.21	4.84	5.45	5.02	48.55	47.55	89.02	88.40
30	3	4.59	4.45	5.44	5.29	65.66	65.44	98.20	98.16
50	4	4.66	4.51	5.20	5.12	86.61	86.37	99.99	99.99
n	m	χ_2^2		χ_3^2		W(1.5, 1)		W(1, 2)	
		$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$
10	2	22.59	21.77	13.45	12.93	15.53	14.78	13.14	12.45
20	3	47.50	46.49	28.13	27.20	32.84	31.89	27.53	26.50
30	3	65.44	65.07	41.29	40.88	47.58	47.05	39.01	38.65
50	4	86.35	86.13	61.18	60.84	69.27	68.81	61.14	60.76
n	m	G(1.5, 1)		G(2.5, 1)		LN(0, 1)		LN(0, 4)	
		$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$
10	2	13.08	12.41	8.64	8.05	5.65	5.24	15.10	14.43
20	3	28.15	27.36	15.16	14.37	6.16	5.66	35.05	34.19
30	3	41.71	41.32	21.65	21.20	7.21	6.95	53.31	52.83
50	4	62.13	61.70	32.68	32.13	9.10	8.93	75.81	75.39

통계량 $K_{m,n}^*$ 와 $K_{m,n}$ 의 값을 계산하는 과정을 10000번 반복했다. 검정통계량의 계산에서 사용한 윈도크기 m 의 값은 Vasicek (1976)의 추천에 따라 $n = 10$ 에 대해서는 $m = 2$, $n = 20, 30$ 에 대해서는 $m = 3$, $n = 50$ 에 대해서는 $m = 4$ 를 사용했다. 검정력은 10000개의 $K_{m,n}^*$ 의 값들 중 표 2.1에서 제시된 m, n 에서 해당하는 기각값보다 작게 나온 값들의 수를 세어서 얻은 기각비율을 백분율로 계산했다. $K_{m,n}$ 의 경우도 동일한 방식으로 검정력을 계산했으며 기각값은 Mudholkar와 Tian (2002)에 제시된 표에서 찾아서 사용했다.

표 3.1은 유의수준 5%에 대한 $K_{m,n}^*$ 검정의 추정된 검정력을 보여주고 있다. IG(1, 0.5)와 IG(1, 2)의 경우에는 영가설이 사실이기 때문에 추정된 검정력은 모든 표본크기에서 5% 근처에 있는 값으로 나타나야만 한다. 만일 아주 작거나 크게 나온다면 표 3.1에 주어진 기각값의 정확성에 문제가 있음을 나타내는 것이므로 이 표를 이용해서 얻은 $K_{m,n}^*$ 검정의 결과를 신뢰할 수가 없게 된다. 10000번 반복을 한 모의실험에서 영가설의 기각비율에 대한 표준오차를 계산해보면 $\sigma_p = \sqrt{0.05(1 - 0.05)/10000} \approx 0.0022$ 가 되므로 기각비율에 대한 95% 신뢰구간은 대략 (0.0456, 0.0544), 즉 (4.56%, 5.44%)가 된다. 표 3.1의 IG(1, 0.5)와 IG(1, 2)에 대한 값들은 4.59%부터 5.45%에 걸쳐 분포하므로 $K_{m,n}^*$ 검정은 유의수준을 잘 통제함을 알 수 있다.

$K_{m,n}^*$ 검정은 제시된 결과에서 보는 바와 같이 모든 대립분포와 표본크기에서 $K_{m,n}$ 검정보다는 더 좋은 성능을 가짐을 알 수 있다. $K_{m,n}$ 에 대한 $K_{m,n}^*$ 의 검정력 차이는 로그정규분포, 감마분포, 와이블분포의 순으로 크게 나타나며 균일분포에서는 가장 작음을 볼 수 있다. 또한 카이제곱분포와 표준지수분포에서도 두 검정력의 차이는 와이블분포의 경우보다는 작지만 다소 크게 되는 경향을 보인다. 표본크기의 경우 10과 20에서의 $K_{m,n}^*$ 검정력은 $K_{m,n}$ 의 것보다 더 좋게 나타나지만 30과 50으로 증가함에 따라 두 검정력은 거의 같게 되며 이런 현상을 얻게 된 것은 2절에서 밝힌 바와 같이 대표본에서 두 검정은 동일한 검정결과를 주기 때문이다.

표 3.2는 윈도크기에 따라 $K_{m,n}^*$ 의 검정력이 어떻게 변하는지를 알아보기 위해 표본크기가 20일 때 10000번의 반복을 통해 유의수준 5%에서 추정한 검정력을 백분율로 나타낸 것이다. 결과에서 보는

표 3.2. 윈도크기에 따른 유의수준 5%에 대한 $K_{m,n}^*$ 검정의 추정된 검정력

대립분포	m					
	2		3		4	
	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$	$K_{m,n}^*$	$K_{m,n}$
$G(1, 1)$	46.65	46.02	47.17	46.34	44.00	42.53
$G(2, 1)$	47.57	46.88	48.01	47.11	44.86	43.37
$W(1, 1)$	46.60	45.74	47.55	46.57	44.31	42.91
$W(3, 1)$	47.69	47.15	48.52	47.73	45.09	43.59
$LN(0, 2)$	12.99	12.46	12.33	11.73	10.12	9.41
$LN(0, 3)$	23.38	22.91	22.99	22.22	19.64	18.56

표 4.1. 볼 베어링 자료에 대한 기초통계량

평균	최소값	최대값	표준편차	왜도	첨도
72.225	17.88	173.40	37.489	1.008	3.926

바와 같이 $K_{m,n}^*$ 검정은 $K_{m,n}$ 검정보다 일관되게 더 좋은 검정력을 가지는 것으로 나타난다. 또한 $K_{m,n}^*$ 의 검정력은 대체적으로 윈도크기가 3일 때 가장 높게 나타나며, 표에서는 제시하지 않았지만 5 이상의 윈도크기에 대한 결과를 보면 윈도크기가 커짐에 따라 검정력은 점점 낮아지게 되는 것을 발견할 수 있다. 이와 같은 현상은 $K_{m,n}$ 에 대해서도 동일하게 나타난다.

4. 응용 사례

$K_{m,n}^*$ 검정은 아래의 절차에 따라 수행을 하면 된다.

1. 표본 X_1, X_2, \dots, X_n 로부터 $Y = 1/\sqrt{X}$ 에 의해 변환된 표본 Y_1, Y_2, \dots, Y_n 을 얻는다.
2. 변환된 표본을 오름차순으로 나열한 $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ 를 얻는다.
3. 윈도크기를 선택해서 검정통계량 $K_{m,n}^*$ 를 계산한다.
4. 표본크기와 검정통계량의 계산에서 사용한 윈도크기에 해당하는 기각값 $K_{m,n}^*(0.05)$ 를 표 2.1에서 찾는다. 표에 없는 경우에는 적절한 보간법을 이용하여 근사적인 기각값을 구한다.
5. 계산된 검정통계량의 값이 기각값 $K_{m,n}^*(0.05)$ 보다 작으면 유의수준 5%에서 영가설을 기각한다.

이제, $K_{m,n}^*$ 검정을 Lieblein과 Zelen (1956)에 의해 분석되어진 볼 베어링 자료에 적용해 보기로 한다. 이 자료는 깊은 홈이 파여져 있는 볼 베어링의 내구력을 조사한 것으로 수명검사에 사용된 23개 볼 베어링 각각에 대하여 파손되기전 회전수를 백만 단위로 측정된 것이다. 검정에 앞서 자료에 대한 몇 개의 기초통계량을 구해서 정리해 보면 표 4.1과 같고 왜도와 첨도는 각각 1.008과 3.926으로 계산이 되었음을 볼 수 있다. 역가우스분포에 대한 이론적인 왜도와 첨도는 각각 $\sqrt{\beta_1} = 3\sqrt{\mu/\lambda}$ 와 $\beta_2 = 15\mu/\lambda$ 가 되고 자료로부터 계산한 $\hat{\mu} = 72.225$ 와 $\hat{\lambda} = 231.671$ 을 사용해서 추정치를 구해 보면 각각 $\sqrt{b_1} = 1.675$ 와 $b_2 = 4.676$ 이 된다. 이들 $\sqrt{b_1}$ 과 b_2 를 표 4.1에 제시된 왜도와 첨도를 비교해 보면 큰 차이가 나지 않음을 알 수가 있다. 그림 4.1은 자료분포의 형상을 좀 더 명확히 보기 위해서 막대그래프와 역가우스분포의 적합선을 그린 것으로 실선은 추정된 IG(72.225, 231.671)의 확률곡선을 나타낸다. 이들 결과를 종합해보면 볼 베어링 자료는 아마도 역가우스분포를 따를 것으로 짐작해 볼 수가 있다.

검정을 통해 판단하기 위하여 $K_{m,n}^*$ 검정을 수행해 보기로 한다. 윈도크기 $m = 3$ 을 사용해서 검정통계량을 계산해 보면 $K_{3,23}^* = 2.405$ 가 된다. 기각값은 표 2.1에 나와 있지 않으므로 근사값을 구해야만 하

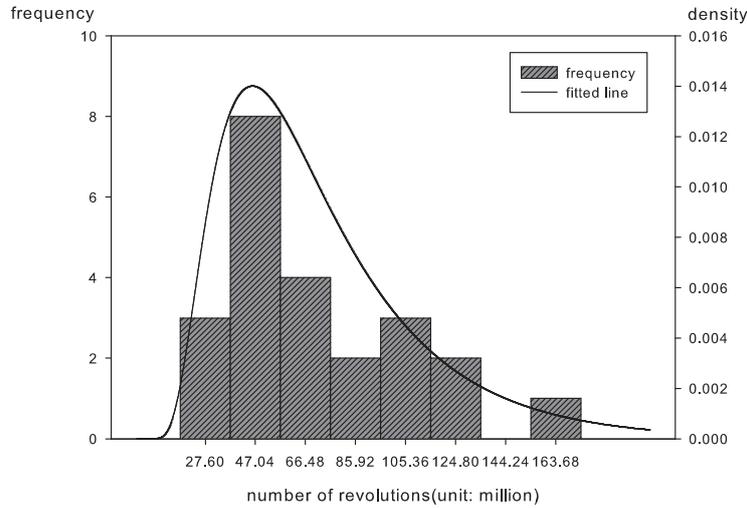


그림 4.1. 볼 베어링 자료에 대한 막대그래프

며, $n = 20, m = 3$ 일 때 기각값 $K_{3,20}^*(0.05) = 1.890$ 과 $n = 25, m = 3$ 일 때 기각값 $K_{3,25}^*(0.05) = 2.008$ 을 이용하여 선형보간시킨 기각값을 구해보면 $K_{3,23}^*(0.05) = 1.9608$ 이 된다. 따라서 $K_{3,23}^* = 2.405$ 는 $K_{3,23}^*(0.05) = 1.9608$ 보다 크기 때문에 유의수준 5%에서 영가설을 채택하게 되므로 볼 베어링 자료는 역가우스분포를 하는 것으로 볼 수 있다.

5. 결론

본 논문에서는 역가우스분포에 대한 적합을 알아보기 위한 변형된 엔트로피 기반 검정을 소개했다. 이 검정은 표본크기와 윈도우크기가 커짐에 따라 모든 대립가설에 대하여 일치성을 가짐을 보였으며 대표본에서는 Mudholkar와 Tian (2002)이 제시한 기존의 엔트로피 기반 검정과 동일하게 됨을 밝혔다. 검정력 분석을 위해서 모의실험을 수행했으며 실험에서 고려한 모든 대립분포와 표본크기에서 제안한 검정이 Mudholkar와 Tian (2002)의 검정보다 더 좋은 성능을 보이는 것으로 나타났다. 변형된 엔트로피 기반 검정은 영가설하에서의 검정통계량의 분포가 알려져 있지 않아서 기각값 또는 유의확률을 모의실험을 통해서 계산해야만 하는 제한이 있기는 하지만, 응용에서 검정력 이득을 성취하기 위해 기존 엔트로피 기반 검정의 대안으로 사용될 수 있을 것으로 판단된다.

참고문헌

Ahmed, N. A. and Gokhale, D. V. (1989). Entropy expressions and their estimators for multivariate distributions, *IEEE Transactions on Information Theory*, **35**, 688–692.

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

Choi, B. (2006). Minimum variance unbiased estimation for the maximum entropy of the transformed inverse Gaussian random variable by $Y = X^{-1/2}$, *The Korean Communications in Statistics*, **13**, 657–667.

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

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

- Dudewicz, E. J. and van der Meulen, E. C. (1981). Entropy-based test for uniformity, *Journal of the American Statistical Association*, **76**, 967–974.
- 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.
- Edgeman, R. L., Scott, R. C. and Pavur, R. J. (1992). Quadratic statistics for the goodness-of-fit test for the inverse Gaussian distribution, *IEEE Transactions on Reliability*, **41**, 118–123.
- Gradshteyn, I. S. and Ryzhik, I. M. (2000). *Table of Integrals, Series, and Products*, Academic Press, San Diego.
- Grzegorzewski, P. and Wieczorkowski, P. (1999). Entropy-based test goodness-of-fit test for exponentiality, *Communications in Statistics-Theory and Methods*, **28**, 1183–1202.
- Kapur, J. N. and Kesavan, H. K. (1992). *Entropy Optimization Principles with Applications*, Academic Press, San Diego.
- Lieblein, J. and Zelen, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearings, *Journal of Research of the National Bureau of Standards*, **57**, 273–316.
- 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.
- O'Reilly, F. J. and Rueda, R. (1992). Goodness of fit for the inverse Gaussian distribution, *The Canadian Journal of Statistics*, **20**, 387–397.
- 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.
- 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*, **B38**, 54–59.

A Modified Entropy-Based Goodness-of-Fit Test for Inverse Gaussian Distribution

Byungjin Choi¹

¹Department of Applied Information Statistics, Kyonggi University

(Received February 2011; accepted March 2011)

Abstract

This paper presents a modified entropy-based test of fit for the inverse Gaussian distribution. The test is based on the entropy difference of the unknown data-generating distribution and the inverse Gaussian distribution. The entropy difference estimator used as the test statistic is obtained by employing Vasicek's sample entropy as an entropy estimator for the data-generating distribution and the uniformly minimum variance unbiased estimator as an entropy estimator for the inverse Gaussian distribution. The critical values of the test statistic empirically determined are provided in a tabular form. Monte Carlo simulations are performed to compare the proposed test with the previous entropy-based test in terms of power.

Keywords: Inverse Gaussian distribution, entropy, entropy characterization, entropy estimator, entropy-based test, power.

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