Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 2, pp. 387 ~ 395

엔트로피 추정에 기초한 적합도 검정

김 종태 [1] · **차 영준** [2] · **김 영훈** [3] · **이 재만** [4] · **강 상길** [5]

요약

본 연구의 목적은 엔트로피 추정량에 기초한 적합도검정 통계량들을 제시하고 경험적 분포에 기초한 적합도 검정 통계량들과의 검정력을 비교하여 향후 적합도 검정에 사용될 검정력을 선정하는데 그 목적이 있다. 또한 모란 (Moran)의 검정통계량과 엔트로피의 추정을 연구함으로 그 일치성을 알 수 있다

주제어: 적합도 검정, 엔트로피

1. 서론

$X_1, X_2, \cdots, X_n$이 독립인 연속확률변수로서 확률밀도함수 $f(x)$과 분포함수 $F(x)$을 가진다고 하자. 확률표본이 주어진 확률밀도함수 $f_0(x;\theta)$를 따르는지를 검정하기 위한 적합도 검정의 가설은 다음과 같다.

$$H_0 : f(x) = f_0(x;\theta)$$

이러한 적합도 검정 문제는 분포함수 $F(x)$에 대하여 모수의 벡터 θ를 가지는 모수모형 $F(x;\theta)$의 적합성을 검정을 하는 것이다. 확률변수 X의 백분율 함수 $Q(u) = F^{-1}(u)$라 하면 백분율 밀도함수는 $q(u) = Q'(u) = 1/fQ(u)$이고 밀도 백분율함수는 $fQ(u) = f(Q(u))$이다. 확률밀도함수 f를 갖는 분포함수 F에 대한 엔트로피는 다음과 같다.

$$H(f) \quad = \quad -\int_{-\infty}^{\infty} f(x)\ln f(x)dx = \int_{-\infty}^{\infty} \ln\{q(u)\}du.$$

$H(f)$는 Shannon의 엔트로피 혹은 엔트로피라고 부른다. 엔트로피 정보함수의 추정량에 기초한 많은 검정통계량들을 고찰함으로 그 유형을 분석하는 일은 매우 중요하다. 2절에서는

[1]경북 경산시 진량면 대구대학교 자연과학대학 통계학과, 부교수
[2]경북 안동시 송천동 안동대학교 자연과학대학 통계학과, 교수
[3]경북 안동시 송천동 안동대학교 자연과학대학 통계학과,교수
[4]경북 안동시 송천동 안동대학교 자연과학대학 통계학과, 부교수
[5]대구시 복현동 경북대학교 자연과학대학 통계학과, 강사

Moran의 통계량과 엔트로피와의 관계를 조사하고, 3절과 4절에서는 엔트로피의 추정과 엔트로피에만 기초한 검정통계량들의 흐름을 고찰할 것이다. 5절에서는 기각역을 제시하고 모의실험을 위한 대립모형을 제시한다. 마지막으로 6절에서는 검정력비교와 결론 을 제시한다.

2. Moran의 검정통계량

$X_1, X_2, \cdots, X_n$이 독립인 연속확률변수로서 확률밀도함수 $f(x)$과 분포함수 $F(x)$을 가지고, $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$은 표본크기 n을 가지는 확률표본에 기초한 순서통계량이라 하자. 확률누적변환 (probability integral transformation), $U = F(X;\theta)$를 이용하여 U의 순서통계량은 $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$이고 $U_{(0)} = 0$이고 $U_{(n+1)} = 1$이다. Cheng과 Stephens (1989)는 Moran의 통계량을 다음과 같이 정의했다.

$$M(\theta) = -\sum_{i=1}^{n+1} \log D_i(\theta)$$

여기서 $D_i(\theta) = U_{(i)} - U_{(i-1)}$이고, $i = 1, 2, \cdots, n+1$이다. 그들은 모수 θ에 대한 효율적인 추정량 $\hat{\theta}$가 주어질 때, $M(\theta)$에 대한 점근적 분포를 연구하였다. 또한 정규성에 대한 검정력의 비교에 있어서 기존의 고전적 검정통계량인 Kolmogorov-Smirnov와 Cramer von Mises의 검정통계량들의 검정력에 비하여 우수함을 보였다. Parzen (1982)은 Moran의 통계량을 정규화시켜 다음과 같은 형태의 통계량을 제시했다.

$$M(\theta) = -\frac{\sum_{i=1}^{n+1} \log\{(n+1)D_i(\theta)\}}{n+1} = -\int_0^1 \log \tilde{d}(u;\theta)du.$$

여기서 $\tilde{d}(u;\theta) = (n+1)D_i(\theta)$로서 $\frac{i-1}{n+1} < u < \frac{i}{n+1}$이다. Moran의 통계량 $M(\theta)$은 $U = F(X;\theta)$의 음의 엔트로피 $-H(U)$로서 그것의 추정량은 $\tilde{M}(\theta)$이 된다.

3. 엔트로피의 추정

엔트로피 $H(f)$에 대한 추정을 위한 많은 연구들은 Dmitriev와 Tarasenko (1973), Vasicek (1976), Ahmad와 Lin (1976), Mack (1988), van Es (1992), Correa (1995)등에 의해 제시되어 왔다. 확률표본 $X_1, X_2, \cdots, X_n$ $(n \ge 3)$이 절대연속 확률밀도함수 $f(x)$를 가지는 모집단으로부터 주어지고, $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$은 표본크기 n을 가지는 확률표본에 기초한 순서통계량이다. 만약 $i < 1$이면 $X_{(i)} = X_{(1)}$이고, 만약 $i > n$이면 $X_{(i)} = X_{(n)}$이고 윈도우 크기 m은 $n/2$보다 작은 양의 상수라고 하자.

Vasicek (1976)에 의해 개발된 엔트로피 추정량은 다음과 같다.

$$HV_{mn} = \frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\right\}$$

Vasicek은 $n \to \infty$, $m \to \infty$이고 $\frac{m}{n} \to \infty$이면 $HV_{mn} \to H(f)$임을 증명하였다.

van Es (1992)는 확률표본의 차이에 기초한 엔트로피 추정량을 다음과 같이 제시하였다.

$$HE_{mn} = \frac{1}{n-m}\sum_{i=1}^{n-m} \ln\left\{\frac{n+1}{m}(X_{(i+m)} - X_{(i)})\right\} + \sum_{k=m}^{n}\frac{1}{k} + \log(m) - \log(n+1).$$

그는 이 추정량에 대한 점근적 정규분포의 성질을 증명하였다.

Correa (1995)는 Vasiseck의 추정량을 변형하여 다음과 같은 엔트로피 추정량을 제시하였다.

$$HC_{mn} = -\frac{1}{n}\sum_{i=1}^{n} ln(b_i).$$

여기서 $b_i = \frac{\sum_{k=i-m}^{i+m}(X_{(k)} - \bar{X}_{(i)})(k-i)}{n\sum_{k=i-m}^{i+m}(X_{(k)} - \bar{X}_{(i)})^2}$, 이고, $\bar{X}_{(i)} = \frac{1}{2m+1}\sum_{k=i-m}^{i+m} X_{(k)}$.

실제로 Vasicek의 엔트로피 추정량은 엔트로피를 이용한 적합도 검정 뿐 아니라 쿨백-레이블러 정보함수 (Kullback-Leibler Information Function)를 이용한 적합도 검정에도 많은 영향을 주었다. van Es는 확률표본의 차이에 기초한 엔트로피 추정량을 제시하였고, Correa는 Vasicek의 엔트로피 추정량을 변형시킨 엔트로피 추정량을 제시하였다.

4. 엔트로피에 기초한 적합도검정

Dudewicz와 van Der Meulen (1981)은 Vasicek의 엔트로피 추정량을 이용하여 적합도 검정의 귀무가설에서, $n \to \infty$, $m \to \infty$이고 $m/n \to 0$이면 HV_{mn}이 0에 확률적 수렴임을 보였다. 절대연속함수 $f(x)$의 구간 $[0,1]$을 가지는 대립함수에서는, $n \to \infty$, $m \to \infty$이고 $m/n \to 0$이면 HV_{mn}이 음의 실수 (유한실수 혹은 $-\infty$)에 확률적 수렴함을 보였다. 또한 $[0,1]$ 구간에 속하는 모든 f에 대하여 $HV_{mn} \leq 0$임을 보였다.

이러한 성질들을 기초로 Vasicek과 van Es와 Correa의 엔트로피 추정량을 단조변환 시킴으로서 가설에 대한 적합도검정의 다음의 기각영역을 가능하게 한다.

$$\begin{aligned} TV_{mn} &= \exp(-HV_{mn}) \leq TV^*{}_{mn}(\alpha), \\ TE_{mn} &= \exp(-HE_{mn}) \leq TE^*{}_{mn}(\alpha), \\ TC_{mn} &= \exp(-HC_{mn}) \leq TC^*{}_{mn}(\alpha). \end{aligned}$$

여기서 $TV^*{}_{mn}(\alpha)$, $TE^*{}_{mn}(\alpha)$, $TC^*{}_{mn}(\alpha)$는 주어진 m과 n에 대하여 수준 α에서 기각값들이다.

5. 기각영역과 모의실험

엔트로피 추정량의 값들이 적을수록 귀무가설을 기각할 확률이 높아지고 이에 반해 3절에서 제시한 검정통계량들, $TV^*_{mn}(\alpha)$, $TE^*_{mn}(\alpha)$, $TC^*_{mn}(\alpha)$은 값이 클수록 귀무가설을 기각시킬 확률이 높아짐을 알 수 있다. 그러므로 기각값 $T(\cdot)^*_{mn}(\alpha)$에 대하여, 만약 $T(\cdot)_{mn} < T(\cdot)^*_{mn}(\alpha)$이면, H_0을 기각한다.

여기서 기각영역을 구하는 방법을 다음과 같이 생각 할 수 있다.

(1) 표본의 크기 n에 관련되는 윈도우 크기 $m < \frac{n}{2}$에서 정규분포 난수를 20,000번 반복횟수로 기각값들 중 가장 큰 기각값을 가지는 m을 구한다

(2) (1)에서 구한 m을 이용하여 각각의 유의수준 α에서 $T(\cdot)^*_{mn}(\alpha)$을 기각 영역으로 결정한다.

(3) m에 대하여 $T(\cdot)_{mn}$을 가장 크게 하는 값을 구하여 검정한다.

위의 방법에 대하여 $5 \le n \le 250$의 표본의 크기들에 대하여 20,000번 반복횟수를 사용하여 유의수준 α가 0.05와 0.01이 되는 기각 영역을 구하였다.

엔트로피에 기초한통계량들, TV_{mn}과 TE_{mn}과 TC_{mn}에 대하여 경험적 분포에 기초한 기존의 검정통계량들, Kolmogorov-Smirnov (D), Kuiper (V), Cramer-von Mises (W^2), Watson (U^2), Andeson-Darling (A^2) Moran (M)의 검정력을 모의실험을 이용하여 비교분석 할 것이다. 검정력에 대한 비교를 위하여 대립가설의 분포로 다음과 같은 대립모형의 분포들을 사용하였다.

[**모형 A**] **(균등분포에서의 변환)**

만약 $0 \le x \le 0.5$인 경우, $F(X) = 2^{k-1}x^k$ 이고, 만약 $0.5 < x \le 1$ 인 경우, $F(x) = 1 - 2^{k-1}(1-x)^k$에 대하여 $k = 1.5, 2.0, 3.0$인 경우와

[**모형 B**] **(베타분포에서의 변환)**

$f(x) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}x^{a-1}(1-x)^{b-1}$, $a > 0, b > 0, 0 < x < 1$대하여 $a = b = 0.5, 1.0, \cdots, 4.0$인 경우.

각각의 대립분포들의 모형들에 대하여 표본의 크기 n은 5, 10, 20, 30, 40, 50을 정하였고 모의실험의 반복횟수는 10,000번을 하였다.

6. 검정력비교와 결론

대립가설 [모형 A]의 경우는 균등분포에서 파생된 것으로 확률밀도함수가 단조 증가하는 모양을 가졌다. 이러한 모형에 대한 검정력 비교에 있어서 〈표 1〉에서 보는 것과 같이 Watson (U^2), Andeson-Darling (A^2)의 통계량이 매우 월등한 검정력을 가짐을 알 수 있다. 다음으로 Kuiper (V)의 통계량이 우수하며, 엔트로피의 추정에 대한 검정통계량들의 검정력들이 뒤를 잇는다. 엔트로피 추정에 의한 검정통계량들 중에서도 Vasicek의 추정량에 의한 검정통계량 (TV)이 가장 우수하며 그 뒤를 Correa와 van Es의 검정통계량 (TC, TE)이

뒤를 잇는다. 비록 모란의 검정통계량은 엔트로피의 추정과 유사점이 많으나 [모형 A]의 검정력 비교에 있어서는 다른 검정통계량들 보다 검정력이 좋지 않음을 알 수 있다.

〈표 2〉 에서는 [모형 B]의 베타분포에 대한 검정력을 비교하였다. $a = b = .5$인 경우는 베타분포의 모양이 U의 형태로 일반적인 밀도함수의 형태가 아니다. 이 경우에 있어서 엔트로피의 추정에 의한 검정통계량들 보다 경험적분포에 기초한 검정통계량들의 검정력이 우수함을 알 수 있다. $a = b = 1.0$인 경우는 베타분포가 균등분포로서 검정력의 값이 유의수준 $\alpha = 0.05$와 일치하는 것을 조사할 수 있다. 대체로 모든 경우에서 만족함을 알 수 있다. $a = b$의 값이 1.0보다 큰 경우는 확률밀도함수의 값이 일반적인 경우로서 엔트로피에 의한 추정량에 기초한 검정통계량의 값이 경험적 분포에 기초한 검정통계량들의 값들 보다 매우 큰 검정력값을 가짐을 알 수 있다.

결론적으로 엔트로피 추정량에 기초한 검정통계량들은 비록 단조 증가 함수나 U자 모양을 가지는 모형에 있어서의 검정력의 값들이 경험적 분포에 의한 검정통계량의 검정력보다 다소 낮지만 일반적인 단일 모드(mode)를 가지는 분포에 있어서는 매우 좋은 검정력을 가짐을 알수 있다.

표 1: 유의수준 $\alpha = 0.05$에서 검정력 비교 (모형 A의 경우)

n	k	TV	TE	TC	D	V	W^2	U^2	A^2	M
	1.5	.03130	.03780	.06120	.09090	.08440	.09320	.08400	.15280	.09340
5	2.0	.03560	.04730	.06750	.13130	.16130	.12950	.16460	.28950	.13090
	3.0	.05440	.08200	.09930	.22060	.33170	.19530	.33550	.51230	.19510
	1.5	.02350	.02390	.02100	.10290	.12880	.09800	.13800	.18120	.09460
10	2.0	.02950	.03310	.02200	.17080	.30350	.16300	.31640	.39120	.15560
	3.0	.07170	.07070	.04380	.30440	.57580	.29150	.60030	.70570	.28240
	1.5	.03850	.01720	.05470	.13530	.23200	.12570	.24660	.25390	.11580
20	2.0	.13330	.03890	.19090	.27380	.55040	.26600	.58260	.60700	.25930
	3.0	.44120	.19100	.53690	.53380	.88230	.57090	.90040	.93010	.57320
	1.5	.09520	.02050	.16300	.15180	.30800	.13490	.33280	.31260	.12980
30	2.0	.40440	.11090	.48210	.36540	.72240	.37270	.75650	.75390	.38250
	3.0	.86990	.53970	.88350	.73670	.97230	.79300	.98070	.98520	.80860
	1.5	.14330	.02150	.16420	.18440	.41210	.17570	.44240	.39680	.17040
40	2.0	.57270	.14300	.59520	.47970	.84450	.51050	.87640	.86100	.52430
	3.0	.96070	.68740	.95840	.87090	.99420	.92020	.99660	.99730	.92920
	1.5	.23700	.03440	.24620	.23420	.51820	.22410	.55840	.49130	.22150
50	2.0	.75860	.26320	.75820	.58330	.91880	.63910	.94200	.93050	.65660
	3.0	.99490	.87610	.99160	.94810	.99870	.97610	.99950	.99940	.97900

표 2:유의수준 $\alpha = 0.05$에서 검정력 비교 (모형 B의 경우)

n	k	TV	TE	TC	D	V	W^2	U^2	A^2	M
	0.5	.03160	.04070	.06520	.10850	.12350	.11600	.12480	.24550	.11810
	1.0	.04730	.05040	.08670	.05010	.04960	.04910	.04780	.04790	.04770
	1.5	.08500	.07750	.13110	.03260	.06480	.02640	.06640	.01740	.02390
5	2.0	.14120	.12370	.19840	.02540	.10510	.01480	.10600	.00610	.01130
	2.5	.18780	.16350	.25130	.01660	.14380	.00910	.14760	.00210	.00520
	3.0	.24830	.21420	.30900	.01770	.19190	.00780	.20130	.00180	.00370
	3.5	.29550	.25060	.35880	.01200	.23040	.00490	.24100	.00060	.00260
	4.0	.35890	.29710	.42030	.01390	.28200	.00500	.29650	.00050	.00220
	0.5	.02550	.02630	.02000	.14000	.21930	.13200	.22440	.30790	.12890
	1.0	.05190	.05210	.05290	.04920	.04980	.05110	.05170	.04940	.04990
	1.5	.14660	.13670	.15080	.03640	.09540	.03010	.09940	.01860	.02640
10	2.0	.27720	.25610	.28820	.03380	.18320	.02350	.19730	.01160	.01860
	2.5	.44270	.38660	.45620	.03930	.28830	.02520	.31580	.01070	.02010
	3.0	.57060	.51240	.58150	.04340	.40190	.02510	.44300	.00910	.02130
	3.5	.68410	.61170	.69930	.05060	.50700	.02910	.55280	.00950	.02650
	4.0	.78690	.71410	.79740	.05770	.61790	.03310	.66100	.01080	.03340
	0.5	.07540	.02420	.11360	.19150	.39330	.18710	.41750	.46700	.18120
	1.0	.05060	.05080	.05180	.05020	.05250	.04990	.05150	.05110	.04900
	1.5	.22940	.24520	.20630	.04840	.15020	.04180	.15990	.03110	.03940
20	2.0	.51200	.53290	.47000	.06900	.35700	.05640	.39150	.04630	.05120
	2.5	.75490	.76740	.70760	.10850	.57690	.09050	.62960	.08590	.09340
	3.0	.89950	.90350	.86560	.15990	.77110	.15880	.82330	.16430	.17490
	3.5	.96640	.96450	.94350	.20870	.88570	.23780	.92040	.26020	.26760
	4.0	.98920	.98830	.98000	.28590	.94570	.34400	.96610	.38690	.39300
	0.5	.24940	.05470	.32340	.25330	.54880	.25080	.58530	.60960	.24800
	1.0	.04640	.04630	.11840	.05080	.04850	.05130	.04760	.05040	.05160
	1.5	.28370	.32400	.47760	.05580	.19820	.04620	.21530	.04080	.04250
30	2.0	.65500	.69630	.82650	.10610	.50920	.09060	.56610	.10710	.09490
	2.5	.89570	.92180	.96960	.19280	.79630	.20580	.85270	.27340	.23290
	3.0	.97610	.98150	.99390	.31560	.93130	.38390	.95730	.50330	.43140
	3.5	.99650	.99740	.99970	.45210	.98210	.57700	.99330	.71020	.63240
	4.0	.99940	.99980	1.00000	.59040	.99710	.74620	.99940	.85460	.79050

n		TV	TE	TC	D	V	W^2	U^2	A^2	M
	0.5	.37030	.06660	.39850	.33150	.68120	.33320	.70870	.72980	.33670
	1.0	.05100	.04820	.04940	.05490	.05090	.05540	.05060	.05460	.05540
	1.5	.36460	.43460	.33330	.07500	.25800	.05820	.28370	.06020	.05660
40	2.0	.80910	.85440	.77030	.16220	.66550	.15660	.71870	.22190	.17430
	2.5	.96910	.98160	.95340	.32340	.90980	.38380	.94230	.52170	.43270
	3.0	.99750	.99880	.99580	.52540	.98710	.65610	.99350	.80780	.70740
	3.5	1.00000	.99980	.99960	.69980	.99810	.85080	.99900	.94290	.88200
	4.0	1.00000	.99990	.99990	.83960	.99950	.94710	.99970	.98590	.96080
	0.5	.55160	.12620	.55860	.40530	.79020	.42230	.81680	.82950	.43510
	1.0	.04620	.04610	.04400	.04770	.05090	.04730	.05070	.04660	.04790
	1.5	.41880	.50120	.38170	.08580	.32220	.06690	.34880	.08240	.06610
50	2.0	.87440	.91230	.84440	.22160	.78020	.23810	.82850	.36280	.26850
	2.5	.99090	.99480	.98300	.46500	.97030	.57860	.98450	.76340	.62720
	3.0	.99930	.99990	.99880	.71110	.99770	.84710	.99880	.95150	.87660
	3.5	1.00000	1.00000	1.00000	.87820	1.00000	.96470	1.00000	.99430	.97420
	4.0	1.00000	1.00000	1.00000	.96090	1.00000	.99400	1.00000	.99960	.99610

참고문헌

1. Ahmad, I. A. and Lin, P. E. (1976). A Nonparametric estimation of the entropy for absolutely continuous distribution. *IEEE Transactions Inference and Theory*, 22 372-375.

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

3. Dmitriev, Y. G. and Tarasenko, F. P. (1973). On the estimation of functional of the probability density and its derivatives. *Theory of Probability and Application,* 18, 628-633.

4. Dudewicz, E. J. and van Der Meulen, E. C. (1981). Entropy-based tests of uniformity. *Journal of American Statistical Association,* 76, 967-974

5. Grezegorzewski, P. and Wieczorkowski, R. (1999). Entropy-based goodness-of-fit test for exponentiality, *Communication in Statistics-Theory and Methed,* 28, 1183-1202.

6. Mack, S. (1988). *A comparative study of entropy estimator and entropy based goodness-of-fit tests.* PhD Dissertation. University of California.

7. Parzen, E. (1982). Maximum entropy interpretation of autoregressive spectral densities. *Statistical and Probability Letters,* 1, 2-6.

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

9. van Es, B. (1992). Estimating functional related to a density by a class of statistics based on spacing. *Scandinavian Journal of Statisitcs,* 19, 61-72.

Goodness-of-fit tests based on Entropy Estimators

Jong-Tae Kim [6] · **Young-Jun Cha** [7] · **Young-Hun Kim** [8]·
Jae-Man Lee [9]· **Sang-Kil Kang** [10]

Abstract

The goal of this paper is to study of the entropy - based on goodness-of-fit tests with several parametric models. It is also to study the relationship of the Moran's test statistic on the view of entropy.

Key Words and Phrases: Goodness-of-fit test, entropy.

[6]Associated Professor, Dept. of Statistics, Taegu University, Kyungbuk 712-714
[7]Professor, Dept. of Statistics, Andong University, Kyungbuk 760-749
[8]Professor, Dept. of Statistics, Andong University, Kyungbuk 760-749
[9]Associated Professor, Dept. of Statistics, Andong University, Kyungbuk 760-749
[10]Lecturer, Dept. of Statistics, Kyungpook University, Taegu 702-701