임의중단모형에서 수명의 엔트로피에 대한 Buckley-James형 추정량

(Buckley-James Type Estimators for Entropy of Lifetimes under Random Censorship Model)

이재만* 차영준* 이우동** 김종태***
(Jae-Man Lee) (Young-Joon Cha) (Woo-Dong Lee) (Jong-Tae Kim)

요 약 이 논문은 임의중단 표본을 이용하여 entropy에 대한 두 개의 버클리-제임스(Buckley-James)형 추정량을 제안한다. 제안된 추정량의 소표본 특징을 조사하기 위하여 컴퓨터 모의실험을 실시하였다. 이 때 사용된 분포는 증가위험률, 상수위험률, 감소위험율을 갖는 분포를 가정하였고, 실제의 자료를 이용하여 분석하는 예를 보인다.

Abstract In this paper, we propose two Buckley-James type nonparametric estimators for entropy of lifetimes under random censorship model. We investigate the small sample behaviors of the proposed estimators when the underlying distribution has decreasing failure rate, constant failure rate, and increasing failure rate. Also some examples are illustrated for analysing data

1. 서 론

확률변수의 엔트로피 (entropy)는 Shannon (1948)에 의해 소개된 이래로 정보이론, 통신, 패턴인식, 통계물리 등의 분야의 기초적인 개념 중에 하나이다. 통계학의 관점에서 Shannon의 엔트로피는 산포도의 일종으로 이용된다. Vasicek (1976)은 분산이 동일한 다른 어떤 분포보다 정규분포의 엔트로피가 최대가 된다는 사실을 이이용하여 정규분포에 대한 적합도 검정을 제시하였다. Vasicek은 이를 위하여 엔트로피에 대한 추정량을 제안하고, 추정량의 약일치성을 밝혔으며, 모의실험을 통하여 제안된 추정량을 기초한 추정량이 검정력 측면에서 기존의 다른 추정량보다 우수함을 밝혔다. van Es(1992)는 엔트로피에 대한 새로운 추정량을 제안하고, 제안된 추정량의 강일치성과 점근적 정규성을 밝혔다. Correa(1995)는 회귀방법을 이용한 추정량을 제안하고, 모의실험을 통하여 제안된 방법이 Vasicek의 추정량보다 평균제곱오차 측면에서 더 우수하다는 것을 보였다. Grezegorzewski와 Wieczorkowski(1999)은 수명의 분포 중에 지수분포의 엔트로피가 최대가 된다는 사실을 바탕으로 한 지수성 검정법을 제시하였다.

이러한 연구에서는 모든 자료값이 관측도중에 중단됨이 없이 완전히 관측되었다는 가정 하에서 얻어진 완전한 자료(complete data)를 이용한 엔트로피에 대한 추정문제를 다루고 있다. 그러나 신뢰성분석 분야에서 이용되는 수명자료는 개별 관측치의 관측이 임의로 중단되는 임의중단모형(random censorship model)하에서 얻어지는 자료가 빈번하게 나타난다. 따라서 신뢰성분석 분야에서의 적합도 검정에 엔트로피를 이용하기 위하여 이러한 임의중단자료(randomly censored data)를 이용한 엔트로피의 추정량이 요구된다. 이러한 관점에서 Cha, Lee, Lee와 Kim (1999)은 Kaplan과 Meier(1958)의 추정량을 기초로 하여 엔트로피에 대한 Vasicek과 Correa의 추정량을 임의중단모형으로 확장하여 제안하고 제안된 추정량의 소표본 특성을 모의실험을 통하여 분석하였다.

한편 Buckley와 James(1979)는 임의중단자료의 선형회귀에서 관측이 중단된 자료값을 관측중단시점이 주어진 반

* 안동대학교 자연과학대학 통계학과
** 경산대학교 자연과학대학 정보과학부
*** 대구대학교 자연과학대학 정보과학부

응변수의 조건부 기대값으로 대치한 변형된 자료를 이용하여 회귀모수의 추정량을 제안하였다.

본 연구에서는 Buckley와 James의 방법을 이용하여 엔트로피에 대한 Vasicek과 Correa의 추정량을 임의중단모형으로 확장하고 제안된 추정량의 특성을 Kaplan과 Meier의 방법으로 임의중단모형으로 확장된 Vasicek과 Correa의 추정량과 비교 · 분석한다.

2. 엔트로피의 Buckley-James형 추정량

분포함수가 F이고, 확률밀도함수가 f인 수명 X에 대한 Shanon의 엔트로피는 다음과 같이 정의된다.

$$H(f) = -\int_0^\infty f(x)\log f(x)dx = \int_0^1 \log\{\frac{d}{dp}F^{-1}(p)\}dp \quad (2.1)$$

여기서 F^{-1}은 연속인 확률변수 X에 대한 분위수함수(quantile function)이다. 즉

$$F^{-1}(p) = \inf\{x|F(x)\geq p\}, 0\leq p\leq 1. \quad (2.2)$$

이 때 평균이 $1/\lambda$인 지수분포를 하는 수명 X의 엔트로피 $H(f)$는 다음과 같은 성질을 갖는다.

$$H(f)\leq 1-\ln\lambda,$$

등호는 수명 X의 분포가 평균이 $1/\lambda$인 지수분포를 따를 때 성립한다.

임의중단모형으로 관측되는 크기 n인 수명자료는 다음과 같이 나타낼 수 있다.

$$(Z_i, \delta_i), i=1,2,\cdots,n$$

여기서, $Z_i = \min(X_i, C_i) = (X_i \wedge C_i)$,

$$\delta_i = \begin{cases} 1, & \text{if } X_i\leq C_i \\ 0, & \text{if } X_i > C_i \end{cases}$$

이고, $X_1, X_2, \cdots, X_n$은 연속인 분포함수 F를 갖는 수명이고, $C_1, C_2, \cdots, C_n$은 분포함수 G를 갖는 중단시간이며, X_i와 C_i는 독립이다. δ_i는 관측된 수명 Z_i의 관측중단여부를 알려주는 중단지시함수(censoring indicator)이다.

임의중단모형에서 관측된 수명자료 (Z_i, δ_i) $, i=1,2,\cdots,n$를 이용하여 엔트로피 $H(f)$의 추정량을 구성하기 위하여 먼저 식 (2.1)의 분위수함수 $F^{-1}(p)$의 추정량이 구성되어야 하고, 이를 위하여 분포함수 F의 추정량이 요구된다.

임의중단자료 $(Z_i, \delta_i), i=1,2,\cdots,n$를 이용한 분포함수 F의 Kaplan과 Meier(1958) 추정량은 다음과 같다.

$$\hat{F}_{KM}(t) = \begin{cases} 1-\prod_{i:Z_{(i)}\leq t}(1-\frac{1}{n-i+1})^{\delta(i)}, & t\leq Z_{(n)} \\ 1, & t> Z_{(n)} \end{cases}$$

여기에서 $Z_{(1)}, Z_{(2)}, \cdots, Z_{(n)}$은 $Z_1, Z_2, \cdots, Z_n$에 대한 순서 통계량(order statistic)이며, $\delta_{(i)}$는 $Z_{(i)}$에 대응되는 중단지시함수이다.

Kaplan과 Meier의 추정량과는 다른 관점에서 Buckley와 James는 관측중단이 일어난 자료값을 관측중단시점보다 크다는 조건이 주어진 수명 X_i의 조건부 기대값으로 대치하여 이용할 것을 제안하였다. 이는 임의중단자료 (Z_i, δ_i) , $i=1,2,\cdots,n$를 다음과 같이 정의되는 가상자료 $(Z_1^*,\cdots,Z_n^*)$로 변환하여 추정에 이용하자는 것이다.

$$Z_i^* = X_i\delta_i + [C_i + \frac{\int_{C_i}^\infty(1-\hat{F}_{KM}(t))dt}{1-\hat{F}_{KM}(C_i)}](1-\delta_i)$$

이와 같이 변환된 수명자료 $(Z_1^*,\cdots,Z_n^*)$의 경험분포함수(empirical distribution function) 를 분포함수 F의 Buckley와 James 형 추정량 $\hat{F}_{BJ}(x)$라고 하자. 즉

$$\hat{F}_{BJ}(x) = \frac{1}{n}\sum_{i=1}^n I(Z_i^*\leq x). \quad (2.3)$$

이제 엔트로피 $H(f)$의 추정량을 구성하기 위하여 먼저 식 (2.1)의 분위수함수 $F^{-1}(p)$의 추정량 $\hat{F}^{-1}(p)$을 $\hat{F}_{BJ}(x)$을 이용하여 다음과 같이 구성할 수 있다.

$$\hat{F}^{-1}(p) = \inf\{x|\ \hat{F}_{BJ}(x)\geq p\}, 0\leq p\leq 1. \quad (2.4)$$

이를 이용하여 (2.1)식의 $\frac{d}{dp}F^{-1}(p)$를 추정하기 위하여 변환된 수명자료 $(Z_1^*,\cdots,Z_n^*)$에서 서로 다른 자료값을 $(Y_1,\cdots,Y_\nu)$라고 하고, $(Y_{(1)},\cdots,Y_{(\nu)})$는 순서통계량이라 하자. 그러면 Vasicek과 같은 방법으로 $\frac{d}{dp}F^{-1}(p)$에 대한 추정량은 두 개의 점, $(\hat{F}_{BJ}(Y_{(i+m)}), Y_{(i+m)})$과 $(\hat{F}_{BJ}(Y_{(i-m)}), Y_{(i-m)})$을 통과하는 직선의 기울기를 사용하여 $\hat{F}_{BJ}(Y_{(i)})\leq p< \hat{F}_{BJ}(Y_{(i+1)})$일 때,

$$\frac{Y_{(i+m)}-Y_{(i-m)}}{\hat{F}_{BJ}(Y_{(i+m)})-\hat{F}_{BJ}(Y_{(i-m)})}$$

와 같이 생각할 수 있다. 여기에서 $m(<\nu/2)$은 미리 고정된 양의 정수로 윈도의 크기이다. 이를 이용하여 엔트로피 H의 추정량을 다음과 같이 구성할 수 있다.

$$\hat{H}_{VC}= \tag{2.5}$$

$$\frac{1}{\nu}\sum_{i=1}^{\nu}\log\{\frac{Y_{(i+m)}-Y_{(i-m)}}{\hat{F}_{BJ}(Y_{(i+m)})-\hat{F}_{BJ}(Y_{(i-m)})}\}$$

한편, Correa의 방법과 같이 두 개의 점만을 이용하는 것이 아니라, $Y_{(i-m)}$과 $Y_{(i+m)}$사이에 있는 $2m+1$개의 점을 이용하여, 국소적 (locall -y)으로

$$F(Y_{(j)})=\alpha+\beta Y_{(j)}+\varepsilon, j=i-m,\cdots,i+m$$

가 만족된다고 가정하면, 기울기 β에 대한 최소제곱추정량을 얻을 수 있고, $\frac{d}{dp}F^{-1}(p)$에 대한 추정량으로 이용할 수 있다. 이를 이용하여 엔트로피에 대한 추정량을 구성하면 다음과 같다.

$$\hat{H}_C=-\frac{1}{\nu}\sum_{i=1}^{\nu}\log\{b_i\}, \tag{2.6}$$

여기서

$$b_i=\frac{\sum_{j=i-m}^{i+m}(Y_{(j)}-\overline{Y}_{(i)})(\hat{F}_{BJ}(Y_{(j)})-\overline{\hat{F}}_{(i)})}{\sum_{j=i-m}^{i+m}(Y_{(j)}-\overline{Y}_{(i)})^2}$$

이며,

$$\overline{Y}_{(i)}=\sum_{j=i-m}^{i+m}Y_{(j)}/(2m+1)$$과

$$\overline{\hat{F}}_{(i)}=\sum_{j=i-m}^{i+m}\hat{F}_{BJ}(Y_{(j)})/(2m+1)$$

이다.

추정량 (2.5)과 (2.6)에서 $j<1$인 경우는 $Y_{(j)}=Y_{(1)}$이고, $j>\nu$인 경우는 $Y_{(j)}=Y_{(\nu)}$이다.

제안된 추정량 $\hat{H}_{VC}$과 $\hat{H}_C$의 점근적 분포와 소표본 특성에 대한 해석학적인 분석은 추정량의 복잡한 구성으로 인하여 접근하기가 어렵기 때문에 이 문제의 실질적인 해결 방안의 하나로 몬테칼로 모의실험을 통하여 제곱평균오차와 편의의 관점에서 기존의 추정량과 비교 · 분석한다.

3. 모의실험

이 절에서는 제안된 추정량 $\hat{H}_{VC}$과 $\hat{H}_C$의 소표본 특성을 몬테칼로방법으로 편의와 평균제곱오차의 관점에서 Cha, Lee, Lee, Kim에 의해 (2.5)과 (2.6)식의 $\hat{F}_{BJ}$ 대신 $\hat{F}_{KM}$을 사용한 추정량 $\hat{H}^*_{VC}$와 $\hat{H}^*_C$의 특성과 비교하려고 한다. 모의실험에서 수명분포의 형태, 관측중단률, 윈도의 크기 m을 실험요인으로 고려하기로 하고, 수명분포의 형태는 와이블분포에서 형태모수 α가 1인 경우, 즉 위험률이 상수(constant failure rate)인 경우 , α가 1보다 작은 경우, 즉 위험률이 감소하는(decreasing failure rate) 경우와 α가 1보다 큰 경우, 즉 위험률이 증가하는 (increasing failure rate) 경우로 나누어 고려하기로 하고, 관측중단률은 10%, 20%, 30%로 하되 관측시간의 분포는 지수분포를 이용하기로 하면 각각의 수명분포와 관측중단률에 따른 관측시간의 분포는 <표3.1>과 같다.

<표3.1>의 수명분포와 관측시간분포의 각 조합으로부터 수명 X_i와 관측시간 C_i에 대응되는 크기 $n=30, 50, 100$ 의 난수를 생성하여 엔트로피 H의 각 추정량 $\hat{H}_{VC}$, $\hat{H}_C$, $\hat{H}^*_{VC}$, $\hat{H}^*_C$ 의 값을 각각의 m에서 구하는 과정을 10000회 반복 시행하여 각 추정량의 평균제곱오차와 편의를 추정한 결과 서로 다른 수명분포의 경우, 수명분포가 Weib(5.0, 1.0)인 경우의 결과를 관측중단률에 따라 정리한 <표3.2>, <표3.3>, <표3.4>과 동일한 경향을 보이고 있어 결과의 보고를 생략하였다.

모의실험의결과를 통하여 다음과 같은 사실을 알 수 있었다.

<표3.1> 수명분포와 관측중단률에 따른 관측시간분포

관측 중단률 / 수명분포	10%	20%	30%
Weib(5.0, 0.5)	Exp(0.336)	Exp(0.905)	Exp(1.87)
Weib(5.0, 1.0)	Exp(0.555)	Exp(1.246)	Exp(2.14)
Weib(5.0, 1.5)	Exp(0.598)	Exp(1.302)	Exp(2.149)

단, Weib(λ,α)의 확률밀도함수 $f(x)=\alpha\lambda(\lambda x)^{\alpha-1}e^{-(\lambda x)^{\alpha}}$이고, Exp($\lambda$)의 확률밀도함수 $g(x)=\lambda e^{-\lambda x}$이다.

(1) 일반적으로 예측할 수 있는 바와 같이 모든 추정량에 있어서 n이 증가할수록 편의와 평균제곱오차가 줄어들고, 관측중단률이 증가할수록 편의와 평균제곱오차는 커진다.

(2) 윈도의 크기 m이 적당한 크기까지 증가할 때 편의와 평균제곱오차가 줄어든다. 이때, $\hat{H}_{VC}$과 $\hat{H}_{C}$의 편의와 평균제곱오차가 $\hat{H}^{*}_{VC}$와 $\hat{H}^{*}_{C}$의 편의와 평균제곱오차보다 더 크게 줄어드는 경향을 볼 수 있다.

(3) 실험에서 고려한 모든 수명분포에서 본 연구에서 제안한 $\hat{H}_{VC}$과 $\hat{H}_{C}$의 편의와 평균제곱오차가 각각 $\hat{H}^{*}_{VC}$와 $\hat{H}^{*}_{C}$의 편의와 평균제곱오차보다 작은 경향을 보였다.

(4) (3)의 경향은 관측중단률이 커질수록 더욱 심화됨을 볼 수 있었다.

결론적으로 모의실험의 결과는 실험에서 고려한 모든 경우에 있어서 특히 관측중단률이 높을수록 편의와 평균제곱오차의 측면에서 $\hat{H}_{C}$을 사용하는 것이 바람직함을 알 수 있다.

<표 3.2> 수명분포 Weib(5.0, 1.0)에서 관측중단률이 10 %인 경우 추정량의 편의와 평균제곱오차의 추정치

		편의				평균제곱오차			
n	m	$\hat{H}^{*}_{VC}$	$\hat{H}^{*}_{C}$	$\hat{H}_{VC}$	$\hat{H}_{C}$	$\hat{H}^{*}_{VC}$	$\hat{H}^{*}_{C}$	$\hat{H}_{VC}$	$\hat{H}_{C}$
30	1	-.4222	-.3283	-.3522	-.2568	.2275	.1564	.1733	.1146
	2	-.2717	-.1690	-.2032	-.0994	.1162	.0712	.0846	.0534
	3	-.2200	-.1298	-.1530	-.0619	.0882	.0570	.0643	.0450
	4	-.1871	-.1078	-.1222	-.0420	.0770	.0544	.0580	.0456
	5	-.1647	-.0923	-.1027	-.0295	.0684	.0511	.0528	.0443
50	1	-.4019	-.3049	-.3237	-.2257	.1898	.1209	.1335	.0793
	2	-.2554	-.1521	-.1794	-.0754	.0902	.0481	.0576	.0311
	3	-.2045	-.1163	-.1294	-.0406	.0655	.0376	.0410	.0261
	4	-.1799	-.1044	-.1058	-.0296	.0563	.0353	.0357	.0257
	5	-.1553	-.0878	-.0831	-.0149	.0488	.0330	.0322	.0260
	6	-.1389	-.0774	-.0686	-.0063	.0437	.0311	.0298	.0258
100	1	-.3872	-.2880	-.3032	-.2033	.1635	.0963	.1057	.0549
	2	-.2443	-.1409	-.1622	-.0585	.0723	.0324	.0390	.0161
	3	-.1963	-.1098	-.1143	-.0275	.0501	.0237	.0249	.0126
	4	-.1702	-.0974	-.0886	-.0154	.0403	.0209	.0194	.0119
	5	-.1531	-.0897	-.0727	-.0089	.0352	.0200	.0172	.0121
	6	-.1417	-.0854	-.0619	-.0051	.0313	.0187	.0155	.0119
	7	-.1306	-.0796	-.0522	-.0006	.0287	.0182	.0147	.0122
	8	-.1212	-.0741	-.0439	.0038	.0264	.0175	.0139	.0122

<표 3.3> 수명분포 Weib(5.0, 1.0)에서 관측중단률이 20 % 인 경우
추정량의 편의와 평균제곱오차의 추정치

		편의				평균제곱오차			
n	m	$\widehat{H}^*_{VC}$	$\widehat{H}^*_{C}$	$\hat{H}_{VC}$	$\hat{H}_{C}$	$\widehat{H}^*_{VC}$	$\widehat{H}^*_{C}$	$\hat{H}_{VC}$	$\hat{H}_{C}$
30	1	-.5323	-.4406	-.4006	-.3044	.3371	.2473	.2152	.1467
	2	-.3821	-.2812	-.2567	-.1531	.1920	.1252	.1146	.0721
	3	-.3241	-.2365	-.2018	-.1117	.1498	.1012	.0874	.0595
	4	-.2921	-.2151	-.1749	-.0956	.1298	.0920	.0769	.0564
	5	-.2691	-.1992	-.1570	-.0850	.1169	.0857	.0718	.0558
50	1	-.5080	-.4125	-.3665	-.2676	.2886	.2002	.1669	.1038
	2	-.3587	-.2570	-.2229	-.1195	.1555	.0928	.0779	.0424
	3	-.3099	-.2232	-.1750	-.0868	.1227	.0768	.0592	.0363
	4	-.2793	-.2050	-.1479	-.0721	.1036	.0681	.0494	.0332
	5	-.2555	-.1898	-.1244	-.0568	.0907	.0620	.0426	.0309
	6	-.2408	-.1812	-.1130	-.0514	.0839	.0597	.0403	.0309
100	1	-.4879	-.3895	-.3370	-.2362	.2528	.1663	.1296	.0716
	2	-.3446	-.2416	-.1970	-.0936	.1316	.0712	.0527	.0226
	3	-.2952	-.2094	-.1489	-.0629	.0998	.0566	.0360	.0177
	4	-.2722	-.2000	-.1272	-.0544	.0869	.0530	.0299	.0168
	5	-.2541	-.1916	-.1103	-.0470	.0769	.0493	.0254	.0156
	6	-.2414	-.1861	-.0989	-.0424	.0706	.0473	.0229	.0151
	7	-.2306	-.1806	-.0897	-.0382	.0660	.0458	.0217	.0154
	8	-.2182	-.1721	-.0784	-.0307	.0604	.0427	.0197	.0149

<표 3.4> 수명분포 Weib(5.0, 1.0)에서 관측중단률이 30 % 인 경우
추정량의 편의와 평균제곱오차의 추정치

		편의				평균제곱오차			
n	m	$\hat{H}_{VC}$	$\hat{H}_{C}$	$\hat{R}_{VC}$	$\hat{R}_{C}$	$\hat{H}_{VC}$	$\hat{H}_{C}$	$\hat{R}_{VC}$	$\hat{R}_{C}$
30	1	-.6415	-.5520	-.4706	-.3722	.4691	.3614	.2859	.2018
	2	-.4882	-.3897	-.3251	-.2215	.2884	.2017	.1599	.1031
	3	-.4325	-.3473	-.2723	-.1819	.2356	.1697	.1271	.0868
	4	-.4001	-.3262	-.2484	-.1696	.2082	.1558	.1151	.0833
	5	-.3706	-.3042	-.2264	-.1548	.1855	.1422	.1037	.0780
50	1	-.6136	-.5195	-.4246	-.3237	.4093	.3021	.2174	.1413
	2	-.4651	-.3646	-.2836	-.1800	.2459	.1624	.1139	.0659
	3	-.4171	-.3323	-.2392	-.1511	.2023	.1391	.0892	.0549
	4	-.3826	-.3098	-.2098	-.1337	.1738	.1239	.0760	.0503
	5	-.3613	-.2977	-.1902	-.1228	.1582	.1170	.0676	.0472
	6	-.3464	-.2890	-.1796	-.1176	.1495	.1139	.0653	.0478
100	1	-.5939	-.4962	-.3906	-.2873	.3688	.2620	.1724	.1020
	2	-.4487	-.3466	-.2506	-.1475	.2154	.1342	.0791	.0379
	3	-.3996	-.3148	-.2044	-.1187	.1733	.1127	.0578	.0302
	4	-.3755	-.3045	-.1835	-.1111	.1548	.1067	.0498	.0285
	5	-.3572	-.2959	-.1672	-.1038	.1408	.1010	.0437	.0267
	6	-.3429	-.2889	-.1533	-.0969	.1312	.0972	.0395	.0256
	7	-.3298	-.2813	-.1445	-.0930	.1225	.0932	.0371	.0252
	8	-.3214	-.2769	-.1372	-.0891	.1172	.0909	.0351	.0246

예제 1의 결과				
m	1	2	3	4
$\widehat{H^*}_{VC}$	4.1372	4.2726	4.2517	4.2529
$\widehat{H^*}_{C}$	4.1867	4.3437	4.3077	4.2849
$\widehat{H}_{VC}$	4.2410	4.3507	4.3468	4.3422
$\widehat{H}_{C}$	4.2848	4.3941	4.3782	4.3656

4. 예 제

이 절에서는 Cha, Lee, Lee, Kim에 의해서 제안된 추정량들과의 비교를 위하여 이 들과 동일한 예제를 통하여 본 연구에서 제안한 추정량 $\widehat{H}_{VC}$과 $\widehat{H}_{C}$의 추정치가 모수모형의 최대우도추정치와 비교해보았다.

예제1. 2, 72*, 51, 60*, 33, 27, 14, 24, 4, 21*는 10개의 기계에 대한 수명을 관측한 자료이다 (Bartholomew, 1957). 여기서 '*'표시는 관측이 중단된 자료값이다. 이때 평균이 $\theta(=1/\lambda)$인 지수분포라는 가정 하에서 평균 θ에 대한 최대우도추정치 $\hat{\theta}=44.0$ 이다. 지수분포의 엔트로피는 평균이 $1/\lambda$인 경우, $H=1-\log\lambda$이고, 엔트로피의 최대우도 추정치는 $\widehat{H}_{MLE}=1+\log 44=4.7842$ 이다.

엔트로피의 Buckley와 James 형 추정량의 값을 계산하기 위하여 자료를 Buckley와 James의 방법으로 변환한 가상자료는 2. 72. 51. 60. 33. 27. 14. 24. 4. 46.5 이다.

m에 따른 엔트로피의 추정치는 예제1의 결과와 같다.

이 결과에서도 알 수 있듯이 근소한 차이가 있으나, $m=2$인 경우에서 $\widehat{H}_{C}$가 가장 $\widehat{H}_{MLE}$에 가깝다는 것을 알 수 있다.

예제 2. 수명자료를 로그변환하면 변환된 자료가 정규분포를 하는 경우가 많다. 이러한 자료를 로그정규분포를 따르는 자료라고 한다. 다음의 자료는 로그로 변환된 자료가 정규분포를 하며 평균이 0이고 분산이 4 경우를 가정하여 인위적으로 추출한 15개의 자료에 중단비율이 50%정도가 되도록 하여 관측하였다.

-3.926252, -3.902384, -3.329395, -2.958726*, -2.769049*,
-2.7482*, -1.6680*, -1.5574*,
-1.347179*, -0.4919527, -0.1194435*, 0.006245852*,
0.5234513*, 1.750694*, 2.216680

위의 표본 자료에 대한 모집단인 평균이 0이고 분산이 4인 정규분포의 엔트로피 $H=\log(\sqrt{2\pi}\sigma)+1/2=2.112085$이다. m에 따른 엔트로피 H의 각 추정치는 다음과 같다.

이 결과에서 $m=5$일 때, 추정량 $\widehat{H}_{C}$의 값이 참값에 가장 가깝다는 것을 알 수 있다.

예제 2의 결과					
m	1	2	3	4	5
$\widehat{H^*}_{VC}$	1.5512	1.9170	1.9711	1.9613	1.9294
$\widehat{H^*}_{C}$	1.5849	1.9483	1.9803	1.9712	1.9487
$\widehat{H}_{VC}$	1.5791	1.8829	1.9482	1.9895	2.0529
$\widehat{H}_{C}$	1.6566	1.9465	2.0115	2.0588	2.1164

참 고 문 헌

[1] Batholomew, D. J. (1957). A Problem in life testing, *Journal of the American Statistical Association*, 52, 350-355.

[2] Buckley, J. and James. I.(1979). Linear regression with censored data. *Biom -etrika*, 66, 89-99.

[3] Correa, J. C. (1995). A New Estimator of Entropy, *Communications in Statisti -cs-Theory and Methods*, 24(10), 2439-2449.

[4] Grezegorzewski, P. and Wieczorkowski, R. (1999) Entropy-based Goodness-of -Fit Test for Exponentiality, *Communi -cations in Statistics-Theory and Method*, 28(5), 1183-1202.

[5] Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from incompl -ete observations, *Journal of the Amer -ican Statistical Association*, 53, 457 -481.

[6] Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Inc.

[7] 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.

[8] Vasicek, O. (1976). A Test for Normality Based on Sample Entropy, *Journal of Royal Statistical Society*, B. 38, 54-59.

[9] Young Joon Cha, Jae Man Lee, Woo Dong Lee and Jong Tae Kim (1999). A Comparative Study of Nonparametric Estimators for Entropy under Random Censorship Model, *The Korean Commu -nication in Statistics* 6(3), 771-779.

이 재 만

경북대학교 문리과대학 통계학과 이학사

경북대학교 대학원 통계학과 이학석사

경북대학교 대학원 통계학과 이학박사

현재 안동대학교 자연과학대학 통계학과 교수

관심분야 : 생존분석/신뢰성분석/ 통계적 품질관리

차 영 준

1980 경북대학교 문리과대학 통계학과 이학사

1982 경북대학교 대학원 통계학과 이학석사

1990 경북대학교 대학원 통계학과 이학박사

현재 안동대학교 자연과학대학 통계학과 교수

관심분야 : biostatistics, 통계소프트웨어

이 우 동

1985년 경북대학교 자연과학대학 통계학과(학사)

1988년 경북대학교 대학원 통계학과 (석사)

1993년 경북대학교 대학원 통계학과 (박사)

1995년 ~ 현재 경산대학교 자연과학대학 정보과학부 조교수

관심분야 : 베이즈 추론, 비모수통계

김 종 태

1985년 경북대학교 자연과학대학 통계학과(이학사)

1987년 경북대학교 대학원 통계학과 (이학석사)

1992년 Texas A&M University(통계학 박사)

현재 대구대학교 자연과학대학 통계학과 부교수

관심분야 :