응용통계연구 (2012)
25(2), 311–319
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.311

척도모수가 미지인 임의중도절단자료의 EDF 통계량을 이용한 지수 검정

김남현[1]

[1]홍익대학교 기초과학과

(2012년 1월 2일 접수, 2012년 2월 29일 수정, 2012년 3월 9일 채택)

요 약

수명시간 분석에서 가장 간단하고 또한 자주 이용되는 분포는 지수분포이다. Koziol과 Green (1976)은 Cramér-von Mises 통계량을 Kaplan-Meier의 product limit 경험분포함수를 이용하여 임의중도절단자료에 대해서 일반화하였다. 그러나 이 통계량은 모수의 값이 주어진 단순귀무가설을 가정하고 있으므로 실제 자료에 적용하기에는 어려운 점이 있다. 본 논문에서는 척도모수가 미지인 지수분포의 적합도 검정에 모수를 추정하여 Koziol-Green 통계량을 적용하였다. 그리고 같은 방법으로, 전통적인 Kolmogorov-Smirnov 검정통계량을 일반화하고 두 가지 통계량의 검정력을 모의실험을 통하여 비교하였다. 그 결과 전반적으로 일반화된 Koziol-Green 통계량이 Kolmogorov-Smirnov 통계량보다 지수분포의 검정에 있어서는 좀 더 좋은 검정력을 보여주었다.

주요용어: 적합도검정, 임의중도절단, Cramér-von Mises 통계량, Kolmogorov-Smirnov 통계량, Kaplan-Meier 추정량.

1. 서론

분포에 대한 적합도 검정은 오랫동안 통계적 추론의 주요 관심사 중의 하나였다. 이를 위한 통계적 방법은 일반적으로 그래프를 이용하는 방법, 왜도나 첨도 등의 적률을 이용하는 방법, 카이제곱 검정통계량을 이용하는 방법, 경험분포함수(empirical distribution function; EDF)에 근거한 방법, 회귀나 상관계수에 근거한 방법 등을 들 수 있다. 임의중도절단자료에 대해서도 위의 방법들을 일반화하기 위한 연구가 진행되어 왔다.

Akritas (1988), Hollander와 Peña (1992)는 카이제곱 검정통계량을 임의중도절단자료로 일반화하였다. Koziol과 Green (1976), Koziol (1980)은 EDF 또는 가중경험과정(weighted empirical process)에 기반한 통계량을 임의중도절단자료로 확장하였다. Nair (1981)도 가중경험과정에 기반한 통계량을 제안하고 이를 P-P 플롯이나 Q-Q 플롯에 적용하는 방법에 대해 연구하였다. Chen (1984)은 상관계수 통계량을 임의중도절단자료에 대해서 일반화하였다.

생존분석에서 가장 간단하고 중요한 분포는 지수분포이다. 따라서 중도절단자료에 대한 지수검정도 오랫동안 연구된 분야 중 하나이다. 그러나 제 1종 중도절단(type I censoring)이나 제 2종 중도절단(type II censoring)에 비해 임의중도절단(random censoring)에 대해서는 상대적으로 연구결과가 많지 않은 편이다. 이는 제 1종이나 2종 중도절단이 이론적으로 좀 더 다루기가 편리하고, 임의중도절단의 경우

이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2009-0072563).
[1](121-791) 서울시 마포구 상수동 72-1, 홍익대학교 기초과학과, 교수. E-mail: nhkim@hongik.ac.kr

에는 중도절단분포에 대한 추가적인 가정이나 모형이 필요하기 때문이라고 생각된다. Koziol과 Green (1976)은 특별한 중도절단모형을 가정하고 EDF에 기반한 Cramér-von Mises 통계량을 임의중도절단자료의 경우로 확장하였다. 그러나 이는 분포의 모수가 주어져 있는 단순귀무가설의 경우에만 이용가능하다. 본 논문에서는 척도모수가 미지인 지수분포의 검정을 위하여 우선 모수의 추정에 대해서 생각해보고 이를 이용하여 Koziol-Green 통계량과 잘 알려진 Kolmogorov-Smirnov 통계량을 일반화하고 이를 비교하였다.

2절에서 모수추정의 과정과 고려한 통계량을 소개한다. 3절에서는 모의실험 결과를, 4절에서는 결론과 함께 좀 더 생각해 보아야 할 문제를 간략히 언급한다.

2. 검정 통계량

우선 Koziol과 Green (1976)의 통계량을 간단히 소개하기로 하자. 이는 product-limit 경험분포함수를 이용하여 Cramér-von Mises 통계량을 임의중도절단자료로 일반화한 것이다. Cramér-von Mises 통계량은 EDF에 근거한 검정통계량 중 가장 널리 쓰이는 것 중 하나이다. 이에 대한 자세한 사항은 Stephens (1986)을 참고로 한다. 중도절단이 없는 경우 이 통계량의 분포에 대해서는 Csörgő와 Faraway (1996)에서 연구하였다. Koziol과 Green (1976)은 특정한 중도절단분포를 가정하고 제안한 통계량의 근사분포를 구하였다. 또한 Koziol-Green 통계량은 모수를 포함하여 분포의 형태가 완전히 주어져 있는 단순귀무가설에서 정의되었으므로 확률적분변환을 이용하면 연속분포의 경우에는 균일분포의 검정으로 바꾸어 이용할 수 있다. 그러나 실제적인 경우에는 가정한 분포의 모수가 미지인 복합귀무가설을 검정하는 경우가 대부분이므로 그 자체로는 실용성이 떨어진다고 볼 수 있다.

본 논문에서는 Koziol-Green 통계량을 척도모수가 미지인 복합귀무가설에서의 지수검정을 위한 형태로 일반화한다. 그리고 전통적인 Kolmogorov-Smirnov 통계량에도 같은 방법을 적용하여 일반화하고 두 통계량을 비교한다. 먼저 Koziol-Green 통계량의 구체적인 형태를 알아보자.

$Z_1^0, \ldots, Z_n^0$는 연속확률분포 F^0에서의 확률표본이고, $T_1, \ldots, T_n$은 $Z_1^0, \ldots, Z_n^0$에 독립인, 연속분포 G에서의 중도절단 확률변수라고 하자. Koziol과 Green (1976)은 $1-G = (1-F^0)^\beta$, β는 양수로 가정하였다. Z_i^0는 T_i에 의해서 우측 중도절단(right censored)되며, 따라서 관측되는 자료는 (Z_i, δ_i), 또는 $(Z_{(i)}, \delta_{(i)})$, $i = 1, \ldots, n$이다. 여기서 $Z_i = \min(Z_i^0, T_i)$이고

$$\delta_i = \begin{cases} 1, & \text{if } Z_i = Z_i^0, \\ 0, & \text{if } Z_i = T_i \end{cases} \tag{2.1}$$

이다. $Z_{(1)} \le \cdots \le Z_{(n)}$은 $Z_1, \ldots, Z_n$의 순서통계량이며 $\delta_{(i)}$도 $Z_{(i)}$에 대해서 식 (2.1)과 유사하게 정의된다. 즉, $\delta_{(i)}$는 i번째 순서통계량이 중도절단자료가 아닐때 1을 갖는다. 검정하는 가설은

$$H_0 : F^0 = F^* \tag{2.2}$$

이며 F^*는 미지의 모수없이 형태가 완전히 주어진 연속분포이다. 따라서 확률적분변환을 이용하면 일반성을 잃지않고 식 (2.2)의 F^*는 균일분포 $U(0,1)$로 생각할 수 있다. 즉, 가설 (2.2)는 $Z_1^0, \ldots, Z_n^0$가 균일분포 $U(0,1)$을 따르는지 검정하는 것과 동일하다. 이러한 임의중도절단자료의 경우 F^0를 추정하기 위해서 일반적인 경험분포함수를 이용하기는 곤란하다. 대신 이를 product-limit 추정량 $\hat{F_n}^0$,

$$1 - \hat{F}_n^0(t) = \begin{cases} 1, & t < Z_{(1)}, \\ \displaystyle\prod_{Z_{(j)} \le t} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, & t < Z_{(n)}, \\ 0, & t \ge Z_{(n)} \end{cases} \tag{2.3}$$

으로 추정한다. 식 (2.3)의 product-limit 추정량은 Kaplan과 Meier (1958), Efron (1967), Meier (1975), Breslow와 Crowley (1974) 등에서 연구되었고, 보통 Kaplan-Meier 추정량이라고 부른다.

식 (2.3)을 이용하면 가장 잘 알려진 EDF에 기반한 적합도 검정통계량인 Kolmogorov-Smirnov 통계량도 임의중도절단자료로 쉽게 확장할 수 있다. 즉, $Z_{(n+1)} = 1$,

$$\hat{p}_i = 1 - \prod_{j \le i} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \quad \hat{p}_0 = 0,\ \hat{p}_{(n+1)} = 1$$

이라고 할 때,

$$D_n = \sup_t \left| \hat{F}_n^0(t) - t \right| = \max_{1 \le i \le n+1,\, \delta_{(i)}=1} \left\{ \hat{p}_i - Z_{(i)}, Z_{(i)} - \hat{p}_{i-1} \right\} \tag{2.4}$$

으로 정의할 수 있다. Koziol (1980)은 가중경험과정(weighted empirical process)에 기반하여 식 (2.4)와 유사한 형태의 통계량을 고려하였다.

Koziol-Green 통계량은

$$\psi_n^2 = n \int_0^1 \left(\hat{F}_n^0(t) - t \right)^2 dt \tag{2.5}$$

로 $\hat{F}_n^0$와 $U(0,1)$과의 차이를 보는 것이다. 식 (2.5)는 $Z_{(0)} = 0$, $Z_{(n+1)} = 1$이라고 정의하면

$$\psi_n^2 = n \sum_{j=1, \delta_{(j)}=1}^{n+1} \hat{F}_n^0 \left(Z_{(j-1)} \right) \left(Z_{(j)} - Z_{(j-1)} \right) \left\{ \hat{F}_n^0 \left(Z_{(j-1)} \right) - \left(Z_{(j)} + Z_{(j-1)} \right) \right\} + \frac{1}{3} n \tag{2.6}$$

이다 (Koziol과 Green, 1976).

식 (2.3)의 $\hat{F}_n^0(t)$는 좀 더 일반적으로

$$\hat{F}_{n,c}^0(t) = 1 - \frac{n-c+1}{n-2c+1} \prod_{Z_{(j)} \le t} \left(\frac{n-j-c+1}{n-j-c+2} \right)^{\delta_{(j)}}, \quad 0 \le c \le 1 \tag{2.7}$$

로 변형되어 사용되기도 한다 (Michael과 Schucany, 1986). 이는 또한

$$\hat{p}_{c,i} = 1 - \frac{n-c+1}{n-2c+1} \prod_{j \le i} \left(\frac{n-j-c+1}{n-j-c+2} \right)^{\delta_{(j)}}, \quad 0 \le c \le 1,\ i = 1, \ldots, n \tag{2.8}$$

으로 표현하기도 한다. 식 (2.8)은 중도절단이 없는 완전표본(complete sample)인 경우에는 $(i-c)/(n-2c+1)$가 됨을 알 수 있다. $c = 0$ 또는 $c = 0.5$가 자주 쓰이는 값이기는 하나, $c = 0.3175$일 때, 균일분포의 순서통계량의 중앙값과 가까우므로 (Filliben, 1975), 중도절단 자료에서도 $c = 0.3175$을 사용하기로 하자. 식 (2.6)에서 식 (2.3) 대신 식 (2.7)의 $c = c^* = 0.3175$를 이용한 통계량을 ψ_{n,c^*}^2이라고 하자.

본 논문에서는 연속확률분포 F^0에서의 확률표본 $X_1^0, \ldots, X_n^0$가 지수분포 $\mathcal{E}(\lambda)$를 따르는지를 검정하고자 한다. 즉, 가설 (2.2)에서 $F^*(x) = 1 - e^{\lambda x}$, $x > 0$일 때이다. 이 경우 중도절단 확률변수를 $C_1, \ldots, C_n$, 관측자료를 $X_i = \min(X_i^0, C_i)$라고 할 때, δ_i는 X_i, X_i^0, C_i에 대해서 식 (2.1)과 마찬가지로 정의할 수 있다. $C_1, \ldots, C_n$의 분포 H도 위와 같이 $\beta > 0$에 대해서

$$1 - H = (1 - F^0)^\beta \tag{2.9}$$

라고 가정한다. Csörgő와 Horváth (1981)는 식 (2.9)를 Koziol-Green 모형이라고 불렀다. 본 논문에서도 이와 같은 명칭을 사용하자. 만일 모수 λ가 주어진 경우이면, $Z_i^0 = F^*(X_i^0) = 1 - e^{\lambda X_i^0}$, $Z_i = F^*(X_i) = \min(F^*(X_i^0), F^*(C_i))$로 하고 식 (2.4)나 식 (2.6)의 검정통계량을 이용하면 된다. Chen 등 (1982)에서는 Koziol-Green 모형의 특징과 그에 대한 실제적인 설명을 하고 있다. Koziol (1978)에 의하면 식 (2.5)의 ψ_n^2은 Koziol-Green 모형에 대해서 로버스트(robust)하며, Chen (1984)도 제안한 통계량이 Koziol-Green 모형에 대해서 점근적으로 로버스트(asymptotically robust)함을 언급하고 있다. 이로 미루어보아 식 (2.9)의 모형이 적합도 검정에 미치는 영향이 그리 강하지 않을 수도 있으나, 자료에 대해서 이 모형이 타당한지는 먼저 생각하고 검증해 보아야 할 문제이다.

실제 상황에서는 대부분

$$H_0 : X_1^0, \ldots, X_n^0\text{의 분포는 미지의 모수 } \lambda\text{에 대해서 지수분포 } \mathcal{E}(\lambda)\text{를 따른다}$$

와 같이 복합귀무가설인 경우가 대부분이다. 이러한 경우에는 식 (2.6)의 검정통계량을 이용하기가 곤란하다. 복합귀무가설의 경우에는 우선 모수 λ를 추정해야 하므로, 모수 λ의 추정에 대해서 간단히 생각해보자.

X_i^0, C_i의 확률밀도함수를 각각 f_{X^0}, f_C, 생존함수(survival function)를 각각 $\bar{F}_{X^0}$, $\bar{F}_C$라고 하면, (X_i, δ_i), $X_i = \min(X_i^0, C_i)$의 우도함수(likelihood function)는

$$\prod_{i=1}^{n} \left(f_{X^0}(x_i)\bar{F}_C(x_i)\right)^{\delta_i} \left(f_C(x_i)\bar{F}_{X^0}(x_i)\right)^{1-\delta_i}$$

이다. 따라서 위와 같이 $f_{X^0}(x) = \lambda e^{-\lambda x}$, $x > 0$, $f_C(x) = \lambda\beta e^{-\lambda\beta x}$, $x > 0$인 경우에는 (X_i, δ_i)의 우도함수 $L(\lambda, \beta)$가

$$L(\lambda, \beta) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i} e^{-\lambda\beta x_i}\right)^{\delta_i} \left(\lambda\beta e^{-\lambda\beta x_i} e^{-\lambda x_i}\right)^{1-\delta_i}$$

이고 로그우도함수 $l(\lambda, \beta)$는

$$l(\lambda, \beta) = \sum_{i=1}^{n} [(\delta_i(\ln\lambda - \lambda x_i - \lambda\beta x_i)) + (1 - \delta_i)(\ln(\lambda\beta) - \lambda\beta x_i - \lambda x_i)]$$

이다. 이를 λ와 β에 대해서 각각 편미분하고 0으로 놓으면,

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i = 0,$$
$$\frac{\partial l}{\partial \beta} = \frac{1}{\beta}\sum_{i=1}^{n}(1 - \delta_i) - \lambda \sum_{i=1}^{n} x_i = 0$$

이다. 이를 연립하면 최대우도추정량(maximum likelihood estimator; MLE) $\hat{\lambda}$, $\hat{\beta}$,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}, \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(1 - \delta_i)}{\sum_{i=1}^{n} \delta_i}$$

를 얻는다. $\sum_{i=1}^{n}\delta_i$는 중도절단되지 않은 자료의 수를 의미하므로 이를 k라 하면

$$\hat{\lambda}=\frac{k}{\sum\limits_{i=1}^{n}x_i},\qquad \hat{\beta}=\frac{n-k}{k} \tag{2.10}$$

이다. 실제로 중도절단 확률변수 C_i의 분포가 제 1종 또는 제 2종 중도절단과 같이 λ에 의존하지 않는 경우에도 λ의 최대우도추정량과은 식 (2.10)과 동일함을 알 수 있다 (Lee와 Wang, 2003, Chapter 7). 또한 중도절단분포를 식 (2.9)와 같이 가정하면, 중도절단 관측의 기대비율 γ는

$$\gamma=P(X_i^0>C_i)=\int_{-\infty}^{\infty}\left(1-F^0(x)\right)dH(x)=\int_0^1\beta(1-x)^{\beta}dx=\frac{\beta}{\beta+1}$$

이다. 즉, 식 (2.10)의 β의 최대우도추정량 $\hat{\beta}$으로부터 γ의 추정량은 $\hat{\gamma}=(n-k)/n$으로 전체자료 중 중도절단 관측자료의 비율로 추정된다. 이는 X_i와 δ_i의 주변분포로부터 얻은 적률추정량(moment estimator)과도 일치한다. 다시 말하면 X_i의 분포는 식 (2.9)의 가정과 X_i^0와 C_i의 독립성으로부터

$$P(X_i>x)=P\left(X_i^0>x,C>x\right)=e^{-\lambda(1+\beta)},\quad x>0$$

이므로 $\mathcal{E}(\lambda(1+\beta))$이고 따라서 $E(X_i)=1/(\lambda(1+\beta))$이다. 그리고 δ_i는 성공률이 $p=1/(1+\beta)$인 베르누이 시행이므로 $E(\delta_i)=1/(1+\beta)$이다. 따라서

$$\bar{X}=\frac{1}{\lambda(1+\beta)},\qquad \sum_{i=1}^{n}\delta_i=\frac{n}{1+\beta}$$

으로부터 역시 식 (2.10)의 결과를 얻는다. 물론 이 결과는 중도절단분포에 대한 식 (2.9)의 가정에 의존한다.

식 (2.10)의 $\hat{\lambda}$에 대해서

$$\hat{Z}_i^0=1-e^{-\hat{\lambda}X_i^0},\qquad \hat{Z}_i=\min\left(1-e^{-\hat{\lambda}X_i^0},1-e^{-\hat{\lambda}C_i}\right)$$

라고 하고 $\hat{Z}_{(1)}\le\cdots\le\hat{Z}_{(n)}$을 $\hat{Z}_1,\ldots,\hat{Z}_n$의 순서통계량이라고 하자. $\hat{Z}_{(i)}$에 대해서 식 (2.4)를 계산한 통계량을 $\hat{D}_n$이라고 하고, $\hat{Z}_{(i)}$와 식 (2.8)의 $c=c^*=0.3175$를 이용하여 식 (2.6)을 계산한 통계량을 $\hat{\psi}^2_{n,c^*}$이라고 하자. 즉,

$$\hat{D}_n=\max_{1\le i\le n+1,\,\delta_{(i)}=1}\left\{\hat{p}_i-\hat{Z}_{(i)},\hat{Z}_{(i)}-\hat{p}_{i-1}\right\}, \tag{2.11}$$

$$\hat{\psi}^2_{n,c^*}=n\sum_{j=1,\,\delta_{(j)}=1}^{n+1}\hat{p}_{c^*,j-1}\left(\hat{Z}_{(j)}-\hat{Z}_{(j-1)}\right)\left\{\hat{p}_{c^*j-1}-\left(\hat{Z}_{(j)}+\hat{Z}_{(j-1)}\right)\right\}+\frac{1}{3}n \tag{2.12}$$

이다. 그러면 식 (2.11), (2.12)는 각각 식 (2.4)의 D_n과 식 (2.6)의 ψ^2_n을 복합귀무가설로 일반화한 통계량이라고 할 수 있다. $\hat{D}_n$, $\hat{\psi}^2_{n,c^*}$의 기각값을 모의실험을 통하여 구한 결과를 표 3.2에 제시하였다. 물론 이 통계량들의 분포는 단순귀무가설과 마찬가지로 중도절단분포와 중도절단 모수 β에 의존한다.

3. 모의실험 결과

우선 식 (2.10)의 추정량 λ의 추정의 정도를 모의실험을 통하여 살펴보았다. 표본크기 n은 $n=50,100$, 중도절단모수 β는 $\beta=0.5,1,1.5$, 표본의 수는 10,000으로 하였다. 결과는 표 3.1에 제시

표 **3.1.** λ의 추정과 RMSE

n	β	MLE	RMSE
$n = 50$	$\beta = 0.5$	1.0233	0.1418
$n = 50$	$\beta = 1.0$	1.0229	0.1659
$n = 50$	$\beta = 1.5$	1.0184	0.1800
$n = 100$	$\beta = 0.5$	1.0123	0.0982
$n = 100$	$\beta = 1.0$	1.0101	0.1140
$n = 100$	$\beta = 1.5$	1.0114	0.1270

표 **3.2.** 통계량의 기각값

통계량	유의수준	$n = 50$			$n = 100$		
		$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$	$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$
D_n	0.01	0.27	0.38	0.51	0.19	0.27	0.38
	0.05	0.22	0.31	0.41	0.16	0.22	0.31
	0.10	0.20	0.27	0.36	0.14	0.20	0.28
$\hat{D}_n$	0.01	0.21	0.37	0.51	0.14	0.26	0.39
	0.05	0.17	0.28	0.40	0.12	0.21	0.31
	0.10	0.16	0.25	0.36	0.11	0.18	0.27
ψ^2_{n,c^*}	0.01	0.98	1.58	2.85	0.96	1.44	2.84
	0.05	0.61	0.98	1.68	0.60	0.91	1.69
	0.10	0.47	0.75	1.24	0.46	0.71	1.28
$\hat{\psi}^2_{n,c^*}$	0.01	0.38	1.05	2.54	0.37	0.87	2.29
	0.05	0.26	0.60	1.39	0.25	0.53	1.30
	0.10	0.22	0.45	1.00	0.21	0.41	0.94

하였다. λ의 실제값은 $\lambda = 1$로 하였다. RMSE는 평균제곱오차(mean squared error)의 제곱근을 평균하여 구하였다. 결과를 살펴보면 RMSE는 예상과 같이 n이 증가함에 따라 감소하고, β가 증가함에 따라 증가하는 경향을 보인다.

다음으로 식 (2.11)의 $\hat{D}_n$과 식 (2.12)의 $\hat{\psi}^2_{n,c^*}$을 비교하고자 한다. 각 통계량의 기각값을 시뮬레이션을 통해 구하였으며 그 결과를 표 3.2에 제시하였다. 마찬가지로 $n = 50, 100$, $\beta = 0.5, 1, 1.5$, 표본의 수는 10,000으로 하였다. 통계량 D_n, ψ^2_{n,c^*}는 모수 λ의 값이 주어진 단순귀무가설에서의 통계량이고 이 통계량들도 참고로 비교에 포함하였다. 이경우 일반성을 잃지 않고 $\lambda = 1$로 가정할 수 있다. 물론 ψ^2_{n,c^*}은 단순귀무가설의 경우에는 검정하려고 하는 분포에 관계없이 동일한 분포를 따르므로 표 3.2의 기각값이 Kim (2011)의 로그정규성 검정의 경우와도 거의 비슷함을 볼 수 있다. $\hat{\psi}^2_{n,c^*}$은 모수 λ가 추정된 통계량으로 기각값이 ψ^2_{n,c^*}의 경우보다 작아졌음을 볼 수 있다. 이러한 현상은 지수분포의 경우, 임의중도절단의 경우뿐만 아니라 제 2종 중도절단이나 완전표본에서도 나타남을 볼 수 있다 (Stephens, 1986, Table 4.5와 Table 4.16, Sec. 4.16.2). $\hat{D}_n$의 경우도 그렇긴 하나 $\hat{\psi}^2_{n,c^*}$에 비해 작아지는 정도가 미약하고 $\beta = 1.5$인 경우에는 거의 차이가 없다. $\hat{\psi}^2_{n,c^*}$의 경우에도 β가 증가할수록 작아지는 정도가 약하다.

또한 몇 가지 대립가설에서 통계량의 검정력을 비교하였다. 표본크기 n과 중도절단모수 β는 위와 같고, 유의수준은 0.1, 표본의 수는 2,500을 이용하였다. 고려한 대립가설은 Weibull 분포 Weib(α) (확률밀도함수 $f(t;\alpha) = \alpha t^{\alpha-1}e^{-t^\alpha}$, $t > 0$)에서 $\alpha = 0.5, 2$인 경우, Gamma 분포 G(α) ($f(t;\alpha) = t^{\alpha-1}e^{-t}/\Gamma(\alpha)$, $t > 0$)에서 $\alpha = 0.5, 2$인 경우, lognormal 분포 lognorm (σ)($f(t) = 1/(t\sigma\sqrt{2\pi})e^{-(\log t)^2/(2\sigma^2)}$)에서 $\sigma = 1$인 경우, log-logistic 분포 log-logis(α) ($f(t;\alpha) = t^{\alpha-1}/(1+t^\alpha)^2$, $t > 0$)에서 $\alpha = 1$인 경우 등

표 **3.3.** 검정력 비교 (유의수준 0.10)

가설	표본크기	$\beta=0.5$				$\beta=1.0$				$\beta=1.5$			
		D_n	$\hat{D}_n$	ψ^2_{n,c^*}	$\hat{\psi}^2_{n,c^*}$	D_n	$\hat{D}_n$	ψ^2_{n,c^*}	$\hat{\psi}^2_{n,c^*}$	D_n	$\hat{D}_n$	ψ^2_{n,c^*}	$\hat{\psi}^2_{n,c^*}$
$\mathcal{E}(1)$	$n=50$	0.09	0.11	0.10	0.10	0.09	0.10	0.10	0.10	0.11	0.11	0.11	0.10
	$n=100$	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.08	0.09	0.11	0.10	0.10
Weib(0.5)	$n=50$	0.81	0.97	0.88	0.99	0.46	0.49	0.71	0.86	0.21	0.17	0.47	0.46
	$n=100$	0.99	*	*	*	0.80	0.84	0.97	*	0.34	0.28	0.81	0.87
Weib(2)	$n=50$	0.78	0.98	0.90	0.99	0.29	0.66	0.62	0.81	0.10	0.38	0.29	0.23
	$n=100$	*	*	*	*	0.79	0.96	0.99	*	0.19	0.69	0.75	0.86
G(0.5)	$n=50$	0.99	0.72	0.99	0.83	0.95	0.21	0.97	0.57	0.78	0.13	0.90	0.25
	$n=100$	*	0.96	*	0.98	*	0.38	*	0.89	0.96	0.18	*	0.57
G(2)	$n=50$	*	0.68	*	0.80	*	0.32	*	0.32	0.86	0.20	*	0.12
	$n=100$	*	0.95	*	0.98	*	0.53	*	0.80	*	0.36	*	0.32
lognorm(1)	$n=50$	0.66	0.23	0.85	0.30	0.33	0.13	0.72	0.15	0.16	0.12	0.48	0.10
	$n=100$	0.94	0.42	0.99	0.54	0.62	0.13	0.95	0.27	0.24	0.13	0.79	0.14
log-logis(1)	$n=50$	0.77	0.82	0.81	0.87	0.52	0.26	0.63	0.48	0.27	0.12	0.43	0.19
	$n=100$	0.97	0.98	0.97	0.98	0.83	0.41	0.91	0.79	0.49	0.16	0.70	0.37

이다. 이는 모두 수명시간의 모형화에 자주 사용되는 분포이다. 표 3.3에서 *는 소수 세째자리에서 반올림하여 검정력이 1이 되었음을 의미한다.

표 3.3의 검정력 결과를 살펴보자. 첫째 전반적으로 단순귀무가설과 복합귀무가설에서 각각 ψ^2_{n,c^*}, $\hat{\psi}^2_{n,c^*}$가 D_n, $\hat{D}_n$보다 우수한 검정력을 보인다. Stephens (1974)에 의하면 중도절단이 없는 경우에도 이와 비슷한 경향을 보인다. 복합귀무가설, $\beta=1.5$, Weib(2)($n=50$)와 G(2)($n=50, n=100$)의 경우에는 예외적으로 $\hat{D}_n$이 더 우수하다. 둘째, 복합귀무가설의 경우 일반적으로 모수의 추정으로 인하여 검정력이 감소하고 lognormal 분포의 경우에는 이러한 현상이 가장 심하게 나타난다. 그러나 Weibull 분포에서는 예외적으로 모수를 추정했을 때 검정력이 대체로 상승함을 볼 수 있다. log-logistic 분포에서도 이러한 현상이 보이나 그 정도가 미약하고 중도절단비율이 증가하면서 사라진다. Weibull 분포의 경우 두 통계량 모두에서 이러한 현상이 나타나는 것으로 보아 이는 통계량보다는 분포의 특징때문에 나타나는 현상으로 보인다. 일반적으로 단순귀무가설의 경우 복합귀무가설보다 더 많은 정보가 주어졌으므로 검정력이 우수할 것이라 생각하기 쉬우나 모수의 추정과 검정력의 관계는 생각보다 그리 단순하지 않다. 이에 관해서는 Stephens (1986, sec 4.16.2)에서도 언급하고 있다. 마지막으로 두 통계량 모두 중도절단자료의 비율이 증가하면서 검정력의 감소가 매우 두드러지게 나타나고 있다. 따라서 완전자료의 비율이 상대적으로 낮은 경우에는 통계량의 이용에 주의가 필요하다.

4. 결론 및 토의

Koziol-Green 통계량은 Cramér-von Mises 통계량을 임의중도절단자료의 경우로 일반화한 것이다. 이는 Kaplan-Meier의 product limit 경험분포함수를 이용한 것으로 Cramér-von Mises 통계량과 마찬가지로 단순귀무가설의 경우에는 확률적분변환을 이용하여 분포에 무관하게 이용가능하다.

본 논문에서는 지수분포의 척도모수가 미지인 경우, 모수를 추정하여 Koziol-Green 통계량을 일반화하고, 같은 방법으로 일반화한 Kolmogorov-Smirnov 통계량과 비교하였다. 그 결과 Kolmogorov-Smirnov 통계량보다는 일반화된 Koziol-Green 통계량이 고려한 대부분의 대립가설에서 더 좋은 검정력을 보여준다. 그러나 모수가 추정된 경우에는 Koziol-Green 통계량의 귀무가설에서의 분포도 단순귀

무가설의 경우와 다를 것이고 이에 대한 이론적인 접근이 필요하다. 또한 모수의 추정과 검정력과의 관계도 그 양상이 그리 단순해 보이지는 않는다. 그리고 실제자료의 경우에는 자료의 중도절단비율을 이용하여 중도절단모수 β를 추정하여야 한다. 따라서 모형에서의 β와는 차이가 있을 것이므로 검정의 과정에서 또 다른 변동성을 만들어 낼 것이다. 또한 모수 추정의 과정은 가정한 중도절단 모형에 의존하다. 중도절단 모형이 달라지면 그에 따라 모수추정의 과정도 달라질 것이고 통계량의 분포도 일반적으로 영향을 받을 것라 생각된다. 이러한 사항들에 대해서는 좀 더 면밀한 연구가 필요하다.

참고문헌

Akritas, M. G. (1988). Pearson type goodness of fit test: The univariate case, *Journal of the American Statistical Association*, **83**, 222–230.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorships, *The Annals of Statistics*, **2**, 437–453.

Chen, C. (1984). A correlation goodness-of-fit test for randomly censored data, *Biometrika*, **71**, 315–322.

Chen, Y. Y., Hollander, M. and Langberg, N. A. (1982). Small-sample results for the Kaplan Meier estimator, *Journal of the American statistical Association*, **77**, 141–144.

Csörgő, S. and Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises Statistics, *Journal of the Royal Statistical Society, Series B*, **58**, 221–234.

Csörgő, S. and Horváth, L. (1981). On the Koziol-Green Model for random censorship, *Biometrika*, **68**, 391–401.

Efron, B. (1967). The two sample problem with censored data, In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 831–853.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality, *Technometrics*, **17**, 111–117.

Hollander, M. and Peña, E. A. (1992). A chi squared goodness of fit test for randomly censored data, *Journal of the American Statistical Association*, **87**, 458–463.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

Kim, N. (2011). Testing log normality for randomly censored data, *The Korean Journal of Applied Statistics*, **24**, 883–891.

Koziol, J. A. (1978). A two sample Cramér-von Mises test for randomly censored data, *Biometrical Journal*, **20**, 603–608.

Koziol, J. A. (1980). Goodness-of-fit tests for randomly censored data,n *Biometrika*, **67**, 693–696.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data, *Biometrika*, **63**, 465–474.

Lee, E.T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*, John Wiley & Sons, Inc. New Jersey.

Meier, P. (1975). Estimation of a distribution function from incomplete observations, In *Perspectives in Probability and Statistics*, Ed. J. Gani, 67–87. London, Academic Press.

Michael, J. R. and Schucany, W. R. (1986). Analysis of data from censored samples, In *Goodness of Fit Techniques*, (Edited by D'Agostino, R. B. and Stephens, M. A.), Chapter 11, Marcel Dekker, New York.

Nair, V. N. (1981). Plots and tests for goodness of fit with randomly censored data, *Biometrika*, **68**, 99–103.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, **69**, 730–737.

Stephens, M. A. (1986). Tests based on EDF statistics, In *Goodness-of-Fit Techniques* (Edited by D'Agostino, R. B. and Stephens, M. A.), Chapter 5, Marcel Dekker, New York.

Testing Exponentiality Based on EDF Statistics for Randomly Censored Data when the Scale Parameter is Unknown

Namhyun Kim[1]

[1]Department of Science, Hongik University

(Received January 2, 2012; Revised February 29, 2012; Accepted March 9, 2012)

Abstract

The simplest and the most important distribution in survival analysis is exponential distribution. Koziol and Green (1976) derived Cramér-von Mises statistic's randomly censored version based on the Kaplan-Meier product limit estimate of the distribution function; however, it could not be practical for a real data set since the statistic is for testing a simple goodness of fit hypothesis. We generalized it to the composite hypothesis for exponentiality with an unknown scale parameter. We also considered the classical Kolmogorov-Smirnov statistic and generalized it by the exact same way. The two statistics are compared through a simulation study. As a result, we can see that the generalized Koziol-Green statistic has better power in most of the alternative distributions considered.

Keywords: Goodness of fit, random censorship, Cramér-von Mises statistic, Kolmogorov-Smirnov statistic, Kaplan-Meier product limit estimate.

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (2009-0072563).

[1]Professor, Department of Science, Hongik University, 72-1 Sangsu-dong, Mapo-gu, Seoul 121-791, Korea. E-mail: nhkim@hongik.ac.kr