응용통계연구 (2011)
24(5), 883–891
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.883

임의중도절단자료에 대한 로그정규성 검정

김남현[1]

[1]홍익대학교 기초과학과

(2011년 6월 접수, 2011년 9월 채택)

요 약

수명시간에 대한 모형으로 로그정규분포가 자주 사용되며, 이는 자료의 변환에 의하여 정규성 검정과 동일한 문제로 생각할 수 있다. 따라서 자료의 로그정규성 검정을 위하여, 정규성 검정에 자주 이용되는 Shapiro-Wilk 형태의 검정통계량을 Kaplan-Meier의 product limit 경험분포함수를 이용하여 임의중도절단자료로 일반화한다. Cramér-von Mises 통계량을 임의중도절단자료로 일반화한 Koziol과 Green (1976)의 통계량과 비교하였으며 이를 위하여 단순귀무가설을 가정하였다. 중도절단분포에 대한 모형으로는 Koziol과 Green (1976)에서 제시한 모형과 이와 유사한 다른 모형 두 가지를 고려하였다. 검정력 비교 결과 제시한 통계량이 로그정규성 또는 정규성 검정에 더 좋은 검정력을 보여주었으며 검정력은 중도절단분포 모형보다는 자료의 중도절단비율에 영향을 받는다는 것을 볼 수 있었다.

주요용어: 적합도검정, 임의중도절단, Kaplan-Meier 추정량.

1. 서론

분포에 대한 적합도 검정은 이론적, 실제적인 측면에서 오랫동안 통계적 추론의 중요한 관심사였다. 이는 중도절단 자료에서도 마찬가지이다 (D'Agostino와 Stephens, 1986). 본 논문에서는 정규성 검정에 자주 쓰이는 Shapiro-Wilk 통계량과 관계 깊은 de Wet과 Venter 통계량을 Kaplan-Meier product limit 경험분포함수를 이용하여 임의중도절단자료(randomly censored data)에 대해서 일반화하고자 한다.

임의중도절단자료에 대한 적합도 검정은 제 1종 중도절단(Type I censoring)이나 제 2종 중도절단(Type II censoring)에 비해서 그리 많은 연구가 되지는 않았다. 대표적인 것으로는 Koziol과 Green (1976), Koziol (1980)을 들 수 있다. 이들은 적합도 검정에 자주 쓰이는 Cramér-von Mises 통계량, Kolmogorov-Smirnov 통계량 등을 임의중도절단 자료로 확장하였다. 이러한 통계량들은 경험분포함수(empirical distribution function)에 기반한 통계량이다. 또한 Chen (1984)은 임의중도절단자료의 복합귀무가설에서 Shapiro-Francia 통계량과 유사한 상관계수 통계량을 일반화하고 주로 지수분포에 적용하였다. Chen 등 (1983)도 임의중도절단자료에 대한 검정을 다루었다.

2절에서는 Koziol-Green 통계량과 제안하는 통계량의 구체적인 형태를 소개하고 3절에서는 특별한 중도절단 모형을 가정하고 Koziol-Green 통계량과 제안한 통계량을 비교한다. 4절에서는 결론과 함께 좀 더 생각해 보아야 할 문제를 간략히 언급한다.

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

2. 로그정규성 검정 통계량

$X_1^0, \ldots, X_n^0$를 연속확률분포 F에서의 확률표본이라고 하고, $C_1, \ldots, C_n$은 X_i^0에 독립이고, 연속분포 G에서의 중도절단 확률변수라고 하자. X_i^0는 C_i에 의해서 우측 중도절단되고(right censored), 관측되는 자료는 (X_i, δ_i), $1 \le i \le n$이다. 즉, $X_i = \min(X_i^0, C_i)$이고

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

이다. 또한 관측자료는 $X_1, \ldots, X_n$의 순서통계량 $X_{(1)} \le \cdots \le X_{(n)}$에 대해서 $(X_{(i)}, \delta_{(i)})$, $1 \le i \le n$으로도 쓸 수 있다. 여기서 $\delta_{(i)}$는 i번째 순서통계량이 중도절단 되지 않았을 때 1을 갖는다. 검정하려는 가설은 특정한 연속분포 F^0에 대해서

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

가 된다. Koziol과 Green (1976)의 통계량은 기본적으로 F^0가 완전히 주어진 단순귀무가설을 가정하고 있다. 따라서 확률적분변환(probability integral transformation)에 의해서 식 (2.2)의 귀무가설은 $X_1^0, \ldots, X_n^0$가 균일분포 $U(0,1)$임을 검정하는 문제로 생각할 수 있다. 즉, $Z_i^0 = F^0(X_i^0)$일 때 Z_i^0가 $U(0,1)$을 따르는지를 검정하면 된다. $Z_i = F^0(X_i) = \min(Z_i^0, F^0(C_i))$이고 $Z_{(1)}, \ldots, Z_{(n)}$을 $Z_1, \ldots, Z_n$의 순서통계량이라고 할 때 $(Z_{(i)}, \delta_{(i)})$의 product-limit 경험분포함수는

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

이고 Koziol과 Green (1976)의 통계량은

$$\psi_n^2 = n \int_0^1 \left(\widehat{F}_n(t) - t \right)^2 dt \tag{2.4}$$

이다. 즉, $\widehat{F}_n(t)$와 균일분포 $U(0,1)$간의 일종의 차이를 보는 것이다. 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 (2.4)의 통계량은 $X_1, \ldots, X_n$의 순서통계량 $(X_{(1)}, \ldots, X_{(n)})$ 중 중도절단되지 않은 관측(uncensored observations) $X_{(i)}$에 대해서 즉, $(X_{(i)}, \delta_{(i)})$ 중에서 $\delta_{(i)} = 1$인 경우, $w_i = \widehat{F}_n(Z_{(i)})$이라고 하고, $Z_{(0)} = 0$, $Z_{(n+1)} = 1$이라고 할 때

$$\psi_n^2 = n \sum_{j=1, \delta_{(j)}=1}^{n+1} w_{j-1} (Z_{(j)} - Z_{(j-1)}) \left\{ w_{j-1} - (Z_{(j)} + Z_{(j-1)}) \right\} + \frac{1}{3} n$$

이 된다 (Koziol과 Green, 1976). 이는 중도절단이 없는 경우 Cramér-von Mises 통계량과 일치한다. ψ_n^2의 귀무가설에서의 분포는 F^0가 완전히 주어진 단순귀무가설의 경우에도 중도절단분포 G에 의존한다.

Koziol과 Green (1976)은 중도절단분포 G에 대해서

$$(1 - G) = (1 - F)^{\beta} \tag{2.5}$$

을 가정하고 $\beta < 2$일 때 ψ_n^2의 점근분포를 구하였다. 여기서 β는 고정된 양수로 중도절단모수(censoring parameter)로 해석될 수 있으므로 이와 같이 부르기로 하자. 즉, $\beta = 0$이면 중도절단이 없는 경우이고, 중도절단관측의 기대비율을 γ_1이라고 하면

$$\gamma_1 = P(X_i^0 > C_i) = \int_{-\infty}^{\infty} (1 - F(x))dG(x) = \int_0^1 \beta(1-x)^\beta dx = \frac{\beta}{\beta+1}$$

로 β가 증가할수록 중도절단관측의 기대비율 γ_1이 높아진다. Csörgő와 Horváth (1981)는 식 (2.5)를 Koziol-Green 모형이라고 불렀다. 본 논문에서도 이와 같은 명칭을 사용하자.

이러한 형태의 중도절단모형의 의미는 Chen 등 (1982)에서 설명하고 있다. 이에 따르면 Koziol-Green 모형은 두 개의 성분을 가진 직렬시스템에 적용할 수 있다. 만일 두 성분이 모두 정상적으로 작동될 때 전체 시스템이 작동한다고 가정하고, X_i^0는 첫 번째 성분의 수명시간, C_i는 두 번째 성분의 수명시간이라고 하자. 그러면 $X_i = \min(X_i^0, C_i)$는 시스템의 수명시간이 되고 δ_i를 식 (2.1)과 같이 정의하면 δ_i는 어떤 성분에서 고장(failure)이 일어났는지를 나타낸다. 만일 첫 번째 성분의 고장분포가 관심의 대상이고 두 번째 성분이 β개의 부성분(subcomponents)으로 직렬 연결되어 있고 각각의 부성분의 분포가 F와 동일하다고 가정하자. 그러면 C_i는 β개의 부성분의 수명 시간 중 최소가 되므로 C_i의 분포 G는 $1 - G = (1 - F)^\beta$의 Koziol-Green 모형을 따른다. 물론 이 경우 β는 양의 정수일 때 의미 있으나 일반적인 Koziol-Green 모형에서는 β를 양수로 가정한다.

만일 두 번째 β개의 성분이 직렬이 아닌 병렬로 연결되어 있고, 한 개 이상이 작동하면 전체 시스템이 정상적으로 작동하는 시스템이라고 하면 중도절단분포 G는

$$G = F^\beta \tag{2.6}$$

를 만족하게 된다. 또한 이 경우 중도절단관측의 기대비율을 γ_2라고 하면

$$\gamma_2 = P(X_i^0 > C_i) = \int_{-\infty}^{\infty} (1 - F(x))dG(x) = \frac{1}{(\beta+1)}$$

이다. 물론 $\beta = 1$일 때는 Koziol-Green 모형과 동일하다. 식 (2.6)의 중도절단모형을 P 모형이라 하고 위의 두 가지 형태의 임의중도절단 모형을 고려하자.

본 연구에서는 귀무가설

$$H_0 : T_1^0, \ldots, T_n^0\text{의 분포는 로그정규분포 lognormal}(0,1)\text{을 따른다}$$

를 고려한다. 로그정규분포는 수명시간의 모형화에 자주 사용되는 대표적인 분포 중 하나이다. 이는 $X_i^0 = \log T_i^0$라고 할 때

$$H_0 : X_1^0, \ldots, X_n^0\text{의 분포는 정규분포 } N(0,1)\text{을 따른다}$$

로 생각할 수 있다. 따라서 고려하는 문제는 단순귀무가설에서의 정규성 검정 문제로 귀결된다. 즉, 식 (2.2)의 귀무가설에서 $F^0 = \Phi$인 경우로 생각할 수 있다. Φ는 표준정규분포 $N(0,1)$의 누적분포함수이다.

정규성 검정을 위한 통계량으로 많이 이용되는 것으로 Shapiro와 Wilk (1965), Shapiro와 Francia (1972) 통계량 등을 들 수 있고 de Wet과 Venter (1972) 통계량도 이들과 관계가 깊은 통계량이다. 우선 중도절단이 없는 경우를 간단히 살펴보자. $Y_1, \ldots, Y_n$이 분포 F에서의 확률표본

이고 $Y_{(1)},\ldots,Y_{(n)}$은 순서통계량이라고 하자. de Wet과 Venter (1972)의 통계량은 단순귀무가설 $H_0 : F = \Phi$에서는

$$L_n^0 = \sum_{j=1}^{n} \left(Y_{(j)} - \Phi^{-1}\left(\frac{j}{n+1}\right)\right)^2 \tag{2.7}$$

이고 복합귀무가설 $F(x) = \Phi\left((x-\mu)/\sigma\right)$ (μ와 σ는 미지)에서는

$$L_n = \sum_{j=1}^{n} \left(\frac{Y_{(j)} - \bar{Y}_n}{\hat{\sigma}_n} - \Phi^{-1}\left(\frac{j}{n+1}\right)\right)^2, \tag{2.8}$$

$\bar{Y}_n = \sum_{j=1}^{n} Y_j/n$, $\hat{\sigma}_n^2 = \sum_{j=1}^{n}(Y_j - \bar{Y}_n)/n$이다. Φ^{-1}은 Φ의 역함수이다. Verrill과 Johnson (1988), Kim (2011)은 L_n 통계량을 제 2종 중도절단자료에 적용하고 이의 극한분포에 대해서 연구하였다.

식 (2.7)과 식 (2.8)을 임의중도절단자료에 대해서 일반화하자. L_n^0와 L_n은 표본에서의 백분위와 가정한 분포에서의 이론적인 백분위의 차이를 보는 것으로 $\Phi^{-1}\left(i/(n+1)\right)$는 근사적으로 $\Phi^{-1}\left(F_n(Y_{(i)})\right) = \Phi^{-1}(\hat{p}_i)$로 생각할 수 있다. $F_n(x)$은 경험적 누적분포함수(empirical cumulative distribution function)이다. $\hat{p}_i$의 좀 더 일반적인 형태로

$$\hat{p}_i = \frac{i-c}{n-2c+1}, \quad 0 \le c \le 1 \tag{2.9}$$

이 이용되기도 한다. 물론 de Wet-Venter 통계량은 $c = 0$인 경우이다. 정규분포의 경우 $c = 3/8$일 때 $\Phi^{-1}(\hat{p}_i)$이 순서통계량의 기댓값과 가깝다는 것이 알려져 있으므로 (Blom, 1958) $c = 3/8$을 자주 이용한다.

순서통계량 $X_{(1)} \le \cdots \le X_{(n)}$에서 일부가 임의절단된 경우, 즉 $(X_{(i)}, \delta_{(i)})$에 대해서는 $\hat{p}_i$이 달라져야 한다. $\hat{p}_i$은 $F(X_{(i)})$의 추정값이며 자료가 임의절단된 경우에는 모집단에서 $X_{(i)}$ 이하의 값을 갖는 비율이 더 이상 (근사적으로도) i/n가 아니기 때문이다. 이 경우 $F(X_{(i)}|\mu,\sigma)$의 추정에 식 (2.3)의 Kaplan과 Meier (1958)의 추정량

$$\widehat{F_n}(x) = 1 - \prod_{X_{(j)} \le x} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$$

을 이용할 수 있다. 이를 약간 변형하여

$$\hat{p}_i^c = 1 - \frac{n-c+1}{n-2c+1} \prod_{j \le i} \left(\frac{n-j-c+1}{n-j-c+2}\right)^{\delta_{(j)}}$$

로 정의하자. 그러면 이는 중도절단이 없는 경우 식 (2.9)의 $\hat{p}_i$와 동일하게 된다. $\hat{p}_i^c$를 이용하여 de Wet-Venter 통계량을

$$L_n^c = \sum_{j:\delta_{(j)}=1} \left(\frac{X_{(j)} - \hat{\mu}_c}{\hat{\sigma}_c} - \Phi^{-1}(\hat{p}_j^c)\right)^2 \tag{2.10}$$

와 같이 일반화할 수 있다. 여기서 $\hat{\mu}_c$, $\hat{\sigma}_c$는 중도절단되지 않은 자료를 이용한 추정량이다. 식 (2.10)은 중도절단이 없고 $c = 0$일 경우 $\hat{p}_i^c = i/(n+1)$이므로 식 (2.8)의 de Wet-Venter 통계량과 일치한다.

본 연구에서는 기존의 Koziol-Green 통계량과의 비교를 위하여 우선 정규분포의 단순귀무가설을 가정

표 **3.1.** Koziol-Green 모형에서의 통계량의 기각값

Koziol-Green 모형		L_n^c			ψ_n^2		
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$n = 20$	$\gamma_1 = 3/5\,(\beta = 3/2)$	5.56	3.43	2.56	2.81	1.78	1.30
	$\gamma_1 = 1/2\,(\beta = 1)$	5.92	3.73	2.85	1.88	1.10	0.81
	$\gamma_1 = 1/3\,(\beta = 1/2)$	6.51	4.26	3.29	1.07	0.65	0.50
$n = 50$	$\gamma_1 = 3/5\,(\beta = 3/2)$	5.65	3.63	2.85	2.99	1.75	1.28
	$\gamma_1 = 1/2\,(\beta = 1)$	6.22	4.04	3.18	1.60	0.97	0.72
	$\gamma_1 = 1/3\,(\beta = 1/2)$	6.95	4.60	3.59	0.98	0.62	0.47
$n = \infty$	$\gamma_1 = 3/5\,(\beta = 3/2)$				1.93	1.35	1.12
	$\gamma_1 = 1/2\,(\beta = 1)$				1.21	0.79	0.62
	$\gamma_1 = 1/3\,(\beta = 1/2)$				0.92	0.58	0.44

표 **3.2.** P 모형에서의 통계량의 기각값

P 모형		L_n^c			ψ_n^2		
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$n = 20$	$\gamma_2 = 1/2\,(\beta = 1)$	6.05	3.68	2.81	1.79	1.07	0.80
	$\gamma_2 = 1/3\,(\beta = 2)$	6.50	4.07	3.18	1.11	0.68	0.52
	$\gamma_2 = 1/4\,(\beta = 3)$	6.83	4.33	3.39	1.01	0.61	0.45
$n = 50$	$\gamma_2 = 1/2\,(\beta = 1)$	6.20	3.98	3.16	1.55	0.95	0.73
	$\gamma_2 = 1/3\,(\beta = 2)$	6.64	4.34	3.46	0.98	0.64	0.49
	$\gamma_2 = 1/4\,(\beta = 3)$	7.13	4.63	3.69	0.87	0.56	0.42

하였다. 이 경우 일반성을 잃지 않고 $\mu = 0$, $\sigma = 1$이라고 가정해도 되고, 모수 추정의 문제는 고려하지 않아도 된다. 따라서 식 (2.10)은

$$L_n^c = \sum_{j:\delta_{(j)}=1} \left(X_{(j)} - \Phi^{-1}\left(\hat{p}_j^c\right)\right)^2 \tag{2.11}$$

이 된다. 이는 물론 식 (2.7)의 L_n^0를 임의절단자료로 일반화한 것이다. c는 중도절단이 없는 경우와 같은 이유로 $c = 3/8$을 선택한다.

3. 모의실험 결과

본 연구에서는 식 (2.11)의 L_n^c 통계량과 식 (2.4)의 Koziol-Green의 ψ_n^2을 비교하고자 한다. 이를 위해서 중도절단분포 G는 식 (2.5)의 Koziol-Green 모형과 식 (2.6)의 P 모형을 가정하였다. 두 가지 중도절단모형에서 각 통계량의 기각값을 시뮬레이션을 통하여 구하였고 그 결과 표 3.1과 표 3.2를 얻었다. 표본의 수는 $N = 10,000$을 사용하였다. 표 3.1에는 Koziol과 Green (1976)에서 구한 ψ_n^2의 극한분포에서의 기각값도 추가하였다. Koziol과 Green (1976)에 따르면 β가 증가할수록 ψ_n^2은 더 기운분포(more skewed distribution)을 가진다. 이에 따라 표 3.1의 ψ_n^2의 경우 β가 증가할수록, 즉 중도절단비율이 높을수록 기각값이 증가하는 것을 볼 수 있다.

또한 몇 가지 대립가설에서 두 통계량의 검정력을 살펴보았다. 표본크기는 $n = 20, 50$, 유의수준은 $\alpha = 0.1$, 표본의 수는 $N = 2,500$을 이용하였다. 고려한 대립가설은 지수분포, Weibull 분포 Weibull(α)(확률밀도함수 $f(t;\alpha) = \alpha t^{\alpha-1}e^{-t^\alpha}$, $t > 0$), Gamma 분포 Gamma $(\alpha, 1)$ ($f(t;\alpha) = 1/\Gamma(\alpha)t^{\alpha-1}e^{-t}$, $t > 0$), log-logistic 분포 ($f(t) = 1/(1+t)^2$, $t > 0$) 등이다. 이는 모두 수명시간의 모형화에 자주 사용

표 **3.3.** Koziol-Gren 모형에서의 검정력 비교 ($\alpha = 0.10$)

Koziol-Green 모형	$\gamma_1(\beta = \beta_1)$	$\gamma_1 = 3/5$ ($\beta_1 = 3/2$)		$\gamma_1 = 1/2$ ($\beta_1 = 1$)		$\gamma_1 = 1/3$ ($\beta_1 = 1/2$)	
가설	표본크기	L_n^c	ψ_n^2	L_n^c	ψ_n^2	L_n^c	ψ_n^2
lognormal$(0, 1)$	$n = 20$	0.10	0.10	0.11	0.10	0.09	0.10
	$n = 50$	0.09	0.11	0.09	0.10	0.10	0.10
exp(1)	$n = 20$	0.81	0.12	0.81	0.18	0.82	0.30
	$n = 50$	0.98	0.37	0.98	0.56	0.98	0.70
Weibull(0.5)	$n = 20$	0.99	0.35	$*$	0.54	$*$	0.72
	$n = 50$	$*$	0.81	$*$	0.96	$*$	0.99
Weibull(2)	$n = 20$	0.16	0.07	0.21	0.06	0.33	0.14
	$n = 50$	0.31	0.12	0.46	0.42	*0.78*	*0.85*
Gamma$(0.5, 1)$	$n = 20$	$*$	0.80	$*$	0.92	$*$	0.97
	$n = 50$	$*$	$*$	$*$	$*$	$*$	$*$
Gamma$(2, 1)$	$n = 20$	0.51	0.34	0.58	0.54	*0.65*	*0.75*
	$n = 50$	0.93	0.76	0.96	0.93	0.98	0.98
log-logistic	$n = 20$	0.76	0.15	0.78	0.21	0.85	0.32
	$n = 50$	0.95	0.24	0.97	0.46	0.99	0.65

표 **3.4.** P 모형에서의 검정력 비교 ($\alpha = 0.10$)

P 모형	$\gamma_2(\beta = \beta_2)$	$\gamma_2 = 1/2$ ($\beta_2 = 1$)		$\gamma_2 = 1/3$ ($\beta_2 = 2$)		$\gamma_2 = 1/4$ ($\beta_2 = 3$)	
가설	표본크기	L_n^c	ψ_n^2	L_n^c	ψ_n^2	L_n^c	ψ_n^2
lognormal$(0, 1)$	$n = 20$	0.11	0.10	0.11	0.11	0.10	0.10
	$n = 50$	0.10	0.11	0.09	0.10	0.10	0.10
exp(1)	$n = 20$	0.82	0.18	0.81	0.29	0.81	0.30
	$n = 50$	0.97	0.57	0.98	0.69	0.98	0.76
Weibull(0.5)	$n = 20$	$*$	0.56	$*$	0.73	$*$	0.76
	$n = 50$	$*$	0.95	$*$	0.99	$*$	$*$
Weibull(2)	$n = 20$	0.22	0.06	0.29	0.10	0.35	0.17
	$n = 50$	0.46	0.40	*0.66*	*0.77*	*0.74*	*0.89*
Gamma$(0.5, 1)$	$n = 20$	$*$	0.92	$*$	0.98	$*$	0.99
	$n = 50$	$*$	$*$	$*$	$*$	$*$	$*$
Gamma$(2, 1)$	$n = 20$	0.59	0.55	*0.69*	*0.72*	*0.71*	*0.78*
	$n = 50$	0.96	0.93	0.98	0.98	0.98	0.99
log-logistic	$n = 20$	0.80	0.22	0.80	0.29	0.83	0.32
	$n = 50$	0.97	0.46	0.98	0.60	0.99	0.69

되는 분포이다. 단순귀무가설을 가정하였으므로 모수의 값에 따라 검정력이 달라지므로 분포에 따라 몇 개의 모수값을 설정하였다. 표 3.3의 Koziol-Green 모형에서는 중도절단비율이 높고($\gamma = 3/5$) 표본크기가 작은 경우($n = 20$)에 모든 자료가 중도절단되는 경우가 발생한다. 특히 대립가설이 Gamma분포와 log-logistic 분포일 때 이런 표본이 1–3개 발생하기도 한다. 이 경우에는 통계량을 계산할 수가 없고 귀무가설에 대한 검정이 곤란하다.

표 3.3과 표 3.4의 결과를 보면 이탤릭체로 표시된 몇 개의 경우를 제외한 대부분의 경우의 대립가설에서 제안한 L_n^c 통계량이 더 좋은 검정력을 보여준다. 이는 ψ_n^2이 정규성 검정만을 위하여 제안된 통계량이 아니기 때문에 나타난 결과 일 수도 있다고 생각된다. 표 3.3과 표 3.4에서 $*$는 소수 셋째 자리에서 반올림하여 검정력이 1이 되었음을 의미한다. 일반적으로 중도절단비율이 높을수록 검정력이 감소

할 것으로 예상된다. 이러한 경향은 주어진 결과에서도 물론 볼 수 있으나 L_n^c 보다는 ψ_n^2의 경우 훨씬 뚜렷하게 나타난다. ψ_n^2의 경우 검정력이 중도절단비율에 크게 영향을 받는 것으로 보인다. 또한 두 가지 중도절단모형의 검정력을 비교하면 Weibull(2), $n = 50$을 제외한 대부분의 경우 중도절단모형이 달라도 중도절단비율이 같으면 검정력도 비슷함을 볼 수 있다. 이로 미루어 보다 임의중도절단자료의 검정력은 중도절단분포보다는 중도절단비율에 더 강한 영향을 받는다고 짐작할 수 있고 이에 대한 구체적인 연구가 필요하다고 판단된다. Chen (1984)은 임의중도절단자료에 대해서 제안한 상관계수 통계량이 지수분포의 경우 중도절단모형에 크게 영향을 받지 않는다는 사실에 대해서 연구하였다. 실제로 표 3.1, 표 3.2의 기각값도 같은 중도절단비율에 대해서는 중도절단분포에 따라 심하게 차이가 나지 않는다는 것을 관찰할 수 있다.

4. 결론 및 토의

본 논문에서의 de Wet-Venter 통계량을 임의중도절단자료의 로그정규성 또는 정규성 검정을 위한 통계량으로 일반화하였다. 이 때 Kaplan-Meier의 product limit 경험분포함수를 이용하였다. 또한 Cramér-von Mises 통계량을 임의중도절단자료로 일반화한 Koziol-Green 통계량과 비교한 결과 고려한 대부분의 대립가설에서 더 좋은 검정력을 보여주었다. 제안한 식 (2.11)의 통계량은 중도절단이 없는 경우나 제 2종 중도절단의 경우 (Kim, 2011)와 마찬가지로, 극한분포가 적절한 공분산 구조를 가진 가우시안 과정(Gaussian process)의 적분의 형태로 나타날 것으로 예상되기는 하나 이에 대한 구체적인 증명이 필요한 상태이다.

본 논문에서는 Koziol-Green 통계량과 비교하기 위해서 Koziol과 Green (1976)과 같이 단순귀무가설을 가정하였다. 제안한 통계량은 복합귀무가설의 경우 식 (2.10)과 같이 일반화된다. 그러나 이 경우에는 모수의 추정이 선행되어야 하므로 임의절단자료의 모수추정의 문제에 대해 먼저 생각해 보아야 한다. 또한 모의실험의 경우에는 중도절단분포 모형을 알고 이에 따른 중도절단관측의 기대비율을 중도절단모수로 조절하였다. 그러나 실제 자료에 대해서는 자료의 중도절단비율로부터 이를 추정하여야 하며 이에 따라 검정의 결과도 영향을 받을 것이라 생각된다. 이에 대해서는 좀 더 자세한 연구가 필요하다.

참고문헌

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*, New York, Wiley.

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.

Chen, Y. Y., Hollander, M. and Langberg, N. A. (1983). Testing whether new is better than used with randomly censored data, *The Annals of Statistics*, **11**, 267–274. Correction(1983), **11**, 1267.

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

D'Agositno, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

de Wet, T. and Venter, J. H. (1972). Asymptotic distributions of certain test criteria of normality, *South African Statistical Journal*, **6**, 135–149.

Efron, B. (1967). The two sample problem with censored data, *Proceeding 5th Berkeley Symposium* , **4**, 831–853.

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

Kim, N. (2011). The limit distribution of a modified Shapiro-Wilk statistic for normality to Type II censored data. *Journal of the Korean Statistical society*, **40**, 257–266.

Koziol, J. A. (1980). Goodness-of-fit tests for randomly censored data, *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.

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

Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215–216,

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, **52**, 591–611.

Verrill, S. and Johnson, R. A. (1988). Tables and large sample distribution theory for censored data correlation statistics for testing normality, *Journal of the American Statistical Association*, **83**, 1192–1197.

Testing Log Normality for Randomly Censored Data

Namhyun Kim[1]

[1]Department of Science, Hongik University

(Received June 2011; accepted September 2011)

Abstract

For survival data we sometimes want to test a log normality hypothesis that can be changed into normality by transforming the survival data. Hence the Shapiro-Wilk type statistic for normality is generalized to randomly censored data based on the Kaplan-Meier product limit estimate of the distribution function. Koziol and Green (1976) derived Cramér-von Mises statistic's randomly censored version under the simple hypothesis. These two test statistics are compared through a simulation study. As for the distribution of censoring variables, we consider Koziol and Green (1976)'s model and other similar models. Through the simulation results, we can see that the power of the proposed statistic is higher than that of Koziol-Green statistic and that the proportion of the censored observations (rather than the distribution of censoring variables) has a strong influence on the power of the proposed statistic.

Keywords: Goodness of fit, random censorship, 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