The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.1.001
27(1), 1–11

A Comparison of Survival Distributions with Unequal Censoring Distributions

Sujeong Song[a] · Jae Won Lee[b,1]

[a]Clinical Research & Pharmacovigilance Team, R&D Center, Yuhan Corporation
[b]Department of Statistics, Korea University

(Received September 10, 2013; Revised November 11, 2013; Accepted November 11, 2013)

Abstract

The Weighted Logrank test and its special case, Logrank test are widely used to compare survival distributions; however, these methods are inappropriate when the sample size is small or censoring distributions are not equal since they use test statistics from approximate distributions. A permutation test can be an alternative for small sample cases; however, this should be used only when censoring distributions are equal. To handle cases with small sample size and unequal censoring distributions, the permutation-imputation method was developed to compare two survival distributions. In this paper, approximate method, permutation method and permutation-imputation method were compared using a Logrank test and Prentice-Wilcoxon test for three or more survival distributions comparison.

Keywords: Permutation test, permutation-imputation test, survival distribution.

1. 서론

임상 시험 분야에서 처리에 변화를 준 세 개 이상의 집단에 대한 생존분포의 비교를 위해 다양한 모수적, 비모수적 방법들이 존재한다. 특히 자료가 중도 절단된 경우에는 일반적으로 가장 흔히 쓰이는 방법이 가중 로그순위 검정법(Weighted Logrank test)이며, 이중에서도 가중치가 모든 개체에 대해 동일한 로그순위 검정법(Logrank test)이 널리 쓰인다 (Mantel, 1966; Cox, 1972).

로그순위 검정법이 가장 널리 쓰이는 이유는 크게 두 가지로 설명될 수 있다. 첫 번째 이유는 로그순위 score인 $\hat{\theta} = \sum(O - E)$는 Mantel-Haenszel 검정 통계량의 중도 절단된 자료에 대한 형태인 동시에 Cox의 비례위험 모형에 대한 부분 우도의 score 통계량 형태라는 점이다. 두 번째 이유는 로그순위 검정통계량이 표준정규분포로 근사적으로 수렴하며, 비례위험 가정이 만족되고 중도절단이 공변량에 의존하지 않는 경우 로그순위 검정법이 근사적으로 최대효율을 갖는다는 점이다.

로그순위 검정법의 경우 결과가 꼬리 부분의 사건에 매우 민감하다는 단점이 있다. 이 점을 보완하고자 하는 경우, 각 시점의 생존확률을 가중치로 하는 가중 로그순위 검정법이 이용되는데 생존확률이 낮은

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A2008686).
[1]Corresponding author: Professor, Department of Statistics, Korea University, 145 Anam-Ro, SeongBuk-Gu, Seoul 136-701, Korea. E-mail: jael@korea.ac.kr

꼬리부분에 적은 가중치가 적용되므로 좀 더 로버스트한 결과를 얻을 수 있다 (Prentice, 1978). 이 검정법은 중도 절단된 자료에 대한 Wilcoxon의 순위합 검정과 동일하므로 Prentice-Wilcoxon 검정법이라고도 불린다.

위 두 방법은 근사적인 분포를 이용한 방법이므로 표본 크기가 작은 경우에는 근사적인 성질이 검정 결과에 반영되지 못할 수 있다. 특히 두 집단의 생존분포 비교에서는 표본 크기가 작아짐에 따라 가중 로그순위 검정법들의 경험적 오류율이 명목 오류율보다 심각하게 큰 것으로 나타났으며, 집단 간의 표본 크기가 다른 경우에 그 차이는 더욱 심각하게 나타났다 (Kellerer과 Chmelevsky, 1983).

일반적으로 표본 크기가 작은 경우에는 순열 검정법(Permutation test) 등의 리샘플링 방법(Resampling procedures)들이 대안으로 이용되는데 이는 분포에 대한 가정이 없이 관찰된 자료만으로 검정통계량의 분포를 추정하고 그 분포를 이용해 검정하는 방법이기 때문이다. 이미 두 집단의 생존 분포를 비교할 때 표본 크기가 작은 경우에는 기존의 로그순위 검정법을 이용하는 것이 유효하지 않은 것으로 나타났으며, 로그순위 검정통계량을 이용한 순열 검정법은 안정적인 범위 내의 경험적 오류율을 보임으로써 표본 크기가 작은 경우에도 이용될 수 있음이 나타났다 (Heller과 Venkatraman, 1996).

기존의 가중 로그순위 검정법의 또 한 가지 문제는 각 집단의 중도절단 분포가 동일하다는 가정이 충족되어야 한다는 것이다. 만약 그렇지 않은 것으로 의심이 되는 자료의 경우에는 기존의 로그순위 검정법뿐만 아니라 그것의 대안인 순열 검정법 또한 유효성을 장담할 수 없다. 두 집단의 생존 분포 비교에서 표본 크기가 작고 각 집단의 중도절단 분포가 동일하지 않은 것으로 의심되는 경우 순열-대치 검정법(Permutation-Imputation test)을 대안으로 생각할 수 있다. 이는 대치 단계(Imputation step)에서 귀무가설 하에서의 생존 확률이 집단에 의존하지 않도록 자료를 조정한 후 순열 검정 단계(Permutation step)를 통해 검정하는 방법이다 (Wang 등, 2010).

본 논문에서는 표본 크기가 작고 각 집단의 중도절단 분포가 다른 경우 두 집단의 생존분포를 비교하고자 할 때 이용될 수 있는 세 가지의 검정법-근사적 방법, 순열 검정법, 순열-대치 검정법-를 세 집단 이상의 생존 분포 비교를 위한 검정법으로 확장하고 각 방법을 로그순위 검정법과 Prentice-Wilcoxon 검정법에 적용해 비교해 보고자 한다. 2장에서는 각 검정법에 대하여 살펴보고 3장에서 모의실험 결과를 정리하였다.

2. 세 개 이상의 집단에 대한 생존분포의 비교

g개 집단의 생존 분포를 $S_1, S_2, \ldots, S_g$라고 할 때, 식 (2.1)의 가설에 대한 검정법들을 설명하기 위해 필요한 몇 가지 기호들을 설명하려고 한다.

$$H_0 : S_1 = S_2 = \cdots = S_g \tag{2.1}$$

$\{(X_i, \delta_i, Z_i); i = 1, \ldots, n\}$을 자료라 하고, 이 때 i번째 개체에 대하여 X_i은 관찰된 시간, δ_i은 중도절단 지시변수, 그리고 Z_i는 개체의 집단을 나타내는 변수이다. 그리고 n_1부터 n_g는 각 집단의 표본크기를 의미한다. 또한 $n = \sum_{j=1}^{g} n_j$라고 하자.

2.1. 가중 로그순위 검정법

g개 집단의 생존 분포를 비교하기 위해 일반적으로 널리 쓰이는 방법인 가중 로그순위 검정법(Weighted Logrank test)은 사건이 일어난 매 시점에서 귀무가설 하에서의 기대되는 사건의 수와 실제 관측된 사건의 수의 차이를 통해 검정하는 방법이다.

Table 2.1. Contingency Table at t_i

집단	사건의 수	남아있는 수	계
집단 1	d_{1i}	$m_{1i} - d_{1i}$	m_{1i}
집단 2	d_{2i}	$m_{2i} - d_{2i}$	m_{2i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
집단 g	d_{gi}	$m_{gi} - d_{gi}$	m_{gi}
계	d_i	$m_i - d_i$	m_i

이 때, 각 시점에서의 가중치를 어떻게 정하느냐에 따라 다양한 검정법으로 나뉠 수 있는데, 각 시점에 대해 동일한 가중치를 주는 검정법이 가중 로그순위 검정법의 특별한 경우인 로그순위 검정법(Logrank test)이다. 또한 각 시점의 생존확률을 가중치로 두어 시간이 갈수록 가중치를 줄이는 Prentice-Wilcoxon 검정법(Prentice-Wilcoxon test)이 널리 쓰이는데, 이는 이 검정법이 후반 시점의 사건에 민감한 로그순위 검정법보다 로버스트한 검정법으로 알려져 있기 때문이다. 이 외에도 다양한 검정법이 존재하지만 본 논문에서는 대표적으로 쓰이는 로그순위 검정법과 Prentice-Wilcoxon 검정법만을 다루려고 한다.

가중 로그순위 검정법의 검정통계량을 살펴보면 다음과 같다. $t_1 < t_2 < \cdots < t_k$을 집단에 상관없이 사건이 일어난 시간을 순서대로 나열한 것이라고 하면 모든 t_i에 대해 Table 2.1의 분할표를 작성할 수 있다.

각 t_i시점에서의 가중치를 w_i라고 할 때, $O_j = \sum_{i=1}^{k} w_i d_{ji}$, $E_j = \sum_{i=1}^{k} w_i m_{ji}(d_i/m_i)$이고, $O = (O_2 O_3 \cdots O_g)$, $E = (E_2 E_3 \cdots E_g)$이며 행렬 $V = \{v_{jj'}\}$, $v_{jj'} = w_i^2 \mathrm{COV}(d_{ji}, d_{j'i})$, $j, j' = 2, 3, \ldots, g$일 때, 검정통계량 T는 다음과 같다.

$$T = (O - E)' V^{-1} (O - E) \sim \chi^2_{g-1}. \tag{2.2}$$

귀무가설 하에서 검정통계량 T는 근사적으로 자유도가 $(g-1)$인 χ^2분포를 따르게 되며 T의 값이 충분히 크면 귀무가설을 기각하게 된다 (Mantel, 1966; Cox, 1972).

2.2. 순열 검정법

귀무가설에 대한 검정을 위해서는 귀무가설 하에서와 대립가설 하에서 매우 다른 값을 보이는 검정통계량을 설정하고 이 검정통계량의 귀무가설 하에서의 표본분포를 정의 내려야 한다. 어떠한 경우에는 표본분포에 대한 정의 자체가 명백하거나 근사적인 표본분포 이론을 이용하는 것이 가능하지만 그렇지 못한 경우 또한 존재한다. 예를 들면, 자료가 표본분포에 대한 가정을 충족하지 못하는 경우나 근사적인 표본분포 이론을 이용하는 것이 가능하나 표본크기가 근사적인 방법을 이용하기에는 너무 작은 경우가 그러하다. 이러한 경우 대안으로 이용될 수 있는 방법이 순열 검정법(Permutation test)이다. 이 방법은 분포에 대한 가정이 없이 관찰된 자료만을 가지고 검정통계량의 표본분포를 추정하고 검정을 수행하는 방법이기 때문이다. 검정 과정은 다음과 같으며 아래 방법은 Neuhaus (1993)의 방법을 따르기로 한다.

먼저 관찰된 자료인 $\{(X_i, \delta_i, Z_i); i = 1, \ldots, n\}$에서 (X_i, δ_i)부분을 n개 무작위로 비 복원 추출한 후, 처음 n_1개는 첫 번째 집단에, 그 다음 n_2개는 두 번째 집단에 할당하는 과정을 거쳐 추출된 n개의 개체를 차례로 각 집단에 모두 할당한다. 이렇게 생성된 새로운 자료인 $\{(X_i^\star, \delta_i^\star, Z_i^\star); i = 1, \ldots, n\}$에서 계산된 가중 로그순위 검정통계량을 $U^\star$라 하자. 위 과정을 B번 반복하면 $\{U_1^\star, U_2^\star, \ldots, U_B^\star\}$가 생성되고,

이것이 $U^\star$의 순열분포(Permutation Distribution)이다. 순열 검정법은 원래 자료인 $\{(X_i, \delta_i, Z_i); i = 1, \ldots, n\}$에서 계산된 U의 값이 $U^\star$의 순열분포의 $(1-\alpha)$분위수보다 크면 식 (2.1)의 가설을 기각하게 된다.

2.3. 순열-대치 검정법

이 논문에서 다뤄지고 있는 순열-대치 검정법(Permutation-Imputation test)은 Wang 등 (2010)의 방법을 확장한 방법이다. 순열-대치 검정법은 각 집단의 표본크기가 작고 집단 간 중도절단 분포가 다른 것으로 의심되는 경우 생존분포의 비교를 위해 개발된 방법으로 순열 검정 단계(Permutation step)와 대치 단계(Imputation step)로 이루어져 있다. 자료가 (X, δ, Z)의 형태이고 각 집단의 중도절단 분포가 $G_1, G_2, \ldots, G_g$라 할 때, 이 분포들이 모두 동일하다면 각 집단에 대하여 (X, δ)의 결합분포가 모두 동일할 것이다. 그러나 중도절단 분포들이 같지 않다면, (X, δ)의 결합분포가 Z에 의존하게 되므로 기존의 근사적방법과 순열 검정법을 이용할 수 없게 된다. 이를 극복하기 위해, 대치법(Imputation)을 이용해 관찰된 자료를 기반으로 한 g개의 새로운 자료를 생성하려고 한다.

i번째 개체에 대한 자료가 (X_i, δ_i)일 때, g개의 새로운 자료인 $(X_{i1}, \delta_{i1}), (X_{i2}, \delta_{i2}), \ldots, (X_{ig}, \delta_{ig})$가 생성되는 과정을 좀 더 구체적으로 살펴보자. i번째 개체가 q번째 집단에 속해 있을 경우, F_q의 생존분포를 따르고, G_j의 중도절단 분포를 따르는 g개의 자료, $\{(X_{ij}, \delta_{ij}); j = 1, \ldots, g\}$을 생성하게 된다. (X_{i1}, δ_{i1})가 생성되는 과정을 예로 들면, i번째 개체가 첫 번째 $Z_i = 1$집단에 속한다면 이며 관찰된 자료가 그대로 이용된다. 즉, $(X_{i1}, \delta_{i1}) = (X_i, \delta_i)$이다. $Z_i = q$, $q = 2, \ldots, g$인 경우, 생존분포 F_q을 따르고 중도절단 분포 G_1을 따르는 자료가 (X_{i1}, δ_{i1})으로 생성된다. 이 때,

$$(X_{i1}, \delta_{i1}) = \begin{cases} (X_i, 1), & \text{if } \delta_i = 1 \text{ and } \min(X_i, C_i^\star) = X_i, \\ (C_i^\star, 0), & \text{if } \delta_i = 1 \text{ and } \min(X_i, C_i^\star) = C_i^\star, \\ (C_i^\star, 0), & \text{if } \delta_i = 0 \text{ and } \min(X_i, C_i^\star) = C_i^\star, \\ (T_i^\star, 1), & \text{if } \delta_i = 0 \text{ and } \min(T_i^\star, C_i^\star) = T_i^\star, \\ (C_i^\star, 0), & \text{if } \delta_i = 0 \text{ and } X_i < C_i^\star < T_i^\star \end{cases} \tag{2.3}$$

이며, r이 uniform$(0,1)$에서 생성된 무작위 수일 때, $C_i^\star = G_1^{-1}(r)$, v가 uniform$(F_q(X_i), 1)$에서 생성된 무작위 수일 때, $T_i^\star = F_q^{-1}(v)$이다. 위 다섯 범주는 상호 배타적이며 전체를 포괄한다. 나머지 (X_{iq}, δ_{iq})도 같은 방식으로 생성될 수 있다. 만약 $F_q(\max(X_i)) < 1$로 F_q가 불완전한 분포이며 $v > F_q(\max(X_i))$이면, $T_i^\star$가 생성될 수 없다. 따라서 이러한 경우에는 $T_i^\star = \max(X_i)$로 하고 이 개체는 중도절단된 것으로 간주한다. 또한 G_j가 불완전한 분포이며 $r > G_j(\max(X \mid Z_i = j))$이면 $C_i^\star = \max(X_i \mid Z_i = j)$로 하고 이 개체는 중도절단된 것으로 간주한다.

생성된 $(X_{i1}, \delta_{i1}), (X_{i2}, \delta_{i2}), \ldots, (X_{ig}, \delta_{ig})$는 귀무가설 하에서 Z_i에 의존하지 않게 된다 (Wang 등, 2010). 이 때, $(X_{i1}, \delta_{i1}), (X_{i2}, \delta_{i2}), \ldots, (X_{ig}, \delta_{ig})$을 고정하고 순열과정(Permutation)을 통해 Z^p을 얻은 후, 다음의 변환을 통해 $(Z^p, X^{\star\star}, \delta^{\star\star})$을 생성한다.

$$(X_i^{\star\star}, \delta_i^{\star\star}) = \begin{cases} (X_i, \delta_i), & \text{if } Z_i^p = Z_i, \\ (X_{i1}, \delta_{i1}), & \text{if } Z_i^p = 1 \text{ and } Z_i^p \neq Z_i, \\ \quad \vdots & \\ (X_{ig}, \delta_{ig}), & \text{if } Z_i^p = g \text{ and } Z_i^p \neq Z_i. \end{cases} \tag{2.4}$$

식 (2.4)에서 생성된 새로운 자료인 $(Z^p,\ X^{\star\star},\ \delta^{\star\star})$에서 계산된 가중 로그순위 검정통계량을 $U^{\star\star}$라 하

자. 위 과정을 B번 반복하면 $\{U_1^{\star\star}, U_2^{\star\star}, \ldots, U_B^{\star\star}\}$가 생성되고, 이것이 $U^{\star\star}$의 순열분포이다. 순열-대치 검정법은 원래 자료인 $\{(X_i, \delta_i, Z_i); i = 1, \ldots, n\}$에서 계산된 U의 값이 $U^{\star\star}$의 순열분포의 $(1-\alpha)$분위수보다 크면 식 (2.1)의 가설을 기각하게 된다.

이 검정 방법은 $(X_{i1}, \delta_{i1}), (X_{i2}, \delta_{i2}), \ldots, (X_{ig}, \delta_{ig})$을 생성하는 과정에서 $(F_1, \ldots, F_g, G_1, \ldots, G_g)$에 대한 정보를 필요로 한다. 하지만 이는 실제로 알 수 없으므로 대신 Kaplan-Meier 추정량인 $(\hat{F}, \hat{G_1}, \ldots, \hat{G_g})$을 이용하기로 한다 (Kaplan과 Meier, 1958). 이 때, $\hat{F}$는 귀무가설 하에서의 생존시간의 누적확률 분포로 모든 $(F_1, \ldots, F_g)$을 대체한다.

3. 모의실험의 구조 및 결과

3.1. 모의실험의 구조

표본 크기가 작고 각 집단의 중도절단 분포가 다른 경우 두 집단의 생존분포를 비교하고자 할 때 이용될 수 있는 세 가지의 방법- 근사적 방법, 순열 검정법, 순열-대치 검정법 -를 세 집단 이상의 생존 분포 비교를 위한 검정법으로 확장하고 각 방법을 로그순위 검정법과 Prentice-Wilcoxon 검정법에 적용해 비교해 보고자 한다. 먼저 귀무가설 하에서의 기각률을 비교해 검정법의 유효성을 보고, 유효한 검정법들 간의 검정력을 비교하려고 한다. 가능한 검정법들은 총 6가지로 각각의 검정법들을 다음과 같이 표기하기로 하자.

1. 로그순위 검정법(Logrank test: LR)
2. 로그순위 순열 검정법(Logrank Permutation test: LRP)
3. 로그순위 순열-대치 검정법(Logrank Permutation-Imputation test: LRPI)
4. Prentice-Wilcoxon 검정법(Prentice-Wilcoxon test: PW)
5. Prentice-Wilcoxon 순열 검정법(Prentice-Wilcoxon Permutation test: PWP)
6. Prentice-Wilcoxon 순열-대치 검정법(Prentice-Wilcoxon Permutation-Imputation test: PWPI)

통계량의 비교를 위한 상황 설정은 Heller과 Venkatraman (1996)을 참고하여 다음의 5가지 상황을 설정하였다. 각 검정법의 유효성 및 검정력에 영향을 미칠 것으로 보이는 요인은 표본 크기, 집단 간 중도절단분포의 동일성 여부, 집단 간 표본 크기의 동일성 여부이며, 각 요인에 변화를 주어 설정된 상황은 다음과 같다.

1. 집단 간 표본 크기가 동일하고 작은 경우
2. 집단 간 표본 크기가 다른 경우
3. 집단 간 표본 크기가 동일하고 크며 중도절단 분포가 다른 경우
4. 집단 간 표본 크기가 동일하고 작으며 중도절단 분포가 다른 경우
5. 집단 간 표본 크기가 다르고 중도절단분포가 다른 경우

각 상황에 대하여 생존분포의 표본추출에는 변동계수가 2, 0.5, 1인 와이블분포를 사용하였고, 중도절단분포의 표본추출에는 균일분포를 사용하였다. 각 상황에 사용된 분포들은 모의실험 결과표에서 결과와 함께 제시하기로 한다. 순열 검정법과 순열-대치 검정법에서의 표본추출 반복은 2000회로 하였고, 모든 검정은 유의수준 0.05하에서 각각 1000번씩의 반복을 거쳐 기각한 횟수를 측정하였다.

Table 3.1. Empirical Type I error and power: Cases with equal sample sizes and equal censoring distributions

	변동계수	LR	LRP	LRPI	PW	PWP	PWPI
중도절단분포 {(3, 12), (3, 12), (3, 12)} 표본크기 (30, 30, 30)	Empirical Type I Error						
	2	0.047	0.047	0.042	0.046	0.046	0.041
	0.5	0.059	0.056	0.052	0.059	0.057	0.052
	1	0.058	0.055	0.042	0.059	0.053	0.047
	Empirical Power						
	2	0.186	0.177	0.172	0.184	0.178	0.173
	0.5	0.427	0.412	0.421	0.403	0.393	0.396
	1	0.180	0.175	0.170	0.164	0.158	0.160
중도절단분포 {(3, 12), (3, 12), (3, 12)} 표본크기 (12, 12, 12)	Empirical Type I Error						
	2	0.054	0.041	0.038	0.054	0.052	0.048
	0.5	0.057	0.050	0.047	0.057	0.048	0.049
	1	0.065	0.057	0.055	0.063	0.057	0.056
	Empirical Power						
	2	0.187	0.178	0.173	0.173	0.170	0.163
	0.5	0.365	0.355	0.358	0.352	0.345	0.350
	1	·	0.194	0.199	0.184	0.179	0.183

3.2. 모의실험의 결과

각 표에 표기된 표본 크기 (a, b, c)는 차례로 집단 1, 집단 2, 집단 3의 표본 크기를 나타내며, 중도절단 분포를 나타내는 균일분포$\{(d, e), (f, g), (h, i)\}$는 차례로 집단 1, 집단 2, 집단 3의 중도절단 분포인 균일분포를 나타낸다. 이 때, 균일분포$(7, 20)$은 약 0.2의 중도절단 비율을 보이는 분포이고, 균일분포$(3, 12)$는 약 0.5, 균일분포$(0, 9)$는 약 0.8의 중도절단 비율을 보이는 분포이다. 유의수준이 0.05일 때, 귀무가설 하에서 0.036이상, 0.064이하의 기각률을 보이지 않는 검정법은 기대되는 제 1종 오류의 확률에서 벗어나 있으므로 검정력 비교에서 제외하기로 한다. 3.2.1–3.2.5의 모의실험 결과와의 비교를 위해 표본 크기가 작거나 중도절단 분포가 이질적이지 않은 모의실험 결과를 Table 3.1에서 제시하였다.

3.2.1. 집단 간 표본 크기가 동일하고 작은 경우 (Table 3.1) 각 집단의 표본 크기가 큰 30인 경우에는 여섯 가지의 모든 검정법에서 유의수준인 0.05 근처의 확률로 기각이 일어남을 확인할 수 있다. 또한 표본 크기가 12로 줄어도 대부분의 상황에서 범위 내의 기각률을 보인다. 그러나 변동계수가 1이고 표본 크기가 작은 경우 LR 검정법의 귀무가설 하에서의 기각률이 0.036이상, 0.064이하의 범위에서 벗어나 있으므로 LR 검정법은 검정력 비교에서 제외된다.

각 집단의 표본 크기가 크고 중도절단 분포가 동일한 경우 근사적인 방법인 LR 검정법이 가장 높은 검정력을 보인다. 각 집단의 표본 크기가 작고 중도절단 분포가 동일한 경우 변동계수가 2일 때에는 LR, 변동계수가 1일 때에는 LR을 제외한 검정법 중 LRPI 검정법이 최강 검정력을 갖는다.

3.2.2. 집단 간 표본 크기가 다른 경우 (Table 3.2) 표본 크기가 (12, 12, 30)인 경우와 (12, 30, 30)인 경우 모두 기각률 0.05 근처의 값을 보임으로써 유효성에 문제가 없는 것으로 나타났다. 따라서 집단 간 표본 크기가 달라도 제시된 검정법을 사용하는 것이 문제가 되지 않음을 알 수 있다.

중도절단 분포가 동일하고 한 집단, 혹은 두 집단의 표본 크기가 작은 경우에는 변동계수와 상관없이 LR 검정법이 가장 높은 검정력을 보였다. 따라서 세 집단 이상의 생존분포를 비교하는 경우에는 표본 크기가 집단마다 이질적이라고 해도 근사적 방법을 이용하는 것이 바람직하다고 할 수 있겠다.

Table 3.2. Empirical Type I error and power: Cases with unequal sample sizes and equal censoring distributions

	변동계수	LR	LRP	LRPI	PW	PWP	PWPI
중도절단분포 {(3, 12), (3,12), (3,12)} 표본크기 (12, 12, 30)	Empirical Type I Error						
	2	0.053	0.039	0.036	0.048	0.043	0.040
	0.5	0.054	0.047	0.045	0.054	0.047	0.047
	1	0.047	0.040	0.042	0.044	0.041	0.039
	Empirical Power						
	2	0.153	0.137	0.141	0.145	0.139	0.138
	0.5	0.323	0.315	0.316	0.323	0.318	0.321
	1	0.167	0.157	0.157	0.158	0.153	0.153
중도절단분포 {(3, 12),(3, 12),(3, 12)} 표본크기 (12, 30, 30)	Empirical Type I Error						
	2	0.056	0.050	0.045	0.051	0.045	0.046
	0.5	0.064	0.056	0.053	0.057	0.052	0.053
	1	0.055	0.048	0.047	0.046	0.039	0.040
	Empirical Power						
	2	0.167	0.158	0.149	0.155	0.147	0.147
	0.5	0.328	0.314	0.317	0.326	0.310	0.319
	1	0.159	0.144	0.152	0.144	0.138	0.138

3.2.3. 집단 간 표본 크기가 동일하고 크며 중도절단 분포가 다른 경우 (Table 3.3) 모든 검정법이 집단 간 중도절단 분포가 다른 모든 상황에서 기각률 0.05 근처의 값을 보임으로써 검정력을 비교하는 데에 문제가 없는 것으로 나타났다. 따라서 집단 간 중도절단 분포가 달라도 표본 크기가 충분히 크면 제시된 검정법을 사용하는데 문제가 없음을 알 수 있다.

이 경우 변동계수가 1, 2일 때에는 LR 검정법이 가장 높은 검정력을 보였으나 변동계수가 0.5일 때에는 LR 검정법과 LRP, LRPI 검정법의 검정력이 같게 나타났다.

3.2.4. 집단 간 표본 크기가 동일하고 작으며 중도절단 분포가 다른 경우 (Table 3.3) 표본 크기가 (12, 12, 12)이며 중도절단 분포가 균일분포{(0, 9), (7, 20), (7, 20)}인 경우 LR, LRP LRPI 모두 범위 내의 기각률을 보이지 않았다. 반면, 이 상황에서도 PW, PWP, PWPI는 0.05 근처의 기각률을 보였다. 따라서 표본 크기가 작고 중도절단 분포가 이질적인 것으로 의심되는 경우에는 PW, PWP, PWPI 방법을 사용하는 것이 바람직하다.

이 경우 대부분의 상황에서 근사적인 방법의 검정력이 가장 높게 나타났지만 변동계수가 0.5인 경우 PW의 검정력보다 PWPI의 검정력이 더욱 높게 나타났다. 귀무가설 하에서의 기각률은 PWPI가 오히려 낮게 나타났으므로 이 경우에는 PWPI가 PW보다 강력한 검정법임을 알 수 있다.

3.2.5. 집단 간 표본 크기가 다르고 중도절단분포가 다른 경우 (Table 3.4) 모든 검정법이 집단 간 표본 크기도 다르고 중도절단 분포 또한 다른 상황에서 기각률 0.05 근처의 값을 보임으로써 검정력을 비교하는 데에 문제가 없는 것으로 나타났다.

이 경우에 대한 상황을 두 가지로 나누어 비교해 보았다. 첫 번째는 표본 크기가 작은 집단의 중도절단 비율이 낮고, 표본 크기가 큰 집단의 중도절단 비율이 높은 상황으로, 이 경우 역시 근사적인 방법의 검정력이 가장 높게 나타났으나 LR보다는 PW의 검정력이 대부분 높게 나타났다. 두 번째는 표본 크기가 작은 집단의 중도절단 비율이 높고, 표본 크기가 큰 집단의 중도절단 비율이 낮은 상황으로, 이 경우에는 LR의 검정력이 항상 높게 나타남을 알 수 있다.

Table 3.3. Empirical Type I error and power: Cases with equal sample sizes and unequal censoring distributions

	변동계수	LR	LRP	LRPI	PW	PWP	PWPI
중도절단분포 {(0, 9), (7, 20), (7, 20)} 표본크기 (30, 30, 30)	Empirical Type I Error						
	2	0.062	0.054	0.041	0.060	0.059	0.045
	0.5	0.062	0.050	0.050	0.057	0.049	0.052
	1	0.062	0.052	0.050	0.054	0.052	0.046
	Empirical Power						
	2	0.208	0.197	0.186	0.193	0.184	0.185
	0.5	0.462	0.453	0.462	0.440	0.438	0.443
	1	0.239	0.237	0.229	0.237	0.235	0.229
중도절단분포 {(0, 9), (3, 12), (7, 20)} 표본크기 (30, 30, 30)	Empirical Type I Error						
	2	0.058	0.053	0.044	0.054	0.052	0.044
	0.5	0.047	0.040	0.047	0.045	0.044	0.042
	1	0.044	0.037	0.039	0.046	0.045	0.042
	Empirical Power						
	2	0.098	0.083	0.083	0.096	0.088	0.091
	0.5	0.164	0.164	0.157	0.159	0.152	0.152
	1	0.106	0.097	0.095	0.098	0.092	0.095
중도절단분포 {(0, 9), (7, 20), (7, 20)} 표본크기 (12, 12, 12)	Empirical Type I Error						
	2	0.030	0.023	0.023	0.039	0.038	0.037
	0.5	0.054	0.045	0.047	0.051	0.044	0.046
	1	0.059	0.049	0.046	0.055	0.054	0.050
	Empirical Power						
	2	.	.	.	0.085	0.077	0.081
	0.5	0.230	0.198	0.208	0.214	0.206	0.201
	1	0.125	0.110	0.110	0.113	0.103	0.102
중도절단분포 {(0, 9), (3, 12), (7, 20)} 표본크기 (12, 12, 12)	Empirical Type I Error						
	2	0.053	0.040	0.039	0.049	0.046	0.043
	0.5	0.042	0.038	0.039	0.046	0.042	0.042
	1	0.057	0.047	0.049	0.059	0.056	0.052
	Empirical Power						
	2	0.121	0.104	0.106	0.101	0.092	0.090
	0.5	0.200	0.182	0.184	0.175	0.169	0.177
	1	0.118	0.108	0.104	0.111	0.104	0.106

4. 결론 및 토의

세 개 이상의 집단에 대한 생존분포의 동일성을 검정하기 위해 가중 로그순위 검정법(Weighted Logrank test)과 그의 특별한 경우인 로그순위 검정법(Logrank test)이 널리 쓰인다. 그러나 이 방법은 근사적인 분포를 이용한 방법이므로 표본 크기가 작은 경우에는 관찰된 자료만으로 검정통계량의 분포를 추정하고 그 분포를 이용해 검정하는 순열 검정법(Permutation test)이 제안되었으며, 또한 중도절단 분포가 동일하지 않은 것으로 의심되는 경우에는 순열 검정법을 향상시킨 검정법인 순열-대치 검정법(Permutation-Imputation test)이 제안되었다. 이 검정법은 대치 단계(Imputation step)에서 귀무가설 하에서의 생존확률이 집단에 의존하지 않도록 자료를 조정한 후 순열 검정 단계(Permutation step)를 통해 검정하는 방법이다.

두 집단의 생존분포의 동일성을 검정하기 위해 제안된 기존의 순열 검정법과 순열-대치 검정법을 세 집

Table 3.4. Empirical Type I error and power: Cases with unequal sample sizes and unequal censoring distributions

	변동계수	LR	LRP	LRPI	PW	PWP	PWPI
중도절단분포 {(7, 20), (7, 20), (0, 9)} 표본크기 (12, 12, 30)	Empirical Type I Error						
	2	0.056	0.039	0.036	0.046	0.036	0.039
	0.5	0.051	0.043	0.042	0.061	0.057	0.052
	1	0.048	0.036	0.038	0.051	0.047	0.048
	Empirical Power						
	2	0.120	0.099	0.102	0.127	0.116	0.121
	0.5	0.236	0.198	0.202	0.226	0.212	0.197
	1	0.139	0.119	0.123	0.143	0.127	0.136
중도절단분포 {(7, 20), (0, 9), (0, 9)} 표본크기 (12, 30, 30)	Empirical Type I Error						
	2	0.049	0.039	0.037	0.048	0.046	0.045
	0.5	0.056	0.044	0.042	0.055	0.047	0.049
	1	0.050	0.046	0.041	0.052	0.048	0.050
	Empirical Power						
	2	0.109	0.095	0.100	0.110	0.104	0.102
	0.5	0.169	0.155	0.143	0.169	0.162	0.153
	1	0.126	0.117	0.114	0.133	0.124	0.121
중도절단분포 {(0, 9), (0, 9), (7, 20)} 표본크기 (12, 12, 30)	Empirical Type I Error						
	2	0.060	0.042	0.039	0.058	0.048	0.051
	0.5	0.058	0.046	0.050	0.055	0.052	0.051
	1	0.054	0.044	0.046	0.057	0.047	0.045
	Empirical Power						
	2	0.164	0.150	0.148	0.154	0.138	0.140
	0.5	0.271	0.254	0.252	0.263	0.253	0.250
	1	0.155	0.139	0.144	0.144	0.138	0.141
중도절단분포 {(0, 9), (7, 20), (7, 20)} 표본크기 (12, 30, 30)	Empirical Type I Error						
	2	0.061	0.054	0.045	0.049	0.048	0.045
	0.5	0.055	0.046	0.046	0.054	0.053	0.053
	1	0.063	0.060	0.057	0.057	0.057	0.057
	Empirical Power						
	2	0.142	0.127	0.118	0.131	0.124	0.123
	0.5	0.362	0.346	0.351	0.350	0.344	0.340
	1	0.165	0.150	0.146	0.148	0.140	0.135

단 이상의 생존분포의 동일성 검정법으로 확장한 후 가중 로그순위 검정법을 포함한 세 가지 검정법의 검정력을 비교하기 위한 모의실험을 수행하였다. 모의실험에는 로그순위 검정법과 Prentice-Wilcoxon 검정법이 적용되었고 그 결과, 모든 집단의 표본 크기가 작을 경우 로그순위 검정법은 사용하지 않는 것이 바람직한 것으로 나타났으며, 이 경우 로그순위 검정법을 제외한 검정법 중 순열-대치 검정법의 검정력이 높게 나타났다. 또한 여러 집단 중 한 집단이라도 표본 크기가 클 경우에는 근사적 방법인 로그순위 검정법과 Prentice-Wilcoxon 검정법의 검정력이 우세했다. 중도절단 분포가 이질적인 경우에는 표본 크기가 큰 경우에도 순열 검정법과 순열-대치 검정법이 근사적 방법과 비등한 검정력을 보였으며 표본 크기가 작은 경우에는 순열-대치 검정법이 우세한 결과를 보인 상황이 발생했다.

두 집단의 생존분포의 비교에서와 마찬가지로 표본 크기가 작은 경우 근사적 방법을 이용하는 것이 바람직하지 않은 것으로 나타났다. 또한 중도절단 분포가 이질적인 경우 순열-대치 검정법의 검정력이 우세하게 나타났으므로 순열-대치 검정법의 이용이 추천된다. 그러나 집단 간 표본 크기가 다르고 중도절단

분포 또한 이질적인 경우 근사적인 방법이 문제를 드러낼 것으로 간주되었지만 모의실험 결과 귀무가설 하에서의 기각률이 모두 안정적이었고, 검정력 또한 높게 나타났다. 이는 세 개 집단 이상의 생존분포의 동일성 검정에서는 한 집단이라도 표본 크기가 크면 근사적 방법을 이용하는데 문제가 없음을 보여준다고 할 수 있겠다.

References

Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Heller, G. and Venkatraman, E. S. (1996). Resampling procedures to compare two survival distributions in the presence of right-censored data, *Biometrics*, **52**, 1204–1213.

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

Kellerer, A. M. and Chmelevsky, D. (1983). Small-Sample properties of censored-data rank tests, *Biometrics*, **39**, 675–682.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its considerations, *Cancer Chemotherapy Reports*, **50**, 163–170.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship, *The Annals of Statistics*, **21**, 1760–1779.

Prentice, R. L. (1978). Linear rank tests with censored data, *Biometrika*, **65**, 167–179.

Wang, R., Lagakos, S. W. and Gray, R. J. (2010). Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring, *Biostatistics*, **11**, 676–692.

이질적인 중도절단분포 하에서 생존분포의 동일성 검정법 비교연구

송수정[a] · 이재원[b,1]

[a]유한양행 중앙연구소 임상PV팀, [b]고려대학교 통계학과

(2013년 9월 10일 접수, 2013년 11월 11일 수정, 2013년 11월 11일 채택)

요 약

세 개 이상의 집단에 대한 생존분포의 비교를 위해 가중 로그순위 검정법(Weighted Logrank test)과 그의 특별한 경우인 로그순위 검정법(Logrank test)이 널리 쓰인다. 그러나 이 방법은 근사적인 분포를 이용한 방법이므로 표본 크기가 작은 경우에는 유효하지 못할 수 있으며, 각 집단의 중도절단 분포가 동일하다는 가정 또한 충족되어야 하기 때문에 이 가정이 충족되지 못할 경우에도 검정법의 유효성을 장담할 수 없다. 표본 크기가 작은 경우에 대한 대안으로, 분포에 대한 가정이 없이 관찰된 자료만으로 검정통계량의 분포를 추정하고 그 분포를 이용해 검정하는 순열 검정법(Permutation test)이 제안되었으나, 순열 검정법 또한 각 집단의 중도절단 분포가 동일하다는 가정이 충족되어야 한다. 따라서 순열 검정법을 향상시킨 순열-대치 검정법(Permutation-Imputation test)이 대안이 될 수 있는데, 이는 대치 단계(Imputation step)에서 귀무가설 하에서의 생존확률이 집단에 의존하지 않도록 자료를 조정한 후 순열 검정 단계(Permutation step)를 통해 검정하는 방법이다. 본 논문에서는 근사적 방법, 순열 검정법, 순열-대치 검정법을 로그순위 검정법과 가중 로그순위 검정법의 한 형태인 Prentice-Wilcoxon 검정법에 적용해 각 검정법의 유효성과 검정력을 비교하였다.

주요용어: 순열검정법, 순열-대치검정법, 생존분석.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2012R1A1A2008686).

[1]교신저자: (136-701) 서울특별시 성북구 안암로 145, 고려대학교 통계학과, 교수. E-mail: jael@korea.ac.kr