응용통계연구 제 8권 2호
1995년 9월 pp. 109-119.

임의로 관측중단된 두 표본 자료에 대한 카이제곱 검정방법[1)]

김 주 한[2)], 김 정 란[3)]

요 약

두 모집단에서 임의로 관측중단된 두 표본을 얻었을 때, 두 모집단의 분포가 같다는 가설을 검정하기 위한 카이제곱 검정방법이 제안되었다. 여기서 제안된 통계량은 대립가설이 두 모집단의 분포가 같지않다는 양측가설일 때 쓰일 수 있다. 귀무가설이 사실일 때 제안된 통계량의 극한분포는 카이제곱 분포가 된다. 두 가지 형태의 카이제곱 검정통계량이 제안되었는데, 하나는 product-limit 추정치로 부터 얻은 관측된 칸(cell) 확률의 차이들의 벡터의 이차형식으로 표현된 것이고, 다른 하나는 간단한 합의 모양으로 표현된 것이다. 두 형태의 검정통계량을 사용하여 암치료를 위한 화학요법 실험으로 부터 얻은 자료를 분석하여 보았다.

1. 서 론

임의로 관측중단된(randomly censored) 두 개의 표본의 분포를 비교하는 방법으로는 일반화된 Wilcoxon 검정법(Gehan(1965), Efron(1967))과 로그 순위 검정법(Mantel(1967), Peto and Peto(1972))이 많이 이용되고 있다. 이러한 검정법들은 대립가설이 한 분포가 다른 분포보다 확률적으로 크다는 단측가설이나 한 분포가 다른 분포보다 확률적으로 크거나 또는 작다는 양측가설에는 사용될 수 있지만, 대립가설이 단순히 두 분포가 같지 않다는 양측가설일 경우는 적당하지 않다. 두 분포가 같다는 가설을 같지 않다는 양측가설에 대하여 검정하는 방법으로 Koziol(1978)은 각 표본의 product-limit 추정치를 이용한 두 표본 Cramer-von Mises 형태의 통계량을 제안하였고, Koziol and Yuh(1982)는 두 product-limit 추정치의 차이를 이용하여 Kolmogorov-Smirnov, Kuiper, Cramer-von Mises 형태의 통계량을 제안하였다. Koziol과 Yuh는 그들이 제안한 통계량들의 극한분포가 관측중단되지 않은 자료에 사용되는 통계량들의 극한분포의 절단된 형태로 됨을 보였고, 두 표본의 관측중단 분포(censoring distribution)가 같은 경우에도 같은 방법이 적용될 수 있음을 보여주었다.

본 논문에서는 이러한 검정 방법들에 대한 대안으로 임의로 관측중단된 두 표본의 분포를 비교하는 카이제곱 검정방법을 제안하고자 한다. 새로이 제안될 검정통계량은 두 분포가 같다는 가정 하에 극한적으로 카이제곱 분포를 하게 되고 계산하기 쉬운 합의 모양으로 표현되기 때문에 다른 통계량보다 손쉽게 사용할 수 있다는 장점이 있다.

---

1) 이 연구는 1992년도 한국과학재단 연구비지원에 의한 결과임(과제번호 : 923-0100-002-1).

2) (305-764) 대전직할시 유성구 궁동 220번지, 충남대학교 자연과학대학 통계학과.

3) (305-764) 대전직할시 유성구 궁동 220번지, 충남대학교 자연과학대학 통계학과.

2. 카이제곱 검정통계량

$X_1^0, X_2^0, \cdots, X_m^0$ 은 서로 독립이고 동일한 연속분포함수 F_0을 따르는 양의 확률변수들이고, $Y_1^0, Y_2^0, \cdots, Y_n^0$ 은 서로 독립이고 동일한 연속분포함수 G_0을 따르는 양의 확률변수들이라 하자. X_i^0 과 Y_j^0 도 서로 독립이다. X_i^0 과 Y_j^0 은 각각 독립인 확률변수 $X_i^c\,(1\le i\le m)$ 와 $Y_j^c\,(1\le j\le n)$ 에 의해 오른쪽으로 부터 관측중단되어 X_i^0 과 Y_j^0 은 모두 직접 관찰될 수 없고, 대신에 우리는 임의로 관측중단된 자료를 얻게 된다. 여기서 양의 확률변수 $X_i^c\,(1\le i\le m)$ 와 $Y_j^c\,(1\le j\le n)$ 는 관측중단변수라 부르고, 서로 독립이며 각각 연속분포함수 F_c와 G_c를 따르는 모집단에서 추출된 확률 표본이다. 우리가 관찰하여 얻을 수 있는 임의로 관측중단된 두 표본 자료는 $(X_i,\ \delta_i),\ i=1,\cdots,m,\ X_i=\min(X_i^0,\ X_i^c),\ \delta_i=I(X_i=X_i^0)$ 와 $(Y_j,\ \varepsilon_j),\ j=1,\ldots,n,\ Y_j=\min(Y_j^0,\ Y_j^c),\ \varepsilon_j=I(Y_j=Y_j^0)$ 이다. 관찰된 X_i 와 Y_j 는 각각 분포함수 $H_X=1-(1-F_0)(1-F_c)$ 와 $H_Y=1-(1-G_0)(1-G_c)$ 를 따르는 모집단에서 얻은 확률 표본이라 생각할 수 있다. 우리가 검정하고자 하는 귀무가설은 $H_0:\ F_0=G_0$ 이며 대립가설은 $H_1:\ F_0\neq G_0$ 이다.

카이제곱 검정통계량을 유도하는 기본적인 방법은 구간 $(0,\ \infty)$ 를 $k+1$ 개의 칸(cell)으로 나누고 각 칸에서 관측된 칸 확률을 비교하는 것이다. 칸의 경계를 $0=a_0<a_1<\cdots<a_k<a_{k+1}=\infty$ 라 하면, i번째 칸에서 관측된 칸 확률은 $\hat{F}_0(a_i)-\hat{F}_0(a_{i-1})$ 과 $\hat{G}_0(a_i)-\hat{G}_0(a_{i-1})$ 이다. 여기서 $\hat{F}_0$ 과 $\hat{G}_0$ 은 각각 F_0 와 G_0 의 product-limit 추정치이다. 두 개의 관측된 칸 확률의 차이를 구해보면

$$\begin{aligned}&(\ \hat{F}_0(a_i)-\hat{F}_0(a_{i-1}))-(\ \hat{G}_0(a_i)-\hat{G}_0(a_{i-1}))\\&=(\ \hat{F}_0(a_i)-\hat{G}_0(a_i))-(\ \hat{F}_0(a_{i-1})-\hat{G}_0(a_{i-1}))\end{aligned}$$

이고, 카이제곱 통계량을 얻기 위해 우리는 k개의 관측된 칸 확률 차이의 결합분포를 알아야 한다.

$X_1, X_2, \cdots, X_m$과 $Y_1, Y_2, \cdots, Y_n$이 크기 순서로 되어 있다고 가정하고 F_0와 G_0의 product-limit 추정치를 다음과 같이 정의하자.

$$\hat{F}_0(t)\ =\ \begin{cases}0 & ,\ t<X_1\\ 1-\prod_{i:X_i\le t}\left[\ (m-i)/(m-i+1)\right]^{\delta_i} & ,\ X_1\le t<X_m\\ 1 & ,\ t\ge X_m\end{cases}$$

$$\hat{G}_0(t) = \begin{cases} 0 & , \ t < Y_1 \\ 1 - \prod_{j:Y_j \le t} [\ (n-j)/(n-j+1)]^{\varepsilon_j} & , \ Y_1 \le t < Y_n \\ 1 & , \ t \ge Y_n \end{cases}$$

Breslow and Crowley(1974)와 Gill(1983)의 결과에 의하면 확률과정 $\sqrt{m}(\ \hat{F}_0(t) - F_0(t))$ 와 $\sqrt{n}(\ \hat{G}_0(t) - G_0(t))$는 각각 m과 n이 ∞로 접근할 때 평균이 0이고 다음과 같은 공분산 함수를 가지는 가우스 확률과정(Gaussian process) $Z_1(t)$와 $Z_2(t)$로 수렴한다.

$$\begin{aligned} Cov(Z_1(s),\ Z_1(t)) &= (1-F_0(s))\ (1-F_0(t))\ \alpha(s)\ , \qquad 0 < s \le t \\ \alpha(s) &= \int_0^s \frac{dF_0}{(1-F_0)^2(1-F_c)} \end{aligned}$$

$$\begin{aligned} Cov(Z_2(s),\ Z_2(t)) &= (1-G_0(s))\ (1-G_0(t))\ \beta(s)\ , \qquad 0 < s \le t \\ \beta(s) &= \int_0^s \frac{dG_0}{(1-G_0)^2(1-G_c)} \end{aligned}$$

이제 관측된 칸 확률 차이의 분포를 알아보기 위해 확률과정

$$Z_{mn}(t) = \sqrt{\frac{mn}{(m+n)}}\ (\ \hat{F}_0(t) - \hat{G}_0(t))$$

를 생각해 보자.

$$\begin{aligned} Z_{mn}(t) &= \sqrt{\frac{mn}{m+n}}\ (\ \hat{F}_0(t) - \hat{G}_0(t)) \\ &= \sqrt{\frac{n}{m+n}}\sqrt{m}\ (\ \hat{F}_0(t) - F_0(t)) - \sqrt{\frac{m}{m+n}}\sqrt{n}\ (\ \hat{G}_0(t) - G_0(t)) \\ &\quad + \sqrt{\frac{mn}{m+n}}\ (F_0(t) - G_0(t)) \end{aligned}$$

이므로 만약 m과 n이 ∞로 접근하고 그들의 비 m/n이 어떤 양수 λ에 가까이 가고 귀무가설 $H_0: F_0 = G_0$가 사실이라면 $Z_{mn}(t)$는 구간 $[0,\ T)$ 에서, 여기서 $T = \sup\{t : F_0(t) < 1\}$, 평균이 0이고 다음과 같은 공분산함수를 가지는 가우스 확률과정 $Z(t)$로 수렴하게 된다.

$$Cov(Z(s),\ Z(t)) = \frac{1}{1+\lambda}\, q(s)\ q(t)\ \alpha(s) + \frac{\lambda}{1+\lambda}\, r(s)\ r(t)\ \beta(s)\ , \quad 0 \le s < t < T$$

여기서 $q(s)=1-F_0(s)$이고 $r(s)=1-G_0(s)$이다.

그러므로 각 칸의 경계에서 $Z_{mn}(t)$의 값 $Z_{mn}(a_i)$를 w_i $(i=1,\cdots,k)$ 라 하면 확률 벡터 $W=(w_1,\cdots,w_k)'$ 은 $Z_{mn}(t)$의 수렴성에 의해 평균이 0이고 분산-공분산 행렬이 Σ_w인 다변량 정규분포로 수렴하게 된다. Σ_w 의 원소는 다음과 같이 표현될 수있다.

$$\sigma_w(i,j) = \frac{1}{1+\lambda} q(a_i)\, q(a_j)\, \alpha(a_i) + \frac{\lambda}{1+\lambda} r(a_i)\, r(a_j)\, \beta(a_i), \quad i \le j .$$

이제 $v_i = Z_{mn}(a_i) - Z_{mn}(a_{i-1}) = w_i - w_{i-1}$, $i=1, \cdots, k$, $w_o=0$, 으로 정의하면 v_i는 i번째 칸에서 표준화된 관측된 칸 확률의 차이가 된다. 확률벡터 $V=(v_1,\cdots,v_k)'$ 은

$$V = AW = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix} W$$

와 같이 표현되고, W의 수렴성에 의해 V는 평균이 0이고 분산-공분산 행렬이 $\Sigma_v = A\,\Sigma_w\,A'$ 인 정규분포로 수렴하게 된다. 그러므로 V의 이차형식 $Q = V'\Sigma_v^{-1}\,V$ 는 자유도가 k인 카이제곱 분포로 수렴하게 된다. Q 를 W 로 표현하면

$$Q = V'\Sigma_v^{-1}\,V = (AW)'(A\,\Sigma_w\,A')^{-1}(AW) = W'\Sigma_w^{-1}\,W \qquad (1)$$

가 된다. Σ_w는 우리가 모르는 F_0, F_c, G_0, G_c, λ 등에 의해 결정되므로, Q 를 검정통계량으로 사용하기 위하여 Σ_w 의 일치추정치 $\hat{\Sigma}_w$을 구하여야 한다. $\hat{\Sigma}_w$는 F_0, F_c, G_0, G_c, λ 대신에, $\hat{F}_0$, $\hat{F}_c$, $\hat{G}_0$, $\hat{G}_c$, $\hat{\lambda}=m/n$ 을 사용하여 구할 수 있다. 즉,

$$\hat{\sigma}_w(i,j) = \frac{n}{m+n} \hat{q}(a_i)\, \hat{q}(a_j)\, \hat{\alpha}(a_i) + \frac{m}{m+n} \hat{r}(a_i)\, \hat{r}(a_j)\, \hat{\beta}(a_i), \quad i \le j$$

$$\hat{q}(a_i) = 1 - \hat{F}_0(a_i), \quad \hat{r}(a_i) = 1 - \hat{G}_0(a_i)$$

$$\hat{\alpha}(a_i) = \int_0^{a_i} \frac{d\,\hat{F}_0}{(1-\hat{F}_0)^2(1-\hat{F}_c)}, \quad \hat{\beta}(a_i) = \int_0^{a_i} \frac{d\,\hat{G}_0}{(1-\hat{G}_0)^2(1-\hat{G}_c)} .$$

$\hat{\alpha}(s)$와 $\hat{\beta}(s)$는 적분으로 표시되었지만 실제로 $\hat{F}_0$과 $\hat{G}_0$이 계단함수(step function)이기 때문에 좀 더 간단한 합의 모양으로 표현될 수 있다.

$$\hat{\alpha}(s) = \int_0^s \frac{d\hat{F}_0}{(1-\hat{F}_c)(1-\hat{F}_0)^2} = \int_0^s \frac{1}{1-\hat{F}_c} d\left(\frac{1}{1-\hat{F}_0}\right)$$
$$= \sum_{i:x_i \le s} \frac{1}{1-\hat{F}_c(x_i)} \left(\frac{1}{1-\hat{F}_0(x_i)} - \frac{1}{1-\hat{F}_0(x_{i-1})} \right)$$
$$= \sum_{i:x_i \le s} \left(\frac{1}{\prod_{j=1}^{i} (\frac{m-j}{m-j+1})} - \frac{1}{\prod_{j=1}^{i-1} (\frac{m-j}{m-j+1})} \cdot \frac{1}{(\frac{m-i}{m-i+1})^{1-\delta_i}} \right)$$

여기서 $\delta_i = 0$이면 합산기호 안의 항이 0이 되므로

$$\hat{\alpha}(s) = \sum_{i:x_i \le s} \left(\frac{m}{m-i} - \frac{m}{m-i+1} \right) \delta_i = \sum_{i:x_i \le s} \frac{m}{(m-i)(m-i+1)} \delta_i$$

가 되고, 같은 방법으로

$$\hat{\beta}(s) = \sum_{i:y_i \le s} \frac{n}{(n-i)(n-i+1)} \varepsilon_i$$

가 된다. 위의 추정치들을 사용하여 얻은 Σ_w의 추정치 $\hat{\Sigma}_w$을 (1)의 식에 대입하면 $H_0: F_0 = G_0$를 검정하기 위한 카이제곱 통계량 $\hat{Q}$는 다음과 같이 표현된다.

$$\hat{Q} = W' \hat{\Sigma}_w^{-1} W$$

$\hat{Q}$ 는 H_0이 사실일 때 근사적으로 자유도가 k인 카이제곱 분포를 하게 된다.

$\hat{Q}$ 를 계산하는 것이 어려운 문제는 아니지만, $H_0: F_0 = G_0$이 사실일 때 좀 더 간단한 합의 모양으로 나타내지는 카이제곱 통계량을 다음과 같이 얻을 수 있다.

H_0이 사실일 때 X_i^0와 Y_j^0의 공통 분포함수를 $F^* = F_0 = G_0$라 하면 $\sigma_w(i,j)$는 다음과 같이 표현될 수 있다.

$$\sigma_w(i,j) = d(a_i)\, d(a_j)\, c(a_i)$$
$$c(a_i) = \frac{1}{1+\lambda} \alpha(a_i) + \frac{\lambda}{1+\lambda} \beta(a_i)$$
$$d(a_i) = 1 - F^*(a_i)$$

$d(a_i) = d_i$, $c(a_i) = c_i$로 간단히 표현하면

$$\Sigma_w = \begin{bmatrix} c_1 d_1^2 & c_1 d_1 d_2 & \cdots & \cdots & c_1 d_1 d_k \\ & c_2 d_2^2 & \cdots & \cdots & c_2 d_2 d_k \\ & \text{대 칭} & & \ddots & \vdots \\ & & & & c_k d_k^2 \end{bmatrix}$$

이 되고, $\Sigma_w = T'T$ 가 되는 삼각행렬 T는 다음과 같이 주어진다.

$$T = \begin{bmatrix} d_1\sqrt{c_1} & d_2\sqrt{c_1} & \cdots & \cdots & d_k\sqrt{c_1} \\ & d_2\sqrt{c_2-c_1} & \cdots & \cdots & d_k\sqrt{c_2-c_1} \\ & 0 & \ddots & & \vdots \\ & & & & d_k\sqrt{c_k-c_{k-1}} \end{bmatrix}$$

그러면

$$Q = W'\Sigma_w^{-1}W = W'(T'T)^{-1}W = (T'^{-1}W)'(T'^{-1}W) = W^{*\prime}W^*$$

가 되므로, 우리는 Q 를 합의 모양으로 나타낼 수 있다.

T 의 역행렬을 구해 보면

$$T^{-1} = \begin{bmatrix} \frac{1}{d_1\sqrt{c_1}} & -\frac{1}{d_1\sqrt{c_2-c_1}} & & \\ & \frac{1}{d_2\sqrt{c_2-c_1}} & \ddots & 0 \\ & & \ddots & \ddots \\ & 0 & \ddots & -\frac{1}{d_{k-1}\sqrt{c_k-c_{k-1}}} \\ & & & \frac{1}{d_k\sqrt{c_k-c_{k-1}}} \end{bmatrix}$$

이므로

$$W^* = T'^{-1}W = \begin{bmatrix} \frac{w_1}{d_1\sqrt{c_1}} \\ \frac{w_2}{d_2\sqrt{c_2-c_1}} - \frac{w_1}{d_1\sqrt{c_2-c_1}} \\ \vdots \\ \frac{w_k}{d_k\sqrt{c_k-c_{k-1}}} - \frac{w_{k-1}}{d_{k-1}\sqrt{c_k-c_{k-1}}} \end{bmatrix}$$

이 되고

$$\begin{aligned} Q &= W^{*\prime}W^* \\ &= \sum_{i=1}^{k}\left(\frac{w_i}{d_i\sqrt{c_i-c_{i-1}}} - \frac{w_{i-1}}{d_{i-1}\sqrt{c_i-c_{i-1}}}\right)^2 \\ &= \sum_{i=1}^{k}\frac{1}{c_i-c_{i-1}}\left(\frac{w_i}{d_i} - \frac{w_{i-1}}{d_{i-1}}\right)^2 \end{aligned} \tag{2}$$

로 나타내진다.

위 식에 c_i와 d_i의 일치추정치를 대입하면 합의 모양으로 표현된 카이제곱 통계량을 얻을 수 있다. c_i의 일치추정치는 λ, $\alpha(a_i)$, $\beta(a_i)$대신에 m/n, $\hat{\alpha}(a_i)$, $\hat{\beta}(a_i)$를 대입하면 다음과 같이 표현된다.

$$\begin{aligned}\hat{c}_i &= \frac{n}{m+n}\hat{\alpha}(a_i)+\frac{m}{m+n}\hat{\beta}(a_i)\\ &= \frac{mn}{m+n}\left(\sum_{j:X_j\le a_i}\frac{\delta_j}{(m-j+1)(m-j)}+\sum_{j:Y_j\le a_i}\frac{\varepsilon_j}{(n-j+1)(n-j)}\right).\end{aligned}$$

$d_i=1-F^*(a_i)$의 일치추정치는 공통 분포함수 F^*의 추정치를 두 표본을 합하여 구하는 것이 바람직하지만, 실제로 두 표본의 관측중단분포들이 다르기 때문에 두 표본을 합하여 F^*의 추정치를 구할 수 없다. 그러므로 각각의 표본에서 구할 수 있는 $1-\hat{F}_0(a_i)$와 $1-\hat{G}_0(a_i)$의 가중평균인

$$\hat{d}_i=\frac{m}{m+n}(1-\hat{F}_0(a_i))+\frac{n}{m+n}(1-\hat{G}_0(a_i))$$

를 d_i의 추정치로 사용한다.

$$w_i=\sqrt{\frac{mn}{m+n}}\,(\hat{F}_0(a_i)-\hat{G}_0(a_i))=\sqrt{\frac{mn}{m+n}}\,(\hat{r}_i-\hat{q}_i)$$
$$\hat{r}_i=\hat{r}(a_i)=1-\hat{G}_0(a_i)\quad,\quad \hat{q}_i=\hat{q}(a_i)=1-\hat{F}_0(a_i)$$

로 표현되고, 이것과 $\hat{d}_i$, $\hat{c}_i$을 (2)의 식에 대입하면 다음과 같은 간단한 모양의 카이제곱 통계량 $\hat{Q}_1$을 얻는다.

$$\hat{Q}_1=\sum_{i=1}^{k}\frac{(f_i-f_{i-1})^2}{b_i-b_{i-1}}$$

여기서, $f_i=(\hat{r}_i-\hat{q}_i)/\hat{d}_i$, $b_i=b_{1i}+b_{2i}$ 이고

$$b_{1i}=\sum_{j:X_j\le a_i}\frac{\delta_j}{(m-j+1)(m-j)}\quad,\quad b_{2i}=\sum_{j:Y_j\le a_i}\frac{\varepsilon_j}{(n-j+1)(n-j)}.$$

3. 예 제

2절에서 구한 카이제곱 검정방법을 이용하여 Koziol(1978)이 사용한 자료를 분석하려고 한다. 자료는 백혈병을 치료하는 새로운 화학요법제의 효과를 보기 위한 연구로 부터 얻어진 자료이다. 백혈병에 걸려 있는 60마리의 쥐를 30마리씩 두 그룹으로 임의로 나눈 후, 한 그룹은 아무

<표 1> 쥐의 생존기간과 PL 추정치

대 조 군		처 리 군	
생 존 기 간	PL 추 정 치	생 존 기 간	PL 추 정 치
15.4	.03333	4.7	.03333
15.4	.06667	5.4	.06667
15.7	.10000	7.1	.10000
16.1*	.10000	7.5	.13333
16.5*	.10000	8.1	.16667
16.6	.13600	8.3*	.16667
16.9	.17200	8.5	.20139
17.9	.20800	8.6	.23611
18.4	.24400	10.0	.27083
18.5	.28000	10.4	.30556
18.9	.31600	11.1	.34028
19.0	.35200	12.1*	.34028
19.1	.38800	13.8	.37693
19.2	.42400	15.0	.41358
19.4	.46000	15.1	.45023
19.7	.49600	15.3	.48688
19.8	.53200	17.6	.52353
20.4*	.53200	21.0	.56019
20.8	.57100	22.7	.59684
20.9*	.57100	23.9	.63349
21.3	.61390	24.1	.67014
21.4	.65680	27.4	.70679
21.4	.69970	31.8	.74344
21.4*	.69970	33.5	.78009
21.5	.74975	34.9	.81674
21.7	.79980	35.5*	.81674
22.0	.84985	35.6*	.81674
22.2	.89990	35.9	.87783
22.5	.94995	37.4	.93891
23.8	1.00000	38.2	1.00000

런 치료를 하지 않고(대조군), 다른 그룹은 새로운 화학요법제로 치료하고(처리군), 쥐들의 생존 기간을 측정하였다.

각 군의 쥐들의 생존 기간과 각 관측치에서 product-limit 추정치는 <표 1>과 같다. 관측중단된 자료는 관측치에 *를 붙여 표시하였다. 처리군에서는 4개, 대조군에서는 5개의 자료가 관측중단되었다.

처리군과 대조군의 생존 분포가 같다는 가설을 검정하기 위해 구간 $(0, \infty)$를 8개의 칸으로 나누고 각 칸의 경계는 9.50, 15.50, 18.00, 19.50, 21.45, 22.30, 33.00으로 하였다. 칸의 수가 정해지면 칸의 경계는 귀무가설이 사실일 때 각 칸의 확률이 같게 되도록 정하는 것이 검정력을 높

이는 방법이라고 알려져 있다. 이 예제에서는 귀무가설이 사실일 때 공통 분포함수 F^*를 알지 못하므로 그것의 추정치로 $F^* = (\hat{F} + \hat{G})/2$를 사용하여 각 칸의 확률이 거의 같도록 정하였다. 칸의 경계와 각 칸의 경계에서 $\hat{q}_i$, $\hat{r}_i$, b_{1i}, b_{2i}를 계산한 결과는 <표 2>와 같다.

<표 2>의 수치를 이용하여 카이제곱 검정통계량의 값을 구해보면 $\hat{Q} = 91.74$이고, $\hat{Q}_1 = 23.99$ 이다. $\hat{Q}$과 $\hat{Q}_1$은 모두 귀무가설이 사실일 때 자유도가 7인 카이제곱 분포를 하므로 p-값은 0.0000과 0.0011 이다. 그러므로 우리는 대조군과 처리군의 생존 분포가 같지 않다는 결론을 내릴 수 있다. 여기서 $\hat{Q}$과 $\hat{Q}_1$의 값이 큰 차이를 보이는 이유는 실제로 대조군과 처리군의 분포에 많은 차이가 있기 때문으로 판단된다. $\hat{Q}$은 공통 분포함수 F^*의 추정치를 사용안하고 $\hat{F}$과 $\hat{G}$을 별개로 사용하여 계산하기 때문에 비교하고자 하는 두 분포함수가 조금만 달라져도 값이 커질 가능성이 있다. 다시말하면 $\hat{Q}$을 사용하면 검정력은 높아질지 모르지만, 귀무가설을 너무 자주 기각함으로써 우리가 정한 유의수준보다 높은 수준의 제 1종의 오류(Type I error)를 범할 가능성이 있다. 그러므로 본인의 견해로는 $\hat{Q}$보다 $\hat{Q}_1$을 사용하는 것이 더 안전하다고 생각된다. 이와같은 견해는 시뮬레이션을 통하여 확인될 수 있을 것이다.

<표 2> 칸의 경계와 계산된 $\hat{q}_i$, $\hat{r}_i$, b_{1i}, b_{2i} 값

칸의 경계	$\hat{q}_i$	$\hat{r}_i$	b_{1i}	b_{2i}
a_1= 9.50	.79861	1.00000	.01045	0.00000
a_2= 15.50	.51312	.93333	.03350	.00238
a_3= 18.00	.47647	.79200	.03900	.00916
a_4= 19.50	.47647	.54000	.03900	.03037
a_5= 21.45	.43981	.30030	.04541	.09106
a_6= 22.30	.43981	.10010	.04541	.42439
a_7= 33.00	.25656	0.00000	.10493	.92439

본인은 본 논문에서 제시한 통계량 $\hat{Q}$와 $\hat{Q}_1$ 뿐만아니라 서론에서 언급한 기존의 여러 검정통계량을 시뮬레이션을 통하여 비교하는 것을 향후 연구과제중 하나로 생각하고 있다.

참고문헌

[1] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and prod-uct limit estimates under random censorship, *Annals of Statististics,* Vol. 2, 437-453.

[2] Efron, B. (1967). The two sample problem with censored data, Proceedings of 5th Berkeley Symposium, Vol. 5, 831-853.

[3] Gehan, E. A.(1965). A generalized Wilcoxen test for comparing arbitrarily single censored samples, *Biometrika,* Vol. 52, 203-223.

[4] Gill, R. (1983). Larger sample behavior of the product-limit estimator on the whole line, *Annals of Statististics,* Vol. 11, 49-58.

[5] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete obser-vations, *Journal of the American Statistical Association,* Vol. 53, 457-481.

[6] Kim, J. H.(1993). Chi-square goodness-of-fit tests for randomly censored data, *Annals of Statististics,* Vol. 21, 1621-1639.

[7] Koziol, J. (1978). A two-sample Cramer-von Mises test for randomly censored data, *Biometrical Journal,* Vol. 20, 603-608.

[8] Koziol, J. and Yuh, Y. S. (1982). Omnibus two-sample test procedures with random-ly censored data, *Biometrical Journal,* Vol. 24, 743-750.

[9] Mantel, N. (1966). Evaluation of survival data and two new rank statistics arising in its consideration, *Cancer Chemotherapy Reports*, Vol. 50, 163-170.

[10] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series* A, Vol 135, 185-207.

Two-Sample Chi-Square Test for Randomly Censored Data[4)]

Joo-Han Kim[5)], Jeong-Ran Kim[6)]

Abstract

A two sample chi-square test is introduced for testing the equality of the distributions of two populations when observations are subject to random censorship. The statistic is appropriate in testing problems where a two-sided alternative is of interest. Under the null hypothesis, the asymptotic distribution of the statistic is a chi-square distribution. We obtain two types of chi-square statistics ; one as a nonnegative definite quadratic form in difference of observed cell probabilities based on the product-limit estimators, the other one as a summation form. Data pertaining to a cancer chemotherapy experiment are examined with these statistics.

4) This research is supported by the KOSEF, 1992.(No. 923-0100-002-1)

5) Department of Statistics, Chungnam National University, 220 Gung-Dong Taejeon, 305-764, KOREA.

6) Department of Statistics, Chungnam National University, 220 Gung-Dong Taejeon, 305-764, KOREA.