The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 441-448

Information Loss from Type I versus Type II Censoring[1)]

Johan Lim[2)], Hyunseok Song[3)] and Sungim Lee[4)]

Abstract

If the completely observed data are assumed to have full information, the censoring causes the loss of information. Previous studies have introduced the indices of information loss via measuring relative changes between the data with censoring and without censoring. In this paper, the comparisons are made for the information loss between type I and type II censoring in two sample problems.

Keywords : Information loss; Kendall's tau; Partial likelihood.

1. 서론

중도절단(censoring)이란 생존자료(survival data)에서 나타나는 자료의 특징으로 생존시간(survival time)을 정확히 알 수 없는 경우를 가리킨다. 이런 경우, 중도절단이 있는 자료의 분석은 정보의 손실을 초래하게 되는데, Lindley (1956)의 제안에 의해 중도절단이 없이 모든 자료가 관측된 경우와 중도절단이 관측된 자료 사이의 상대적인 엔트로피 변화에 기초하여 그 정보 손실의 양을 측정하였다. 중도절단으로부터 발생하는 정보 손실을 측정하기 위한 측도로 Brooks (1982), Turrero (1989), Ebrahimi & Soofi (1990), Chaloner & Verdnelli (1995)등의 연구가 있었는데, 이들 연구는 지수분포로부터의 평균 생존 시간 추정에 따른 엔트로피의 변화를 정보 손실의 양으로 제안하였다. 그러나 실제 응용 분야에서는 생존 자료 분석을 위해 모수적 분포이외에 비모수적 분포가 좀 더 자주 가정되고 이로 인해 좀 더 다양한 통계적 문제에서 중도절단으로 인한 정보 손실의 측도를 필요로 했다. 이에 Lim et al. (2006)은 좀 더 다양한 모형에서 실질적 정보 손실의 의미를 잴 수 있는 측도를 제안하였다.

이에 본 논문에서는 이들 정보 손실 측도를 바탕으로 이변량 자료에서 독립성 검정을 할 때와 비례 위험 모형 추정문제에서 중도절단 형태에 따른 (type I versus

1) This work was partially supported by the Korea Reserach Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2005-003-C00036).
2) Assistant Professor, Department of applied statistics, Yonsei University, Seoul 120-749, Korea.
3) Graduate Student, Department of applied statistics, Yonsei University, Seoul 120-749, Korea.
4) Assistant Professor, Department of Information Statistics, Dankook University, Seoul 140-714, Korea. Correspondence : silee@dankook.ac.kr

type II censoring) 정보 손실의 양을 비교 검토하기로 한다. Type I 중도절단이란 생존시간이 미리 정해진 시간 (예를 들어, t_c)에서 중도절단이 일어나는 것으로 $T_1, T_2, \cdots, T_n$의 생존시간이 있다고 할 때, 관측하는 실제 자료가 $\{\min(T_i, t_c)\}_{i=1}^n$임을 말한다. Type II 중도절단은 n개의 자료 중 미리 정해진 $r(\leq n)$개의 생존 자료가 관측되면 중도절단이 일어나는, 즉, $T_{(1)}, T_{(2)}, \cdots, T_{(r)}$의 자료가 발생한 후 나머지 $(n-r)$자료는 $T_{(r)}$에서 중도절단이 일어난다. 이들 중도절단의 형태는 특히 공학 분야에서 많이 발생하는 자료의 형태를 보여주는데, 관련 분야에서는 좀 더 신뢰성 있는 제품을 생산하기 위해 어떠한 중도 절단 형태가 바람직한 것인지의 논의가 있어 왔다(Tseng & Hsu, 1994). 또한 적절한 표본 추출 방법을 선택하기 위해 자료가 지수분포라 가정하고 Type I 과 Type II 중도절단이 모수에 미치는 영향에 관한 연구가 있었다(Tse & Tso, 1996; Epstein & Sobel 1953). Brooks (1982)는 베이지안 접근법으로 지수분포의 모수에 있어 Type I중도절단 표본과 완전표본사이의 정보량의 손실에 대해 비교하기도 하였다.

본 논문에서는 두 중도 절단 형태에 따른 정보량의 손실에 관하여 기존의 연구가 분포의 모수 에 한정하여 이루어진 것을 생존자료 분석에서 많이 사용되는 통계적 문제를 통해 두 중도절단 형태에 따른 상대적인 정보 손실 정도를 비교할 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 중도절단으로 부터 발생하는 정보 손실을 측정하기 위한 측도를 고찰하고, 3절에서는 중도절단의 형태에 따라 이변량 자료의 독립성 검정 문제와 비례 위험 모형의 추정에서 발생하는 정보 손실의 크기를 비교 평가해 보았다. 마지막으로 4절에서는 그 결과를 토의하기로 한다.

2. 정보 손실의 측도 고찰

Shannon (1948)이 처음 제안한 엔트로피에 기초한 자료의 정보 측정은 Lindley (1956)에 의해 실험 전후에 관심 모수의 정보를 측정함으로써 실험으로 인한 추가적인 정보를 제안하였고 이것은 다음과 같이 정의될 수 있다.

$$\Delta_\epsilon(n;\Upsilon_C,\Upsilon_T) = 1 - \frac{G(n;\Upsilon_C,\theta)}{G(n;\Upsilon_T,\theta)} \tag{2.1}$$

여기서

$$G(n;\Upsilon,\theta) \equiv E_{X_n}\{G(n;\Upsilon,\theta|X_n)\} = \epsilon(p(\theta)) - E_{X_n}\{\epsilon(p(\theta|X_n))\} \tag{2.2}$$

으로 $\epsilon(f) \equiv -\int_\Theta \log f(\theta)\cdot f(\theta)d\theta$ 이고, Θ는 전체모수공간을 의미한다. 또한 $p(\theta)$는 모수 θ의 사전분포 (prior distribution)를 의미하고, $p(\theta|X_n)$은 실험 Υ로부터 n개의 독립적인 관측자료 $X_n = (x_1, x_2, \cdots, x_n)$을 구한 후 얻게 된 θ의 사후분포를 의미한다. 또한 Υ_C는 중도절단이 있는 실험의 자료를 의미하며, Υ_T는 중도절단이 없는 실험의 자료를 의미한다. 한편, Lim et al. (2006)은 정보 손실의 측도로써 중도절단이 가져오는 피셔 정보량 (Fisher information) 의 상대적인 변화로 정보 손실을 (2.3)으로 나

타내고. 중도절단이 있는 자료에 대해 완전자료(complete data)로부터 (2.2)와 같은 값을 필요로 한 자료의 크기를 (상대 효율성을) 계산하여 정보 손실을 아래 (2.4)와 같이 나타낼 수 있음을 보였다. 이 때 이 세 가지 손실 (즉, 엔트로피에 의존하는 손실, 피셔 정보량에 의존하는 손실, 상대효율성에 의존하는 손실) 측도가 근사적으로 서로 밀접한 관계가 있게 된다 (Lim et al., 2006).

$$\Delta_F(n;\Upsilon_C,\Upsilon_T)=1-\frac{FI(n;\Upsilon_C,\theta)}{FI(n;\Upsilon_T,\theta)} \tag{2.3}$$

$$\Delta_R(n;\Upsilon_C,\Upsilon_T)=1-\frac{n}{n^*} \tag{2.4}$$

여기서 n은 중도절단이 없는 실험 Υ_T에서의 자료의 개수이고, n^*는 $G(n;\Upsilon_T,\theta)=G(n^*;\Upsilon_C,\theta)$를 만족하는 값을 의미한다.

3. 정보 손실 비교 평가

3.1 켄달의 독립성 검정

Kendall (1948)의 τ통계량은 두 확률 변수의 독립성을 검정하는 검정 통계량으로, 중도절단이 없을 때 검정 통계량은

$$\tau_n=\sum_{i<j} sgn(X_i-X_j)sgn(Y_i-Y_j)/\binom{n}{2} \tag{3.1}$$

로 정의된다. 이것은 곧 $(X_i-X_j)(Y_i-Y_j)>0$을 만족하는 부합인 쌍(concordant pair)의 개수와 $(X_i-X_j)(Y_i-Y_j)<0$인 비부합인 쌍(discordant pair)의 개수 차이를 나타낸다. 위 검정 통계량은 두 확률 변수가 독립이라는 가정 하에 $E(\tau_n)=0$이고 $Var(\tau_n)=(4n+10)/(9n(n-1))$으로 근사적으로 정규분포를 따른다. 한편 이변량 확률변수 (X,Y)에 대해 중도절단이 있는 n개의 순서쌍 자료 $\{(X_i,U_i),(Y_i,V_i)\}$ (단, $i=1,\cdots,n$)가 독립적으로 관측 되었을 때 Oakes (1982)는 τ_n을 다음과 같이 확장하여 정의하였다. 단, U_i와 V_i는 각각 X_i, Y_i의 중도절단 여부에 따라 0 또는 1의 값을 갖는다.

$$\tau_n^0=\sum_{i<j} L_{ij}M_{ij}/\binom{n}{2} \tag{3.2}$$

$$=\sum_{i<j}\psi_{ij}/\binom{n}{2}$$

식(3.2)에 있는 통계량을 계산하기 위해 중도절단이 없는 X_i에 대해서는 $\widetilde{X}_i=X_i^t$로 정의하고, 중도절단이 있는 경우에는 $\widetilde{X}_i=\infty$로 정의한다. $\widetilde{Y}_i$에 대해서도 마찬가지로 정의한다. 그런 후 $i<j$인 모든 순서쌍 자료에 대해 $\widetilde{X}_i<X_j$이면 $L_{ij}=1$, $\widetilde{X}_j<X_i$이면 $L_{ij}=-1$, 그 외의 경우엔 $L_{ij}=0$으로 정의한다. 확률변수 Y에 대해서도 M_{ij}를 마

찬가지로 정의하면 (i,j) 순서쌍 자료에 대해 부합인 쌍의 경우 $\psi_{ij}=1$, 비부합인 쌍의 경우 $\psi_{ij}=-1$, 그리고 비교가능하지 않은 경우에는 $\psi_{ij}=0$을 주게 된다. Oakes는 이러한 τ_n^0가 근사적으로 평균이 0이고 분산이 $Var(\tau_n^0)=(2\alpha+4(n-2)\gamma)/(n(n-1))$ 인 정규분포임을 증명하였다. 단, $\alpha=E(\psi_{ij}^2)$ 이고 $\gamma=E(\psi_{ij}\psi_{ik})$ (단, $j\neq k$)이다. 따라서 식 (3.1)과 비교하여 식 (3.2)에서 중도절단으로 발생할 수 있는 정보 손실은 식 (2.4)에 의해 측정될 수 있다. 즉,

$$Var(\tau_n^0)=(2\alpha+4(n^*-2)\gamma)n^*(n^*-1)/4 \tag{3.3}$$

$$=(4n+10)n(n-1)/36=Var(\tau_n)$$

을 만족하는 n^*을 계산하여 식 (2.4) 는 다음과 같이 계산된다.

$$\Delta_R(n;\Upsilon_T,\Upsilon_C)\approx 1-(9\gamma)^{1/3}. \tag{3.4}$$

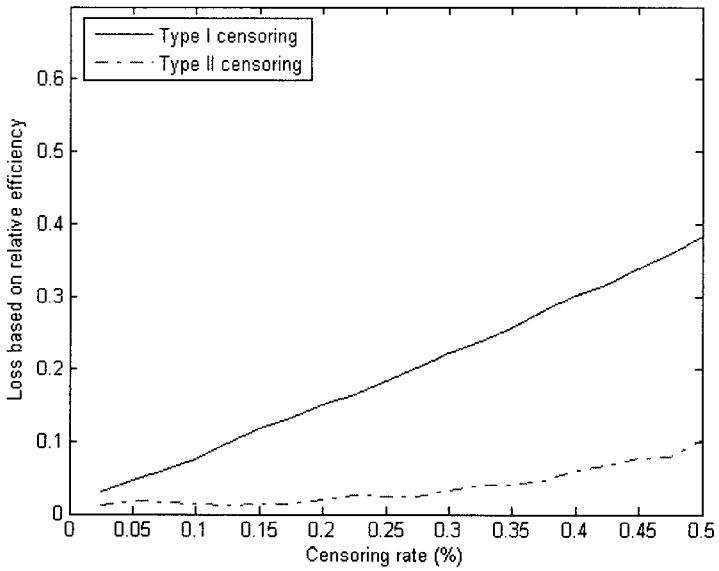
이제 Type I과 Type II 중도절단 형태에 따라 식 (3.4)로 부터 얻어진 정보 손실의 크기가 어떻게 변화하는지 알아보기 위해 다음의 실제 예제를 통해 알아보기로 한다.

예제 1. 확률변수 X는 평균이 10인 지수분포를 따르고, 확률변수 Y 는 평균이 15인 지수분포를 따른다고 하자. 또한 확률변수 Y^C는 평균이 b인 지수분포를 따른다고 할 때, 확률변수 Y의 Type I 중도절단을 위해 50개의 $Z=\min(Y,b)$를 관측하였다. 이때 b는 중도절단비율 즉, $P(Y>Y^c)$의 값이 0-50%사이의 원하는 값이 되도록 고정하였다. Type II 중도절단 자료 50개는 미리 정해진 중도절단 비율을 고려하여 관측개수를 한정하고, 나머지는 중도절단 되었다고 가정 하였다. 각 중도절단 비율에 대해서 $n=50$인 자료를 500번 생성하여 각 자료로부터 식 (3.4)의 값을 평균해서 얻은 결과가 <그림 1>과 같다. 여기서 Type I 중도절단으로 인한 정보 손실은 Type II 중도절단보다 크게 나타나는데 이것은 중도절단 비율이 커짐에 따라 그 차이도 더 크게 나타난다. 이것은 Type I 중도절단의 경우 미리 정해진 시간 t_c까지는 중도절단이 임의로 나타나 비교 가능한 순서쌍의 개수가 훨씬 줄어들어 정보 손실이 상대적으로 크게 나타난 것으로 보여 진다. 반면, Type II 중도절단의 경우에는 오히려 큰 값 쪽에서만 중도절단이 일어나므로(예제1의 경우처럼 오른쪽 중도절단의 경우에는) Type I에 비해 비교가능한 쌍의 개수가 더 많아 정보 손실도 상대적으로 적게 나타난 것으로 여겨진다.

3.2 비례위험회귀모형

위 3.1절 에서는 이변량 확률변수 (X,Y)에 대해 X와 Y의 독립성을 검정하는 문제를 다루었고, 이 절에서는 서로 다른 두 집단의 생존시간 비교 문제를 알아보고자 한다. 이를 위해 생존시간은 (T,δ)로 관측되는데, T는 생존시간을 나타내고, δ는 0 또는 1의 값으로 중도절단 여부를 나타낸다. X는 0 또는 1 의 값을 갖고 두 집단을 나타

내는 공변량이라 하자. 이들 관계를 모형화하기 위해 Cox의 비례 위험 모형을 식 (3.5)와 같이 가정한다.



<그림1> 이변량 독립성 검정할 때 중도절단 형태에 따른 정보 손실 비교

$$\lambda(t_i|x_i) = \lambda_0(t_i)\exp(x_i\beta) \tag{3.5}$$

이때, $\lambda_0(\cdot)$는 기저함수를 나타낸다. Cox (1975)는 위 모형에서 회귀계수 β의 추정을 위해 n개의 서로 다른 자료 (t_i,δ_i,x_i) (단, $i=1,\cdots,n$)로부터 부분 우도 (partial likelihood) 함수를 다음과 같이 정의하였다.

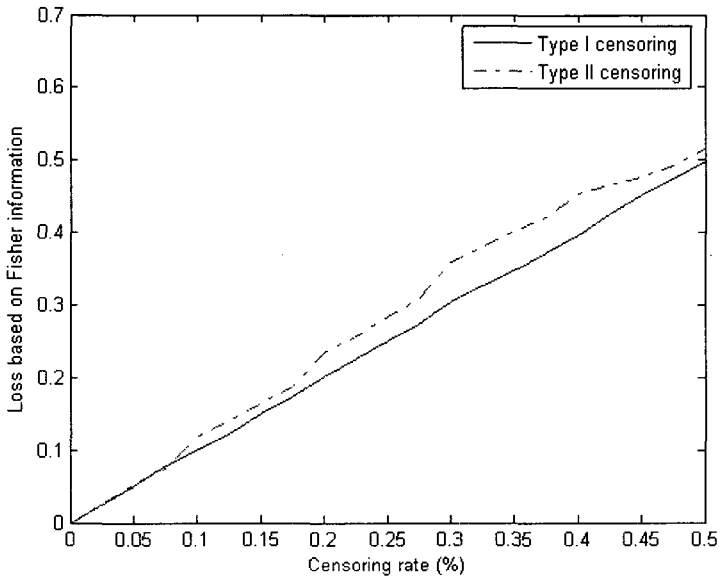
$$PL(\beta) = \prod_{i=1}^{n}\left(\frac{\exp(x_i'\beta)}{\sum_{j\in R_i}\exp(x_j'\beta)}\right)^{\delta_i} \tag{3.6}$$

여기서 R_i는 시간 $t=t_i-$ 에서 생존해 있는 모든 사람들의 집합을 나타낸다. Lim et al. (2006)은 한 집단의 표본크기는 n이고 다른 한 집단의 표본크기는 m인 서로 다른 두 집단 비교문제에서 식 (3.6)의 부분 우도 함수는 $(n+m)$명의 사람들로부터 나타나는 순열의 함수라 하고, 식 (2.3)으로 정보 손실을 평가하여 다음과 같음을 보였다.

$$\Delta_F((n,m);\Upsilon_C,\Upsilon_T,\beta) = 1 - \frac{FI((n,m),\Upsilon_C,\beta)}{FI((n,m),\Upsilon_T,\beta)} = \frac{E_{\Upsilon_C}[\sum_{i\in C}p_i(1-p_i)]}{E_{\Upsilon_T}[\sum_{i=1}^{n+m}p_i(1-p_i)]} \tag{3.7}$$

이때, C는 중도절단된 개체들의 집합을 의미하고, p_i는 i번째로 큰 생존시간에 관측

값들이 그룹 1로부터 나타난 비율을 의미한다. 다음 예제에서는 식 (3.7)을 이용하여 Type I과 Type II 중도절단에 따라 중도절단의 비율에 따라 정보 손실의 크기를 비교하기로 한다.



<그림 2> 비례위험모형에서 서로 다른 두 집단 비교 시
중도절단 형태에 따른 정보 손실 비교

예제 2. 식 (3.5)에서 $\lambda_0(t)=1$, $\beta=\log 2$라 가정하면, 그룹1 ($x=0$)으로부터 나온 개체의 생존시간은 평균이 1인 지수분포를 따르고 그룹2 ($x=1$)에서 나온 개체의 생존시간은 평균 1/2인 지수분포를 따르게 된다. 확률변수 T_1의 Type I 중도절단을 위해 확률변수 T_1^c을 평균 b인 지수 분포로부터 발생시켜 $Z_1=\min(T_1,T_1^c)$의 값을 $n=50$개 생성 하였고, 같은 방법으로 $Z_2=\min(T_2,T_2^c)$의 값을 $m=50$개 생성하여, 중도절단 비율을 0에서 50%로 하여 각 중도절단 비율에 대하여 자료를 500번씩 반복하여 식 (3.7)의 값을 얻었다. 여기서 b의 값은 중도절단 비율, 즉 $P(T_k>T_k^c)(k=1,2)$의 값이 0-50%사이의 원하는 값이 되도록 고정한다. 또한 Type II 중도절단을 위해서는 원하는 중도절단의 비율에 맞추어 관측개수를 한정하고, 나머지는 중도절단되었다고 가정하였다. <그림2>로부터 중도절단 비율에 상관없이 Type II가 Type I보다 정보 손실이 다소 크게 나타났지만 별 차이는 없음을 알 수 있다.

4. 결론

본 논문에서는 중도절단이 있는 이변량 자료에서 두 확률변수의 독립성 검정 문제

와 비례위험모형을 이용한 두 집단의 생존시간 비교 문제를 다루고, Type I과 Type II 중도절단에 따른 정보 손실의 크기를 비교 검토하여 보았다. <그림 1>과 <그림 2>에서 알 수 있듯 중도절단 형태에 따른 정보 손실의 크기는 크게 다르지 않게 나타났다. 다만, Oakes (1982)의 τ 통계량에서는 Type II중도절단이 오른쪽 큰 값 쪽에만 중도절단 되어 나타나 순위를 더 많이 계산할 수 있어 상대적으로 정보 손실이 작게 나타났다.

이러한 비교는 생존시간을 모형화하는 가장 대표적인 지수분포를 사용하여 모의실험 하였는데, 지수분포의 모수 추정에 있어 Type I 과 Type II 중도절단의 피셔 정보량을 살펴보면 다음과 같다. Type I 중도절단은 $FI(\lambda) = n_c(1-\exp(-C\lambda))/\lambda^2$, 그리고 Type II 중도절단은 $FI(\lambda) = n_c/\lambda^2$ 로 나타난다. 이때 n_c는 중도절단이 일어나지 않은 자료의 크기를 나타낸다. 3절의 모의실험에서와 마찬가지로 중도절단의 크기를 같이 한 경우 즉, 두 중도절단 형태에 따라 중도절단이 일어나지 않은 관측 개수를 동일하게 하는 경우 위의 피셔 정보량에서 비교하는 것처럼, 일반적으로 정보량은 Type II 가 크게 됨을 즉, 정보손실이 적음을 알 수 있다. 또한 두 값은 Type I 중도절단 시간이 크게 나타나면 별 차이가 없게 됨을 알 수 있다.

그러나 분포의 모수가 아니라 이변량 자료의 독립성 검정 문제, 그리고 비례위험모형을 통한 비교 문제에서 보았듯이 중도절단의 형태에 따른 상대적인 정보손실은 사용하는 통계량의 성질에 따라 다르게 나타남을 알 수 있었다. 예를 들어, 자료의 순위를 사용하는 문제의 경우, 정해진 관측 개수만큼 자료가 크기 순서대로 관측되는 Type II 중도절단의 경우가 상대적인 정보손실이 다소 작게 나타남을 알 수 있었다.

참고문헌

[1] Brooks, R.J. (1982), On the loss of information from censoring. *Biometrika*, Vol. 69, 137-144.

[2] Chaloner, K. and Verdinelli, I. (1995), Bayesian Experimental Design: A Review. *Statistical Science.*, Vol. 10, 237-304.

[3] Cox, D.R. (1975). Partial likelihood. *Biometrika*, Vol. 62, 269-276.

[4] Ebrahimi, N. and Soofi, E.S. (1990). Relative information loss under type II censored exponential data. *Biometrika*, Vol. 77, 429-435.

[5] Epstein, B. and Sobel, M. (1953). Life Testing. *Journal of the American Statistical Association*, Vol. 48, 486-502.

[6] Kendall, M.(1948). *Rank Correlation Methods*, Charles Griffin & Company Limited.

[7] Lindley, D.V. (1956), On a measure of the information provided by an experiment. *Annals of Mathmatical Statistics.*, Vol. 27, 986-1005.

[8] Lim, J., Lee, S. and Choi, H. (2006). Information loss from censoring in rank based procedures, preprint.

[9] Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics*, Vol. 38, 451-455.

[10] Shannon, C.E. (1948), A mathematical theory of communication. *Bell System Technical Journal*, Vol. 27, 379-423.

[11] Sundberg, R. (2001). Comparison of confidence procedures for Type I censored exponential lifetimes. *Lifetime Data Analysis*, Vol. 7, 393-413.

[12] Tseng, S. and Hsu, C. (1994). Comparison of Type I- & Type-II accelerated life tests for selecting the most reliable product. *IEEE Transactions on reliability,* Vol. 43, 503-510.

[13] Tse, S.k. and Tso, G. (1996). Efficiencies of Maximum Likelihood Estimators under Censoring in Life Testing. *Journal of Applied Statistics,* Vol. 23, 515-524.

[14] Turrero, A. (1989), On the relative efficiency of grouped and censored survival data. *Biometrika*, Vol. 76, 125-131.

[Received March 2006, Accepted June 2006]