Youngnam Statistical Letters
June, 1991, Vol. 2, 45-62

점진적 임의중단법에서 생존함수의 비모수적 추정에 관한 연구*

박 병구** 이 광호***

요 약

신뢰수명의 검정이나 임상실험에서 대상물에 대한 관측치를 충분히 얻기 위해서는 많은 시간과 경비가 필요하거나 현실적인 제한으로 인하여 관측이 불가능한 경우가 흔히 있다. 이러한 이유로 점진적 임의중단법에 의하여 얻어진 관측치를 이용한 생존함수의 추정은 현실적으로 매우 중요하다고 하겠다.

연구에서는 점진적 임의중단된 자료를 기초로 스플라인함수를 이용하여 생존함수의 비모수적 최우추정량을 제안하고 그것의 성질과 효율성을 비교.연구한다.

1. 서 론

여러가지 신뢰성이나 생존분석연구에서 생존함수를 추정하는 문제가 대두되는데, 그것은 필요한 대상물에 대한 임상실험이나 수명검정을 통하여 자료를 수집함으로써 시작된다. 그러나, 현실적인 여러가지 제약조건(시간,경비,대상물의 특징 등)으로 인하여 대상물에 대한 실험 혹은 관측이 중단되는 경우가 많이 있는데, 이런 경우의 자료를 일반적으로 불완전 자료라 부르며 중단되는 특징에 따라 여러가지 다른 이름으로 불리고 있다.

불완전 자료를 이용한 생존함수의 추론은 문제의 다양성과 복잡성으로 인하여 Kaplan과 Meier(1958)의 연구이래 비모수적 방법으로 폭 넓게 연구되고 있다. 특히 임의 중단된 자료를 이용한 생존함수의 추론은 Breslow와 Crowley(1974)를 포함하여 Foldes, Rejto와 Winter(1980), Suzuki(1985), Tsai(1986), Wong(1987) 등에 의해 연구되었다.

실제로 임의 중단에 의한 표본추출은 공학에서의 신뢰성 분석문제나 의학분야에서의

---

* 이 연구는 88년도 한국과학재단 연구비 지원에 의한 결과임 (881-0105-006-1)

** 경북대학교 자연과학대학 통계학과

*** 영남대학교 이과대학 통계학과

생존분석연구 등에서 널리 적용되고 있다. 그러나 이러한 분야에서의 관측치는 대개가 시간이 경과함에 따라 순차적으로 얻어지는 특징을 지니기 때문에 임의중단법으로 표본추출하는 데에도 몇 가지 애로점이 나타날 수 있다. 즉, 현실적인 여건으로 인하여 너무 빨리 중단하게 되면 표본크기가 너무 적게 되어 추정을 어렵게 하거나 신뢰도를 떨어지게 할 수 있으며, 극단적으로는 전혀 관측치를 얻지 못하게 될 수도 있다. 이에 반하여 충분한 관측치를 얻기위하여 중단시간을 충분히 길게 정하면 그 시간에 수반되는 경비가 과다하게 소요될 수 있고 또 필요이상의 실험 단위의 파괴내지는 희생이 수반될 수 있으며, 또한 실험 단위에 대한 사전정보가 전혀 없는 경우에는 이 충분한 시간동안에 실험단위의 특징에 따라 극히적은 관측치만을 얻게 될 지도 모르는 것이다.

이러한 이유로 인하여, 임상실험과 수명검정문제 등에서 점진적 임의 중단법(progressively random censoring scheme)의 필요성이 주장되어 왔다. 이 방법에는 관측치가 실험의 시작부터 수집되고 그 결과를 점차적으로 누적하여 추정 혹은 검정에 이용한다. 만약 실험의 어느 단계에서 현 상태의 누적 관측치가 어떤 분명한 통계적 결정을 보장할 수 있다면 실험은 현 단계에서 마치게 된다. 따라서 점진적 임의 중단법은 실험비용과 실험단위의 희생 모두의 감소와 더불어 비교적 짧은 실험기간이 요구된다.

점진적 임의 중단법으로 Cohen(1963,1965)은 모수모형에 대한 생존함수의 최우추정량을 제안했고 Chatterjee와 Sen(1973)은 일반적인 순위검정법을 제안하였다. Davis(1978)는 생존함수의 비모수적 베이즈추정량을 제안하였다. 점진적 임의 중단법의 다른 응용에 대해 Halperin과 Ware(1974), Sen(1976), Gardiner와 Sen(1978), Gardiner와 Susarla(1982)등이 연구했었다.

본 연구에서는 점진적 임의 중단법으로 생존함수의 새로운 비모수적추정량을 제안하고 그 성질을 연구하고자 한다. 즉, 스플라인함수를 사용하여 위험률함수를 근사적으로 구하고 그것과 생존함수의 관계를 이용하여 생존함수의 비모수적 최우추정량을 구하고자 한다. 만약 실험종결단계(m)가 완전한 표본(n)과 일치하면 제안된 통계량은 Whittmore와 Keller(1986)의 통계량과 일치한다. 끝으로, 점진적 임의 중단법하에서 제안된 통계량의 성질을 모의 실험을 통해 비교.연구하며 실제의 수치적 예를 들어 적용시켜 보고자 한다.

2. 점진적 임의 중단법

이 절에서는 점진적 임의 중단법에서의 표본추출과정과 생존함수에 대한 정의를 한다. 생존시간 T와 중단시간 C는 각각 연속분포함수 $F(\cdot)$와 $G(\cdot)$, 그리고 확률밀도함수 $f(\cdot)$와 $g(\cdot)$를 각 각 가지며 T와 C는 서로 독립이라 한다. 실제 관측시간 Y와 중단 지시함수 δ

는 각각의 실험단위에 대하여 다음의 식으로 표시된다.

$$Y = \min (T, C),$$

$$\delta = I[\, T \le C] = \begin{cases} 1, & \text{if } T \le C \quad (\text{사망}), \\ 0, & \text{if } T > C \quad (\text{중단}). \end{cases}$$

따라서, (Y_1, δ_1), (Y_2, δ_2),..., (Y_n, δ_n) 는 n개의 항목을 실험할 때의 관측 가능한 자료이다. $Y_{(1)}$, $Y_{(2)}$,..., $Y_{(n)}$은 Y_1, Y_2,...,Y_n 의 순서통계량이고 δ_i^*은 $Y_{(i)}$, i=1,2,...,n, 과 연관된 중단지시함수라 하자. 지금 m단계에서 실험이 종결되었다 하자. 여기에서 $m \in \{1,2,...,n\}$이다. 그러면 실제로 기록된 자료는 $(Y_{(1)}, \delta_1^*)$, $(Y_{(2)}, \delta_2^*)$,..., $(Y_{(m)}, \delta_m^*)$와 나머지 n-m개의 항목의 중단시간과 생존시간이 모두 $Y_{(m)}$을 초과한다는 것뿐이다. 즉 $Y_{(j)} > Y_{(m)}$, j=m+1,...,n,이다. 식을 간단하게 하기 위해 $Y_{(i)}$와 δ_i^*를 Y_i와 δ_i로 쓰기로 한다.

이와 같이 관측된 자료를 이용하여 다음과 같이 정의된 생존함수와 위험률함수, 누적위험률함수를 추정하고자 한다.

$$\text{생존함수} \quad : S(t) = P(T > t) = 1 - F(t). \tag{2.1}$$

$$\text{위험률함수} \quad : \lambda(t) = \frac{f(t)}{1-F(t)} \tag{2.2}$$

$$\text{누적위험률함수} \quad : \Lambda(t) = \int_0^t \lambda(\mu)d\mu = -\int_0^t d\log(1-F(\mu)). \tag{2.3}$$

여기서 생존함수와 누적위험률함수의 관계는 다음과 같다.

$$S(t) = \exp\{-\Lambda(t)\}. \tag{2.4}$$

Kaplan과 Meier(1958), Nelson(1969)는 생존함수 S(t)의 비모수적 추정량을 임의 중단법(random censoring)된 정보를 기초로 다음과 같은 S_{KM} 과 S_{NA}로 각 각 제안하고 그 성질을 연구 하였다.

$$\hat{S}_{KM}(t) = \prod_{y \le t} \left(\frac{n-i}{n-i+1}\right)^{\delta}$$

$$\hat{S}_{NA}(t) = \exp\{ - \hat{\Lambda}_{NA}(t) \}.$$

단 $\Lambda_{NA}(t) = \sum_{y_i \le t} \frac{\delta}{n-i+1}$ 이다.

[비고 2.1] 시각 t에서 n-i가 클 때는 $S_{KM}(t)$와 $S_{NA}(t)$는 거의 일치하며 모두 다음과 같은 단점을 내포하고 있다.

(1) 위험률함수의 추정량을 직접 구할 수 없다.

(2) 대단히 유용한 정보가 될 지도 모르는 정확한 중단시간을 사용하고 있지 않고 있다.

3. 생존함수의 비모수적 최우추정량

이 절에서는 스플라인함수를 사용하여 생존함수의 비모수적 최우추정량을 구하고자 한다. 먼저 [0,T] 상에서 생존함수를 추정하기 위해 점진적 임의 중단법으로 얻어진 자료 (y_i, δ_i) , i=1,2,···,n 의 우도함수를 이용하고자 한다. 각 표본의 확률은 다음과 같다.

$$\begin{aligned} \Pr(y_i=t,\ \delta_i=0) &= \Pr(C_i=t,\ T_i > C_i) \\ &= g(t)\,(1-F(t)), \\ \Pr(y_i=t,\ \delta_i=1) &= \Pr(T_i=t,\ T_i \le C_i) \\ &= f(t)\,(1-G(t)). \end{aligned}$$

위의 식들은 다음과 같이 표현할 수 있다.

$$\Pr(y_i=t,\delta_i)=[f(t)\{1-G(t)\}]^{\delta_i}[g(t)\{1-F(t)\}]^{1-\delta_i}$$

$m < i \le n$ 에 대한 자료로부터 이용 가능한 유일한 정보는 그들의 중단시간과 생존시간이 모두 y_m을 초과한다는 것뿐이다. 즉 $y_i > y_m$, $i = m+1, \cdots, n$, 이다. 따라서 (y_i, δ_i), $i = 1,2,\cdots,n$, 의 우도함수는 다음과 같다.

$$\begin{aligned} &\prod_{i=1}^{m} \{1-G(y_i)\}^{\delta_i} g(y_i)^{1-\delta_i} \{1-G(y_m)\}^{n-m} \\ &\cdot \prod_{i=1}^{m} f(y_i)^{\delta_i} \{1-F(y_i)\}^{1-\delta_i} \{1-F(y_m)\}^{n-m} \end{aligned} \tag{3.1}$$

F 와 G 는 서로 독립이므로 $G(\cdot)$와 $g(\cdot)$에는 관심의 대상인 어떤 모수도 포함되어 있지 않는다. 따라서 식 (3.1)의 첫 항은 무시할 수 있으며 다음의 결과를 얻는다.

[정리 3.1] 점진적 임의중단법에서 자료 (Y_1,δ_1), (Y_2,δ_2),..., (Y_m,δ_m)과 $Y_i > Y_m$, $i= m+1,\ldots,n$,의 우도함수는

$$L = \prod_{i=1}^{m} f(y_i)^{\delta_i} \{1-F(y_i)\}^{1-\delta_i} \{1-F(y_m)\}^{n-m}$$

이다.

식 (2.1)과 (2.2)에서의 생존함수와 위험률함수의 관계식으로 부터 정리 2.1의 우도함수는

$$L(\lambda) = \prod_{i=1}^{m} \lambda(y_i)^{\delta_i} S(y_i)\, S(y_m)^{n-m}$$

으로 나타낼 수 있으며 식 (2.3)와 (2.4)으로부터 $L(\lambda)$의 로그-우도함수는

$$\log L(\lambda) = \sum_{i=1}^{m} \{ \delta_i \log \lambda(y_i) - \int_0^{y_i} \lambda(u)du \} - (n-m)\int_0^{y_i} \lambda(u)du \qquad (3.2)$$

가 된다.

이제부터 식(3.2)에서의 위험률함수 $\lambda(\cdot)$를 스플라인함수를 이용하여 근사적으로 구해 보자.

[정의 3.1] $0 = t_0 < t_1 < , \ldots , < t_k < t_{k+1} = T$ 라 하고 $\theta_j(t)$, $j = 0,1,\ldots,k+1$,는 알려진 t의 함수라 하면 t의 스플라인함수 λ_Q 를

$$\lambda_Q(t) = \sum_{j=0}^{k+1} \theta_j(t)\lambda_j \qquad (3.3)$$

로 정의한다. 단 $\lambda_j \equiv \lambda(t_j)$ 이다. 실제로 스플라인함수 λ_Q는 위험률함수를 보간하고 있으며 알려진 함수 $\theta_j(t)$ 에 의해 결정되어진다.

[정리 3.2] 정리 3.1의 조건에서 위험률함수와 누적위험률함수, 생존함수의 최우추정량은 각 각 다음과 같이 주어진다.

$$\hat{\lambda}_Q(t) = \sum_{j=1}^{k} \theta_j(t)d_j\gamma_j^{-1}, \qquad (3.4)$$

$$\hat{\Lambda}_Q(t) = \sum_{j=1}^{k} \alpha_j(t)d_j\gamma_j^{-1}, \qquad (3.5)$$

$$S_Q(t) = \exp\{-\Lambda_Q(t)\}, \tag{3.6}$$

단 $j = 1,2, \ldots ,k$ 대해서 d_j는 시각 t_j에서의 사망한 항목의 갯수이며 $\theta_j(t)$는 시각 t의 알려진 함수이고, $\alpha_j(t)$와 γ_j는 식 (3.8)와 (3.10)에서 각 각 주어진다.

[증명] 식 (3.3)에서 주어진 λ_Q의 정의로 부터 함수 λ_Q의 적분값은 근사적인 누적위험률함수 $\Lambda(t)$가 된다. 즉, λ_Q 함수의 적분은

$$\int_0^t \lambda_Q(u)du = \sum_{j=0}^{k+1} \lambda_j \int_0^t \theta_j(u)du$$

$$= \sum_{j=0}^{k+1} \alpha_j(t)\lambda_j \tag{3.7}$$

이다. 여기에서 계수

$$\alpha_j(t) = \int_0^t \theta_j(u)du \tag{3.8}$$

이며 $\lambda(.)$에 의존하지 않는 알려진 함수이다.

여기에서 t_j, $j = 1,2, \ldots ,k$,는 k개의 서로 다른 사망시간(dead time)으로 잡자. 만약, 로그-우도함수 식 (3.2)에서 $\lambda(\cdot)$의 적분을 λ_Q의 적분으로 치환하여 식 (3.8)을 이용하면

$$L(\lambda_Q) = \sum_{j=o}^{k+1} (d_j \log\lambda_j - \gamma_j\lambda_j) \tag{3.9}$$

가 된다. 단

$$\gamma_j = \sum_{i=1}^{m} \alpha_j(y_i) + (n-m)\alpha_j(y_m) \tag{3.10}$$

이다. 식 (3.9)은 평균이 $\gamma_j\lambda_j$, $j = 0,1, \ldots ,k+1$, 인 k+2개의 독립인 포아송 변수 d_j의 로그-우도함수이며 그 해는

$$\hat{\lambda}_j = d_j\gamma_j^{-1} \tag{3.11}$$

이 된다. 따라서, 이 값 $\hat{\lambda}_j$을 식 (3.3)과 (3.7)에서 사용하면 위험률함수와 누적위험률함수, 생존함수에 대한 최우추정량을 각 각 구할 수 있다.

[비고 3.1] 정리 3.2에서 얻어진 추정량들은 식 (3.3)에 주어진 $\theta_j(t)$에 전적으로 의존한다. 즉 특별한 함수 $\theta_j(t)$를 선택함에 따라 위험률함수와 생존함수의 추정량들이 결정된다.

[정의 3.2] 임의의 측정가능한(measurable) 함수 h에 대해 함수 $\delta(\cdot)$가

$$\int h(x)\delta(\tau-x)dx = h(\tau) \tag{3.12}$$

을 만족할 때 함수 $\delta(\cdot)$를 Dirac-delta함수라 한다.

[따름정리 3.1] 정리 3.2의 조건에서 $\theta_j(t) = \delta(t-t_j)$, $j = 0,1, \ldots, k+1$, 이면 누적위험률함수와 생존함수의 최우추정량은 다음과 같이 주어진다.

$$\hat{\Lambda}_{QO}(t) = \sum_{t_j \le t} d_j \gamma_j^{-1} ,$$

$$\hat{S}_{QO}(t) = \exp\{ - \hat{\Lambda}_{QO}(t)\}. \tag{3.13}$$

단

$$\gamma_j = \sum_{i=1}^{m} I(y_i \ge t_j) + (n-m)I(y_m \ge t_j)$$

이다.

[증명] 식 (3.5)에서의 스플라인함수는 주어진 조건에 의해

$$\lambda_{QO}(t) = \sum_{j=0}^{k+1} \delta(t-t_j)\lambda_j$$

이고, 식 (3.8)과 (3.12)로 부터

$$\alpha_j(t) = \int_0^t \delta(u-t_j)du$$

$$= \int I(t \ge u) \cdot \delta(u-t_j)du$$

$$= I(t \ge t_j)$$

가 된다. 더우기 식 (3.7)으로부터 γ_j을 얻을 수 있다. 그러므로 식 (3.4) 및 (3.5), (3.6), (3.12)로 부터 본 정리의 결과를 얻을 수 있다.

[비고 3.2] 식 (3.13) 을 자세히 보면 $\hat{S}_{Q0}$는 실험종결단계(m)가 n과 같고 대등한 생존시간(tied survival time)이 없을 때는 2절에서 소개된 Nelson의 추정량 $\hat{S}_{AN}$이 됨을 알 수 있다.

만약, $\theta_j(t)$를

$$\theta_j(t) = \begin{cases} I(t_{j-1}<t\leq t_j), & j = 1,2, \ldots ,k+1 \\ 0, & j = 0 \end{cases} \tag{3.14}$$

과 같이 정의하면 [따름정리 3.1]의 증명에서와 유사한 방법으로 다음의 결과를 얻을수 있다.

[따름정리 3.2] 정리 3.2의 조건에서 $\theta_j(t)$를 식 (3.5)와 같이 하면 누적위험률함수와 생존함수의 최우추정량은 각 각

$$\hat{\Lambda}_{Q1}(t) = \sum_{j=1}^{k} a_j(t)d_j\gamma_j^{-1}$$

와

$$\hat{S}_{Q1}(t) = \exp\{ - \hat{\Lambda}_{Q1}(t)\}$$

이다. 단 $a_0(t) \equiv 0$ 이고 $a_j(t)$와 γ_j는 j=1,2, ... ,k+1 에 대해 각 각

$$a_j(t) = \begin{cases} 0, & 0\leq t\leq t_{j-1}. \\ t-t_{j-1}, & t_{j-1} < t\leq t_j. \\ t_j-t_{j-1}\equiv\Delta_j. & t_j<t\leq T, \end{cases} \tag{3.15}$$

와

$$\gamma_j = \sum_{i=1}^{m} a_j(y_i) + (n-m)a_j(y_m) \tag{3.16}$$

이다.

다음으로, $\theta_j(t)$를

$$\theta_j(t) = \begin{cases} I(t_j<t\leq t_{j+1}), & j = 0,1, \ldots ,k, \\ 0, & j = k+1, \end{cases} \tag{3.17}$$

과 같이 정의하면 다음의 결과를 얻는다.

[따름정리 3.3] 정리 3.2의 조건에서 $\theta_j(t)$를 (3.17)식과 같이 하면 누적위험률함수와 생존함수의 최우추정량은 각 각

$$\hat{\Lambda}_{Q2}(t) = \sum_{j=1}^{k} a_{j+1}(t)d_j\gamma_j^{-1} \quad \text{와} \quad \hat{S}_{Q2}(t) = \exp\{ - \hat{\Lambda}_{Q2}(t)\}$$

이다. 단 $j=1,2,\ldots,k$ 에 대해서 $a_j(t)$는 (3.15)에 주어진 것이고, $a_{k+1}(t)=0$ 이며 γ_j는 다음과 같다.

$$\gamma_j = \sum_{i=1}^{m} a_{j+1}(y_i) + (n-m)a_{j+1}(y_m).$$

여기에서 $\hat{\Lambda}_{Q2}$는 우측 연속 계단함수이고 높이는 $d_j\lambda_j^{-1}$을 가진다.

마지막으로, $\theta_j(t)$를

$$\theta_j(t) = \begin{cases} I[t_0<t<t_i]\cdot[\Delta_i - a_i(t)]/\Delta_i, & j=0, \\ I[t_{j-1}\le t\le t_j]\cdot a_j(t)/\Delta_j & \\ \quad + I[t_j\le t\le t_{j+1}]\cdot[\Delta_{j+1}-a_{j+1}(t)]/\Delta_{j+1}, & j=1,2,\ldots,k, \\ I[t_k\le t\le t_{k+1}]\cdot a_{k+1}(t)/\Delta_{k+1}, & j=k+1 \end{cases} \tag{3.18}$$

와 같이 정의하면 다음의 결과를 얻을 수 있다.

[따름정리 3.4] 정리 3.2의 조건에서 식 $\theta_j(t)$를 (3.18)와 같이 정의하면 누적위험률함수와 생존함수의 최우추정량은 각각 다음과 같이 주어진다.

$$\hat{\Lambda}_{Q3}(t) = \sum_{j=1}^{k} \alpha_j(t)\, d_j\, \gamma_j^{-1},$$

$$\hat{S}_{Q3}(t) = \exp\{-\hat{\Lambda}_{Q3}(t)\}.$$

단 $\alpha_j(t)$와 γ_j는 각 각

$$\alpha_j(t) = a_{j+1}(t) + \frac{1}{2}\left\{\frac{a_j(t)^2}{\Delta_j} - \frac{a_{j+1}(t)^2}{\Delta_{j+1}}\right\}$$

와

$$\gamma_j = \sum_{i=1}^{m}\left[a_{j+1}(y_i) + \frac{1}{2}\left\{\frac{a_j(y_i)^2}{\Delta_j} - \frac{a_{j+1}(y_i)^2}{\Delta_{j+1}}\right\}\right]$$

$$+ (n-m)\left[a_{j+1}(y_m) + \frac{1}{2}\left\{\frac{a_j(y_m)^2}{\Delta_j} - \frac{a_{j+1}(y_m)^2}{\Delta_{j+1}}\right\}\right]$$

이다.

4. 적용사례

이 절에서는 부분적인 자료 $\{(y_i,\delta_i);\ 1\le i\le m,\ y_i\ge y_m\ ;j=m+1,\ldots,n\}$을 가지고 점진적 임의 중단하에서 제안한 생존함수의 추정값을 계산한다. 다음의 실제 자료는 Gardiner와 Susarla(1982)에서 인용했다. 위스콘신대학교 중앙종양학그룹에서 흑색종

(melanoma)환자 81명의 생존시간(단위 week)을 측정한 결과이다. "-" 가 있는 자료는 중단된 수명시간을 표시한다.

136	58	55-	181-	21	23	190-	65
234	194	14	90	20	130	213-	215-
124	108-	54	98	193-	138	141	110
67-	50	26	103	59	134-	147-	152-
65	40	34	57	81-	152-	125-	151-
34	158-	27	148-	27	132-	140-	32
130-	38	85	129-	100-	19	118	53
120-	66	46	37	50-	114-	124-	26-
102	93-	80-	60	86-	21-	44-	23
70	73-	19	38	31	25	76-	13
16-							

이 자료로부터 시각 t = 25, 44, 56, 65, 76, 100, 148, 190에 대해 실험종결단계 m이 65, 73, 81일때 제안된 추정량 $\hat{S}_{Q0}(t)$, $\hat{S}_{Q1}(t)$, $\hat{S}_{Q3}(t)$을 계산하였다. 종결단계 m이 완전한 표본크기(81)와 같을 때는 Kaplan과 Meier(1958)의 KM추정량이 된다.

m	t	25	44	54	65	76	100	148	190
	Q0	.9325	.7982	.7309	.6624	.6486	.6055	.4244	.4244
65	Q1	.9146	.7891	.7309	.6624	.6412	.5985	.4251	.4116
	Q3	.9239	.7936	.7385	.6632	.6450	.6002	.4332	.4077
	Q0	.9101	.7816	.7162	.6362	.6092	.5652	.3713	.3713
73	Q1	.8929	.7729	.7164	.6364	.6039	.5584	.3724	.3549
	Q3	.9021	.7767	.7232	.6373	.6076	.5594	.3801	.3476
	Q0	.8765	.7280	.6649	.5877	.5616	.5166	.3124	.3124
81	Q1	.8766	.7197	.6652	.5880	.5564	.5099	.3131	.2876
	Q3	.8795	.7226	.6701	.5881	.5592	.5102	.3203	.2725
81	KM	.8753	.7252	.6616	.5838	.5575	.5123	.3045	.3045

위의 도표로부터 각 경우에서의 제안된 추정량의 값은 서로 비슷하다. 그리고 KM추정량과 제안된 추정량의 값의 차이도 거의 없음을 볼 수 있다. 즉, 점진적 임의중단법에서 종결단계(m)을 줄여 표본위 크기가 작아도 제안된 추정량들은 이용가치가 있음을 보여주고 있다.

5. 몬테칼로 연구

이제까지 점진적 임의 중단된 자료를 기초로 스플라인함수를 이용하여 생존함수의 몇 가지 추정량을 제안하였으며, 이 절에서는 몬테칼로 방법을 이용하여 제안된 추정량들 $\hat{S}_{Q0}$, $\hat{S}_{Q1}$, $\hat{S}_{Q2}$, $\hat{S}_{Q3}$과 KM추정량의 효율성을 비교하고자 한다.

단순하게 나타내기 위해 추정량 $\hat{S}_{Q0}$, $\hat{S}_{Q1}$, $\hat{S}_{Q2}$, $\hat{S}_{Q3}$을 각 각 Q0, Q1, Q2, Q3이라 하자. 추정량 Q2는 우측연속계단함수이고 Q1은 좌측연속계단함수인 점을 제외하고는 두 추정량은 거의 같은 형태이다. 즉, Q1이 생존함수를 과대 추정한다면 Q2는 과소 추정할 것이고 반대로 Q1이 과소 추정 한다면 Q2는 과대 추정하게 될 것이다. 따라서 여기서는 Q2추정량은 고려하지 않겠다.

(표 1)에서는 이 절에서 사용되는 세 가지 형태의 생존분포(F)와 중단분포(G)를 나타냈다. 생존분포는 각각 위험률함수가 증가, 감소 그리고 상수함수인 경우가 되도록 선택된것이고, 중단분포는 중단되는 관측치가 많은 경우(70%)와 보통인 경우(50%)가 되도록 고려된 것이다.

각 경우에서 표본크기(n)가 20이고 실험종결단계(m)가 16, 18, 20일 때 500번 시행하여 표 2, 3, 4에 나타내었으며 표본크기가 30과 40일 때도 유사한 결과를 얻을 수 있었다. 표의 수치들은 5개 시간구간 (0.0, 0.5], (0.5, 1.0], (1.0, 1.5], (1.5, 2.0], 0.0, 2.0]에서 시간이 0.005단위로 변할 때마다 편의와 평균제곱오차를 계산하여 평균한 것이다.

표2를 통하여 다음의 결과를 얻을수 있다.

(1) 생존함수의 참값이 큰 경우에는 점진적 중단법이 대단히 유익한 표본추출법이 되는 것 같다.

(2) 중단된 관측치가 많을 때는 (1.0,2.0]구간의 모든 경우에 대해여

Q3추정량이 Q0와 Q1 보다 더 작은 편의와 평균제곱오차를 가지는 것 같다. 그러나 (0.0,0.5]구간에서는 Q1은 더 적은 편의와 평균제곱오차를 가진다.

(3) 중단된 관측치가 보통일 때는 Q3추정량이 (1.0,2.0]구간에서 다른 모든 추정량보다 더 작은 편의를 가지고 (0.0,1.0]구간에서는 모든 경우에 대해 Q3가 KM추정량보다 더 작은 편의와 평균제곱오차를 가지는 것 같다.

(4) (1.0, 2.0]구간에서 중단된 관측치가 많을 때는 보다 더 작은 평균제곱오차를 가지고, 중단된 관측치가 보통일때는 Q3가 KM추정량보다 더 작은 편의를 가진다.

(5) 시간구간 (0.0,0.5]에서는 실험종결 단계(m)가 작더라도, 즉 자료를 부분적으로 이용하더라도 추정량들 모두는 유사한 편의와 평균제곱오차를 가지고 있다.

표3 과 표4를 통하여서도 표2에서 얻은 결과와 유사한 결과들을 찾을 수 있다.

표 1. 수명분포와 중단분포

수명 분포	중단 분포	중단율
Weibull (1, 0.5)	Exponential (0.5) Exponential (1.0)	70 50
Weibull (1, 1.0)	Exponential (0.5) Exponential (1.0)	70 50
Weibull (1, 1.5)	Exponential (0.5) Exponential (1.0)	70 50

표 2. Q0, Q1, Q3와 KM추정량들의 편의와 평균오차제곱(위험률함수가 상수)

(1) 중단된 관측치가 많은경우 (70 %)

S(t)		(1.0 .13]		(1.0 .60]		(0.6 .36]		(.36 .22]		(.22 .13]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
	Q0	.2903	.1199	.0795	.0150	.2303	.0674	.3594	.1454	.5212	.2639
16	Q1	.2952	.1285	.0588	.0115	.2242	.0651	.3815	.1621	.5459	.2882
	Q3	.2298	.0857	.0977	.0182	.2111	.0639	.2770	.1093	.3562	.1602
	Q0	.2202	.0815	.0516	.0120	.1610	.0441	.2690	.0937	.4214	.1841
18	Q1	.2311	.0927	.0297	.0093	.1561	.0424	.3027	.1135	.4593	.2149
	Q3	.1477	.0514	.0678	.0139	.1322	.0407	.1722	.0637	.2334	.0926
	Q0	.1165	.0493	.0219	.0114	.0674	.0330	.1255	.0537	.2626	.1038
20	Q1	.1416	.0615	.0016	.0096	.0640	.0302	.1866	.0722	.3315	.1401
	Q3	.0280	.0287	.0345	.0122	.0183	.0306	.0083	.0356	.0538	.0392
20	KM	.1220	.0515	.0137	.0122	.0546	.0318	.1458	.0556	.2862	.1114

(2) 중단된 관측치가 보통인경우 (50 %)

S(t)		(1.0 .13]		(1.0 .60]		(0.6 .36]		(.36 .22]		(.22 .13]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
	Q0	.2277	.0721	.0848	.0148	.1946	.0496	.2745	.0876	.3794	.1436
16	Q1	.2104	.0648	.0608	.0104	.1709	.0397	.2603	.0788	.3707	.1368
	Q3	.2166	.0679	.1019	.0176	.1992	.0523	.2551	.0830	.3317	.1253
	Q0	.1546	.0410	.0631	.0119	.1313	.0312	.1798	.0477	.2596	.0774
18	Q1	.1351	.0346	.0382	.0084	.1054	.0236	.1628	.0400	.2476	.0698
	Q3	.1391	.0367	.0786	.0137	.1305	.0320	.1540	.0429	.2073	.0617
	Q0	.0619	.0218	.0389	.0107	.0589	.0220	.0587	.0251	.0974	.0318
20	Q1	.0352	.0162	.0133	.0084	.0277	.0172	.0315	.0186	.0716	.0222
	Q3	.0374	.0188	.0528	.0118	.0498	.0221	.0172	.0228	.0336	.0203
20	KM	.0511	.0205	.0327	.0108	.0384	.0221	.0369	.0222	.1015	.0290

표 3. Q1, Q2, Q3와 KM추정량들의 편의와 평균오차제곱(위험률함수가 증가함수)

(1) 중단된 관측치가 많은경우 (70 %)

S(t)		(1.0 .06]		(1.0 .70]		(.70 .36]		(.36 .16]		(.16 .06]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
	Q0	.3062	.1443	.0526	.0093	.2077	.0608	.3959	.1763	.5991	.3451
16	Q1	.3104	.1556	.0272	.0062	.1988	.0576	.4204	.1974	.6263	.3766
	Q3	.2517	.1096	.0692	.0109	.2002	.0623	.3218	.1436	.4408	.2324
	Q0	.2291	.0961	.0348	.0085	.1325	.0397	.2898	.1093	.4822	.2365
18	Q1	.2386	.1103	.0083	.0060	.1235	.0373	.3253	.1338	.5216	.2753
	Q3	.1652	.0646	.0506	.0094	.1179	.0403	.2036	.0815	.3052	.1337
	Q0	.1121	.0539	.0156	.0091	.0260	.0353	.1214	.0550	.2867	.1216
20	Q1	.1305	.0691	-.0132	.0075	.0122	.0335	.1774	.0776	.3587	.1649
	Q3	.0360	.0296	.0298	.0091	-.0061	.0351	.0190	.0350	.1051	.0420
20	KM	.1175	.0563	.0099	.0099	.0127	.0342	.1406	.0567	.3183	.1300

(2) 중단된 관측치가 많은경우 (50 %)

S(t)		(1.0 .06]		(1.0 .70]		(.70 .36]		(.36 .16]		(.16 .06]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
	Q0	.2184	.0743	.0448	.0073	.1522	.0363	.2781	.0895	.4205	.1717
16	Q1	.2018	.0686	.0191	.0048	.1277	.0279	.2655	.0816	.4149	.1671
	Q3	.2067	.0676	.0604	.0087	.1595	.0386	.2581	.0831	.3697	.1469
	Q0	.1413	.0395	.0291	.0067	.0873	.0231	.1743	.0444	.2888	.0877
18	Q1	.1226	.0348	.0029	.0049	.0599	.0172	.1586	.0374	.2811	.0830
	Q3	.1248	.0335	.0442	.0074	.0883	.0237	.1453	.0384	.2337	.0680
	Q0	.0423	.0178	.0121	.0070	.0120	.0198	.0407	.0198	.1085	.0261
20	Q1	.0174	.0147	-.0147	.0061	-.0202	.0182	.0140	.0151	.0924	.0210
	Q3	.0174	.0141	.0266	.0072	.0038	.0206	-.0061	.0173	.0472	.0127
20	KM	.0373	.0182	.0077	.0074	-.0092	.0215	.0233	.0171	.1311	.0287

표 4. Q0, Q1, Q3와 KM추정량들의 편의와 평균오차제곱(위험률함수가 감소함수)

(1) 중단된 관측치가 많은경우 (70 %)

S(t)		(1.0 .24]		(1.0 .49]		(.49 .36]		(.36 .29]		(.29 .24]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
16	Q0	.2187	.0671	.0969	.0205	.1948	.0503	.2567	.0792	.3483	.1250
	Q1	.2210	.0704	.0805	.0171	.1909	.0492	.2698	.0865	.3650	.1360
	Q3	.1654	.0485	.1122	.0237	.1736	.0468	.1840	.0575	.2083	.0709
18	Q0	.1488	.0417	.0535	.0149	.1260	.0306	.1743	.0472	.2564	.0784
	Q1	.1551	.0460	.0361	.0126	.1226	.0299	.1949	.0549	.2824	.0911
	Q3	.0843	.0277	.0658	.0162	.0957	.0274	.0884	.0322	.0959	.0380
20	Q0	.0573	.0272	.0084	.0141	.0417	.0220	.0586	.0306	.1261	.0449
	Q1	.0732	.0302	-.0108	.0130	.0387	.0211	.0974	.0345	.1747	.0554
	Q3	-.0289	.0233	.0152	.0143	-.0120	.0241	-.0556	.0284	-.0660	.0288
20	KM	.0520	.0263	-.0040	.0151	.0237	.0219	.0613	.0281	.1320	.0427

(2) 중단된 관측치가 많은경우 (50 %)

S(t)		(1.0 .24]		(1.0 .49]		(.49 .36]		(.36 .29]		(.29 .24]	
t		(0.0 2.0]		(0.0 0.5]		(0.5 1.0]		(1.0 1.5]		(1.5 2.0]	
m		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
16	Q0	.1693	.0419	.0856	.0167	.1511	.0327	.1919	.0469	.2656	.0756
	Q1	.1645	.0417	.0654	.0129	.1364	.0279	.1943	.0482	.2781	.0822
	Q3	.1562	.0393	.1000	.0194	.1526	.0344	.1759	.0453	.2121	.0620
18	Q0	.0995	.0236	.0431	.0123	.0831	.0187	.1105	.0248	.1715	.0412
	Q1	.0960	.0242	.0221	.0101	.0671	.0156	.1151	.0262	.1894	.0475
	Q3	.0823	.0216	.0551	.0136	.0801	.0192	.0901	.0237	.1123	.0321
20	Q0	.0177	.0169	.0002	.0124	.0093	.0152	.0112	.0180	.0516	.0236
	Q1	.0175	.0174	-.0216	.0120	-.0092	.0142	.0202	.0183	.0822	.0270
	Q3	-.0115	.0180	.0091	.0130	-.0029	.0166	-.0247	.0205	-.0285	.0235
20	KM	.0039	.0165	-.0108	.0133	-.0101	.0162	-.0058	.0170	.0428	.0209

6. 결론

본 연구에서는 점진적 임의중단법으로 신뢰함수(생존함수)의 비모수적 추정문제를 다루었다. 스플라인함수를 이용하여 위험함수를 근사적으로 구하고, 생존함수와 위험률함수와의 관계식으로 부터 생존함수의 비모수적 최우추정량을 제안하였다. 즉, 몇 몇의 특정한 스플라인함수를 가정하여, Nelson형 추정량 및 두 종류의 계단함수형 추정량, 한 종류의 연속형 추정량을 얻었다.

4 절에서 실제의 자료를 가지고 제안된 추정량과 Kaplan-Meier(KM)추정량의 값을 구해본 결과, 연속형 추정량(Q3)의 값은 실험종결단계(m)가 완전한 자료의 크기(n)보다 작더라도 KM추정량과 차이가 거의 없음을 알 수 있었다.

5 절에서 제안한 추정량 (Q0, Q1, Q3)과 KM추정량의 효율성을 몬테칼로 방법으로 비교하여 다음의 결과를 얻을 수 있었다.

(1) 연속형 추정량(Q3)는 다를 추정량보다 편의와 최소제곱오차 면에서 대체로 더 유효하였다.

(2) 실험종결단계(m)가 전체 자료의 크기(n)와 같을 때라도 연속형 추정량 (Q3)은 KM추정량보다 대체로 더 유효하였다.

(3) 생존함수의 참 값이 비교적 클 때 점진적 임의중단법에서의 추정량들은 실험중단 단계(m)가 자료의 크기(n)보다 작을 때라도 편의와 평균제곱오차들은 거의 변화가 없었다.

참 고 문 헌

1. Breslow, N. and Crowley, J. (1974). A Large Sample Study of Life Table and Product Limit Estimates under Random Censorship. *The Annals of Statistics*, 2, 437-453.
2. Chatterjee, S. K. and Sen, P. K. (1973). Norparametric Testing under progressive Censoring. *Calcutta Statististical Association, Bulletion.*,22, 13-50.
3. Cohen, A. C. (1963). Progressively Censored Samples in Life Testing. *Technometrics*, 5, 327-339.
4. Cohen, A. C. (1965). Maximum Likelihood Estumation in the Weibull Distribution based on Complete and on Censored Samples. *Technometrics*, 7, 579-588.

5. Davis, C. E. (1978). A Two Sample Wilcoxon test for Progressively Censored Data. *Communications in Statistics - Theory and Method*, A7, 389-398.
6. Földes, A., Rejtö, L. and Winnnter, B. B. (1980). Strong Consistency Properties of Nonparametric Estimators for Randomly Censored data, I: The Product Limit Estimator. *Periodica Mathemetica Hungarica*, 11, 233-250.
7. Gardiner, J. C. and Sen, P. K. (1978). Asymptotic Normality of a Class of Time-sequential Statistics and Applications. *Communications in Statistics-Theory and Method*, A7, 373-388.
8. Gardiner, J. C. and Susarla V. (1982). A Nonparametric Estimator of the Survival Function under Progressive Censoring. *IMS, Lecture notes: Special Topics Meeting on Survival Analysis*, 26-40.
9. Halperin, M. and Ware, J. (1974). Early Decision in a Censored Wilcoxon two Sample Test for Accumulation Survival Data. *Journal of the American Statistical Association*, 69, 414-422.
10. Kaplan, E. L. and Meier P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.
11. Nelson, W. B. (1969). Harzard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1, 27-52.
12. Sen, P. K. (1976). Weak Convergence of Progressively Censored Likelihood Ratio Statstic and Its Role in Asymptotic Theory of Life Testing. *Annals of Statistics*, 4, 1247-1257.
13. Suzuki,K. (1985). Nonparametric Estimation of Lifetime Distributions from a Record of Failures and Follow-ups. *Journal of the American Statistical Association*, 80, 68-72.
14. Tsai, W. Y. (1986). Estimation of Survival Curves from Dependent Censorship Models via a Generalized self-consistent Property whit Nonparametric Bayesian Estimation Application. *The Annals of Statistics*, 14, 238-249.
15. Wang, M. (1987). Product Limit Estimates : A Generalized Maximum Likelihood Study. *Communications in Statistics- Theory and Method*, 16, 3117-3132.
16. Whittemore, A. S. and Keller, J. B.(1986). Survival Estimation using Splines. *Biometrics.*, 42, 496-506.
17. Wong, W. H. (1986). Theory of partial likelihood. *The Annals of Statistics*, 14, 88-123.