응용통계연구 (2010)
23(3), 521–531

구간중도절단자료에서 생존함수와 중간생존시간에 대한 추정

윤은영[1] · 김충락[2]

[1]신라대학교 의생명과학대학 간호학과, [2]부산대학교 통계학과

(2010년 4월 접수, 2010년 5월 채택)

요 약

구간중도절단은 중도절단의 가장 일반적인 개념으로 구간중도절단자료는 의학 및 역학분야의 연구에서 흔히 관찰된다. 본 연구에서는 구간중도절단의 상황에서 생존함수와 중간생존시간을 추정하는 방법으로 평균대치법과 자기일치법을 비교 연구하고, 실제 자료로 혈우병환자에서 선천성면역결핍바이러스 감염시점을 추정하였다. 또한 구간중도절단자료를 생성하는 새로운 방법을 제시하였으며, 생성된 구간중도절단자료를 이용한 모의실험을 통하여 두 추정치에 대한 다양한 비교연구를 시행하였다. 구간중도절단자료에서 생존함수와 중간생존시간을 추정할 경우 중도절단율이 크지 않다면 평균대치법이 자기일치법보다 더 우수한 추정치로 판명되었다.

주요용어: 구간중도절단자료, 평균대치법, 생존함수, 자기일치법, 중간생존시간.

1. 서론

구간중도절단(interval censoring)은 중도절단의 가장 일반적인 개념으로 우측중도절단(right censoring)과 좌측중도절단(left censoring)은 구간중도절단의 특별한 형태이며, 구간중도절단자료는 의학 또는 역학연구에서 흔히 발생한다. 생존분석에서 그간 개발되고 연구된 내용의 대부분은 우측중도절단자료에 대한 것으로 다양한 모수적, 비모수적인 추정법들이 존재한다. 그러나 구간중도절단자료는 중도절단의 특별한 메커니즘으로 인하여 우측중도절단자료를 분석하기 위한 추정법을 그대로 사용할 수 없으며, 구간중도절단자료를 분석하기위한 방법은 그리 많지 않다. 예를 들어, 우측중도절단자료의 경우 생존함수의 비모수적 최우추정치(NPMLE; nonparametric maximum likelihood estimator)인 카플란-마이어 추정치(Kaplan-Meier estimator; Kaplan과 Meier, 1958)가 간단한 식으로 표현되지만 구간중도절단의 경우 생존함수의 일반화 최우추정치는 간단한 식으로 표현되지 않고 반복법을 이용하여 구할 수밖에 없다.

구간중도절단자료는 사건이 발생한 정확한 시점을 알 수 없으며, 다만 알려진 두 개의 시점 사이에서 발생하였다는 사실만이 알려져 있다. 구간중도절단자료의 예로는 종양이 재발하는 시점을 들 수 있다. 종양이 재발한 시점은 대개 정확히 알 수 없으며, 환자가 병원을 규칙적 혹은 비규칙적으로 방문할 경우 각 방문 사이에 발생하였다는 사실만을 알 수 있다. 이러한 경우 종양재발시점은 구간중도절단 되었다고 한다. 우측중도절단과 마찬가지로 구간중도절단자료에서 사건의 발생시점을 추정하는 것은 매우 중요하며 이는 생존함수의 추정과 중간생존시간(median survival time)의 추정을 요구한다.

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
[2]교신저자: (609-735) 부산광역시 금정구 장전동 산 30, 부산대학교 자연과학대학 통계학과, 교수.
E-mail: crkim@pusan.ac.kr

지금까지 구간중도절단자료에 대하여 여러 학자들의 연구가 있었다 (Turnbull, 1976; Finkelstein, 1986; Odell 등, 1992; Taylor와 Kim, 1994). 그러나 구간중도절단자료의 발생빈도가 매우 높음에도 불구하고, 계산상의 복잡성 등으로 인하여 관련 연구는 부족한 것이 사실이다. 이처럼 구간중도절단의 상황에서 연구의 가장 큰 어려움은 우도함수의 기여도를 구간에서 계산하는 것이 매우 어렵고 복잡하기 때문이다.

본 연구에서는 구간중도절단자료를 분석하기 위해 매우 자주 사용되는 생존함수 추정법과 이를 바탕으로 중간생존시간을 추정하기 위한 방법으로 평균대치법(mean imputation method)과 자기일치법(self-consistency method)을 이용하고 두 추정치를 비교 연구한다. 평균대치법은 구간중도절단자료를 우측중도절단자료의 형태로 변환하여 우측중도절단자료의 분석방법을 사용하는 것으로 매우 간단하고 편리한 방법이다. 한편, 자기일치법은 구간중도절단의 상황에서 비모수적 최우추정치를 반복법으로 계산하여 생존함수를 추정하는 방법으로 계산과정이 복잡하지 않고 정확하다는 장점이 있다.

본 논문의 구성은 다음과 같다. 2절에서는 구간중도절단의 개념과 정의를 소개하고 생존함수와 중간생존시간의 추정치로 평균대치법(mean imputation method)과 자기일치법(self consistency method)을 소개하며 제시한 두 가지 추정법을 실제 자료인 혈우병환자의 자료에 적용하여 HIV(Human Immunodeficiency Virus)-1감염시점을 추정하였고 각각의 방법을 비교 분석하였다. 3절에서는 모의실험을 위한 설계와 구간중도절단자료를 생성하는 방법을 제시하였다. 또한, 여러 상황에서 생존함수와 중간생존시간에 대한 두 가지 추정법을 비교하였다. 마지막으로 결론을 4절에 제시하였다.

2. 구간중도절단자료의 추정

$T_1, T_2, \ldots, T_n$은 생존시간을 나타내는 확률변수로서 서로 독립이고 같은 분포를 가지며 T의 분포함수는 F, 밀도함수는 f라 하자. 또한, 생존함수를 $S(t) = 1 - F(t)$로 표현하고 $t_{.5}$를 중간생존시간이라 하자. 구간중도절단자료란 T_i를 정확하게 관찰할 수 없으며 다만 $(L_i, R_i]$라는 구간 안에 발생하였다는 것만을 알 수 있다. 즉,

$$T_i \in (L_i, R_i]$$

이며, $L_i \leq R_i$이다. 여기서 L_i는 구간의 왼쪽값(left endpoint)을 나타내며, R_i는 구간의 오른쪽값(right endpoint)을 나타낸다. 본 논문에서는 구간중도절단 된 자료 $(L_i, R_i]$, $i = 1, \ldots, n$를 이용하여 생존함수와 중간생존시간을 추정하는 문제를 다루고자 한다.

구간중도절단자료의 형태는 크게 네 가지로 나뉜다. 첫째, 제 1종 구간중도절단자료(Case I interval-censored data)는 각 관측치가 $L = 0$ 또는 $R = \infty$로만 나타나는 경우이다. 즉, 실제 발생시간은 각 관측치의 전 또는 후를 나타내는 것으로 흔히 현상자료(current state data)라고도 한다. 둘째, 제 2종 구간중도절단자료(Case II interval-censored data)는 R이 유한 또는 무한의 값을 가질 수 있는 경우로 가장 일반적인 구간중도절단의 개념으로 사용된다. 예를 들어, $L = 0$이고 R이 유한이면 좌측중도절단을 나타내고 L이 유한이며 $R = \infty$면 우측중도절단을 나타낸다. 즉, 제 1종 구간중도절단, 좌측중도절단, 우측중도절단 등은 모두 제 2종 구간중도절단의 특별한 경우라고 할 수 있다. 셋째, 중복중도절단자료(doubly censored data)는 두 가지의 생존시간이 모두 중도절단 된 경우다. 예를 들어, HIV(human immunodeficiency virus) 감염 후 AIDS(acquired immunodeficiency syndrome) 발병 때까지의 시간에 관심이 있는데 HIV 감염시기와 AIDS 발병시기 모두 중도절단 된 경우다. 마지막으로 패널 카운트 자료(panel count data) 등이 있다. 본 연구에서는 구간중도절단자료 중 가장 일반적인 형태인 제 2종 구간중도절단자료를 대상으로 생존함수를 추정하고 이를 바탕으로 중간생존시간을 추정한다. 본 연구

에서 예제로 사용하게 될 혈우병환자 자료는 제2종 구간중도절단자료의 전형적인 예이며, 이에 대한 대표적 연구로는 Groeneboom과 Wellner (1992), Huang과 Wellner (1997)와 Sun (1998, 2005, 2006) 등이 있다.

2.1. 평균대치법

대치법은 결측자료를 처리하는 방법으로 널리 사용되고 있다 (Rubin, 1987). 대치법은 크게 단일 대치법(single imputation method)과 다중 대치법(multiple imputation method) 등의 두 가지로 구분하며, 구간중도절단자료에서 대치법의 사용 목적은 자료를 우측중도절단자료의 형태로 만드는 것이다. 본 연구에서는 단일 대치법만을 고려하는데 이는 다중 대치법에 비해 효율성은 크게 뒤지지 않지만 계산이 편리하고 빠르다는 장점이 있기 때문이다. 단일 대치법에는 대치방법에 따라 좌대치법(left imputation), 우대치법(right imputation), 평균대치법(mean imputation) 등이 있는데 이 중에서 평균대치법이 가장 뛰어난 것으로 알려져 있다.

평균대치법이란 구간중도절단 된 자료 $(L_i, R_i], i = 1, \ldots, n$ 대신 각 구간의 평균값을 관측치로 사용하여 구간중도절단자료를 우측중도절단자료로 변환하는 방법이다. 우측중도절단은 생존시간 $T_1, T_2, \ldots, T_n$ 대신 (Y_i, δ_i), $i = 1, 2, \ldots, n$을 관측하는데 여기서 $Y_i = \min(T_i, C_i)$이고 C_i는 중도절단을 나타내는 변수, $\delta_i = I(T_i \leq C_i)$는 중도절단을 나타내는 지시함수이다. 평균대치법은 구간중도절단 된 자료에서 $R_i < \infty$이면 $(L_i, R_i]$ 대신 구간의 평균인 $Y_i = (L_i + R_i)/2$를 중도절단 되지 않은 관측치 (즉, $\delta_i = 1$)로 간주하고 $R_i = \infty$이면 (즉, $(L_i, \infty]$) $Y_i = L_i$로 간주하고 이를 중도절단 되었다고 (즉, $\delta_i = 0$) 취급한다. 따라서, 구간중도절단자료는 평균대치법에 의해 우측중도절단자료로 변환되는 것이다. 평균대치법의 장점은 과정이 복잡하지 않고 비교적 간단하며, 우측중도절단자료를 위해 개발된 기존의 다양한 추정치들을 사용하여 추론할 수 있으므로 편리하다는 점이다.

2.2. 자기일치법

우측중도절단자료에서 생존함수의 비모수적 최대우도추정치(NPMLE)는 바로 카플란-마이어 추정치이다. 구간중도절단자료에서 비모수적 추론은 우측중도절단자료에서보다 더 복잡하다. 실제로 구간중도절단자료에서 비모수적 최대우도추정치는 식으로 표현되지 않으며 반복알고리즘을 통해 수리적 값만을 계산할 수 있다.

구간중도절단자료에서 생존함수의 NPMLE를 계산하는 방법에는 세 가지가 있다. 첫 번째는 Turnbull (1976)이 제안한 자기일치법(self-consistency method)이며, 두 번째는 iterative convex minorant(ICM) 알고리즘 (Groeneboom과 Wellner, 1992; Jongbloed, 1998), 세 번째는 자기일치 알고리즘과 ICM 알고리즘의 혼합방법인 EM-ICM 알고리즘 (Wellner과 Zhan, 1997)이 있다. 이 세 가지 알고리즘 중 ICM과 EM-ICM은 자기일치 알고리즘보다 좀 더 빠른 장점이 있으나, 자기일치 알고리즘이 간단하면서 정확하기 때문에 지금까지도 자주 쓰이고 있다. 그러므로 본 연구에서도 생존함수의 NPMLE를 추정하는 방법으로 자기일치 알고리즘을 사용하였다.

자기일치 알고리즘 (Gentleman과 Geyer, 1994; Turnbull, 1976)을 소개하기 위해 다음과 같은 기호를 정의한다. 먼저, $\{s_j\}_{j=0}^m$를 집합 $\{0, L_i, R_i \ i = 1, \ldots, n\}$에서 중복되지 않는 값들을 순서대로 배열한 벡터라 하고,

$$\alpha_{ij} = I(s_j \in (L_i, R_i]), \qquad p_j = S(s_{j-1}) - S(s_j), \quad i = 1, \ldots, n, \ j = 1, \ldots, m$$

으로 정의할 때, 우도함수는

$$L_s(\boldsymbol{p}) = \prod_{i=1}^{n}[S(L_i) - S(R_i)] = \prod_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}p_j$$

이며, 이 때 $\boldsymbol{p} = (p_1, \ldots, p_m)'$이다. 따라서, 생존함수 $S(\cdot)$에 대한 NPMLE는 $L_s(\boldsymbol{p})$를 $\boldsymbol{p}$에 대해 최대화하되 다음의 제약조건

$$\sum_{j=1}^{m} p_j = 1, \quad p_j \geq 0,\ j = 1, \ldots, m$$

을 만족해야 한다. 이를 바탕으로 자기일치 알고리즘은 다음과 같이 주어진다.

단계 0: 먼저 $\boldsymbol{p}$의 초기값 $\hat{\boldsymbol{p}}^{(0)}$을 정한다.

단계 1: l번째 반복에서 $\boldsymbol{p}$의 추정치 $\hat{\boldsymbol{p}}^{(l)} = (\hat{p}_1^{(l)}, \ldots, \hat{p}_m^{(l)})'$을 다음과 같이 계산한다.

$$\hat{p}_j^{(l)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_{ij}\hat{p}_j^{(l-1)}}{\sum_{k=1}^{m}\alpha_{ik}\hat{p}_k^{(l-1)}}, \quad j = 1, \ldots, m.$$

단계 2: 단계 1을

$$\left|\left| \hat{\boldsymbol{p}}^{(l)} - \hat{\boldsymbol{p}}^{(l-1)} \right|\right| < \gamma$$

이 이루어질 때 까지 반복한다. 이 때 γ는 미리 정해진 상수이다.

2.3. 예제

앞서 제시된 두 가지 생존함수의 추정방법을 혈우병(hemophilia) 환자의 자료 (Goedert 등, 1989)에 적용하였다. 본 자료는 혈우병이 있는 환자를 대상으로 HIV-1의 감염에 영향을 미치는 위험요인을 규명하기 위한 전향적 연구로부터 수집되었으며, 총 16개 병원으로부터 수집되었다 (표 2.1 참조). 혈우병이 있는 환자는 치료를 위해 수천 명의 수혈자들로부터 생성되는 혈청(plasma)으로부터 만들어진 혈액제재(농축 factor VIII와 factor IX)를 사용해야 되기 때문에 HIV에 감염될 위험이 높다. 혈우병 환자가 HIV에 감염된 시점을 파악하는 것은 매우 중요한데, 정확한 감염시기를 알 수는 없고 단지 혈액제제를 주사맞기 시작한 시점 이후부터 HIV에 감염이 된 것을 진단한 시점 사이라는 것만을 알 수 있다. 연구대상자는 총 368명이었으며, 환자들은 투여 받은 혈액제재(농축 factor VIII)의 평균 용량에 따라 두 군으로 나뉘어졌는데, 본 연구에서는 그 중 저용량(low dose)을 사용한 총 132명의 환자의 자료에 대해 분석을 시행하였다.

생존함수 $S(t)$의 추정치로 평균대치법과 자기일치법에 의한 것을 각각 $\hat{S}_M(t)$와 $\hat{S}_S(t)$로 표현하였으며 이를 바탕으로 한 중간생존시간의 추정치는 각각 $\hat{t}_M$와 $\hat{t}_S$로 표현하였다. 그림 2.1에 주어진 것처럼 $\hat{t}_M$는 26.5일이었으며, $\hat{t}_S$는 26.0일이었다.

3. 모의실험

3.1. 모의실험 설계

구간중도절단 자료에서 생존함수 및 중간생존시간에 대한 추정법으로 평균대치법과 자기일치법을 고려하였으며 두 추정법의 비교를 위해 폭 넓은 모의실험을 시행하였다. 모의실험에서 사용된 표본의 크기

표 2.1. 16개 병원으로 부터 수집된 저용량 혈액제재를 투여받은 132명의 혈우병 환자가 HIV-1에 감염된 시간에 대한 구간중도절단 자료.

환자	L_i	R_i	환자	L_i	R_i	환자	L_i	R_i
1	7	20	45	25	34	89	30	–
2	9	20	46	53	–	90	45	–
3	0	25	47	41	–	91	21	26
4	57	–	48	50	–	92	16	32
5	23	26	49	0	36	93	17	24
6	8	21	50	0	29	94	49	–
7	20	26	51	55	–	95	0	37
8	25	27	52	0	55	96	0	41
9	24	29	53	10	16	97	0	30
10	12	21	54	13	29	98	56	–
11	26	29	55	14	19	99	0	30
12	54	–	56	0	16	100	55	–
13	18	22	57	11	29	101	51	–
14	14	22	58	11	20	102	0	30
15	11	17	59	31	–	103	50	–
16	55	–	60	40	–	104	45	–
17	8	15	61	53	–	105	8	30
18	29	31	62	11	15	106	5	30
19	55	–	63	20	24	107	53	–
20	57	–	64	15	20	108	11	41
21	15	20	65	32	–	109	52	–
22	18	22	66	54	–	110	3	33
23	14	22	67	51	–	111	0	47
24	56	–	68	33	–	112	7	49
25	23	30	69	17	26	113	56	–
26	17	21	70	14	17	114	57	–
27	54	–	71	41	–	115	6	29
28	20	31	72	42	–	116	7	29
29	56	–	73	53	–	117	55	–
30	23	27	74	0	26	118	8	29
31	56	–	75	49	–	119	7	29
32	53	–	76	39	–	120	36	–
33	15	19	77	18	29	121	7	28
34	52	–	78	22	25	122	55	–
35	0	17	79	50	–	123	49	–
36	0	21	80	54	–	124	46	–
37	46	–	81	38	–	125	0	30
38	16	23	82	20	30	126	57	–
39	24	32	83	46	–	127	30	–
40	16	24	84	51	–	128	53	–
41	53	–	85	6	30	129	12	21
42	12	20	86	53	–	130	56	–
43	18	22	87	0	30	131	38	–
44	0	33	88	45	–	132	0	44

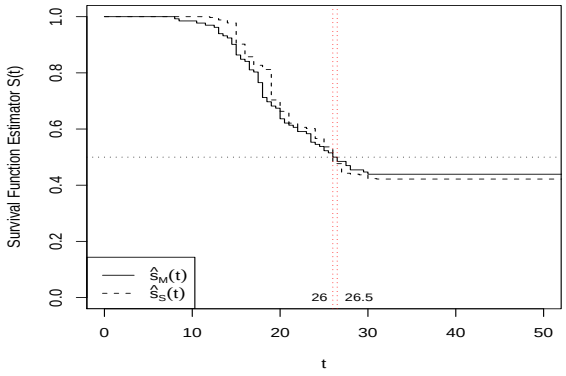

그림 2.1. 혈우병 자료(표 2.1)를 이용한 생존함수 및 중간생존시간에 대한 평균대치법과 자기일치법 추정치.

는 $n = 30, 50$이었고 반복횟수는 100번이었다. 또한, 평균이 1인 지수분포로 부터 난수를 생성하였다. 앞 절의 예제에서처럼 생존함수 $S(t)$의 추정치로 평균대치법과 자기일치법에 의한 것을 각각 $\hat{S}_M(t)$와 $\hat{S}_S(t)$로 표현하였으며 이를 바탕으로 한 중간생존시간의 추정치는 각각 $\hat{t}_M$와 $\hat{t}_S$로 표현하였다.

두 가지 추정치를 비교하기 위하여, 중간생존시간은 평균오차제곱(mean squared error; MSE)으로, 생존함수의 추정은 적분평균오차제곱(mean integrated squared error; MISE)으로 평가하였으며 이들은 다음과 같이 정의된다.

$\hat{\theta}(t)$를 t가 주어졌을 때 $\theta(t)$의 추정치라 하면, $\hat{\theta}(t)$의 MSE는

$$\mathrm{MSE}\left(\hat{\theta}(t)\right) = E\left[\left(\hat{\theta}(t) - \theta(t)\right)^2\right]$$

으로 정의되고, MISE는 생존시간을 나타내는 확률변수 T의 전 범위에서 다음과 같이 정의된다.

$$\mathrm{MISE}\left(\hat{\theta}\right) = \int_0^\infty E\left[\left(\hat{\theta}(t) - \theta(t)\right)^2\right] dt.$$

MSE와 MISE를 모의실험 자료로 추정하기 위해 i번째 발생된 난수를 이용하여 계산한 $t_{.5}$에 대한 추정치를 $\hat{t}_i$이라 하면 r번 반복하여 구한 MSE에 대한 추정치는 다음과 같다.

$$\widehat{\mathrm{MSE}}\left(\hat{t}\right) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{t}_i - t_{.5}\right)^2.$$

한편, MSE는 분산(variance; V)과 편의제곱(squared bias; SB)의 합으로 표현할 수 있다. 즉,

$$\begin{aligned}\widehat{\mathrm{MSE}}\left(\hat{t}\right) &= V + \mathrm{SB} \\ &= \frac{1}{r}\sum_{i=1}^{r}\left(\hat{t}_i - \bar{t}\right)^2 + \left(\bar{t} - t_{.5}\right)^2,\end{aligned}$$

단, 여기서

$$\bar{t} = \sum_{i=1}^{r}\frac{\hat{t}_i}{r}$$

는 r번 반복으로 구한 추정치의 산술평균이다.

이와 비슷한 방법으로, t가 구간 (a,b)상의 어떤 시점일 때 생존함수 $S(t)$의 추정치에 대하여 MISE를 정의할 수 있다. 우선 $\Delta=(b-a)/m$를 구간 (a,b)를 m등분한 m개의 구간 I_j, $j=1,\ldots,m$에 대한 길이라 하자. 또한, $\hat{S}_i(t)$를 i번째 난수를 이용하여 구한 $S(t)$에 대한 추정치라 하고 r번 반복하여 구한 경우 MISE에 대한 몬테칼로 추정치는

$$\widehat{\mathrm{MISE}}\left(\hat{S}\right)=\frac{1}{r}\sum_{i=1}^{r}\sum_{t\in I_1}^{I_m}\left(\hat{S}_i(t)-S(t)\right)^2\Delta.$$

또한, MISE는 적분분산(integrated variance; IV)과 적분편의제곱(integrated squared bias; ISB)의 합으로 나타낼 수 있다. 즉,

$$\begin{aligned}\widehat{\mathrm{MISE}}\left(\hat{S}\right)&=\mathrm{IV}+\mathrm{ISB}\\&=\frac{1}{r}\sum_{i=1}^{r}\sum_{t\in I_1}^{I_m}\left(\hat{S}_i(t)-\bar{S}(t)\right)^2\Delta+\sum_{t\in I_1}^{I_m}\left(\bar{S}(t)-S(t)\right)^2\Delta,\end{aligned}$$

이 때,

$$\bar{S}(t)=\sum_{i=1}^{r}\frac{\hat{S}_i(t)}{r}$$

이다.

3.2. 구간중도절단자료의 생성

구간중도절단자료를 이용한 모의실험을 시행함에 있어서 가장 어려운 점은 구간중도절단자료의 생성이다. 우측중도절단자료를 생성하는 것은 절단(censoring)의 비율을 정하기만 하면 크게 어렵지 않으나, 구간중도절단자료의 생성은 여러 관점에서 더 복잡하다. 첫째, 중도절단율과 깊은 관련이 있는 구간의 길이를 결정하는 것이 쉽지 않다. 만약 구간의 길이를 길게 할 경우 대부분의 자료가 중도절단이 되고, 구간의 길이를 짧게 할 경우 중도절단율이 너무 작아지게 된다. 둘째, 생성된 구간 중 우측값, 즉 R_i를 ∞로 만들어 줄 갯수를 적절히 조절해야한다. 이는 평균대치법에서 중도절단율에 해당되기 때문이다. 본 연구에서 구간중도절단자료의 구간, 즉 $(L,R]$의 생성방법을 다음과 같이 제안한다.

단계 1: 특정분포(예: Weibull(λ,γ))로부터 T를 생성하였다.

단계 2: 주어진 상수 a에 대해 일양분포 $U(0,a)$로 부터 U를 생성하고 이를 바탕으로 $L=\max(0,T-U)$과 $R=T+U$을 생성한다.

단계 3: 생성된 구간자료들 중에서 구간의 오른쪽 값인 R이 상위 $k\%$, $k=10,30$에 해당되면 이를 ∞로 변환한다.

3.3. 모의실험 결과

앞에서 언급했듯이 본 모의실험에서 사용된 표본의 크기는 $n=30,50$이고 반복횟수는 100번이다. 또한, 평균이 1인 지수분포로 부터 난수를 생성하였다. 난수 생성과정에서 일양분포의 범위를 결정하는 a값으로 0.5와 1을 사용했으며 구간의 오른쪽 값이 ∞가 되는 비율은 각각 10%와 30%를 고려하였다.

먼저 생존함수 추정치를 비교해 보면(표 3.1 참조) 무한대의 비율이 10%인 경우는 a의 값에 관계없이 항상 평균대치법이 자기일치법보다 더 좋았으며 무한대의 비율이 30%인 경우는 정반대로 a의 값에 관

표 **3.1.** 생존함수 $S(t)$에 대한 두 추정치 $\hat{S}_M$와 $\hat{S}_S$에 대한 MISE. 난수는 평균이 1인 지수분포로 부터 생성되었으며 표본크기는 $n = 30$와 $n = 50$에 대해 실시하였고 구간의 폭을 나타내는 일양분포의 크기 a는 0.5와 1에 대해 실시하였다 (괄호안의 값은 각각 적분분산과 적분편의오차를 나타냄).

% of ∞	a	$n = 30$		$n = 50$	
		$\hat{S}_M$	$\hat{S}_S$	$\hat{S}_M$	$\hat{S}_S$
10%	0.5	.0203 (.0182 + .0021)	.0288 (.0271 + .0017)	.0150 (.0122 + .0028)	.1849 (.0159 + .0025)
	1	.0242 (.0182 + .0060)	.0367 (.0339 + .0027)	.0182 (.0116 + .0066)	.0199 (.0179 + .0019)
30%	0.5	.0782 (.0422 + .0359)	.0733 (.0479 + .0253)	.0799 (.0335 + .0464)	.0709 (.0348 + .0361)
	1	.0682 (.0432 + .0249)	.0666 (.0542 + .0124)	.0709 (.0351 + .0358)	.0632 (.0382 + .0250)

표 **3.2.** 중간생존시간 $t_{.5}$에 대한 두 추정치 $\hat{t}_M$와 $\hat{t}_S$에 대한 MSE. 난수는 평균이 1인 지수분포로 부터 생성되었으며 표본크기는 $n = 30$와 $n = 50$에 대해 실시하였고 구간의 폭을 나타내는 일양분포의 크기 a는 0.5와 1에 대해 실시하였다 (괄호안의 값은 각각 분산과 편의오차를 나타냄).

% of ∞	a	$n = 30$		$n = 50$	
		$\hat{t}_M$	$\hat{t}_S$	$\hat{t}_M$	$\hat{t}_S$
10%	0.5	.0374 (.0359 + .0015)	.0458 (.0458 + .0000)	.0239 (.0238 + .0000)	.0318 (.0313 + .0004)
	1	.0359 (.0298 + .0061)	.0536 (.0526 + .0010)	.0175 (.0160 + .0014)	.0308 (.0302 + .0005)
30%	0.5	.0328 (.0311 + .0016)	.0399 (.0399 + .0000)	.0323 (.0323 + .0000)	.0353 (.0351 + .0001)
	1	.0255 (.0209 + .0046)	.0409 (.0407 + .0002)	.0145 (.0145 + .0000)	.0216 (.0183 + .0033)

계없이 항상 자기일치법이 평균대치법보다 더 좋았다. 즉, 무한대의 비율이 높아지면 평균대치법에서 중도절단율이 커지므로 이런 현상이 발생한 것으로 생각된다. 한편, 그림 3.1과 3.2는 각각 생존함수에 대한 두 추정치를 한 번의 난수생성에 근거한 것($r = 1$)과 100번의 추정치를 평균한 것을 나타내고 있다.

다음으로 중간생존시간의 추정치를 비교해 보면(표 3.2 참조) 무한대의 비율에 관계없이 항상 평균대치법이 자기일치법보다 더 좋았다. 즉, 중간생존시간 근처에서는 평균대치법이 자기일치법보다 참 생존곡선에 더 가깝다는 사실을 보여준다.

4. 결론

본 논문에서는 구간중도절단자료의 생존함수를 추정하는 두 가지 방법으로 평균대치법과 자기일치법을 소개하고, 혈우병 환자의 HIV-1 감염시점을 추정하는데 적용하였다. 평균대치법이란 구간중도절단 된 자료 대신 각 구간의 평균값을 관측치로 사용하여 구간중도절단자료를 우측중도절단자료로 변환함으로써 기존의 우측중도절단자료를 분석하는 기법을 이용할 수 있다. 한편, 자기일치법은 구간중도절단자료에서 비모수적 최대우도추정치를 구하는 알고리즘 중의 하나로 계산식이 매우 간단하고 효율성이 뛰어난 것으로 알려져 있다. 또한, 구간중도절단자료를 생성하는 방법을 제안하였으며 이를 토대로 여러 상황에서의 모의실험을 실시하였다.

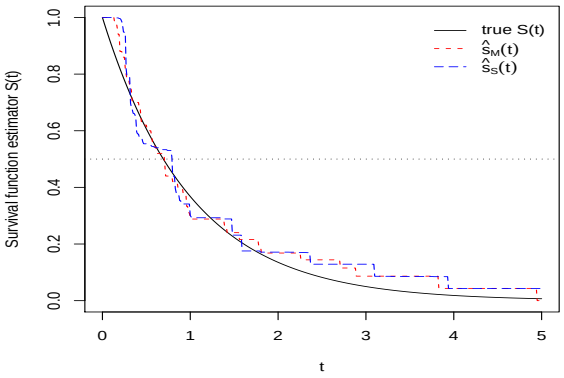

그림 3.1. 1회의 난수생성($r = 1$)을 이용하여 그린 생존함수에 대한 두 추정치($n = 50,\ a = 1$)의 경우.

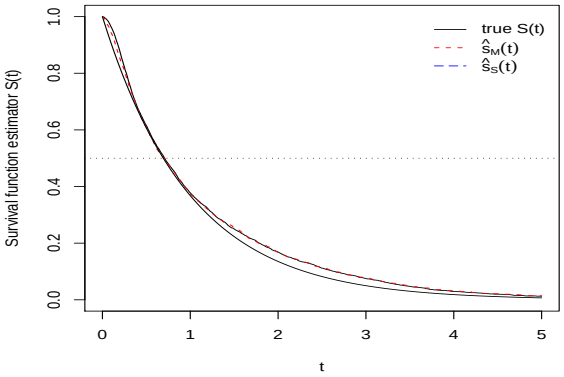

그림 3.2. 100회의 난수생성($r = 100$)을 이용하여 그린 생존함수에 대한 두 추정치($n = 50,\ a = 1$)의 경우 - 매회 생성된 난수를 이용하여 구한 추정치들의 평균을 나타낸 것임.

생존함수를 추정하는 관점에서는 중도절단율이 작은 경우 평균대치법이 자기일치법보다 더 좋았으며 중도절단율이 큰 경우는 정반대로 나타났다. 한편, 중간생존시간의 추정치를 비교해 보면 중도절단율에 관계없이 항상 평균대치법이 자기일치법보다 더 좋았다. 결론적으로 구간중도절단자료에서 중도절단율이 심각하지 않다면 생존함수와 중간생존시간의 추정치로 평균대치법이 매우 안정적인 추정치임을 알 수 있었다.

참고문헌

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics*, **42**, 845–854.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation, *Biometrika*, **81**, 618–623.

Goedert, J. J., Kessler, C. M., Aledort, L. M., Biggar, R. J., Andes, W. A., White, G. C., Drummond, J. E., Vaidya, K., Mann, D. L., Eyster, M. E., Lederman, M. M., Hilgartner, M. W., Ragni, M. V., Cohen, A. R., Gordon, L. B., Rosenberg, P. S., Friedman, R. M., Blattner, W. A., Kroner, B. and Gail, M. H. (1989). A prospective-study of human immunodeficiency virus type-1 infection and the development of AIDS in subjects with hemophilia, *New England Journal of Medicine*, **321**, 1141–1148.

Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation, *DMV Seminar, Band 19*, Birkhauser, New York.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: A review of recent progress, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis,* eds. Lin, D. and Fleming, T. Springer Verlag, New York, 123–169.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation, *Journal of Computational and Graphical Statistics*, **7**, 310–321.

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

Odell, P. M., Anderson, K. M. and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model, *Biometrics*, **48**, 951–959.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley, New York.

Sun, J. (1998). *Interval Censoring. Encyclopedia of Biostatistics*, John Wiley, First Edition, New York, 2090–2095.

Sun, J. (2005). *Interval Censoring. Encyclopedia of Biostatistics*, John Wiley, Second Edition, New York, 2603–2609.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer Verlag, New York.

Taylor, J. M. G. and Kim, D. K. (1994). Marker values at the time of an AIDS diagnosis, *Statistics in Medicine*, **13**, 2059–2066.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B*, **38**, 290–295.

Wellner, J. and Zhan, Y. (1997). A hybird algorithm for computation of the nonparametric maximum likelihood estimator from censored data, *Journal of the American Statistical Association*, **92**, 945–959.

Estimation of Survival Function and Median Survival Time in Interval-Censored Data

Eunyoung Yun[1] · Choongrak Kim[2]

[1]Department of Nursing, College of Medical and Life Science, Silla University
[2]Department of Statistics, Pusan National University

(Received April 2010; accepted May 2010)

Abstract

Interval-censored observations are common in medical and epidemiologic studies; however, limited studies exist due to the complexity and special structure of interval-censoring. This paper introduces the imputation method and the self consistency method in the interval-censored data. We propose a new method of generating random numbers under an interval-censoring set-up. Through simulation studies we compare two methods under various simulation schemes in the sense of the mean squared error for estimating the median survival time and the mean integrated squared error for estimating the survival function. Under a moderate censoring percentage, the mean imputation method showed a better performance than the self-consistency method in estimating the median survival time and the survival function.

Keywords: Interval-censored data, mean imputation, median survival time, self consistency, survival function.

This work was supported for two years by a Pusan National University Research Grant.
[2]Corresponding author: Professor, Department of Statistics, Pusan National University, Pusan 609-735, Korea. E-mail: crkim@pusan.ac.kr