고급생존분석

3장 중도절단 및 우도함수

2020년 가을학기

전북대학교 통계학과

중도절단의 종류

중도절단의 종류

- **제1유형 우중도절단** : 미리 정해놓은 시간에서 관측을 중단하는 경우
  공학연구에서 주로 발생, 모든 실험 개체의 우중도절단시간이 미리 정해놓은 시간으로 동일

  예1) 연봉조사: 5600만원인데 5000만원으로 기입

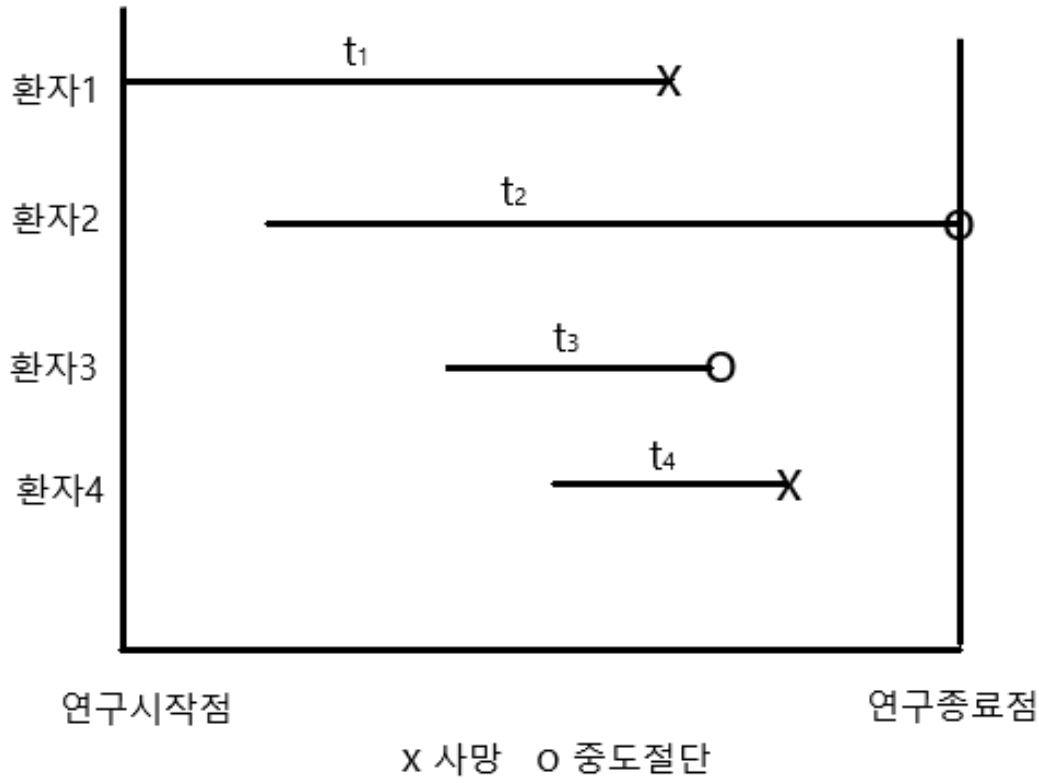
  예2) 장기이식수술 후 63개월 생존했으나 연구기간은 60개월인 경우

- **제2유형 우중도절단** : 전체 실험 개체들 중 미리 정해놓은 사건 발생률까지 관측 후 중지하는 경우
  이 경우에도 모든 관측개체는 같은 중도절단시간을 가짐

  예) 전구 수명 실험: 전구 100개 켜놓은 상태에서 5개가 꺼질 때 관측 종료

중도절단의 종류

- **임의 우중도절단 (random right censoring)** : 서로 다른 시점에서 연구에 참여하여 연구종료 후에 사건이 발생하거나 다른 이유로 연구종료이전에 중단된 경우



중도절단의 종류

- **좌중도절단 (left censoring)** : 사건이 연구시작 이전에 발생하는 경우

  예) 고등학생 흡연조사에서 "중학교 때 피웠는데 언제인지 기억이 나지 않는다" 라는 학생의 대답

- **구간중도절단 (interval censoring)** : 사건이 어떤 구간 내에서 발생하는 경우

  예1) 임상시험에서 주기적으로 환자가 방문하여 검사할 경우 환자의 사건발생시간은 구간 $(L_i, R_i]$내라는 정보만 있음

  예2) 산업분야에서 장비에 대한 점검이 주기적으로 이루어지는 경우

중도절단 이유

중도절단 이유

- 연구종료
- 사건사고:연구문제와 다른 이유로 사망

  예) 폐암환자 연구에서 연구기간 내 교통사고로 사망
- 추적실패
- 경쟁위험모형: 다른 종류의 사건발생으로 인해 관심사건 중도절단

  예) 폐암환자 연구 중 골수암 발생으로 골수암 연구대상자로 관리 필요
- 측정기구의 한계

중도절단 데이터를 포함한 우도함수

(중도절단이 없는 경우) 우도 함수 (likelihood function): $L(\theta)$

$Y_1, Y_2, \ldots, Y_n$: $f(y|\theta)$를 확률밀도함수로 갖는 모집단으로부터의 랜덤샘플

- 결합확률밀도함수

$$f(\boldsymbol{y}|\theta) = f(y_1|\theta)f(y_2|\theta)\cdots f(y_n|\theta)$$

  - 주어진 θ에 대하여 $\boldsymbol{y}$를 관측할 확률밀도를 나타냄
  - 확률밀도함수는 고정된 모수 θ의 값에서 $\boldsymbol{y}$의 함수

- 우도 함수 (likelihood function)

$$L(\theta) = f(y_1|\theta)f(y_2|\theta)\cdots f(y_n|\theta)$$

  - $\boldsymbol{Y} = \boldsymbol{y}$를 관측한 후 $\boldsymbol{y}$는 고정되어 있는 것으로 생각하고 $f(\boldsymbol{y}|\theta)$를 θ의 함수로 간주.

제1유형 임의 우중도절단 데이터

용어

- Y_i: i번째 개체의 생존시간
- C_i: i번째 개체의 중도절단시간
- $T_i = \min(Y_i, C_i)$: i번째 개체의 관측시간
- $\Delta_i = I(Y_i \leq C_i)$: i번째 개체의 중도절단여부
  - 사건을 실제 관측한 경우 (즉 중도절단이 아닌 경우)=1, (if $T_i = Y_i$)
  - 사건을 실제 관측하지 않은 경우 (즉 중도절단인 경우)=0 (if $T_i < Y_i$)

NOTE

- 잠재적인 데이터(potential data): $\{(Y_1, C_1), (Y_2, C_2), \ldots, (Y_n, C_n)\}$
- 실제 관측하는 데이터(actual observed data):
  $\{(T_1, \Delta_1), (T_2, \Delta_2), \ldots, (T_n, \Delta_n)\}$
- 생존시간 Y의 분포에 관심, 즉 $f(t), F(t), S(t), h(t)$ 추정
- 중도절단시간 C의 해당 함수들 : $g(t), G(t), U(t), v(t)$

제1유형 임의 우중도절단 데이터를 포함한 우도함수

- 우도함수 구할 때 중도절단의 유형을 고려
- **가정**: 생존시간과 중도절단시간은 서로 독립

(T, Δ)의 확률밀도함수

$$f(t, \delta) = \lim_{a \to 0} \frac{P(t \leq T < t + a, \Delta = \delta)}{a}, \ t \geq 0, \delta = 0, 1$$

NOTE: Y의 $f(y)$와 (T, Δ)의 $f(t, \delta)$를 혼동하지 말 것

제1유형 임의 우중도절단 데이터를 포함한 우도함수

Case 1 : $\delta = 1$, i.e. $Y \le C$, $T = \min(Y, C) = Y$

$$P(t \le T < t + a, \Delta = 1) = P(t \le Y < t + a, Y \le C)$$

$$\approx P(t \le Y < t + a, t \le C) = P(t \le Y < t + a) \times P(C \ge t)$$

$$f(t, \delta = 1) = \lim_{a \to 0} \frac{P(t \le Y < t + a) \times P(t \le C)}{a} = f(t) \times U(t)$$

Case 2 : $\delta = 0$, i.e. $Y > C$, $T = \min(Y, C) = C$

$$P(t \le T < t + a, \Delta = 0) = P(t \le C < t + a, Y > C)$$

$$\approx P(t \le C < t + a, Y \ge t) = P(t \le C < t + a) \times P(Y \ge t)$$

$$f(t, \delta = 0) = \lim_{a \to 0} \frac{P(t \le C < t + a) \times P(Y \ge t)}{a} = g(t) \times S(t)$$

(T, Δ)의 확률밀도함수

$$f(t, \delta) = [f(t) \times U(t)]^{\delta} [g(t) \times S(t)]^{1-\delta} = \{f(t)^{\delta} S(t)^{1-\delta}\}\{g(t)^{1-\delta} U(t)^{\delta}\}$$

제1유형 임의 우중도절단 데이터를 포함한 우도함수

- 생존시간 Y의 확률밀도함수는 $f(y;\theta)$ 이고
- 실제 관측하는 데이터 $\{(T_i, \Delta_i),\ i = 1, \ldots, n\}$가 주어져 있을 때

θ의 우도함수 $L(\theta)$

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(t_i, \delta_i) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} g(t_i)^{1-\delta_i} U(t_i)^{\delta_i} \\
&\propto \prod_{i}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\frac{f(t_i)}{S(t_i)} \right)^{\delta_i} S(t_i) = \prod_{i=1}^{n} (h(t_i))^{\delta_i} \exp[-H(t_i)]
\end{aligned}
$$

'$\propto$'대신에 '='사용, 즉 $L(\theta) = \prod_{i}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} (h(t_i))^{\delta_i} \exp[-H(t_i)]$

일반적인 중도절단의 경우 우도함수

: $L(\theta) = \prod_{d \in D} f(t_d) \prod_{r \in R} S(t_r)$

D: 사건발생시간의 집합, R: 우중도절단시간의 집합

- 생존시간 관측값의 경우, 즉 $Y = t_d$일 때 우도함수 기여값: $f(t_d)$
- 우중도절단 관측값의 경우, 즉 $Y \geq t_r$일 때 우도함수 기여값: $S(t_r)$

일반적인 중도절단의 경우 우도함수

- 좌중도절단 관측값의 경우, 즉 $t_l \geq Y$일 때 우도함수 기여값: $1 - S(t_l)$
- 구간중도절단 관측값의 경우, 즉 $t_l < Y \leq t_r$일 때 우도함수 기여값: $S(t_l) - S(t_r)$

$$P(t_l < Y \leq t_r) = P(Y > t_l) - P(Y > t_r) = S(t_l) - S(t_r)$$

$$L(\theta) = \prod_{d \in D} f(t_d) \prod_{r \in R} S(t_r) \prod_{l \in L} [1 - S(t_l)] \prod_{i \in I} [S(t_{il}) - S(t_{ir})]$$

L: 좌중도절단시간의 집합, I:구간중도절단시간의 집합

제1유형 임의 우중도절단 데이터를 포함한 우도함수

예제 3.4(a)

A 병원의 5명의 환자 표본에 대한 생존시간 데이터가 다음과 같을 때 우도함수식을 쓰시오

번호	이름	생존시간	(a) 우도함수 기여값 L_i	(b) 우도함수 기여값 L_i
1	김	$t = 2$		
2	위	$t > 8$ (우중도절단)		
3	현	$t = 6$		
4	신	$t > 2$(우중도절단)		
5	민	$4 < t < 6$ (구간중도절단)		

예제 3.4(b)

생존시간이 평균 $1/\lambda$인 지수분포를 따를 때 위 데이터에 대한 우도함수를 구하시오

제2유형 임의 우중도절단 데이터를 포함한 우도함수

전체 실험 개체들 중 미리 정한 사건발생률까지 관측 후 중지

제2유형 임의 우중도절단 데이터에 대한 우도함수

r번째 사건발생 시 실험을 멈춘다고 할 때 생존시간의 순서통계량

$t_{(1)} < t_{(2)} < \cdots < t_{(r)}$

$$L(\theta) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f(t_{(i)})\right][S(t_{(r)})]^{n-r}$$

최대 우도 추정 (maximum likelihood estimation)

최대 우도 추정량(maximum likelihood estimator)
$\hat{\theta} = \text{argmax}_\theta \, L(\theta)$: $L(\theta)$를 최대로 하는 θ

로그 우도함수 (log-likelihood function) : $\log L(\theta)$, $l(\theta)$

$l(\theta)$를 이용한 MLE

- 로그함수는 증가함수이므로 $\hat{\theta} = \text{argmax}_\theta \, l(\theta)$
- $l(\theta)$: 일반적으로 concave function (위로볼록), θ에 대하여 미분 가능

$$\hat{\theta} : \frac{d}{d\theta} l(\theta) = 0 \text{ 의 해}$$

최대 우도 추정 (maximum likelihood estimation)

최대우능도 추정량의 점근적 성질

- $\hat{\theta}$는 θ의 일치추정량: $\hat{\theta} \xrightarrow{P} \theta$
- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$

모수 θ에 대한 근사적 신뢰구간

방법 1: 피셔정보수 $I(\theta_0)$에서 미지의 θ_0 을 추정값 $\hat{\theta}$으로 대체

$$\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}, \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}\right)$$

방법 2: 피셔정보수$I(\theta_0)$ 대신 $-\frac{1}{n}\frac{d^2}{d\theta^2}\log L(\theta)\Big|_{\theta=\hat{\theta}}$ 사용

$$\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{-\frac{d^2}{d\theta^2}\log L(\theta)\Big|_{\theta=\hat{\theta}}}}, \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{-\frac{d^2}{d\theta^2}\log L(\theta)\Big|_{\theta=\hat{\theta}}}}\right)$$

연습문제 10번

사망까지의 시간 (time to death) X는 위험률 λ인 지수분포를 따른다.

우중도절단 시간 (right-censoring time) C는 위험률 θ인 지수분포를 따른다. X와 C는 서로 독립이라고 가정하자. 생존데이터는

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

$$\text{만약 } X \leq C \text{ 이면 } \delta = 1$$

$$\text{만약 } X > C \text{ 이면 } \delta = 0$$

다음을 구하시오

(1) $P(\delta = 1)$

(2) T의 분포

(3) δ와 T가 서로 독립임을 보이시오

(4) 랜덤표본 $(T_1, \delta_1), \ldots, (T_n, \delta_n)$에서 λ에 대한 최대우도추정량을 구하시오

(5) 랜덤표본 $(T_1, \delta_1), \ldots, (T_n, \delta_n)$에서 θ에 대한 최대우도추정량을 구하시오