Summary

Lecture 3

LTE : From aggregated data (on intervals $[\,,\,)$ (or $(\,,\,]$))

→ have info on
- # people alive @ beginning of interval $n(x)$
- # dying in interval $d(x)$
- # censored or withdrawn in interval $w(x)$

⇒ Naive estimates ignoring censoring

1) $$\hat{S}(t_j) = 1 - \frac{\#\text{ deaths in first } j \text{ intervals}}{\#\text{ people in study}}$$

= assuming all censored data = survivors!

⇒ overly optimistic estimate

2) $$\hat{S}(t_j) = 1 - \frac{\#\text{ deaths in first } j \text{ intervals}}{\#\text{ people in study} - \#\text{ people censored in first } j \text{ intervals}}$$

= Assumes those censored @ t_j were all censored @ onset of study!

⇒ pessimistic estimate, since for some of those censored we have confirmation of them surviving until $t_{j'}$, $j' < j$ at least

3) Really naive $\Rightarrow$ let's just ignore all censored data

$$\hat{S}(t_j) = 1 - \frac{\#\text{ deaths in } j \text{ first intervals}}{\#\text{ people in study} - \#\text{ censored people in study}}$$

= gross underestimate of survival possible outcome

[More in lab 2]

LTE

Write $P(T \geq 5)$ as a product of prior survivals. If alive @ 5 means you were alive @ 4, meaning <u>did not</u> die in interval $[4, 5)$!

$$\Rightarrow P(T \geq 5) = P(T \geq 5, T \geq 4) = P(T \geq 4)\, P(T \geq 5 \mid T \geq 4)$$

$$= P(T \geq 4)\{1 - P(4 \leq T < 5 \mid T \geq 4)\}$$

$\vdots$

$$= \prod_{t_j \leq 5} \{1 - \underbrace{P(t_{j-1} \leq T < t_j \mid T \geq t_{j-1})}_{m(t_{j-1})}\}$$

$m(t_{j-1})$ = mortality rate @ year 4 closing

An estimate of the mortality rate $\boxed{\hat{m}(t_{j-1}) = \dfrac{d(t_{j-1})}{n(t_{j-1})}}$

1) If we assume censoring in interval $[t_{j-1}, t_j)$ occurs @ end of interval $\hat{m}(t_{j-1}) = \dfrac{d(t_{j-1})}{n(t_{j-1})}$ natural estimate

2) If we assume censoring occurs @ beginning of interval, then $\hat{m}(t_{j-1}) = \dfrac{d(t_{j-1})}{n(t_{j-1}) - w(t_{j-1})}$

3) If we assume censoring can happen anywhere in interval $U[t_{j-1}, t_j)$, then $\hat{m}(t_{j-1}) = \dfrac{d(t_{j-1})}{n(t_{j-1}) - w(t_{j-1})/2}$

natural [Expected censoring time is halfpoint of interval].

Confidence intervals for the LTE

$$\hat{S}(t) = \prod_{x \leq t} \{1 - \hat{m}(x)\}$$

so building block for LTE estimate are components $1 - \hat{m}(x)$.

However, product makes this expression cumbersome to work with $\Rightarrow$ let's focus on $\ln \hat{S}(t)$ instead.

- $\ln \hat{S}(t) = \sum_{x \le t} \ln\{1 - \hat{m}(x)\}$

$$\Rightarrow \mathrm{Var}(\ln \hat{S}(t)) = \mathrm{Var}\left(\sum_{x \le t} \ln\{1-\hat{m}(x)\}\right)$$

$\Downarrow$

What do we need to work out this expression?

→ Well, $V(X+Y) = V(X) + V(Y)$ if $X \perp Y$ (uncorr)

→ $V(f(X)) \approx \left(f'(X)\big|_{\mu}\right)^2 V(X)$ Delta-method.

(Linearization around $\mu = EX$ of function $f(X)$.)

$\Downarrow$

So, we need to a) find out what $V(1-\hat{m}(x))$ is.

b) find out if the components in $\sum_{x \le t}$ are uncorrelated

b) find out what $V(\ln(1-\hat{m}(x)))$ is

Appeal to results from <u>Counting Processes</u>

– Can show that process $\sum_{x \leq t} d(x)$ is made up of uncorrelated increments

$\Rightarrow$ $d(x) \mid$ History up until time point x (right before interval in question)

$\sim \text{Bin}\,(n(x), m(x))$

people at risk @ time x

mortality rate @ time x

$\Rightarrow$ $\hat{m}(x) = \dfrac{d(x)}{n(x)}$ natural estimate in Binomial setting.

Step a)

$$\Rightarrow V(1 - \hat{m}(x)) = V(\hat{m}(x)) = E\Big[\underbrace{V(\hat{m}(x) \mid \mathcal{H}(x))}_{(1)}\Big] + V\Big[\underbrace{E(\hat{m}(x) \mid \mathcal{H}(x))}_{(2)}\Big]$$

[Thm of iterated conditional expectation and variance
$E(X) = E(E(X|Y))$, $V(X) = V(E(X|Y)) + E(V(X|Y))$]

$$= E\Big[\underbrace{V\Big(\frac{d(x)}{n(x)} \,\Big|\, \mathcal{H}(x)\Big)}_{(1)}\Big] + V\Big[\underbrace{E\Big(\frac{d(x)}{n(x)} \,\Big|\, \mathcal{H}(x)\Big)}_{(2)}\Big]$$

$$= E\left(\frac{m(x)(1-m(x))}{n(x)}\right) + V(\underbrace{m(x)}_{\text{true mortalityrate @x}})$$

$$\hat{=} \frac{\hat{m}(x)(1-\hat{m}(x))}{n(x)}$$

Step b) $V(\ln(1-\hat{m}(x)) = ?$

$$\left[f(x) = \ln(1-x) \quad , f'(x) = \frac{-1}{1-x} \right]$$

$$V(\ln(1-\hat{m}(x)) \approx \frac{1}{(1-m(x))^2} V(\hat{m}(x))$$

$$\hat{=} \frac{\hat{m}(x)\cancel{(1-\hat{m}(x))}}{n(x)} \frac{1}{(1-\hat{m}(x))^{\cancel{2}}} = \frac{\hat{m}(x)}{n(x)(1-\hat{m}(x))}$$

$$= \frac{d(x)}{n(x)(n(x)-d(x))}$$

Step c) Need to figure out :) $V\left(\sum_{x \le t} \ln(1-\hat{m}(x))\right)$

$$= \sum_{x \le t} V(\ln(1-\hat{m}(x)))$$

Since working w. building blocks $\hat{m}(x)$, need only figure out if $\mathrm{Cov}(\hat{m}(x), \hat{m}(x+1)) = 0$ or not.

$$\mathrm{Cov}(\hat{m}(x), \hat{m}(x+1)) = \underbrace{E(\hat{m}(x)\hat{m}(x+1))} - E(\hat{m}(x))E(\hat{m}(x+1))$$

$$E(\hat{m}(x)\hat{m}(x+1)) = E\left[E(\hat{m}(x)\hat{m}(x+1) \mid \mathcal{H}(x) = \{n(x), d(x), w(x), n(x-1), d(x-1), w(x-1) \ldots\})\right]$$

$$= E\Bigg[\underbrace{\frac{d(x)}{n(x)}}_{\hat{m}(x)} E(\underbrace{\hat{m}(x+1)}_{\mathrm{Bin}(n(x+1),\, m(x+1))} \mid \mathcal{H}(x))\Bigg] = E\left(\hat{m}(x)\, m(x+1)\right) = m(x+1)\, E(\hat{m}(x))$$

$$= E(\hat{m}(x))\, E(\hat{m}(x+1))$$

$$\mathrm{Cov}(\hat{m}(x), \hat{m}(x+1)) = 0$$

$\Rightarrow$ All taken together

$$V(\ln \hat{S}(t)) = V\Big(\sum_{x \le t} \ln\{1 - \hat{m}(x)\}\Big)$$

$$= \sum_{x \le t} V(\ln\{1 - \hat{m}(x)\})$$

$$= \sum_{x \le t} V(\hat{m}(x)) \frac{1}{(1-\hat{m}(x))^2}$$

$$= \sum_{x \le t} \frac{d(x)}{n(x)(n(x)-d(x))}$$

$\Rightarrow$ Last step $\Rightarrow$ $V(\hat{S}(t))$ $\begin{bmatrix} f(x) = e^x \\ f'(x) = e^x \end{bmatrix}$

$$\cong (\hat{S}(t))^2 V(\ln \hat{S}(t))$$

$$\Rightarrow \boxed{V(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{x \le t} \frac{d(x)}{n(x)(n(x)-d(x))}}$$

(If use other LTE, adjust $n(x)$ to $n(x) - w(x)$ or $n(x) - \frac{w(x)}{2}$ in the above expression).

NOTE :

- Approx of $V(\hat{S})$ via the Delta-method
- Appeal to results from Counting Processes

  → uncorrelated increments

Other interesting estimates often given in Life Tables

Density function : $\hat{f}(\text{mid pt of interval}) = \left[\dfrac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{t_j - t_{j-1}} \right]$

(remember $-\dfrac{dS(t)}{dt} = f(t)$)

: $\hat{h}(\text{mid pt}) = \dfrac{\hat{f}(\text{mid pt})}{\left[\hat{S}(t_j) + \left[\dfrac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{2} \right] \right]}$

(remember - $h(t) = \dfrac{f(t)}{S(t)}$

assumes $S(t)$ piecewise linear).