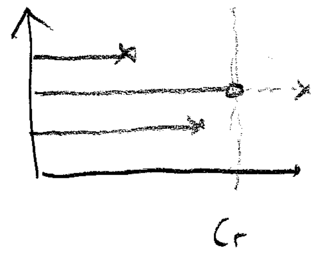
Lecture 2 Life-table estimators, Kaplan-Meier, Nelson-Aalen

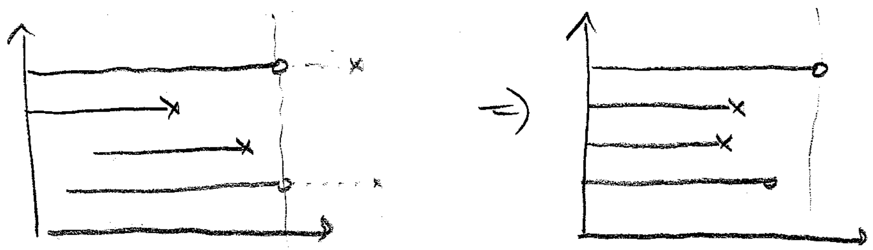
①

Censoring

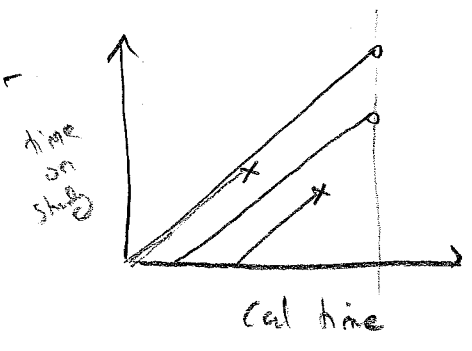
- Type I
  - Same, fixed censoring time for everyone
  - Common for experiments



- Generalized type I
  - delayed entry, recruitment or individual censoring times
  - Calendar time vs. study time



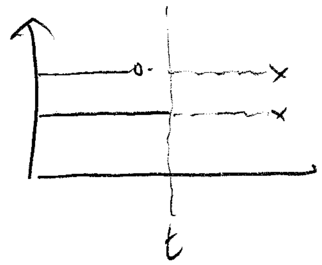
Or



lexis diagram

- Progressive type I
  - several different, but fixed, censoring times (e.g. experiments w. multiple sacrifice time points)

Censoring : assume that survival same @ t for those censored and those not



Key to use censored data to infer something about survival

- **Type II** – study n individuals until rth failure

  – progressive $n \to r_1$

  $n_1 - r_1 \to r_2$

  $n_2 - r_2 \to r_3$

- Left censoring – only know event happened before time C_l

- Interval censoring – only know event happened in interval $[C_R, C_L]$

Truncation

Left truncation – entry in study depends on some event before event of interest

(age @ death among retirement home residents

event before – entering retirement home (live long enough to do this

(if T = time in home ⇒ not truncated

if T = chronological age ⇒ truncation) → conditional probs)

Right truncation — entry depends on event of interest

AIDS study, all have AIDS – time since blood transfusion

Example Myocardial Infarction (entry event) (Right censored data)

Year since entry into study	# alive @ beginning of interval	# dying in interval	# censored or withdrawn
[0,1)	146	27	3
[1,2)	116	18	10
[2,3)	88	21	10
(3,4)	57	9	3
[4,5)	45	1	3
[5,6)	41	2	11
⋮	⋮	⋮	⋮
			$\Sigma = 60$

(Can also define (,] — keep track of which you use)

Want to estimate survival $S(5) = P(T \geq 5)$; 5 year survival

Alt 1 $\hat{S}(5) = 1 - \hat{F}(5) = 1 - \dfrac{76 \text{ deaths in 5 years}}{146 \text{ in study}} = 47.9\%$

Assumes all censored data = survivors, or all censoring happened @ year 5. Overly optimistic!

Alt 2 $\hat{S}(5) = 1 - \hat{F}(5) = 1 - \dfrac{76 \text{ deaths}}{146 - 29 \text{ (withdrawn in 5 years)}} = 35\%$

Assumes all were censored @ onset.

Pessimistic estimate (since these people did survive 1-5 years at least)

Alt 3 Using complete data only $\hat{S}(5) = 1 - \dfrac{76 \text{ deaths}}{146 - 60 \text{ (all censored)}} = 11.6\%$

Pessimistic!

Life-table estimates

Go back to def of S in discrete case.

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

$$= P(T \geq 4) \cdot \{1 - \underbrace{P(4 \leq T < 5 \mid T \geq 4)}_{= \text{mortality rate } m(4) \text{ @ year 4 closing}}\}$$

⋮

$$= \prod_{t_j \leq 5} \{1 - m(t_{j-1})\}$$

	(t_{i-1}, t_i)	$n(x)$	$d(x)$	$w(x)$	$\hat{m}(x) = \frac{d(x)}{n(x)}$ natural estimate	$1 - \hat{m}(x)$	$\hat{S}(t_i) = \prod(1 - m(x))$
$x=0$	$[0, 1)$	146	27	3	0.185	0.815	0.815
$x=1$	$[1, 2)$	116	18	10	0.155	0.845	0.689
$x=2$	$[2, 3)$	88	21	10	0.239	0.761	0.524
	$[3, 4)$	57	9	3	0.158	0.842	0.441
	$[4, 5)$	45	1	3	0.022	0.972	0.432

($n(x) = \#$ at risk)

What did we do here? We assumed that censoring occurred @ end of interval

Also $d(0) \sim Bin(n(0), m(0))$ $\Rightarrow \hat{m}(0) = \frac{27}{146}$

$$d(1) \mid \underbrace{n(0), d(0), m(0)}_{\text{history of data}} \sim Bin(n(1), m(1)) \Rightarrow \hat{m}(1) = \frac{18}{116}$$

⋮

__Alt2__ – assume censoring occurred @ beginning of interval

$\Rightarrow \hat{m}(x) = \dfrac{d(x)}{n(x) - w(x)}$

$\Rightarrow$ in this example $\hat{S}(5) = 0.40$

__Alt3__ – assume censoring occurred anywhere in interval $\Rightarrow$ uniform

– use "effective sample size" $n(x) - \frac{w(x)}{2}$

$\Rightarrow \hat{S}(5) = 0.417$

So; $\hat{S}(5)$ ranges from 40% to 43.2% using life-table estimates

Comp. naive estimates 11.6%, 37.3 – 47.9%

Setting up confidence intervals for $\hat{S}$

- Note, $\hat{S} = \prod (1-\hat{m}(x)) \Rightarrow \ln \hat{S} = \sum \ln\{1-\hat{m}(x)\}$

- So, building block here is $1-\hat{m}(x) \Rightarrow$ what is the variance / co-variance structure of $\hat{m}$?

Remember $d(0) \sim \text{Bin}(n(0), m(0))$

$d(1) \mid n(0), d(0), m(0) \sim \text{Bin}(n(1), m(1))$

$d(2) \mid \overset{\text{not needed}}{\cancel{n(0), d(0), m(0)}},\ n(1), d(1), m(1) \sim \text{Bin}(n(2), m(2))$

$$V(1-\hat{m}(i-1)) = V(\hat{m}(i-1)) = E\left[V(\hat{m}(i-1) \mid H)\right] + V\left[E(\hat{m}(i-1) \mid H)\right]$$

$$= \left[\underset{\text{is Bin}}{\hat{m} = \frac{d}{n}}\right] = \frac{m(i-1)\,(1-m(i-1))}{n(i-1)} + \cancel{V[m(i-1)]}$$

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

- Now, appeal to approx of $V(\ln\{1-m(x)\})$ $\left[\begin{array}{c} V(f(X)) \approx f'(X)^2\, V(X) \\ \text{Delta method} \end{array}\right]$

$$\Rightarrow V\left(\ln\{1-\hat{m}(x)\}\right) = \frac{1}{(1-m(x))^2}\, V(\hat{m}(x))$$

$$= \frac{\hat{m}(i-1)\,\cancel{(1-\hat{m}(i-1))}}{n(i-1)} \cdot \frac{1}{(1-\hat{m}(i-1))^{\cancel{2}}} = \frac{d(i-1)}{n(i-1)\,(n(i-1)-d(i-1))}$$

Now - one can also show that

$$\mathrm{Cov}(\hat{m}(i), \hat{m}(i-1)) = 0$$

$$\Rightarrow V(\ln \hat{S}(5)) = \sum_{i=1}^{5} V(\ln\{1-\hat{m}(i-1)\})$$

$$\Rightarrow V(\hat{S}(5)) = \begin{bmatrix} f(x) = e^x \\ V(f(x)) = (e^x)^2 V(x) \end{bmatrix} = \hat{S}(5)^2 \left[\sum_{i=1}^{5} V(\ln\{1-\hat{m}(i-1)\}) \right]$$

$$= \hat{S}(5)^2 \sum_{i=1}^{5} \frac{d(i-1)}{n(i-1)(n(i-1)-d(i-1))}$$

[Note, if use other LTE, adjust sample size accordingly $n(i) - \frac{w(i)}{2}$ etc]

In our example $\hat{S}(5) = 0.417$, $SE(\hat{S}(5)) = \sqrt{} = 0.0446$

NOTE

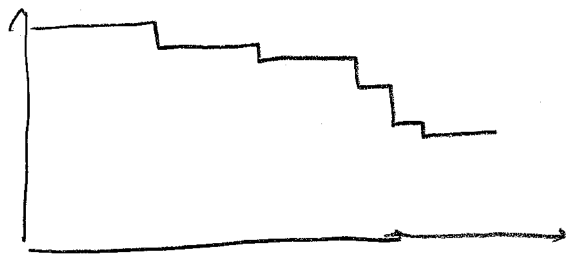
- Approximation of $V(\hat{S})$ via δ-method
- Appeal to results from <u>Counting Processes</u>
  - $\rightarrow$ uncorrelated estimates $\hat{m}(i-1)$ and $\hat{m}(i)$ given current history.

Other info in life tables

- $\hat{f}(\text{midpt of interval}) \hat{=} \left[\frac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{t_j - t_{j-1}} \right]$ (deriv of S)
- $\hat{h}(\text{midpt}) \hat{=} \hat{f}(\text{midpt}) \Big/ \left[\hat{S}(t_j) + \left[\frac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{2} \right] \right]$ (assumes linear segments of S)

Kaplan–Meier

product limit estimator = limit of LTE when interval length so small only at most one distinct event occurs.



Step size $\frac{1}{n(x)}$

$$\hat{S}(t) = \prod \{1 - \hat{m}(x)\}$$ as before

where $\hat{m}(x) = \dfrac{d(x)}{\text{persons @ risk } n(x)}$

$$\boxed{KM(t) = \prod_{x \le t} \left(1 - \frac{d(x)}{n(x)}\right)}$$

Variance of the KM-estimate

Greenwood's formula

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

(see LTE variance estimate & motivation for)