ISyE 6404 – Derivation of K-M Estimator and its Variance. (in Large-Sampl

---

[1] ~~Likelihood of~~ data: d_i: death at $t_{(i)}$

c_i: censored at $(t_{(i)}, t_{(i+1)})$

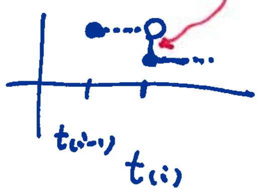
where $t_{(1)} \le t_{(2)} \le \cdots \le t_{(m)}$ are the m cases of death-data.

[2] Likelihood:

i) For death-data (complete-sample):

$$p_{(i)} = Pr(T = t_{(i)}) = S[t_{(i-1)}] - S[t_{(i)}]$$

$$\Rightarrow L = [p_{(i)}]^{d_i}$$



ii) For censored data:

$$\Rightarrow L = Pr(T \ge t_{(i)}]^{c_i} = [S(t_{(i)})]^{c_i}$$

thus,

$$L = \prod_{\substack{i=1 \\ \text{indep.} \\ \text{cases}}}^{m} [S[t_{(i-1)}] - S(t_{(i)})]^{d_i} [S(t_{(i)})]^{c_i}$$

Re-formulation.

3 **Likelihood. (Simplification):**

Let $\pi_i = \dfrac{S(t_{(i)})}{S(t_{(i-1)})}$.

Thus, $S(t_{(i)}) = \pi_1 \cdot \pi_2 \cdots \pi_i$ with $S(t_{(0)}) = 1$, $t_{(0)} = 0$

$$= \frac{\cancel{S(t_{(1)})}}{1} \frac{\cancel{S(t_2)}}{\cancel{S(t_{(1)})}} \cdots \boxed{\frac{S(t_{(i)})}{\cancel{S(t_{(i-1)})}}}$$

$$\Rightarrow \cancel{L = \prod_{i=1}^{n}} \; S[t_{(i-1)}] - S[t_{(i)}]$$

$$= S[t_{(i-1)}]\left[1 - \frac{S[t_{(i)}]}{S[t_{(i-1)}]}\right]$$

$$= (\pi_1 \cdots \pi_{i-1})(1 - \pi_i)$$

$$\Rightarrow \text{Likelihood} \equiv L = \prod_{i=1}^{m} \left[(\pi_1 \pi_2 \cdots \pi_{i-1})(1 - \pi_i)\right]^{d_i} (\pi_1 \pi_2 \cdots \pi_i)^{c_i}$$

$$= \prod_{i=1}^{m} (1 - \pi_i)^{d_i} \boxed{\pi_i^{c_i} \cdot (\pi_1 \pi_2 \cdots \pi_{i-1})^{d_i + c_i}}$$

— Eq (1)

4 Further Reformulation of the Likelihood.

Define. $n_i = \sum_{m \geq j \geq i} (d_j + c_j)$

↑ up to m.

then, P4 shows that

Eq. (1) $\Rightarrow$ $L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \boxed{\pi_i^{n_i - d_i}}$

This is a "binomial likelihood."
with parameters (n_i, π_i)

(~~event~~ even that all π_i are "cond. prob.")

5 Based on the "binomial likelihood",

the MLE for π_i is

$$\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}$$

$\Downarrow$ Thus,

$$\hat{S}_{KM}(t_{(i)}) = \hat{\pi}_1 \hat{\pi}_2 \cdots \hat{\pi}_i = \prod_{i=1}^{m} \left(1 - \frac{d_i}{n_i}\right)$$

Next, 6 Asymptotic Variance ~~properties~~ for the ~~$\hat{\pi}_i$~~ est. $\hat{S}(t_{(i)})$

P4

$$L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{c_i} (\pi_1 \pi_2 \cdots \pi_{i-1})^{d_i+c_i}$$

$$L^* = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{n_i-d_i}$$

$m=1$

$i=1$ $\quad L = (1-\pi_1)^{d_1} \pi_1^{c_1} \cdot 1$

$m=1$

$i=1$ $\quad n_1 = \sum_{j \geq 1}^{m=1} (d_j + c_j)$

$$n_1 = d_1 + c_1$$

$$\Rightarrow n_1 - d_1 = (d_1 + c_1) - d_1 = c_1$$

$$\Rightarrow L^* = (1-\pi_1)^{d_1} \pi_1^{c_1}$$

OK

$m=2$

$i=1$ $\quad L_1 = \left[(1-\pi_1)^{d_1} \pi_1^{c_1}\right] \cancel{(\pi_1)^{d_1+c_1}}$

$i=2$ $\quad L_2 = (1-\pi_2)^{d_2} \pi_2^{c_2} (\pi_1)^{d_2+c_2}$

$L = L_1 \times L_2$

$$L = (1-\pi_1)^{d_1} (1-\pi_2)^{d_2} \pi_1^{c_1+d_2+c_2} \cdot \pi_2^{c_2}$$

$m=2$

$i=1$ $\quad L_1 = (1-\pi_1)^{d_1} \pi_1^{n_1-d_1}$

where

$$n_1 = (d_1+c_1) + (d_2+c_2)$$

$$\Rightarrow n_1 - d_1 = c_1 + d_2 + c_2$$

$i=2$ $\quad L_2 = (1-\pi_2)^{d_2} \pi_2^{n_2-d_2}$

where $n_2 = d_2 + c_2$

$$n_2 - d_2 = c_2$$

$$L_* = L_1 \times L_2$$

$$= (1-\pi_1)^{d_1} (1-\pi_2)^{d_2} \times \pi_1^{c_1+d_2+c_2} \pi_2^{c_2}$$

should be correct for all other cases.

$\boxed{6}$ Next, Asy. variance for $\hat{\pi}_i$.

(Greenwood's procedure).

Note that based on the Binomial (n_i, d_i) distribution.

$$Var(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n_i}$$

we need to use the following Delta-method twice to derive the asy. variance for

$$Var(\hat{S}(t_{(i)}))$$

$\boxed{7}$ Delta-method:

Large-sample

$$Var[f(X)] = (f'(X))^2 \, Var(X)$$

Now $\boxed{1}$ Let us work on $\log \hat{S}(t_{(i)})$ first.

Denoted by, $k_i = \uparrow = \log(\hat{\pi}_1 \cdots \hat{\pi}_i)$

$$= \sum_{j=1}^{i} \log \hat{\pi}_j$$

Since we have $\mathrm{Var}(\pi_i)$

$$\mathrm{Var}\left[\log(\hat{\pi}_i)\right] = \left(\frac{1}{(\hat{\pi}_i)}\right)^2 \mathrm{Var}(\hat{\pi}_i)$$

note

$$[\log(x)]' = \frac{1}{x}$$

$$= \frac{1-\pi_i}{n_i \pi_i}$$

Now $\cancel{\boxed{2}}$

$$\mathrm{Var}\left[\log\left[\hat{S}(t_{(i)})\right]\right]$$

$$= \sum_{j=1}^{i} \frac{1-\pi_j}{n_j \pi_j} = \cancel{\sum_{j=1}^{i} \frac{d_j}{n_j(n_j-d_j)}} \quad - \text{Eq}(3)$$

This is the formula given in the text book.

Eq (10.4) $\hat{\sigma}_{KM}(t_i)$

New $\boxed{2}$

P7

consider $g(x) = e^x$. when $x = \log(\hat{S}(t_{(i)}))$

$g'(x) = e^x$

thus,

$$Var\left[e^{\log \hat{S}(t_{(i)})}\right] = Var(\hat{S}(t_{(i)}))$$

$$= \left[\hat{S}(t_{(i)})\right]^2 \underbrace{Var(\log(\hat{S}(t_{(i)})))}_{\substack{\| \\ \text{Eq}(3)}}$$

$$= \left[\hat{S}(t_{(i)})\right]^2 \sum_{j=1}^{i} \frac{1-\hat{\pi}_j}{n_j \hat{\pi}_j}$$