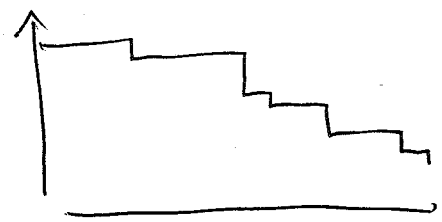
Kaplan-Meier

Product limit estimator = limit of LTE when interval length so small that at most one distinct event occurs (no ties).



Step size $\frac{1}{n(x)}$ ⟹ Note! More in a bit.

$$KM(t) = \hat{S}(t) = \prod_{x \le t} \{1 - \hat{m}(x)\} \quad \text{where } \hat{m}(x) = \frac{d(x)}{n(x)}$$

← persons @ risk @ time pt x.

Definition hazard rate $P(x \le T < x + \Delta x \mid T \ge x) \approx h(x)\Delta x$, $\Delta x \to 0$

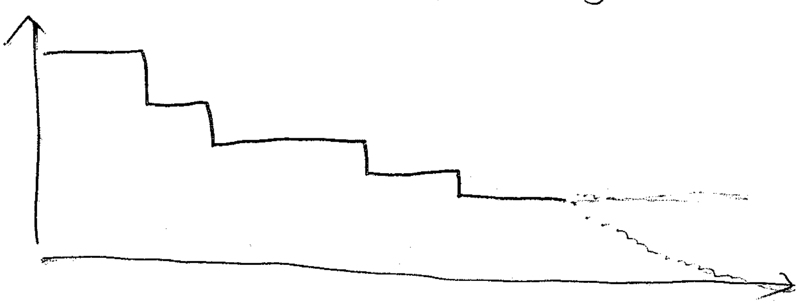
NOTE

- If t_{max} such that $\delta = 1$ (real event time)

  $\Rightarrow KM(t) = 0 \quad t > t_{max}$

  If t_{max} s.t. $\delta = 0$ (censored), then $KM(t)$ not defined beyond t_{max} $\xrightarrow{\text{Alts}}$ extend $\hat{S}(t) = \hat{S}(t_{max})$ (infinite survival for censored last)

  $\rightarrow \hat{S}(t) = \exp\left[\frac{t}{t_{max}} \ln \hat{S}(t_{max})\right]$



KM and redistribution to the right

Remember the LTE; we obtain estimates of $S(t)$ by adjusting / accounting for censoring @ end / beginning or mid-interval.

Another way of looking at this.

$\Rightarrow$ Back to basics; if no censoring

$$\hat{S}(t) = \sum_{i=1}^{n} \frac{1}{n} 1\{t_i > t\}$$

$$= 1 - \sum_{i=1}^{n} \frac{1}{n} 1\{t_i \leq t\}$$

stepsize $\frac{1}{n}$.

but what if we now find out that a t_i is a censored observation?

$\Rightarrow$ adjust denominator to reflect this

This is called the "redistribute to the right" algorithm.

That is, @ t_i: $(T_i, \delta_i = 0)$, take the mass $\frac{1}{n}$ and redistribute equally over all $t_j > t_i$ (remaining time points). From now on, the step size $\frac{1}{n(x)} > \frac{1}{n}$.

If the next t you consider is $\delta = 1$ (real event), step size $\frac{1}{n(t)}$ is used, o/w you again push $\frac{1}{n(t)}$ mass equally to the right...

The "redistribute to the right" = KM.

Since KM(t) is the limit of the LTE, variance formulae look very similar

Greenwood's formula

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

Nelson-Aalen estimator:

If $\frac{d(x)}{n(x)}$ is small $\Rightarrow \exp\left\{-\frac{d(x)}{n(x)}\right\} \approx \left\{1-\frac{d(x)}{n(x)}\right\}$

and so $KM(t) = \hat{S}_{KM}(t) = \prod_{x \le t}\left\{1-\frac{d(x)}{n(x)}\right\} \approx \prod_{x \le t} \exp\left(-\frac{d(x)}{n(x)}\right)$

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

Definition of Basic Concepts

Cumulative Hazard $H(t) = -\ln S(t)$

Alt 1: $\hat{H}_{KM}(t) = -\ln \hat{S}_{KM}(t)$

Alt 2: $\hat{H}_{NA}(t) = \sum_{x \le t} \frac{d(x)}{n(x)}$

$\hat{S}_{NA}(t) = \exp\left(-\hat{H}_{NA}(t)\right)$

Turns out, the second alternative

$$\hat{H}_{NA}(t) = \sum_{x \le t} \frac{d(x)}{n(x)}$$ has better small-sample properties.

So, we have two alternative estimates of survival

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

Variance approx. by Greenwood's formula

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

② Nelson - Aalen

$$\hat{H}_{NA}(t) = \sum_{x \le t} \frac{d(x)}{n(x)} \Rightarrow \hat{S}_{NA}(t) = \exp(-\hat{H}_{NA}(t))$$

To get variance approx for $\hat{S}_{NA}(t)$, we again rely on results from counting processes (uncorrelated increments), and the delta-method.

$$\underset{\text{derivation}}{\overset{\text{after some}}{\Longrightarrow}} \quad \hat{V}(\hat{H}_{NA}(t)) = \sum_{x \le t} \frac{\frac{d(x)}{n(x)}\left(1 - \frac{d(x)}{n(x)}\right)}{n(x) - 1} \qquad \left(\begin{array}{l}\text{unbiased} \\ \text{estimator of} \\ V(\hat{H}_{NA}(t))\end{array}\right) \quad (14)$$

$$\underset{\substack{\text{that } d(x) = 0 \text{ or } 1 \\ \text{only} \\ (\text{no ties})}}{\overset{\text{if } \Delta x \text{ so small}}{=}} \quad \sum_{x \le t} \frac{d(x)}{(n(x))^2}$$

Summary:

(1)
$$\begin{cases} \hat{S}_{KM}(t) = KM(t) = \prod_{x \le t} \left\{1 - \frac{d(x)}{n(x)}\right\} \\ \hat{V}(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t)^2 \sum_{x \le t} \frac{d(x)}{n(x)(n(x) - d(x))} \end{cases}$$

(2)
$$\begin{cases} \hat{H}_{NA}(t) = \sum_{x \le t} \frac{d(x)}{n(x)}, \quad \hat{S}_{NA}(t) = \exp(-\hat{H}_{NA}(t)) \\ \hat{V}(\hat{H}_{NA}(t)) = \sum_{x \le t} \frac{d(x)}{n(x)^2} \\ \longrightarrow \hat{V}(\hat{S}_{NA}(t)) = \hat{S}_{NA}(t)^2 \sum_{x \le t} \frac{d(x)}{n(x)^2} \end{cases}$$

Note, estimates of variance only really OK for large samples. Also, be cautious when $n(x)$ is small (tail of data).

Confidence intervals for $S(t)$

Counting process theory $\Rightarrow$ $KM(t)$ and $\hat{S}_{NA}(t) \xrightarrow{n\to\infty}$ Gaussian Processes

but remember this is a large-sample result.

4 alternative methods for setting up CI for $S(t)$

1. [Most commonly used]

KM + Greenwood

$$\left[\hat{S}_{KM}(t_0) \pm z_{1-\alpha/2}\sqrt{\hat{V}_{Greenwood}(\hat{S}_{KM}(t_0))}\right]$$

2. Use KM to ~~set up~~ estimate $\hat{S}_{KM}(t_0)$ and $\hat{H}_{KM}(t_0) = -\ln \hat{S}(t_0)$,

$$\Rightarrow \hat{V}(\hat{H}_{KM}(t_0)) = \sum_{x \le t_0} \frac{d(x)}{n(x)(n(x)-d(x))}$$

$\Rightarrow$ CI for $S(t_0)$ by exponentiating CI bounds of $\hat{H}_{KM}(t_0)$

$$\exp\left[\hat{H}_{KM}(t_0) \pm z^*_{1-\alpha/2}\sqrt{\hat{V}(\hat{H}_{KM}(t_0))}\right]$$

(3) Use the Nelson-Aalen estimator of $H(t)$, transform to get $\hat{S}_{NA}(t)$ and use the delta-method

$$\left[\hat{S}_{NA}(t_0) \pm z_{1-\alpha/2}\sqrt{\hat{S}_{NA}(t_0)^2 \sum_{x \le t_0} \frac{d(x)}{n(x)^2}}\right]$$

(4) [Usually works well in practise]

Use the Nelson-Aalen estimator of $H(t)$, establish CI for $H(t)$, exponentiate the CI bounds

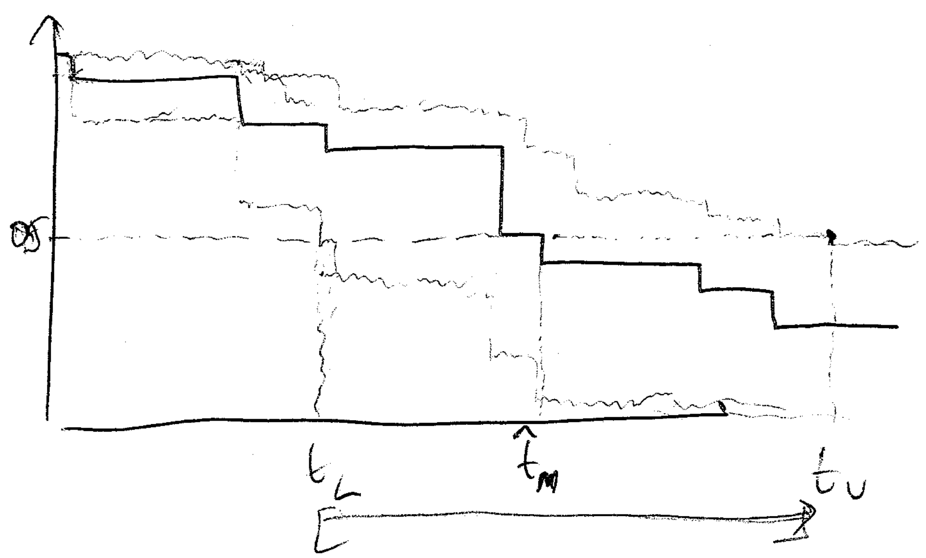
$$\exp\left[\hat{H}_{NA}(t_0) \pm z_{1-\alpha/2}\sqrt{\sum_{x \le t_0} \frac{d(x)}{n(x)^2}}\right]$$

<u>NOTE</u>: These CI are <u>pointwise</u>

→ another lecture (part of) we'll talk about confidence <u>bands</u> (simultaneous CI).

Another quantity of interest – median survival time

$$t_M = \{ t : S(t) = \tfrac{1}{2} \}$$



Find $t_{M,U}$ s.t. $\hat{S}(t_{M,U}) + z_{\alpha/2}\, SE(\hat{S}(t_{M,U})) = \frac{1}{2}$

and $t_{M,L}$ s.t. $\hat{S}(t_{M,L}) - z_{\alpha/2}\, SE(\hat{S}(t_{M,L})) = \frac{1}{2}$

$\Rightarrow$ An approx $1-\alpha$ CI for t_M is thus

$$[\hat{t}_{M,L} \,,\, \hat{t}_{M,U}]$$

(Note; $P(\hat{t}_{M,L} < t_M < \hat{t}_{M,U}) = P(S(\hat{t}_{M,U}) < \frac{1}{2} < S(\hat{t}_{M,L}))$

$= 1 - \left(P\left(S(\hat{t}_{M,U}) > \frac{1}{2}\right) + P\left(S(\hat{t}_{M,L}) < \frac{1}{2}\right)\right)$ (S is a decreasing function and by def $S(t_M) = \frac{1}{2}$)

(Since; $S(\hat{t}_{M,L}) > S(\hat{t}_{M,U})$ for $\hat{t}_L < \hat{t}_U$
and $P(S(\hat{t}_U) > \frac{1}{2}) = P(S(\hat{t}_U) > \frac{1}{2}, S(\hat{t}_L) > \frac{1}{2})$
$P(S(\hat{t}_L) < \frac{1}{2}) = P(S(\hat{t}_L) < \frac{1}{2}, S(\hat{t}_U) < \frac{1}{2})$ $\Rightarrow$

$$\Rightarrow P(S(\hat{t}_U) > \tfrac{1}{2}) + P(S(\hat{t}_L) < \tfrac{1}{2})$$

$$= P(S(\hat{t}_U) > \tfrac{1}{2},\ S(\hat{t}_L) > \tfrac{1}{2}) + P(S(\hat{t}_L) < \tfrac{1}{2},\ S(\hat{t}_U) < \tfrac{1}{2})$$

$$= P(\text{both above}) + P(\text{both below})$$

$$= 1 - P(\text{one above, one below})$$

$$= 1 - P(S(\hat{t}_U) < \tfrac{1}{2} < S(\hat{t}_L))$$

Ok, so $P(\hat{t}_L < t_M < \hat{t}_U) = 1 - \left(P\left(S(\hat{t}_U) > \tfrac{1}{2}\right) + P(S(\hat{t}_L) < \tfrac{1}{2})\right)$

Now, if t_U is the solution to $S(t_U) + z_{\alpha/2} SE(\hat{S}(t_U)) = \frac{1}{2}$

then $\hat{t}_U \approx t_U \Rightarrow P(S(\hat{t}_U) > \tfrac{1}{2}) = P(S(\hat{t}_U) > \hat{S}(\hat{t}_U) + z_{\alpha/2} SE(\hat{S}(\hat{t}_U)))$

$$= P\left(\frac{\hat{S}(\hat{t}_U) - S(\hat{t}_U)}{SE(\hat{S}(\hat{t}_U))} < -z_{\alpha/2}\right)$$

$$= P(Z < -z_{\alpha/2}) \approx \alpha/2 \qquad (\hat{S} \to \text{Gaussian process})$$

and similarly

$$P(S(\hat{t}_L) < \tfrac{1}{2}) \approx \alpha/2$$

$$\Rightarrow P(\hat{t}_L < t_M < \hat{t}_U) = 1 - \alpha$$