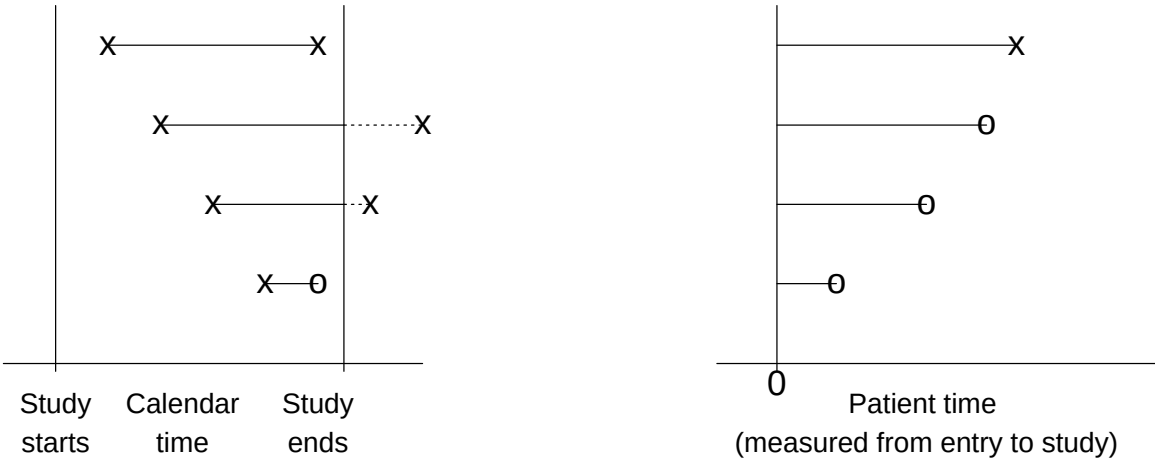
3. Nonparametric Estimation for Censored Survival Data

In biomedical applications, especially in clinical trials, two important issues arise when studying "time-to-event" data (we will assume the event to be "death". But it can be any event of interest):

- Some individuals are still alive at the end of the study or at the time of analysis. So the event of interest, namely death, has not occurred. Therefore we have right censored data.
- Length of follow-up varies due to staggered entry. So we cannot observe the event for those individuals with insufficient follow-up time.

Note: It is important to distinguish calendar time and patient time

Figure 1: *Illustration of censored data*



In addition to censoring because of insufficient follow-up (*i.e.*, end of study censoring due to staggered entry), other reasons for censoring includes

- loss to follow-up: patients stop coming to clinic or move away.
- deaths from other causes: competing risks.

Censoring from these types of causes may be inherently different from censoring due to staggered entry. We will discuss this in more detail later.

Censoring and differential follow-up create certain difficulties in the analysis of such data as is illustrated by the following example taken from a clinical trial of 146 patients treated after they had a myocardial infarction (MI).

The data have been grouped into one year intervals and all time is measured in terms of patient time.

Table 1: *Data from a clinical trial on myocardial infarction (MI)*

Year since entry into study	Number alive and under observation at beginning of interval	Number dying during interval	Number censored or withdrawn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
$[6, 7)$	28	3	5
$[7, 8)$	20	1	8
$[8, 9)$	11	2	1
$[9, 10)$	8	2	6

Question: Estimate the 5 year survival rate, *i.e.*, $S(5) = P[T \geq 5]$.

Two naive and incorrect answers are given by

1. $\widehat{F}(5) = P[T < 5] = \frac{\text{76 deaths in 5 years}}{\text{146 individuals}} = 52.1\%$, $\widehat{S}(5) = 1 - \widehat{F}(5) = 47.9\%$.
2. $\widehat{F}(5) = P[T < 5] = \frac{\text{76 deaths in 5 years}}{\text{146 -29 (withdrawn in 5 years)}} = 65\%$, $\widehat{S}(5) = 1 - \widehat{F}(5) = 35\%$.

Obviously, we can observe the following

1. The first estimate would be correct if all censoring occurred after 5 years. Of cause, this was not the case leading to overly **optimistic** estimate (*i.e.*, overestimates $S(5)$).
2. The second estimate would be correct if all individuals censored in the 5 years were censored immediately upon entering the study. This was not the case either, leading to overly **pessimistic** estimate (*i.e.*, underestimates $S(5)$).

Our clinical colleagues have suggested eliminating all individuals who are censored and use the remaining "complete" data. This would lead to the following estimate

$$\widehat{F}(5) = P[T < 5] = \frac{\text{76 deaths in 5 years}}{\text{146 -60 (censored)}} = 88.4\%, \quad \widehat{S}(5) = 1 - \widehat{F}(5) = 11.6\%.$$

This is even **more pessimistic** than the estimate given by (2).

3.1 Life-table estimate

More appropriate methods use **life-table** or **actuarial** method. The problem with the above two estimates is that they both ignore the fact that each one-year interval experienced censoring (or withdrawing). Obviously we need to take this information into account in order to reduce bias. If we can express $S(5)$ as a function of quantities related to each interval and get a very good estimate for each quantity, then intuitively, we will get a very good estimate of $S(5)$. By

the definition of $S(5)$, we have:

$$\begin{aligned} S(5) &= P[T \geq 5] = P[(T \geq 5) \cap (T \geq 4)] = P[T \geq 4] \cdot P[T \geq 5 | T \geq 4] \\ &= P[T \geq 4] \cdot \{1 - P[4 \leq T < 5 | T \geq 4]\} = P[T \geq 4] \cdot q_5 \\ &= P[T \geq 3] \cdot P[T \geq 4 | T \geq 3] \cdot q_5 = P[T \geq 3] \cdot \{1 - P[3 \leq T < 4 | T \geq 3]\} \cdot q_5 \\ &= P[T \geq 3] \cdot q_4 \cdot q_5 \\ &= \cdots = q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot q_5 \end{aligned}$$

where $q_i = 1 - P[i - 1 \leq T < i | T \geq i - 1], i = 1, 2, ..., 5$. So if we can estimate q_i well, then we will get a very good estimate of $S(5)$. Note that $1 - q_i$ is the mortality rate $m(x)$ at year $x = i - 1$ by our definition.

Table 2: *Life-table estimate of* $S(5)$ *assuming censoring occurred at the end of interval*

duration $[t_{i-1}, t_i)$	$n(x)$	$d(x)$	$w(x)$	$\widehat{m}(x) = \frac{d(x)}{n(x)}$	$1 - \widehat{m}(x)$	$\widehat{S}^R(t_i) = \prod(1 - \widehat{m}(x))$
$[0, 1)$	146	27	3	0.185	0.815	0.815
$[1, 2)$	116	18	10	0.155	0.845	0.689
$[2, 3)$	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

- **Case 1**: Let us first assume that anyone censored in an interval of time is censored at the end of that interval. Then we can estimate each $q_i = 1 - m(i - 1)$ in the following way:

  $$d(0) \sim \text{Bin}(n(0), m(0)) \Longrightarrow \widehat{m}(0) = \frac{d(0)}{n(0)} = \frac{27}{146} = 0.185, \quad \widehat{q}_1 = 1 - \widehat{m}(0) = 0.815$$

  $$d(1)|H \sim \text{Bin}(n(1), m(1)) \Longrightarrow \widehat{m}(1) = \frac{d(1)}{n(1)} = \frac{18}{116} = 0.155, \quad \widehat{q}_2 = 1 - \widehat{m}(1) = 0.845$$

  $\cdots$

  where H means data history (i.e, data before the second interval).

  The life table estimate would be computed as shown in Table 2. So the 5 year survival probability estimate $\widehat{S}^R(5) = 0.432$. (If the assumption that anyone censored in an in-

terval of time is censored at the end of that interval is true, then the estimator $\widehat{S}^R(5)$ is approximately unbiased for $S(5)$.)

Of course, this estimate $\widehat{S}^R(5)$ will have variation since it was calculated from a sample. We need to estimate its variation in order to make inference on $S(5)$ (for example, construct a 95% CI for $S(5)$).

However, $\widehat{S}^R(5)$ is a product of 5 estimates $(\widehat{q}_1, \cdots, \widehat{q}_5)$, whose variance is not easy to find. But we have

$$\log(\widehat{S}^R(5)) = \log(\widehat{q}_1) + \log(\widehat{q}_2) + \log(\widehat{q}_3) + \log(\widehat{q}_4) + \log(\widehat{q}_5).$$

So if we can find out the variance of each $\log(\widehat{q}_i)$, we might be able to find out the variance of $\log(\widehat{S}^R(5))$ and hence the variance of $\widehat{S}^R(5)$.

For this purpose, let us first introduce a very popular method in statistics: **delta method**:

Delta Method:

$$\text{If} \quad \widehat{\theta} \overset{a}{\sim} \mathrm{N}(\theta, \sigma^2)$$

$$\text{then} \quad f(\widehat{\theta}) \overset{a}{\sim} \mathrm{N}(f(\theta), [f'(\theta)]^2\sigma^2)$$

Proof of delta method: If σ^2 is small, $\widehat{\theta}$ will be close to θ with high probability. We hence can expand $f(\widehat{\theta})$ about θ using Taylor expansion:

$$f(\widehat{\theta}) \approx f(\theta) + f'(\theta)(\widehat{\theta} - \theta).$$

We immediately get the (asymptotic) distribution of $f(\widehat{\theta})$ from this expansion.

Returning to our problem. Let $\widehat{\phi}_i = \log(\widehat{q}_i)$. Using the delta method, the variance of $\widehat{\phi}_i$ is approximately equal to

$$\mathrm{var}(\widehat{\phi}_i) = \left(\frac{1}{q_i}\right)^2 \mathrm{var}(\widehat{q}_i).$$

Therefore we need to find out and estimate $\text{var}(\widehat{q}_i)$. Of course, we also need to find out the covariances among $\widehat{\phi}_i$ and $\widehat{\phi}_j$ $(i \neq j)$. For this purpose, we need the following theorem:

Double expectation theorem (Law of iterated conditional expectation and variance): If X and Y are any two random variables (or vectors), then

$$\text{E}(X) = \text{E}[\text{E}(X|Y)]$$

$$\text{Var}(X) = \text{Var}[\text{E}(X|Y)] + \text{E}[\text{Var}(X|Y)]$$

Since $\widehat{q}_i = 1 - \widehat{m}(i-1)$, we have

$$\begin{aligned}
\text{var}(\widehat{q}_i) &= \text{var}(\widehat{m}(i-1)) \\
&= \text{E}[\text{var}(\widehat{m}(i-1)|H)] + \text{var}[\text{E}(\widehat{m}(i-1)|H)] \\
&= \text{E}\left[\frac{m(i-1)[1-m(i-1)]}{n(i-1)}\right] + \text{var}[m(i-1)] \\
&= m(i-1)[1-m(i-1)]\text{E}\left[\frac{1}{n(i-1)}\right],
\end{aligned}$$

which can be estimated by

$$\frac{\widehat{m}(i-1)[1-\widehat{m}(i-1)]}{n(i-1)}.$$

Hence the variance of $\widehat{\phi}_i = \log(\widehat{q}_i)$ can be approximately estimated by

$$\left(\frac{1}{\widehat{q}_i}\right)^2 \frac{\widehat{m}(i-1)[1-\widehat{m}(i-1)]}{n(i-1)} = \frac{\widehat{m}(i-1)}{[1-\widehat{m}(i-1)]n(i-1)} = \frac{d(i-1)}{(n(i-1)-d(i-1))n(i-1)}.$$

Now let us look at the covariances among $\widehat{\phi}_i$ and $\widehat{\phi}_j$ $(i \neq i)$. It is very amazing that they are all approximately equal to zero!

For example, let us consider the covariance between $\widehat{\phi}_1$ and $\widehat{\phi}_2$. Since $\widehat{\phi}_1 = \log(\widehat{q}_1)$ and $\widehat{\phi}_2 = \log(\widehat{q}_2)$, using the same argument for the delta method, we know that we only need to find out the covariance between $\widehat{q}_1$ and $\widehat{q}_2$, or equivalently, the covariance between $\widehat{m}(0)$

and $\widehat{m}(1)$. This can be seen from the following:

$$\begin{aligned}
\mathrm{E}[\widehat{m}(0)\widehat{m}(1)] &= \mathrm{E}[\mathrm{E}[\widehat{m}(0)\widehat{m}(1)|n(0), d(0), w(0)]] \\
&= \mathrm{E}[\widehat{m}(0)\mathrm{E}[\widehat{m}(1)|n(0), d(0), w(0)]] \\
&= \mathrm{E}[\widehat{m}(0)m(1)] \\
&= m(1)\mathrm{E}[\widehat{m}(0)] \\
&= m(1)m(0) = \mathrm{E}[\widehat{m}(0)]\mathrm{E}[\widehat{m}(1)].
\end{aligned}$$

Therefore, the covariance between $\widehat{m}(0)$ and $\widehat{m}(1)$ is zero. Similarly, we can show other covariances are zero. Hence,

$$\mathrm{var}(\log(\widehat{S}^R(5))) = \mathrm{var}(\widehat{\phi}_1) + \mathrm{var}(\widehat{\phi}_2) + \mathrm{var}(\widehat{\phi}_3) + \mathrm{var}(\widehat{\phi}_4) + \mathrm{var}(\widehat{\phi}_5).$$

Let $\widehat{\theta} = \log(\widehat{S}^R(5))$. Then $\widehat{S}^R(5) = e^{\widehat{\theta}}$. So

$$\mathrm{var}(\widehat{S}^R(5)) = (e^{\theta})^2\mathrm{var}(\log(\widehat{S}^R(5))) = (S(5))^2[\mathrm{var}(\widehat{\phi}_1)+\mathrm{var}(\widehat{\phi}_2)+\mathrm{var}(\widehat{\phi}_3)+\mathrm{var}(\widehat{\phi}_4)+\mathrm{var}(\widehat{\phi}_5)],$$

which can be estimated by (Greenwood's formula)

$$\begin{aligned}
\widehat{\mathrm{var}}(\widehat{S}^R(5)) &= (\widehat{S}^R(5))^2 \left[\frac{d(0)}{(n(0)-d(0))n(0)} + \frac{d(1)}{(n(1)-d(1))n(1)} + \frac{d(2)}{(n(2)-d(2))n(2)} \right. \\
&\qquad \left. + \frac{d(3)}{(n(3)-d(3))n(3)} + \frac{d(4)}{(n(4)-d(4))n(4)}\right] \\
&= (\widehat{S}^R(5))^2 \sum_{i=0}^{4} \frac{d(i)}{[n(i)-d(i)]n(i)}. \qquad (1)
\end{aligned}$$

- **Case 2**: Let us assume that anyone censored in an interval of time is censored right at the beginning of that interval. Then the life table estimate would be computed as shown in Table 3. So the 5 year survival probability estimate = 0.400. (In this case, the estimator $\widehat{S}^L(5)$ is approximately unbiased to $S(5)$.)

  The variance estimate of $\widehat{S}^L(5)$ is similar to that of $\widehat{S}^R(5)$ except that we need to change the "sample size" for each mortality estimate to $n - w$ in equation (1).

The two naive estimates range from 35% to 47.9% for the five year survival probability. And the "complete case" (*i.e.*, eliminating anyone censored) estimator gives 11.6%.

Table 3: *Life-table estimate of* $S(5)$ *assuming censoring occurred at the beginning of interval*

duration $[t_{i-1}, t_i)$	$n(x)$	$d(x)$	$w(x)$	$\widehat{m}(x) = \frac{d(x)}{n(x)-w(x)}$	$1-\widehat{m}(x)$	$\widehat{S}^L(t_i) = \prod(1-\widehat{m}(x))$
$[0,1)$	146	27	3	0.189	0.811	0.811
$[1,2)$	116	18	10	0.170	0.830	0.673
$[2,3)$	88	21	10	0.269	0.731	0.492
$[3,4)$	57	9	3	0.167	0.833	0.410
$[4,5)$	45	1	3	0.024	0.976	0.400

The life-table estimate ranged from 40% to 43.2% depending on whether we assume censoring occurred at the left (*i.e.*, beginning) or right (*i.e.*, end) of each interval.

More than likely censoring occurs during the interval. Thus $\widehat{S}^L$ and $\widehat{S}^R$ are not correct. A compromise is to use the following modification:

Table 4: *Life-table estimate of* $S(5)$ *assuming censoring occurred during the interval*

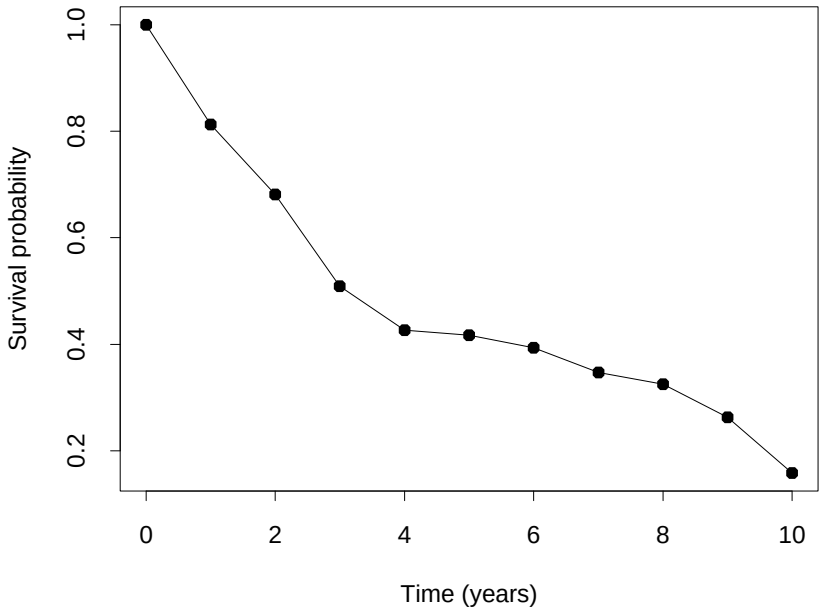
duration $[t_{i-1}, t_i)$	$n(x)$	$d(x)$	$w(x)$	$\widehat{m}(x) = \frac{d(x)}{n(x)-w(x)/2}$	$1-\widehat{m}(x)$	$\widehat{S}^{LT}(t_i) = \prod(1-\widehat{m}(x))$
$[0,1)$	146	27	3	0.187	0.813	0.813
$[1,2)$	116	18	10	0.162	0.838	0.681
$[2,3)$	88	21	10	0.253	0.747	0.509
$[3,4)$	57	9	3	0.162	0.838	0.426
$[4,5)$	45	1	3	0.023	0.977	0.417

That is, when calculating the mortality estimate in each interval, we use $(n(x) - w(x)/2)$ as the "sample size". This number is often referred to as the *effective sample size*.

So the 5 year survival probability estimate $\widehat{S}^{LT}(5) = 0.417$, which is between $\widehat{S}^L = 0.400$ and $\widehat{S}^R = 0.432$.

Figure 2 shows the life-table estimate of the survival probability assuming censoring occurred

Figure 2: *Life-table estimate of the survival probability for MI data*



during interval. Here the estimates were connected using straight lines. No special significance should be given to this. From this figure, the median survival time is estimated to be about 3 years.

The variance estimate of the life-tabble estimate $\widehat{S}^{LT}(5)$ is similar to equation (1) except that the sample size $n(i)$ is changed to $n(i) - w(i)/2$. That is

$$\widehat{\text{var}}(\widehat{S}^{LT}(5)) = (\widehat{S}^{LT}(5))^2 \sum_{i=0}^{4} \frac{d(i)}{[n(i) - w(i)/2 - d(i)][n(i) - w(i)/2]}. \quad (2)$$

Of course, we can also use the above formula to calculate the variance of $\widehat{S}^{LT}(t)$ at other time points. For example:

$$\begin{aligned}\widehat{\text{var}}(\widehat{S}^{LT}(1)) &= (\widehat{S}^{LT}(1))^2 \left\{ \frac{d(0)}{[n(0) - w(0)/2 - d(0)][n(0) - w(0)/2]} \right\} \\ &= 0.813^2 \times \frac{27}{(146 - 3/2 - 27)(146 - 3/2)} = 0.813^2 \times 0.001590223 = 0.001051088.\end{aligned}$$

Therefore $SE(\widehat{S}^{LT}(1)) = \sqrt{0.001051088} = 0.0324$.

The calculation presented in Table 4 can be implemented using `Proc Lifetest` in `SAS`:

```
options ls=72 ps=60;

Data mi;
  input survtime number status;
  cards;
  0 27 1
  0  3 0
  1 18 1
  1 10 0
  2 21 1
  2 10 0
  3  9 1
  3  3 0
  4  1 1
  4  3 0
  5  2 1
  5 11 0
  6  3 1
  6  5 0
  7  1 1
  7  8 0
  8  2 1
  8  1 0
  9  2 1
  9  6 0
;

proc lifetest method=life intervals=(0 to 10 by 1);
  time survtime*status(0);
  freq number;
run;
```

Note that the number of observed events and withdrawals in $[t_{i-1}, t_i)$ were entered after t_{i-1} instead of t_i. Part of the output of the above `SAS` program is

The LIFETEST Procedure

Life Table Survival Estimates

Interval [Lower, Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure
0 1	27	3	144.5	0.1869
1 2	18	10	111.0	0.1622
2 3	21	10	83.0	0.2530
3 4	9	3	55.5	0.1622
4 5	1	3	43.5	0.0230
5 6	2	11	35.5	0.0563
6 7	3	5	25.5	0.1176
7 8	1	8	16.0	0.0625

```
       8        9       2          1        10.5         0.1905
       9       10       2          6         5.0         0.4000
```

```
                  Conditional
                  Probability                                  Survival   Median
     Interval       Standard                                  Standard  Residual
  [Lower,  Upper)     Error     Survival   Failure    Error   Lifetime

       0        1     0.0324     1.0000         0         0    3.1080
       1        2     0.0350     0.8131    0.1869    0.0324    4.4265
       2        3     0.0477     0.6813    0.3187    0.0393    5.2870
       3        4     0.0495     0.5089    0.4911    0.0438         .
       4        5     0.0227     0.4264    0.5736    0.0445         .
       5        6     0.0387     0.4166    0.5834    0.0446         .
       6        7     0.0638     0.3931    0.6069    0.0450         .
       7        8     0.0605     0.3469    0.6531    0.0470         .
       8        9     0.1212     0.3252    0.6748    0.0488         .
       9       10     0.2191     0.2632    0.7368    0.0558         .
```

Here the numbers in the column under `Conditional Probability of Failure` are the estimated mortality $\widehat{m}(x) = d(x)/(n(x) - w(x)/2)$.

The above lifetable estimation can also be implemented using R. Here is the R code:

```
> tis <- 0:10
> ninit <- 146
> nlost <- c(3,10,10,3,3,11,5,8,1,6)
> nevent <- c(27,18,21,9,1,2,3,1,2,2)
> lifetab(tis, ninit, nlost, nevent)
```

The output from the above R function is

```
     nsubs nlost nrisk nevent      surv         pdf     hazard    se.surv
0-1    146     3 144.5     27 1.0000000 0.186851211 0.20610687 0.00000000
1-2    116    10 111.0     18 0.8131488 0.131861966 0.17647059 0.03242642
2-3     88    10  83.0     21 0.6812868 0.172373775 0.28965517 0.03933747
3-4     57     3  55.5      9 0.5089130 0.082526440 0.17647059 0.04382194
4-5     45     3  43.5      1 0.4263866 0.009801991 0.02325581 0.04452036
5-6     41    11  35.5      2 0.4165846 0.023469556 0.05797101 0.04456288
6-7     28     5  25.5      3 0.3931151 0.046248831 0.12500000 0.04503654
7-8     20     8  16.0      1 0.3468662 0.021679139 0.06451613 0.04699173
8-9     11     1  10.5      2 0.3251871 0.061940398 0.21052632 0.04879991
9-10     8     6   5.0      2 0.2632467          NA         NA 0.05579906

          se.pdf  se.hazard
0-1  0.032426423 0.03945410
1-2  0.028930638 0.04143228
2-3  0.033999501 0.06254153
3-4  0.026163333 0.05859410
4-5  0.009742575 0.02325424
```

```
5-6  0.016315545 0.04097447
6-7  0.025635472 0.07202769
7-8  0.021195209 0.06448255
8-9  0.040488466 0.14803755
9-10          NA         NA
```

Note: Here the numbers in the column of `hazard` are the estimated hazard rates at the midpoint of each interval by assuming the true survival function $S(t)$ is a straight line in each interval. You can find an explicit expression for this estimator using the relation

$$\lambda(t) = \frac{f(t)}{S(t)},$$

and the assumption that the true survival function $S(t)$ is a straight line in $[t_{i-1}, t_i)$:

$$S(t) = S(t_{i-1}) + \frac{S(t_i) - S(t_{i-1})}{t_i - t_{i-1}}(t - t_{i-1}), \quad \text{for } t \in [t_{i-1}, t_i).$$

These estimates are very close to the mortality estimates we obtained before (the column under `Conditional Probability of Failure` in the `SAS` output.)

3.2 Kaplan-Meier estimator

The **Kaplan-Meier** or **product limit** estimator is the limit of the life-table estimator when intervals are taken so small that only at most one distinct observation occurs within an interval. Kaplan and Meier demonstrated in a paper in JASA (1958) that this estimator is "nonparametric maximum likelihood estimate".

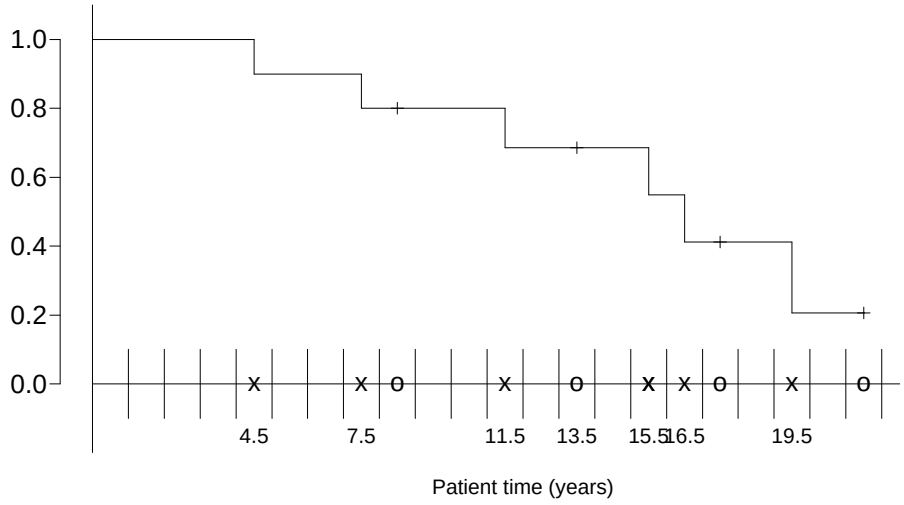
We will illustrate through a simple example shown in Figure 3 how the Kaplan-Meier estimator is constructed. By convention, the Kaplan-Meier estimate is a **right continuous** step function which takes jumps only at the death time. The calculation of the above KM estimate can be implemented using `Proc Lifetest` in `SAS` as follows:

```
Data example;
  input survtime censcode;
  cards;
  4.5 1
  7.5 1
  8.5 0
  11.5 1
```

```
  13.5 0
  15.5 1
  16.5 1
  17.5 0
  19.5 1
  21.5 0
;

Proc lifetest;
  time survtime*censcode(0);
run;
```

Figure 3: *An illustrative example of Kaplan-Meier estimator*



$1-\widehat{m}(x)$:	1	1	1	1	$\frac{9}{10}$	1	1	$\frac{8}{9}$	1	1	1	$\frac{6}{7}$	1	1	1	$\frac{4}{5}$	$\frac{3}{4}$	1	1	$\frac{1}{2}$	1	1
$\widehat{S}(t)$:	1	1	1	1	$\frac{9}{10}$	.	.	$\frac{8}{10}$	.	.	.	$\frac{48}{70}$	.	.	.	$\frac{192}{350}$	$\frac{144}{350}$	.	.	$\frac{144}{700}$	.	.

And part of the output from the above program is

```
                    The LIFETEST Procedure

                Product-Limit Survival Estimates

                                        Survival
                                        Standard      Number      Number
 SURVTIME      Survival      Failure      Error      Failed        Left
```

```
  0.0000       1.0000          0          0       0          10
  4.5000       0.9000     0.1000     0.0949       1           9
  7.5000       0.8000     0.2000     0.1265       2           8
  8.5000*           .          .          .       2           7
 11.5000       0.6857     0.3143     0.1515       3           6
 13.5000*          .          .          .       3           5
 15.5000       0.5486     0.4514     0.1724       4           4
 16.5000       0.4114     0.5886     0.1756       5           3
 17.5000*          .          .          .       5           2
 19.5000       0.2057     0.7943     0.1699       6           1
 21.5000*          .          .          .       6           0
                        * Censored Observation
```

The above Kaplan-Meier estimate can also be obtained using R function `survfit()`. The code is given in the following:

```
> survtime <- c(4.5, 7.5, 8.5, 11.5, 13.5, 15.5, 16.5, 17.5, 19.5, 21.5)
> status <- c(1, 1, 0, 1, 0, 1, 1, 0, 1, 0)
> fit <- survfit(Surv(survtime, status), conf.type=c("plain"))
```

Then we can use R function `summary()` to see the output:

```
> summary(fit)
Call: survfit(formula = Surv(survtime, status), conf.type = c("plain"))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  4.5     10       1    0.900  0.0949       0.7141        1.000
  7.5      9       1    0.800  0.1265       0.5521        1.000
 11.5      7       1    0.686  0.1515       0.3888        0.983
 15.5      5       1    0.549  0.1724       0.2106        0.887
 16.5      4       1    0.411  0.1756       0.0673        0.756
 19.5      2       1    0.206  0.1699       0.0000        0.539
```

Let $d(x)$ denote the number of deaths at time x. Generally $d(x)$ is either zero or one if the death times are continuous, but we allow the possibility of tied survival times in which case $d(x)$ may be greater than one. Let $n(x)$ denote the number of individuals at risk just prior to time x; *i.e.*, number of individuals in the sample who neither died nor were censored prior to time x. Then the Kaplan-Meier estimate for $S(t) = P[T > t]$ can be expressed as

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

Note: The Kaplan-Meier estimates $\widehat{S}_{KM}(t)$ is a right continuous step functions, which jumps only at the observed distinct death times. Suppose that the death times are continuous, i.e no

ties among survival times. Let $T_{(1)} < T_{(2)} < ... < T_{(K)}$ denote the K ordered distinct death times in n samples. Then

$$\widehat{S}_{KM}(t) = \prod_{k=1, T_{(k)} \leq t}^{K} \left(1 - \frac{1}{n(T_{(k)})}\right).$$

Non-informative Censoring

In order that the life-table estimates give unbiased results there is an important assumption that individuals who are censored are at the same risk of subsequent failure as those who are still alive and uncensored. The risk set at any time point (the individuals still alive and uncensored) should be representative of the entire population alive at the same time. If this is the case, the censoring process is called **non-informative**. Statistically, if the censoring process is **independent** of the survival time, then we will automatically have non-informative censoring. Actually, we almost always mean independent censoring by non-informative censoring.

If censoring only occurs because of staggered entry, then the assumption of non-informative censoring seems plausible. However, when censoring results from loss to follow-up or death from a competing risk, then this assumption is more suspect. If at all possible censoring from these later situations should be kept to a minimum.

3.2.1 Variance estimation for the Kaplan-Meier and Nelson-Aalan estimates

The derivation given below is heuristic in nature but will try to capture some of the salient feature of the more rigorous treatments given in the theoretical literature on survival analysis. For this reason, we will use some of the notation that is associated with the "counting process" approach to survival analysis. In fact we have seen it when we discussed the life-table estimator.

It is useful when considering the product limit estimator to partition time into many small intervals, say $(x - \Delta x, x]$, with interval length equal to Δx where Δx is small. Let "x" denote some arbitrary time point on the grid and define the at-risk and counting process as

$$Y(x) = \sum_{i=1}^{n} I(\tilde{T}_i \geq x) \equiv \sum_{i=1}^{n} Y_i(x) \text{ and } N(x) = \sum_{i=1}^{n} \delta_i I(\tilde{T}_i \leq x) \equiv \sum_{i=1}^{n} N_i(x).$$

Here

- $Y(x)$ denotes the number of individuals at risk (*i.e.*, alive and uncensored) just prior to time x.
- $N(x)$ counts the number of deaths that have occurred in the interval $(0, x]$.
- $\Delta N(x)$ denotes the number of observed deaths occurring in $(x - \Delta x, x]$, *i.e.* $\Delta N(x) = N(x) - N(x - \Delta x)$.

Note: Previously, $Y(x)$ is denoted by $n(x)$, $\Delta N(x)$ is denoted by $d(x)$ and $w(x)$ denotes the number of individuals censored in the interval $(x - \Delta x, x]$. In theory when both the death and censoring times are continuous, we should be able to choose Δx small enough so that $\{\Delta N(x) > 0 \text{ and } w(x) > 0\}$ should never occur. In practice, however, data may not be collected in that fashion, in which case, approximations such as those given with life-table estimators may be necessary.

With these definitions, the Kaplan-Meier estimator can be written as

$$\widehat{S}_{KM}(t) = \prod_{\text{all grid points } x \text{ such that } x - \Delta x < t} \left\{1 - \frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\} = \prod_{x \le t} \left\{1 - \frac{dN(x)}{Y(x)}\right\},$$

as $\Delta x \to 0$, where $dN(x) = N(x) - N(x - dx)$ counts the number of deaths occurred at time point x. If the sample size is large and Δx is small, then $\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$ is a small number (*i.e.*, close to zero) and as long as x is not close to the right tail of the survival distribution (where $Y\{(x - \Delta x)^+\}$ may be very small). If this is the case, then

$$\exp\left\{-\frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\} \approx \left\{1 - \frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\}.$$

Here we used the approximation $e^x \approx 1 + x$ when x is close to zero.

Therefore, the Kaplan-Meier estimator can be approximated by

$$\begin{aligned}
\widehat{S}_{KM}(t) &= \prod_{x - \Delta x < t} \left\{1 - \frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\} \approx \prod_{x - \Delta x < t} \exp\left\{-\frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\} \\
&= \exp\left\{-\sum_{x - \Delta x < t} \frac{\Delta N(x)}{Y\{(x - \Delta x)^+\}}\right\} = \exp\left\{-\int_0^t \frac{dN(x)}{Y(x)}\right\}, \text{ as } \Delta x \to 0.
\end{aligned}$$

Here and hereafter, the notations $\prod_{x-\Delta x<t}$ or $\sum_{x-\Delta x<t}$ means that the product or summation is taken over all the grid points x such that $x-\Delta x<t$. Since if Δx is taken to be small enough so that all distinct times (either death times or withdrawal times) are represented at most once in any time interval, then the estimator $\sum_{x-\Delta x<t}\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$ will be uniquely defined and will not be altered by choosing a *finer* partition for the grid of time points. In such a case, the limit of $\sum_{x-\Delta x<t}\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$ is represented as $\int_0^t \frac{dN(x)}{Y(x)}$.

Nelson-Aalen estimator for the cumulative hazard function $\Lambda(t)=\int_0^t \lambda(x)dx$:

$$\widehat{\Lambda}(t)=\int_0^t \frac{dN(x)}{Y(x)}.$$

Recall that $S(t)=\exp\{-\Lambda(t)\}$. Suppose that the death times are continuous and the K ordered distinct failure times in n samples are $T_{(1)}<T_{(2)}<...<T_{(K)}$. Then

$$\widehat{\Lambda}(t)=\sum_{1\le k\le K,\, T_{(k)}\le t}\frac{1}{\sum_{i=1}^n I(\tilde{T}_i\ge T_{(k)})}.$$

Basically, the Nelson-Aalen estimator takes the sum over all the number of deaths divided by the number of individuals at risk at each of those distinct death times up to time t.

Another derivation of Nelson-Aalen estimator:

By the definition of an integral,

$$\Lambda(t)=\int_0^t \lambda(x)dx \approx \sum_{\text{grid points } x \text{ such that } x-\Delta x<t} \lambda\{(x-\Delta x)^+\}\Delta x.$$

In addition,

$$\lambda\{(x-\Delta x)^+\}\Delta x \approx P[x-\Delta x<T\le x|T>x-\Delta x].$$

With independent censoring, it would seem reasonable to estimate $\lambda\{(x-\Delta x)^+\}\Delta x$, *i.e.*, "the conditional probability of dying in $(x-\Delta x,x]$ given being alive at time $x-\Delta$" by $\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$. As Δx goes to zero, we obtain the Nelson-Aalen estimator

$$\widehat{\Lambda}(t)=\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}=\int_0^t\frac{dN(x)}{Y(x)}.$$

We will now show how to estimate the variance of the Nelson-Aalen estimator and then show how this will be used to estimate the variance of the Kaplan-Meier estimator. For a grid point x, let $\mathcal{H}(x)$ denote the history of all deaths and censoring occurring before and at time $x - \Delta x$.

$$\mathcal{H}(x) = \{\Delta N(u), w(u); \text{for all values } u \text{ on our grid of points for } u < x - \Delta x\}.$$

Note the following

- Conditional on $\mathcal{H}(x)$, we would know the value of $Y\{(x-\Delta x)^+\}$ (*i.e.*, the number of individuals at risk just prior to time x, $Y\{(x-\Delta x)^+\} = \sum_{i=1}^n I(\tilde{T}_i > x - \Delta x))$ and that $\Delta N(x)$ would follow a binomial distribution denoted as

  $$\Delta N(x)|\mathcal{H}(x) \sim \text{Bin}(Y\{(x-\Delta x)^+\}, \pi(x)),$$

  where $\pi(x)$ is the conditional probability of an individual dying in $(x - \Delta x, x]$ given that the individual was at risk at time $x - \Delta x$ (*i.e.*, $\pi(x) = P[x - \Delta x < T \leq x | T > x - \Delta x]$). Recall that this probability can be approximated by $\pi(x) \approx \lambda\{(x-\Delta x)^+\}\Delta x$.

- The following are standard results for a binomially distributed random variable.

  (*a*) $\text{E}[\Delta N(x)|\mathcal{H}(x)] = Y\{(x-\Delta x)^+\}\pi(x)$,

  (*b*) $\text{Var}[\Delta N(x)|\mathcal{H}(x)] = Y\{(x-\Delta x)^+\}\pi(x)[1-\pi(x)]$,

  (*c*) $\text{E}\left[\dfrac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\middle|\mathcal{H}(x)\right] = \pi(x)$,

  (*d*) $\text{E}\left\{\left[\dfrac{Y\{(x-\Delta x)^+\}}{Y\{(x-\Delta x)^+\}-1}\right]\left[\dfrac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]\left[\dfrac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]\middle|\mathcal{H}(x)\right\}$
  $= \pi(x)[1-\pi(x)]$.

Consider the Nelson-Aalen estimator $\widehat{\Lambda}(t) = \int_0^t \frac{dN(x)}{Y(x)} = \sum_{x-\Delta x<t} \frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$ (when Δx is small enough). We have

$$\begin{aligned}
\text{E}[\widehat{\Lambda}(t)] &= \text{E}\left[\sum_{x-\Delta x<t} \frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right] = \sum_{x-\Delta x<t} \text{E}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right] \\
&= \sum_{x-\Delta x<t} \text{E}\left[\text{E}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\middle|\mathcal{H}(x)\right]\right] = \sum_{x-\Delta x<t} \pi(x) \\
&\approx \sum_{x-\Delta x<t} \lambda\{(x-\Delta x)^+\}\Delta x \approx \int_0^t \lambda(x)dx = \Lambda(t).
\end{aligned}$$

Hence

- $\mathrm{E}[\widehat{\Lambda}(t)] = \sum_{x-\Delta x<t} \pi(x)$.
- If we take Δx smaller and smaller, then in the limit $\sum_{x-\Delta x<t} \pi(x)$ goes to $\Lambda(t)$. Namely $\widehat{\Lambda}(t)$ is nearly unbiased to $\Lambda(t)$.

How to Estimate the Variance of $\widehat{\Lambda}(t)$

The definition of variance is given by

$$\begin{aligned}
\mathrm{Var}(\widehat{\Lambda}(t)) &= \mathrm{E}[\widehat{\Lambda}(t) - \mathrm{E}(\widehat{\Lambda}(t))]^2 \\
&= \mathrm{E}\left[\sum_{x-\Delta x<t} \frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \sum_{x-\Delta x<t} \pi(x)\right]^2 \\
&= \mathrm{E}\left[\sum_{x-\Delta x<t} \left\{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \pi(x)\right\}\right]^2.
\end{aligned}$$

Note: The square of a sum of terms is equal to the sum of the squares plus the sum of all cross product terms. So the above expectation is equal to

$$\begin{aligned}
&\sum_{x-\Delta x\neq x'-\Delta x<t} \mathrm{E}\left[\left\{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \pi(x)\right\}\left\{\frac{\Delta N(x')}{Y\{(x'-\Delta x)^+\}} - \pi(x')\right\}\right] \\
+ &\sum_{x-\Delta x<t} \mathrm{E}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \pi(x)\right]^2
\end{aligned}$$

We will first demonstrate that the cross product terms have expectation equal to zero. Let us take one such term and let us say, without loss of generality, that $x < x'$.

$$\begin{aligned}
&\mathrm{E}\left[\left\{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \pi(x)\right\}\left\{\frac{\Delta N(x')}{Y\{(x'-\Delta x)^+\}} - \pi(x')\right\}\right] \\
= \ &\mathrm{E}\left(\mathrm{E}\left[\left\{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}} - \pi(x)\right\}\left\{\frac{\Delta N(x')}{Y\{(x'-\Delta x)^+\}} - \pi(x')\right\}|\mathcal{H}(x')\right]\right)
\end{aligned}$$

Note: Conditional on $\mathcal{H}(x')$, $\Delta N(x), Y(x)$ and $\pi(x)$ are constant since $x < x'$. Therefore the above expectation is equal to

$$\mathrm{E}\left(\left\{\frac{\Delta N(x)}{Y\{(x'-\Delta x)^+\}} - \pi(x)\right\}\mathrm{E}\left[\left\{\frac{\Delta N(x')}{Y\{(x'-\Delta x)^+\}} - \pi(x')\right\}\middle|\,\mathcal{H}(x')\right]\right)$$

The inner conditional expectation is zero since

$$\mathrm{E}\left\{\frac{\Delta N(x')}{Y\{(x'-\Delta x)^+\}}\middle|\mathcal{H}(x')\right\}=\pi(x')$$

by $(2.c)$. Therefore we show that

$$\mathrm{E}\left[\left\{\frac{\Delta N(x)}{Y(x)}-\pi(x)\right\}\left\{\frac{\Delta N(x')}{Y(x')}-\pi(x')\right\}\right]=0.$$

Since the cross product terms have expectation equal to zero, this implies that

$$\mathrm{Var}(\widehat{\Lambda}(t))=\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\mathrm{E}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}-\pi(x)\right]^2$$

Using the double expectation again, we get that

$$\begin{aligned}
&\mathrm{E}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}-\pi(x)\right]^2\\
=\;&\mathrm{E}\left(\mathrm{E}\left[\left\{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}-\pi(x)\right\}^2\middle|\mathcal{H}(x)\right]\right)\\
=\;&\mathrm{E}\left(\mathrm{Var}\left[\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\middle|\mathcal{H}(x)\right]\right)\\
=\;&\mathrm{E}\left[\frac{\pi(x)[1-\pi(x)]}{Y\{(x-\Delta x)^+\}}\right].
\end{aligned}$$

Therefore, we have that

$$\mathrm{Var}(\widehat{\Lambda}(t))=\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\mathrm{E}\left[\frac{\pi(x)[1-\pi(x)]}{Y\{(x-\Delta x)^+\}}\right].$$

If we wanted to estimate $\frac{\pi(x)[1-\pi(x)]}{Y\{(x-\Delta x)^+\}}$, then using $(2.d)$ we might think that

$$\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1}$$

may be reasonable. In fact, we would estimate $\mathrm{Var}(\widehat{\Lambda}(t))$ by the following quantity:

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(t))=\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1},$$

In fact, the above variance estimator is unbiased for $\mathrm{Var}(\widehat{\Lambda}(t))$, which can be seen using the following argument:

$$
\begin{aligned}
& \mathrm{E}\left[\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1}\right] \\
= & \lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\mathrm{E}\left[\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1}\right] \\
= & \lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\mathrm{E}\left(\left[\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1}\right]|\mathcal{H}(x)\right) \quad \text{(double expectation again)} \\
= & \lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\mathrm{E}\left[\frac{\pi(x)[1-\pi(x)]}{Y\{(x-\Delta x)^+\}}\right] \quad \text{(by (2.}d\text{))} \\
= & \mathrm{Var}[\widehat{\Lambda}(t)].
\end{aligned}
$$

Note: If the survival data are continuous (*i.e.*, no ties) and Δx is taken small enough, then $\Delta N(x)$ would take on the values 0 or 1 only. In this case

$$\frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1}=\frac{\Delta N(x)}{Y^2\{(x-\Delta x)^+\}},$$

and

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(t))=\lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\frac{\Delta N(x)}{Y^2\{(x-\Delta x)^+\}}=\int_0^t\frac{dN(x)}{Y^2(x)}.$$

Remark:

- We proved that the Nelson-Aalen estimator $\widehat{\Lambda}(t)=\int_0^t\frac{dN(x)}{Y(x)}$ is unbiased for $\Lambda(t)=\int_0^t\lambda(s)ds$. (However, this result is not rigorous. Actually, $E[\widehat{\Lambda}(t)]=E[\int_0^t I(Y(s)>0)\lambda(s)ds]$)

- Since $\widehat{\Lambda}(t)=\sum_{x-\Delta x<t}\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}$ is made up of a sum of random variables that are conditionally uncorrelated, they have a "martingale" structure

  $$\widehat{\Lambda}(t)-\Lambda(t)=\int_0^t\left[\frac{dN(x)}{Y(x)}-d\Lambda(x)\right]=\int_0^t\frac{dN(x)-Y(x)d\Lambda(x)}{Y(x)}=\int_0^t\frac{dM(x)}{Y(x)},$$

  for which there exists a body of theory that enables us to show that

$\widehat{\Lambda}(t)$ is asymptotically normal with mean $\Lambda(t)$ and variance $\mathrm{Var}[\widehat{\Lambda}(t)]$, which can be estimated unbiasedly by

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(t)) = \lim_{\Delta x \to 0} \sum_{x-\Delta x<t} \frac{\frac{\Delta N(x)}{Y\{(x-\Delta x)^+\}}\left[\frac{Y\{(x-\Delta x)^+\}-\Delta N(x)}{Y\{(x-\Delta x)^+\}}\right]}{Y\{(x-\Delta x)^+\}-1};$$

and in the case of no ties, by

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(t)) = \lim_{\Delta x \to 0} \sum_{x-\Delta x<t} \frac{\Delta N(x)}{Y^2\{(x-\Delta x)^+\}} = \int_0^t \frac{dN(x)}{Y^2(x)} = \sum_{1\le k\le K,\, T_{(k)}\le t} \frac{1}{\{\sum_{i=1}^n I(\tilde{T}_i \ge T_{(k)})\}^2}.$$

Let us refer to the estimated standard error of $\widehat{\Lambda}(t)$ by

$$\mathrm{se}[\widehat{\Lambda}(t)] = \sqrt{\widehat{\mathrm{Var}}(\widehat{\Lambda}(t))}.$$

The unbiasedness and asymptotic normality of $\widehat{\Lambda}(t)$ about $\Lambda(t)$ allow us to form confidence intervals for $\Lambda(t)$ (at time t). Specifically, the $(1-\alpha)$th confidence interval for $\Lambda(t)$ is given by

$$\widehat{\Lambda}(t) \pm z_{\alpha/2} * \mathrm{se}(\widehat{\Lambda}(t)),$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of a standard normal distribution. That is, the random interval

$$[\widehat{\Lambda}(t) - z_{\alpha/2} * \mathrm{se}(\widehat{\Lambda}(t)), \widehat{\Lambda}(t) + z_{\alpha/2} * \mathrm{se}(\widehat{\Lambda}(t))]$$

covers the true value $\Lambda(t)$ with probability $1-\alpha$.

This result could also be used to construct confidence intervals for the survival function $S(t)$. This is seen by realizing that

$$S(t) = \mathrm{e}^{-\Lambda(t)},$$

in which case the confidence interval is given by

$$[\mathrm{e}^{-\widehat{\Lambda}(t)-z_{\alpha/2}*\mathrm{se}(-\widehat{\Lambda}(t))}, \mathrm{e}^{-\widehat{\Lambda}(t)+z_{\alpha/2}*\mathrm{se}(\widehat{\Lambda}(t))}],$$

meaning that this random interval will cover the true value $S(t)$ with probability $1-\alpha$.

An example: We will use the hypothetical data shown in Figure 3 to illustrate the calculation of $\widehat{\Lambda}(t)$, $\widehat{\mathrm{Var}}(\widehat{\Lambda}(t))$, and confidence intervals for $\Lambda(t)$ and $S(t)$. For illustration, let us take $t = 17$. Note that there are no ties in this example. So

$$\begin{aligned}
\widehat{\Lambda}(17) &= \widehat{\Lambda}(16.5) = \int_0^t \frac{dN(x)}{Y(x)} = \frac{1}{10} + \frac{1}{9} + \frac{1}{7} + \frac{1}{5} + \frac{1}{4} = 0.804, \\
\widehat{\mathrm{Var}}[\widehat{\Lambda}(17)] &= \widehat{\mathrm{Var}}[\widehat{\Lambda}(16.5)] = \int_0^t \frac{dN(x)}{Y^2(x)} = \frac{1}{10^2} + \frac{1}{9^2} + \frac{1}{7^2} + \frac{1}{5^2} + \frac{1}{4^2} = 0.145, \\
\widehat{\mathrm{se}}[\widehat{\Lambda}(17)] &= \widehat{\mathrm{se}}[\widehat{\Lambda}(16.5)] = \sqrt{0.145} = 0.381.
\end{aligned}$$

So the 95% confidence interval for $\Lambda(t)$ is

$$0.804 \pm 1.96 \cdot 0.381 = [0.0572, 1.551].$$

and the estimate of $S(t)$ based on the Nelson-Aalen estimate is

$$\widehat{S}(t) = \mathrm{e}^{-\widehat{\Lambda}(t)} = \mathrm{e}^{-0.804} = 0.448.$$

The 95% confidence interval for $S(t)$ is

$$[\mathrm{e}^{-1.551}, \mathrm{e}^{-0.0572}] = [0.212, 0.944].$$

Note The above Nelson-Aalen estimate gives $\widehat{S}(17) = 0.448$, which is different from (but close to) the Kaplan-Meier estimate $\widehat{S}_{KM}(17) = 0.411$. It should also be noted that above confidence interval for the survival probability $S(t)$ is not symmetric about the estimator $\widehat{S}(t)$. Another way of getting approximate confidence intervals for $S(t) = \mathrm{e}^{-\Lambda(t)}$ is by using the **delta** method. This method guarantees symmetric confidence intervals.

Hence a $(1-\alpha)$th confidence interval for $f(\theta)$ is given by

$$f(\widehat{\theta}) \pm z_{\alpha/2} |f'(\widehat{\theta})| \widehat{\sigma}.$$

In our case, $\Lambda(t)$ takes on the role of θ, $\widehat{\Lambda}(t)$ takes on the role of $\widehat{\theta}$, $f(\theta) = \mathrm{e}^{-\theta}$ so that $S(t) = f\{\Lambda(t)\}$. Since

$$|f'(\theta) = |-\mathrm{e}^{-\theta}| = \mathrm{e}^{-\theta}, \text{ and } \widehat{S}(t) = \mathrm{e}^{-\widehat{\Lambda}(t)}.$$

Consequently, using the delta method we get

$$\widehat{S}(t) \overset{a}{\sim} \mathrm{N}(S(t), [S(t)]^2 \mathrm{Var}[\widehat{\Lambda}(t)]),$$

and a $(1-\alpha)$th confidence interval for $S(t)$ is given by

$$\widehat{S}(t) \pm z_{\alpha/2}\{\widehat{S}(t) * \mathrm{se}[\widehat{\Lambda}(t)]\}.$$

Now we get that a 95% CI for $S(t)$ (where t=17) using the delta-method approximation based on the Nelson-Aalen estimator

$$\mathrm{e}^{-\widehat{\Lambda}(t)} \pm 1.96 * \mathrm{e}^{-\widehat{\Lambda}(t)} \mathrm{se}[\widehat{\Lambda}(t)] = e^{-0.804} \pm 1.96 * e^{-0.804} * 0.381 = [0.114, 0.784].$$

The estimated $\mathrm{se}[\widehat{S}(t)] = \mathrm{e}^{-\widehat{\Lambda}(t)} \mathrm{se}[\widehat{\Lambda}(t)] = 0.171$.

Remark: Note that $[S(t)]^2 \mathrm{Var}[\widehat{\Lambda}(t)]$ is an estimate of $\mathrm{Var}[\widehat{S}(t)]$, where $\widehat{S}(t) = \exp[-\widehat{\Lambda}(t)]$. Previously, we showed that the Kaplan-Meier estimator

$$\widehat{S}_{KM}(t) = \prod_{x \le t} \left[1 - \frac{dN(x)}{Y(x)}\right]$$

was well approximated by $\widehat{S}(t) = \exp[-\widehat{\Lambda}(t)]$.

Thus a reasonable estimator of $\mathrm{Var}(\widehat{S}_{KM}(t))$ would be to use the estimator of $\mathrm{Var}[\exp(-\widehat{\Lambda}(t))]$, or (by using the delta method)

$$[\widehat{S}_{KM}(t)]^2 \widehat{\mathrm{Var}}[\widehat{\Lambda}(t))] = [\widehat{S}_{KM}(t)]^2 \int_0^t \frac{dN(x)}{Y^2(x)} = [\widehat{S}_{KM}(t)]^2 \lim_{\Delta x \to 0} \sum_{x - \Delta x < t} \frac{\Delta N(x)}{Y^2\{(x-\Delta x)^+\}}.$$

This is very close (asymptotically the same) as the estimator for the variance of the Kaplan-Meier estimator given by Greenwood. Namely

$$\begin{aligned} & \widehat{\mathrm{Var}}\{\widehat{S}_{KM}(t)\} \\ = \; & \{\widehat{S}_{KM}(t)\}^2 \left[\lim_{\Delta x \to 0} \sum_{x - \Delta x < t} \frac{\Delta N(x)}{[Y\{(x-\Delta x)^+\} - w(x)/2][Y\{(x-\Delta x)^+\} - \Delta N(x) - w(x)/2]}\right]. \end{aligned}$$

Note: `SAS` uses the above formula to calculate the estimated variance for the life-table estimate of the survival function, by replacing $\widehat{S}_{KM}(t)$ on both sides by $\widehat{S}^{LT}(t)$.

Note: The summation in the above equation can be viewed as the variance estimate for the cumulative hazard estimator defined by $\widehat{\Lambda}_{KM}(t) = -\log[\widehat{S}_{KM}(t)]$. Namely,

$$\begin{aligned}\mathrm{Var}\{\widehat{\Lambda}_{KM}(t)\} &= \lim_{\Delta x\to 0}\sum_{x-\Delta x<t}\frac{\Delta N(x)}{[Y\{(x-\Delta x)^+\}-w(x)/2][Y\{(x-\Delta x)^+\}-\Delta N(x)-w(x)/2]}\\ &= \int_0^t \frac{dN(x)}{Y(x)\{Y(x)-1\}}\end{aligned}$$

The last equality holds due to the assumption that both the death and censoring times are continuous, *i.e.* no ties among survival times. In such cases, when Δx is small enough, $\Delta N(x) = 0/1$ and when $\Delta N(x) = 1$, $w(x) = 0$.

If we use the Kaplan-Meier estimator, together with Greenwood's formula for estimating the variance, to construct a 95% confidence interval for $S(t)$, we would get

$$\begin{aligned}\widehat{S}_{KM}(t) &= \left[1-\frac{1}{10}\right]\left[1-\frac{1}{9}\right]\left[1-\frac{1}{7}\right]\left[1-\frac{1}{5}\right]\left[1-\frac{1}{4}\right] = 0.411\\ \widehat{\mathrm{Var}}[\widehat{S}_{KM}(t)] &= 0.411^2\left\{\frac{1}{10*9}+\frac{1}{9*8}+\frac{1}{7*6}+\frac{1}{5*4}+\frac{1}{4*3}\right\} = 0.03077\\ \widehat{\mathrm{se}}[\widehat{S}_{KM}(t)] &= \sqrt{0.03077} = 0.175\\ \widehat{\mathrm{Var}}[\widehat{\Lambda}_{KM}(t)] &= \frac{1}{10*9}+\frac{1}{9*8}+\frac{1}{7*6}+\frac{1}{5*4}+\frac{1}{4*3} = 0.182\\ \mathrm{se}[\widehat{\Lambda}_{KM}(t)] &= 0.427.\end{aligned}$$

Thus a 95% confidence interval for $S(t)$ is given by

$$\widehat{S}_{KM}(t) \pm 1.96 * \widehat{\mathrm{se}}[\widehat{S}_{KM}(t)] = 0.411 \pm 1.96 * 0.175 = [0.068, 0.754],$$

which is close to the confidence interval using delta method based on the Nelson-Aalan estimator, considering the sample size is only 10. In fact the estimated standard errors for $\widehat{S}(t)$ and $\widehat{S}_{KM}(t)$ using delta method and Greenwood's formula are 0.171 and 0.175 respectively, which agree with each other very well.

Note: If we want to use *R* function `survfit()` to construct a confidence interval for $S(t)$ with the form $\widehat{S}_{KM}(t) \pm z_{\alpha/2} * \widehat{\mathrm{se}}[\widehat{S}_{KM}(t)]$, we have to specify the argument `conf.type=c("plain")` in `survfit()`. The default constructs the confidence interval for $S(t)$ by exponentiating the

confidence interval for the cumulative hazard using the Kaplan-Meier estimator. For example, a 95% CI for $S(t)$ is $\widehat{S}_{KM}(t) * [e^{-1.96*\text{se}[\widehat{\Lambda}_{KM}(t)]}, e^{1.96*\text{se}[\widehat{\Lambda}_{KM}(t)]}] = 0.411 * [e^{-1.96*0.427}, e^{1.96*0.427}] = [0.178, 0.949]$.

Comparison of confidence intervals for $S(t)$

1. exponentiating the 95% CI for cumulative hazard using Nelson-Aalen estimator: $[0.212, 0.944]$.
2. Delta-method using Nelson-Aalen estimator: $[0.114, 0.784]$.
3. exponentiating the 95% CI for cumulative hazard using Kaplan-Meier estimator: $[0.178, 0.949]$.
4. Kaplan-Meier estimator together with Greenwood's formula for variance: $[0.068, 0.754]$.

These are relatively close and the approximations become better with larger sample sizes.

Of the different methods for constructing confidence intervals, "usually" the most accurate is based on exponentiating the confidence intervals for the cumulative hazard function based on Nelson-Aalen estimator. We don't feel that symmetry is necessarily an important feature that confidence interval need have.

Summary

1. We first estimate $S(t)$ by

$$\widehat{S}_{KM}(t) = \prod_{x \le t} \left(1 - \frac{dN(x)}{Y(x)}\right) = \prod_{\text{all the distinct death times } x \le t} \left(1 - \frac{d(x)}{n(x)}\right),$$

then estimate $\Lambda(t)$ by $\widehat{\Lambda}_{KM}(t) = -\log[\widehat{S}_{KM}(t)]$. Their variance estimates are

$$\begin{aligned} \widehat{\text{Var}}\{\widehat{\Lambda}_{KM}(t)\} &= \int_0^t \frac{dN(x)}{Y(x)\{Y(x)-1\}} \\ \widehat{\text{Var}}\{\widehat{S}_{KM}(t)\} &= \{\widehat{S}_{KM}(t)\}^2 * \widehat{\text{Var}}\{\widehat{\Lambda}_{KM}(t)\}. \end{aligned}$$

The confidence intervals for $S(t)$ can be constructed in two ways:

$$\widehat{S}_{KM}(t) \pm z_{\alpha/2} * \text{se}[\widehat{S}^2_{KM}(t)], \quad \text{or} \quad \text{e}^{-\widehat{\Lambda}_{KM}(t) \pm z_{\alpha/2} * \text{se}[\widehat{\Lambda}_{KM}(t)]} = \widehat{S}_{KM}(t) * \text{e}^{\pm z_{\alpha/2} * \text{se}[\widehat{\Lambda}_{KM}(t)]}$$

2. We first estimate $\Lambda(t)$ by Nelson-Aalen estimator

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN(x)}{Y(x)} = \sum_{\text{all the distinct death times } x \le t} \frac{d(x)}{n(x)},$$

then estimate $S(t)$ by $\widehat{S}(t) = \mathrm{e}^{-\widehat{\Lambda}(t)}$. Their variance estimates are given by

$$\begin{aligned} \widehat{\mathrm{Var}}\{\widehat{\Lambda}(t)\} &= \int_0^t \frac{dN(x)}{Y^2(x)} \\ \widehat{\mathrm{Var}}\{\widehat{S}(t)\} &= \{\widehat{S}(t)\}^2 * \widehat{\mathrm{Var}}\{\widehat{\Lambda}(t)\}. \end{aligned}$$

The confidence intervals for $S(t)$ can also be constructed in two ways:

$$\widehat{S}(t) \pm z_{\alpha/2} * \mathrm{se}[\widehat{S}(t)], \quad \text{or} \quad \mathrm{e}^{-\widehat{\Lambda}(t) \pm z_{\alpha/2} * \mathrm{se}[\widehat{\Lambda}(t)]} = \widehat{S}(t) * \mathrm{e}^{\pm z_{\alpha/2} * \mathrm{se}[\widehat{\Lambda}(t)]}.$$

Estimators of quantiles (such as median, first and third quartiles) of a distribution can be obtained by inverse relationships. This is most easily illustrated through an example.

Suppose we want to estimate the median $S^{-1}(0.5)$ or any other quantile $\varphi = S^{-1}(\theta)$; $0 < \theta < 1$. Then the point estimate of φ is obtained (using the Kaplan-Meier estimator of $S(t)$)

$$\widehat{\varphi} = (\widehat{S}_{KM})^{-1}(\theta), \quad i.e., \quad \widehat{S}_{KM}(\widehat{\varphi}) = \theta.$$

An approximate $(1-\alpha)$th confidence interval for φ if given by $[\widehat{\varphi}_L, \widehat{\varphi}_U]$, where $\widehat{\varphi}_L$ satisfies

$$\widehat{S}_{KM}(\widehat{\varphi}_L) - z_{\alpha/2} * \mathrm{se}[\widehat{S}_{KM}(\widehat{\varphi}_L)] = \theta$$

and $\widehat{\varphi}_U$ satisfies

$$\widehat{S}_{KM}(\widehat{\varphi}_U) + z_{\alpha/2} * \mathrm{se}[\widehat{S}_{KM}(\widehat{\varphi}_U)] = \theta.$$

Proof: We prove this argument for a general estimator $\widehat{S}(t)$. So if we use the Kaplan-Meier estimator, then $\widehat{S}(t)$ is $\widehat{S}_{KM}(t)$. Then

$$\begin{aligned} P[\widehat{\varphi}_L < \varphi < \widehat{\varphi}_U] &= P[S(\widehat{\varphi}_U) < \theta < S(\widehat{\varphi}_L)] \ \text{(note that } S(t) \text{ is decreasing and } S(\varphi) = \theta) \\ &= 1 - (P[S(\widehat{\varphi}_U) > \theta] + P[S(\widehat{\varphi}_L) < \theta]). \end{aligned}$$

Denote φ_U the solution to the equation

$$S(\varphi_U) + z_{\alpha/2} * \text{se}[\widehat{S}(\varphi_U)] = \theta.$$

Then $\widehat{\varphi}_U$ will be close to φ_U. Therefore,

$$\begin{aligned} P[S(\widehat{\varphi}_U) > \theta] &= P[S(\widehat{\varphi}_U) > \widehat{S}(\widehat{\varphi}_U) + z_{\alpha/2} * \text{se}[\widehat{S}(\widehat{\varphi}_U)]] \\ &= P\left[\frac{\widehat{S}(\widehat{\varphi}_U) - S(\widehat{\varphi}_U)}{\text{se}[\widehat{S}(\widehat{\varphi}_U)]} < -z_{\alpha/2}\right] \\ &\approx P\left[\frac{\widehat{S}(\varphi_U) - S(\varphi_U)}{\text{se}[\widehat{S}(\varphi_U)]} < -z_{\alpha/2}\right] \\ &\approx P[Z < -z_{\alpha/2}] \quad (Z \sim \text{N}(0,1)) \\ &= \frac{\alpha}{2}. \end{aligned}$$

Similarly, we can show that

$$P[S(\widehat{\varphi}_L) < \theta] \approx \frac{\alpha}{2}.$$

Therefore,

$$P[\widehat{\varphi}_L < \varphi < \widehat{\varphi}_U] \approx 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1 - \alpha.$$

We illustrate this practice using a simulated data set generated using the following R commands

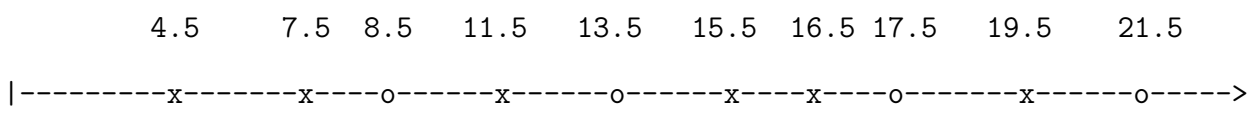
```
> survtime <- rexp(50, 0.2)
> censtime <- rexp(50, 0.1)
> status <- (survtime <= censtime)
> obstime <- survtime*status + censtime*(1-status)
> fit <- survfit(Surv(obstime, status))
> summary(fit)
Call: survfit(formula = Surv(obstime, status))

    time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.0747     50       1    0.980  0.0198       0.9420        1.000
  0.0908     49       1    0.960  0.0277       0.9072        1.000
  0.4332     46       1    0.939  0.0341       0.8747        1.000
  0.4420     45       1    0.918  0.0392       0.8446        0.998
  0.5454     44       1    0.897  0.0435       0.8161        0.987
  0.6126     43       1    0.877  0.0472       0.7887        0.974
  0.7238     42       1    0.856  0.0505       0.7622        0.961
  1.1662     40       1    0.834  0.0536       0.7356        0.946
  1.2901     39       1    0.813  0.0563       0.7097        0.931
```

1.3516	38	1	0.791	0.0588	0.6843	0.915
1.4490	37	1	0.770	0.0609	0.6594	0.899
1.6287	35	1	0.748	0.0630	0.6342	0.882
1.8344	34	1	0.726	0.0649	0.6094	0.865
1.9828	33	1	0.704	0.0666	0.5850	0.847
2.1467	32	1	0.682	0.0680	0.5610	0.829
2.3481	31	1	0.660	0.0693	0.5373	0.811
2.4668	30	1	0.638	0.0704	0.5140	0.792
2.5135	29	1	0.616	0.0713	0.4910	0.773
2.5999	28	1	0.594	0.0721	0.4683	0.754
2.9147	27	1	0.572	0.0727	0.4459	0.734
2.9351	25	1	0.549	0.0733	0.4228	0.713
3.2168	24	1	0.526	0.0737	0.3999	0.693
3.4501	22	1	0.502	0.0742	0.3762	0.671
3.5620	21	1	0.478	0.0744	0.3528	0.649
3.6795	20	1	0.455	0.0744	0.3298	0.627
3.8475	18	1	0.429	0.0744	0.3056	0.603
4.8888	16	1	0.402	0.0745	0.2800	0.578
5.3910	15	1	0.376	0.0742	0.2551	0.553
6.1186	14	1	0.349	0.0736	0.2307	0.527
6.1812	13	1	0.322	0.0726	0.2069	0.501
6.1957	12	1	0.295	0.0714	0.1837	0.474
6.2686	10	1	0.266	0.0701	0.1584	0.445
6.3252	9	1	0.236	0.0682	0.1340	0.416
6.5206	7	1	0.202	0.0663	0.1065	0.385
7.1127	6	1	0.169	0.0632	0.0809	0.352
9.3017	3	1	0.112	0.0623	0.0379	0.333
11.1589	1	1	0.000	NA	NA	NA

3.2.2 Redistributed-to-the-right algorithm and self-consistency estimator (optional)

Redistributed-to-the-right algorithm: (Efron, 1967, pp 831-853; Miller, 1980, pp 52)



Step 1: 1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10

Step 2: 1/10 1/10 0 1/10+1/7(1/10) . . . . . .

1/10+1/7(1/10)

```
Step 3:  1/10   1/10   0     8/70    0  8/70+1/5(8/70)   .  .  .
8/70+1/5(8/70)
```

.

.

Self-consistency estimator: (Miller, 1980, pp 54-57)

For $t < \tilde{T}_{(n)}$, where $\tilde{T}_{(n)}$ is the largest observed event time (either a death or censoring), the Kaplan-Meier estimate $\widehat{S}_{KM}(t)$ is the unique estimator satisfying the following self-consistency equation:

$$\widehat{S}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} I(\tilde{T}_i > t) + \sum_{\tilde{T}_i \le t}(1-\delta_i)\frac{\widehat{S}(t)}{\widehat{S}(\tilde{T}_i)}\right] \tag{3}$$