5.2 Estimation of survival function:

1. Parametric approach

Suppose $t_1, t_2, \cdots, t_n$ are failure times corresponding to censor indicators $w_1, w_2, \cdots, w_n$ ($w_i = 1$, death; $w_i = 0$, censoring). Then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} [f(t_i)]^{w_i} [S(t_i)]^{1-w_i} = \prod_{i=1}^{n} \left[\frac{f(t_i)}{S(t_i)}\right]^{w_i} S(t_i) = \prod_{i=1}^{n} [\lambda(t_i)]^{w_i} S(t_i),$$

where

$$w_i = 1 \Longrightarrow [f(t_i)]^{w_i} [S(t_i)]^{1-w_i} = f(t_i)$$

and

$$w_i = 0 \Longrightarrow [f(t_i)]^{w_i} [S(t_i)]^{1-w_i} = S(t_i) = P(T \geq t_i).$$

where $\lambda(t), S(t)$ depends on some parameter θ. Then, the parameter estimate $\hat{\theta}$ can be obtained by solving

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

and $\hat{\lambda}(t)$ and $\hat{S}(t)$ can be obtained by evaluating θ at $\hat{\theta}$.

Example:

Let T have exponential density. Then, $f(t) = \lambda \cdot exp(-\lambda t), S(t) = exp(-\lambda t)$, and $\lambda(t) = \lambda$. Then,

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{w_i} \cdot exp(-\lambda t_i)$$

and further

$$l(\lambda) = log[L(\lambda)] = \sum_{i=1}^{n} [w_i \cdot log(\lambda) - \lambda t_i] = \left(\sum_{i=1}^{n} w_i\right) \cdot log(\lambda) - \lambda \left(\sum_{i=1}^{n} t_i\right).$$

Thus,

$$\frac{dl(\lambda)}{d\lambda} = \frac{\sum_{i=1}^{n} w_i}{\lambda} - \sum_{i=1}^{n} t_i = 0 \Longrightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} t_i}.$$

$$\hat{S}(t) = exp\left(-\hat{\lambda} t\right).$$

For example, in the motivating example,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} t_i} = \frac{1+1+1+1}{6+7+9.5+7+10+7+6+11} = \frac{4}{63.5}.$$

Then,

$$\hat{S}(t) = exp\left(-\frac{4t}{63.5}\right).$$

Note:

Intuitively,

$$E(T) = \frac{1}{\lambda},$$

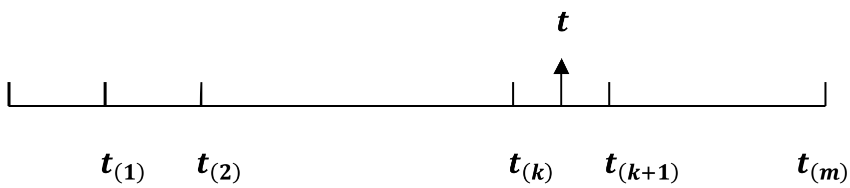
i.e., the estimate for mean survival time is

$$\frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} w_i}.$$

2. Nonparametric approach

Let $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ be death times. The number of individuals who alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted n_j, for $j = 1, \cdots, m$, and d_j will denote the number who die at this time. Thus, we have the following table:

$t_{(1)}$	$t_{(2)}$	$\cdots$	$t_{(m)}$
n_1	n_2	$\cdots$	n_m
d_1	d_2	$\cdots$	d_m



Then, for $t_{(k)} \le t < t_{(k+1)}$,

$$\begin{aligned}
\hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \\
&= \left(\frac{n_1 - d_1}{n_1}\right)\left(\frac{n_2 - d_2}{n_2}\right)\cdots\left(\frac{n_k - d_k}{n_k}\right) \\
&= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)\cdots\left(1 - \frac{d_k}{n_k}\right) \\
&= [1 - \hat{\lambda}(t_{(1)})][1 - \hat{\lambda}(t_{(2)})]\cdots[1 - \hat{\lambda}(t_{(k)})]
\end{aligned}$$

and $\hat{S}(t)$ is referred to as Kaplan-Meier estimate.

Note:

Intuitively, if T is a discrete random variable taking values $t_{(1)} < t_{(2)} < \cdots$ with associated probability function

$$P(T = t_{(j)}), j = 1, 2, \cdots,$$

then

$$\lambda(t_{(j)}) = P(T = t_{(j)} | T \geq t_{(j)}) = \frac{f(t_{(j)})}{S(t_{(j)})}.$$

Therefore, for $t_{(k)} < t < t_{(k+1)}$,

$$\begin{aligned}
S(t) = P(T \geq t) &= P(T \geq t_{(k+1)}) \\
&= \frac{P(T \geq t_{(2)})}{P(T \geq t_{(1)})} \cdot \frac{P(T \geq t_{(3)})}{P(T \geq t_{(2)})} \cdot \cdots \cdot \frac{P(T \geq t_{(k+1)})}{P(T \geq t_{(k)})} \\
&= \left[\frac{P(T \geq t_{(1)}) - P(T = t_{(1)})}{P(T \geq t_{(1)})}\right] \cdots \left[\frac{P(T \geq t_{(k)}) - P(T = t_{(k)})}{P(T \geq t_{(k)})}\right] \\
&= \left[1 - \frac{P(T = t_{(1)})}{P(T \geq t_{(1)})}\right]\left[1 - \frac{P(T = t_{(2)})}{P(T \geq t_{(2)})}\right] \cdots \left[1 - \frac{P(T = t_{(k)})}{P(T \geq t_{(k)})}\right] \\
&= [1 - \lambda(t_{(1)})][1 - \lambda(t_{(2)})] \cdots [1 - \lambda(t_{(k)})]
\end{aligned}$$

since $P(T \geq t_{(1)}) = 1$.

Example (continue):

In the motivating example, we have

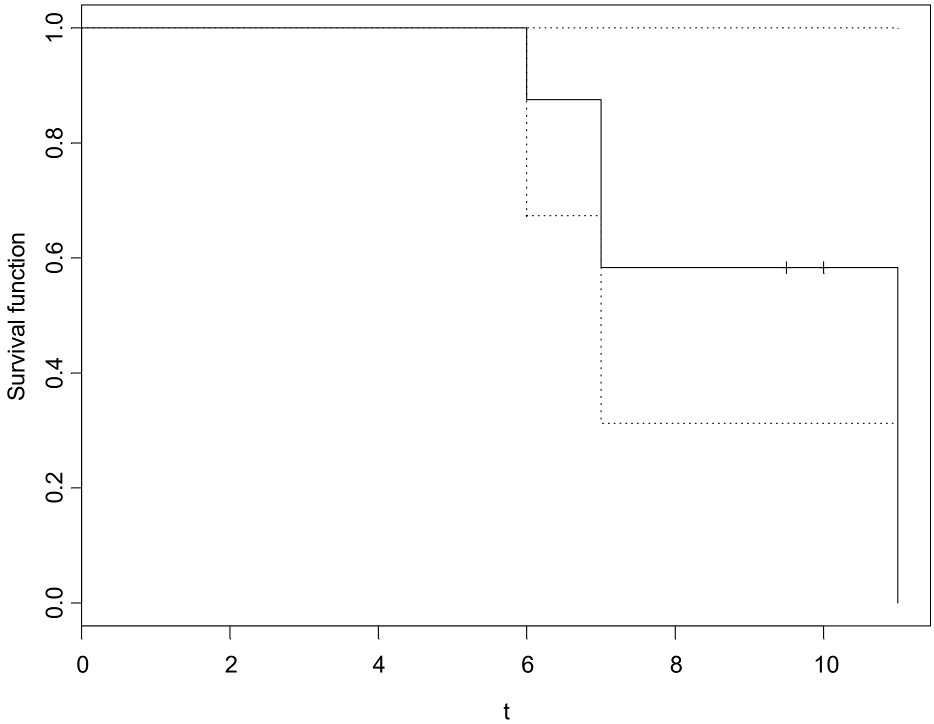
$t_{(j)}$	6	7	11
n_j	8	6	1
d_j	1	2	1

Thus,

$$\hat{S}(t)$$

$$= \begin{cases} 1, 0 \leq t < 6 \\ \dfrac{n_1 - d_1}{n_1} = \dfrac{8 - 1}{8} = 0.875, 6 \leq t < 7 \\ \left(\dfrac{n_1 - d_1}{n_1}\right)\left(\dfrac{n_2 - d_2}{n_2}\right) = \dfrac{(8 - 1)}{8} \cdot \dfrac{(6 - 2)}{6} = \dfrac{7}{12}, 7 \leq t < 11 \\ \left(\dfrac{n_1 - d_1}{n_1}\right)\left(\dfrac{n_2 - d_2}{n_2}\right)\left(\dfrac{n_3 - d_3}{n_3}\right) = \dfrac{(8 - 1)}{8} \cdot \dfrac{(6 - 2)}{6} \cdot \dfrac{(1 - 1)}{1} = 0, t \geq 11 \end{cases}$$

The plot of the survival function is


Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
t