Notes for Biostatistics 140.641.01 'Survival Analysis'

First Term, 2022

Chapter 3. One Sample Estimation

Treating the target population as a group, one sample estimation considers problems such as how to estimate survival function or hazard function or pdf for survival time, T, of subjects in the group. In this chapter we address the question of how to estimate the survival function based on either complete survival data or possibly censored survival data.

Consider estimation of the survival function

$$\begin{aligned} S(t) &= \mathrm{P}(T \geq t) \\ &= \text{Population fraction surviving beyond } t \end{aligned}$$

3.1 Nonparametric Estimation from Complete Survival Data

The set of the complete data $t_1, t_2, \ldots, t_n$ reflects the structure of population survival times.

Thus, we estimate $S(t)$ by the sample fraction surviving beyond t:

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

$\hat{S}(t)$ is also called the **empirical survival distribution**.

Next question: How to derive confidence interval for $S(t)$?

Note that the empirical survival distribution $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq t)$ is a 'sample mean estimator'. Define the binary variable $B_i(t) = I(T_i \geq t)$. Note that $B_i(t)$ is a Bernoulli variable with mean $S(t)$ and variance $S(t)(1 - S(t))$. By the central limit theorem, **when n is large**,

$$\sqrt{n}\left(\hat{S}(t) - S(t)\right) \xrightarrow{\text{d}} \text{Normal}\left(0, \text{var}[B_i(t)]\right)$$

Since $\text{var}[B_i(t)] = S(t)(1 - S(t))$, we derive

$$\sqrt{n}\left(\hat{S}(t) - S(t)\right) \xrightarrow{\text{d}} \text{Normal}\left(0, S(t)(1 - S(t))\right) \ ,$$

or equivalently, when n is large,

$$\hat{S}(t) \overset{\text{approx}}{\sim} \text{Normal}\left(S(t), \frac{S(t)(1 - S(t))}{n}\right) \ .$$

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

$$\left(\hat{S}(t) - 1.96\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}},\ \hat{S}(t) + 1.96\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}} \right).$$

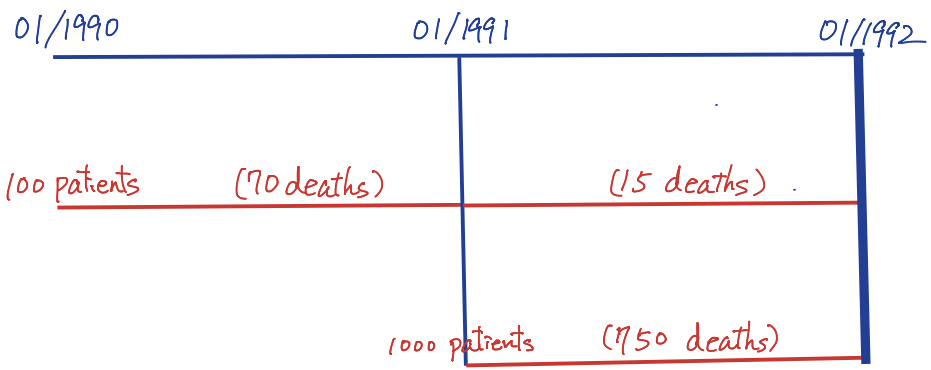
Remarks

- If n is small $(n < 20)$, it is more appropriate to find confidence intervals using the binomial distribution tables.
- The normal approximation works better when $S(t)$ is not close to 0 or 1). When $S(t)$ is close to 0 or 1, the Poisson approximation technique is better.

3.2 Nonparametric Estimation from Censored Survival Times

Without parametric assumption on the distribution of T, how do we estimate the survival function $S(t)$? First consider a simple example.

Example. A prospective study recruited 100 patients in January 1990 and recruited 1000 patients in January 1991. The study ended in January 1992. Survival time $T =$ time from treatment/enrollment to death. The time unit for the survival time is discrete (by years), $T = 1, 2, 3, \ldots$. Suppose 70 patients died in year 1 and 15 patients died in year 2 from the first cohort (recruited in 1990), and 750 patients died in year 1 from the second cohort (recruited in 1991).



It is important that the two cohorts were sampled from the same target population, since this would implicitly imply independent censoring (why?).

How to estimate 3-year survival rate $S(3) = P(T \geq 3 \text{ years})$?

Approach 1. Reduced sample estimate
Only use information from individuals who had been followed for at least two years. That is, use only group 1 data to derive

$$\hat{S}(3) = \frac{100 - 70 - 15}{100} = \frac{15}{100} = 0.15$$

This estimate is statistically appropriate but inefficient. It is appropriate in the sense that $\hat{S}(3)$ is very close to $S(3)$ when n_1 is large. It is inefficient because only part of the data is used. Here

$$\hat{\text{var}}(\hat{S}(3)) = \frac{\hat{S}(3)(1 - \hat{S}(3))}{100} = 0.001275.$$

Approach 2. (Statistically inappropriate approaches)

- Assume 250 individuals from group 2 died in year 2,

$$\hat{S}(3) = \frac{15}{1100} = 0.014$$

- Assume 250 individuals from group 2 remained alive in year 2

$$\hat{S}(3) = \frac{15 + 250}{1100} = 0.241$$

- Exclude 250 patients from the analyzed data (Watch out! A common mistake!)

$$\hat{S}(3) = \frac{15}{1100 - 250} = 0.018.$$

Approach 3. (A simple case of the Kaplan-Meier estimate). Decompose the survival function into conditional probabilities.

$$
\begin{aligned}
S(3) = \mathrm{P}(T \geq 3) &= \frac{Pr(T \geq 2)}{Pr(T \geq 1)} \cdot \frac{Pr(T \geq 3)}{Pr(T \geq 2)} \\
&= Pr(T \geq 2 | T \geq 1) \cdot Pr(T \geq 3 | T \geq 2) \\
&= \left\{ 1 - \frac{Pr(T = 1)}{Pr(T \geq 1)} \right\} \cdot \left\{ 1 - \frac{Pr(T = 2)}{Pr(T \geq 2)} \right\} \\
\frac{Pr(T = 1)}{Pr(T \geq 1)} &= \lambda(1) = \frac{70 + 750}{1100} \\
\frac{Pr(T = 2)}{Pr(T \geq 2)} &= \lambda(2) = \frac{15}{30}
\end{aligned}
$$

Thus

$$\hat{S}(3) = \frac{280}{1100} \cdot \frac{15}{30} = 0.127$$

This estimator is more efficient than the reduced sample estimator.

This estimator is a simple example of the Kaplan-Meier estimator.

Kaplan-Meier Estimator

Now consider the Kaplan-Meier estimator in its general form. The Kaplan-Meier estimator (1958, *JASA*) is a nonparametric estimator for the survival function S. Consider the random censoring mechanism. **Assume T_i and C_i are independent (independent censoring).** That is, assume that T_i is independent of C_i for each i. The data are

$$(y_1, \delta_1),\ (y_2, \delta_2),\ \ldots,\ (y_n, \delta_n).$$

Let $y_{(1)} < y_{(2)} < \ldots < y_{(k)},\ k \leq n$, be the distinct, uncensored and ordered survival times.

Example. Data: $3, 2^+, 0, 1, 5^+, 3, 5$

$$(y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}) = (0, 1, 3, 5).$$

Suppose $y_{(i-1)} < t \leq y_{(i)}$. A principle of nonparametric estimation of S is to assign positive probability to and only to uncensored survival times. Therefore, we try to estimate

$$S(t) \approx \frac{Pr(T \geq y_{(2)})}{Pr(T \geq y_{(1)})} \cdot \frac{Pr(T \geq y_{(3)})}{Pr(T \geq y_{(2)})} \cdots \frac{Pr(T \geq y_{(i)})}{Pr(T \geq y_{(i-1)})}.$$

How to estimate $S(t)$? Define

$R(t) = \{k : \ y_k \geq t, k = 1, 2, ..., n\}$ - the so-called 'risk set' at t

$d_{(j)} =$ no. of failures at $y_{(j)}$

$N_{(j)} =$ no. of individuals in $R(y_{(j)})$;

Example Using the previous example $3 \ 2^+ \ 0 \ 1 \ 5^+ \ 3 \ 5$

$$N_{(1)} = 7, \ N_{(2)} = 6, \ N_{(3)} = 4, \ N_{(4)} = 2$$

$$d_{(1)} = 1, \ d_{(2)} = 1, \ d_{(3)} = 2, \ d_{(4)} = 1.$$

Now estimate $\frac{Pr(T=y_{(j)})}{Pr(T\geq y_{(j)})}$ by $\frac{d_{(j)}}{N_{(j)}}$, $j = 1, 2, \ldots, i-1$. The Kaplan-Meier estimate can be constructed as

$$\hat{S}(t) = \left(1 - \frac{d_{(1)}}{N_{(1)}}\right)\left(1 - \frac{d_{(2)}}{N_{(2)}}\right)\ldots\left(1 - \frac{d_{(i-1)}}{N_{(i-1)}}\right)$$

$$= \boxed{\prod_{y_{(j)} < t}\left(1 - \frac{d_{(j)}}{N_{(j)}}\right)}$$

Question: Why estimate $\frac{Pr(T=y_{(j)})}{Pr(T\geq y_{(j)})}$ by $\frac{d_{(j)}}{N_{(j)}}$?

An important property: Under the independent censoring assumption, the probability for uncensored Y occurring at y conditioning on $Y \geq y$ is the same as the probability of T occurring at y conditioning on $T \geq y$ (i.e., the hazard probability at y):

$$\boldsymbol{P(y \leq Y < y + dy, \Delta = 1 \mid Y \geq y) = \lambda(y)dy}$$

$$\begin{aligned}
&P(y \leq Y < y + dy, \Delta = 1 \mid Y \geq y) \\
&= P(T < C, y \leq T < y + dy \mid T \geq y, C \geq y) \\
&\approx \frac{S_C(y^+) \cdot f_T(y)dy}{S_T(y)S_C(y)} = \frac{f_T(y)dy}{S_T(y)} = \lambda(y)dy \qquad \text{(assuming } C \text{ is continuous)}
\end{aligned}$$

Thus, we can use $\frac{\text{no. of uncensored failures at } y}{\text{no. of individuals in risk set } R(y)}$ to estimate $\lambda(y)dy$.

Remark: This is the base for the use of risk sets in many nonparametric and semiparametric models when analyzing survival data.

Example $3,\ 2^+,\ 0, 1,\ 5^+,\ 3,\ 5$

uncensored times	0	1	3	5
$d_{(i)}$	1	1	2	1
$N_{(i)}$	7	6	4	2

$$\hat{S}(0) = 1$$

$$\hat{S}(1) = \left(1 - \frac{1}{7}\right) = \frac{6}{7} = 0.86$$

$$\hat{S}(3) = \left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right) = \frac{5}{7} = 0.71$$

$$\hat{S}(5) = \left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{2}{4}\right) = \frac{5}{14} = 0.36$$

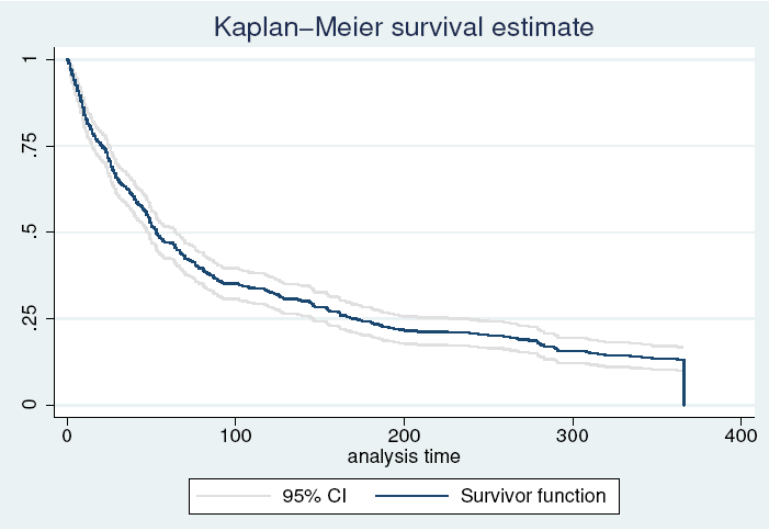
$$\hat{S}(5^+) = \hat{P}(T > 5) = \left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{2}{4}\right)\left(1 - \frac{1}{2}\right) = \frac{5}{28} = 0.18$$

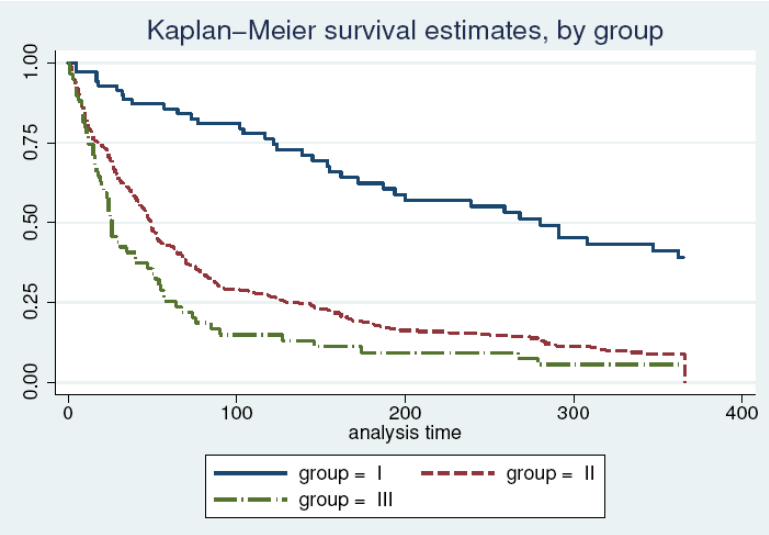
Remark:

1. If all of the largest observed times are uncensored, the Kaplan-Meier estimate will eventually reach the value 0.

2. If one or more than one of the largest observed times are censored, the Kaplan-Meier estimate will not reach 0 and is undetermined for $t >$ largest uncensored y_i.

Typical Kaplan-Meier curves





- Features
  - Always starts at S(0)=1
  - Monotonic decreasing (non-increasing)
  - Step functions
  - May not go down to zero all the way when time progresses
  - Shows time-varying profile of absolute risk

3.3 Properties of the Kaplan-Meier Estimator

Greenwood's formula.

The next question is how to identify the variance of the Kaplan-Meier estimator. For continuous survival data, rigorous discussion is usually provided in advanced survival-analysis courses.

Property. When n is large

$$\hat{S}(t) \overset{\text{approx}}{\sim} \text{Normal}(S(t), \sigma(t)^2)$$

where $\sigma(t)^2$ can be estimated by the Greenwood's variance estimate:

$$\boxed{\hat{\sigma}(t)^2 = [\hat{S}(t)]^2 \sum_{y_{(j)}<t} \frac{d_{(j)}}{N_{(j)}(N_{(j)}-d_{(j)})}}$$

Greenwood's formula

It is important to find an approach to derive an estimator for the large sample variance of the Kaplan-Meier estimator. The discussion here is a sketch for grouped/discrete survival data with finitely many values.

First, group the data using the uncensored times $y_{(1)} < y_{(2)} < \ldots < y_{(k)}$. For each risk set $R_{(j)} = \{y_i \ : \ y_i \geq y_{(j)}\}$, counting the number of failures is a binomial experiment. Thus $d_{(j)} \sim$ Binomial $(N_{(j)}, \lambda_{(j)})$, where $\lambda_{(j)}$ is the hazard at $y_{(j)}$.

Let $q_{(j)} = 1 - \lambda_{(j)}$. For $y_{(i-1)} < t \leq y_{(i)}$,

$$\begin{aligned} \text{var}(\log\ \hat{S}(t)) &= \text{var}(\log\{\hat{q}_{(1)}\hat{q}_{(2)}, \ldots, \hat{q}_{(i-1)}\}) \\ &= \text{var}(\log \hat{q}_{(1)} + \log \hat{q}_{(2)} + \ldots + \log \hat{q}_{(i-1)}) \\ &= \sum_{j=1}^{i} \text{var}(\log \hat{q}_{(j)}) \end{aligned}$$

The variances are additive because the risk sets at $y_{(1)}$, $y_{(2)}$, ..., $y_{(k)}$ are nested ($R_{(1)} \supset R_{(2)} \supset \ldots$). Thus, by statistical theory (see last page of Appendix), we can treat $\log \hat{q}_{(1)}, \log \hat{q}_{(2)} \ldots$ as uncorrelated terms.

Use the Delta method, for a transformation ϕ of an estimate $\hat{\theta}$, we have

$$\text{var}(\phi(\hat{\theta})) \approx [\phi'(\theta)]^2 \text{var}(\hat{\theta}).$$

Thus

$$\begin{aligned}
\text{var}(\log \hat{q}_{(j)}) &\approx \left[\frac{1}{q_{(j)}}\right]^2 \text{var}(\hat{q}_{(j)}) = \frac{1}{q_{(j)}^2} \cdot \frac{\lambda_{(j)} q_{(j)}}{N_{(j)}} = \frac{\lambda_{(j)}}{q_{(j)} N_{(j)}}, \\
\text{var}(\log \hat{S}(t)) &= \sum_{j=1}^{i} \text{var}(\log \hat{q}_{(j)}) \approx \sum_{y_{(j)} < t} \left(\frac{\lambda_{(j)}}{q_{(j)} N_{(j)}} \right)
\end{aligned}$$

Use the Delta method again,

$$\begin{aligned}
\sigma(t)^2 = \text{var}(\hat{S}(t)) &= \text{var}(\underset{\phi}{\exp}(\underset{\hat{\theta}}{\log \hat{S}(t)})) \\
&\approx [S(t)]^2 \cdot \text{var}(\log \hat{S}(t))
\end{aligned}$$

Plug in $\hat{\lambda}_{(j)} = d_{(j)}/N_{(j)}$ and $\hat{q}_{(j)} = \frac{N_{(j)}-d_{(j)}}{N_{(j)}}$. The Greenwood's formula, for estimating the variance of the Kaplan-Meier estimate, is

$$\hat{\text{var}}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{y_{(j)}<t} \frac{d_{(j)}}{N_{(j)}(N_{(j)}-d_{(j)})}$$

Remark: The asymptotic normality property and the Greenwood's formula hold for both discrete and continuous survival data, but more advanced techniques are required for the proofs. A more formal approach which allows for theoretical developments of continuous survival data is through a stochastic representation of $S(t) = e^{-\int_0^t \lambda(v)dv}$ using empirical processes or Martingale theory.

Example. (Lee, p29) Forty-two patients with acute leukemia were randomized into a treatment group and a placebo group to assess the treatment effect to maintain remission. T : remission time.

- 6-MP (6-mercaptopurine) group, $n_1 = 21$

  Ordered observations:

  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+, 17^+,$

  $19^+, 20^+, 25^+, 32^+, 32^+, 34^+, 35^+$ (months)

- Placebo group, $n_2 = 21$

  Ordered observations:

  $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15,$

  $17, 22, 23$ (months)

The empirical survival function from the placebo group is

$$\begin{aligned}
\hat{S}(0) &= 1 \\
\hat{S}(1) &= \frac{21}{21} = 1 \\
\hat{S}(2) &= \frac{19}{21} = 0.90 \\
\hat{S}(3) &= \frac{17}{21} = 0.81 \\
\hat{S}(4) &= \frac{16}{21} = 0.76 \\
\hat{S}(5) &= \frac{14}{21} = 0.67 \\
&\vdots \\
\hat{\text{var}}(\hat{S}(5)) &= \frac{(0.67)(0.33)}{21} = 0.01053 \\
S.E. = \hat{SD}(\hat{S}(5)) &= \sqrt{\frac{(0.67)(0.33)}{21}} = 0.103
\end{aligned}$$

A 95% confidence interval for $S(5)$ is

$$(0.67 - 1.96 \times 0.103, \quad 0.67 + 1.96 \times 0.103) = (0.47,\ 0.87).$$

Warning: The sample size $n_2 = 21$ may not be large enough for the normal approximation!

For the 6MP group, use the K-M estimate to derive

$$\begin{aligned}
\hat{S}(6) &= 1 \\
\hat{S}(7) &= \left(1 - \frac{3}{21}\right) = 0.86 \\
\hat{S}(10) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right) = 0.81 \\
\hat{S}(13) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right)\left(1 - \frac{1}{15}\right) = 0.753
\end{aligned}$$

.....

Apply the Greenwood's formula to get

$$\begin{aligned}
\widehat{\text{var}}(\hat{S}(13)) &= (0.753)^2 \left(\frac{3}{21 \times 18} + \frac{1}{17 \times 16} + \frac{1}{15 \times 14}\right) \\
&= 0.0093
\end{aligned}$$

A 95% confidence interval for $S(13)$ is

$$(0.753 - 1.96\sqrt{0.0093}\ ,\ 0.753 + 1.96\sqrt{0.0093}) = (0.564\ ,\ 0.942)$$

What about $\hat{S}(11)$ and $\hat{\text{var}}(\hat{S}(11))$?
— Same as ($\hat{S}(13)$ and $\hat{\text{var}}(\hat{S}(13))$.

Remark 1. The K-M estimate is a nonparametric method which can be applied to either discrete or continuous data.

Remark 2. The accuracy of the K-M estimate and Greenwood's formula relies on large sample size of uncensored data. Make sure that you have at least, say, 20 or 30 uncensored survival times in your data set before using the methods.

Remark 3. Greenwood's formula is more appropriate when $0 << S(t) << 1$. Using Greenwood's formula, the confidence interval limits could be above 1 or below 0. In these cases, we usually replace these limit points by 1 or 0. For example, a 95% confidence interval could be $(0.845, 1.130)$, we will use $(0.845, 1)$ instead.

Nonparametric MLE.

Kaplan and Meier (1958, JASA) showed that the K-M estimate is the unique nonparametric mle from the likelihood function

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right],$$

where the likelihood maximization is subject to the class of probability distributions which assign probability to, and only to uncensored survival times. See more details in the Appendix.

Self-consistency of the KM estimator

KM estimator can be derived as a special case of the EM Algorithm (Dempster, Rubin and Laird, 1977 JRSS-B)

Maximization-Step: With complete survival data, the NPMLE of $S(t)$ is $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq t)$.

Expectation-Step: With survival data with censoring $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$, the expected values of complete-data-MLE is

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} I(T_i \geq t) \mid (Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[I(T_i \geq t) \mid Y_i, \Delta_i]$$

Note that

$$\begin{aligned}
\mathrm{E}[I(T_i \geq t) \mid Y_i, \Delta_i = 1] &= \mathrm{E}[I(T_i \geq t) \mid T_i = Y_i, \Delta_i = 1] = I(Y_i \geq t) \\
\mathrm{E}[I(T_i \geq t) \mid Y_i, \Delta_i = 0] &= \mathrm{E}[I(T_i \geq t) \mid T_i \geq C_i = Y_i, \Delta_i = 0] \\
&= I(Y_i \geq t) + I(Y_i < t)\frac{S(t)}{S(Y_i)}
\end{aligned}$$

The MLE based on incomplete data likelihood satisfies

$$\begin{aligned} S(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[I(T_i \geq t) \mid Y_i, \Delta_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{ \Delta_i \cdot I(Y_i \geq t) + (1-\Delta_i) \cdot [I(Y_i \geq t) + I(Y_i < t)\frac{S(t)}{S(Y_i)}] \right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{ I(Y_i \geq t) + (1-\Delta_i) I(Y_i < t)\frac{S(t)}{S(Y_i)} \right\} \end{aligned}$$

Efron (1967) showed that the MLE coinsides with the Kaplan-Meier estimator. This is called the self-consistency property of the Kaplan-Meier estimator. To calculate the solution using the EM-algorithm:

1. Use an initial estimate $S^{(0)}(t)$.
2. Calculate the updated estimate $S^{(k+1)}(t)$ from the previous estimate $S^{(k)}(t)$:
$S^{(k+1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \left\{ I(Y_i \geq t) + (1-\Delta_i) I(Y_i < t)\frac{S^{(k)}(t)}{S^{(k)}(Y_i)} \right\}$.
3. Iterate until convergence.

Remark: The self-consistency algorithm can be extended to derive NPMLE for other type of incomplete data (such as interval-censored survival data).

Dependent censoring

- The Kaplan-Meier estimator is a nonparametric estimator whose validity does not rely on parameteric distributional assumption. However, its validity relies on the independent censoring assumption that T_i is independent of C_i. This assumption could be violated when, for example, the censoring mechanism is correlated with an individual's heath and the health condition is correlated with the length of survival time.

- Usually the dependent censoring problem could be much relaxed in a regression model, since in a regression model setting the required independent censoring becomes: T_i and C_i are independent conditionally on covariates X_i.

- In many applications, the problem of dependent censoring remains a challenge because marginal models or marginal data analysis are desired for clinical trials or exploratory data analysis.

*Appendix

Kaplan-Meier estimator as the NPMLE

To see the Kaplan-Meier estimator is the unique mle of the likelihood function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} &\propto \prod_{i=1}^{n}\left[f(y_i)^{\delta_i}S(y_i)^{1-\delta_i}\right] = \prod_{i=1}^{n}\left\{\frac{f(y_i)}{S(y_i)}\right\}^{\delta_i}\{S(y_i)\} \\ &= \left\{\prod_{(i)}\lambda_{(i)}^{d_{(i)}}\right\}\left\{\prod_{i=1}^{n}\prod_{y_{(j)}<y_i}(1-\lambda_{(j)})\right\} = \prod_{(i)}\lambda_{(i)}^{d_{(i)}}(1-\lambda_{(i)})^{N_{(i)}-d_{(i)}}\end{aligned}$$

Thus, by the invariance property, the unique mle of $\lambda_{(i)}$ is $d_{(i)}/N_{(i)}$ and the Kaplan-Meier estimate

$$\hat{S}(t) = \left(1-\frac{d_{(1)}}{N_{(1)}}\right)\left(1-\frac{d_{(2)}}{N_{(2)}}\right)\cdots\left(1-\frac{d_{(i-1)}}{N_{(i-1)}}\right)$$

is the unique nonparametric mle of S.

Kaplan-Meier Estimator for grouped survival data

For survival data grouped into time intervals $1, 2, \ldots, K$ (such as year $1, 2, \ldots, K$), we treat censored survival times during the jth time interval at risk for half of the count, and take a modified approach: Define $N^*_{(j)} = N_{(j)} - \frac{1}{2}w_{(j)}$, where $w_{(j)}$ is the number of censored times during the jth time interval.

Now estimate $\frac{Pr(T \geq y_{(j+1)})}{Pr(T \geq y_{(j)})}$ by $\frac{N^*_{(j)} - d_{(j)}}{N^*_{(j)}} = 1 - \frac{d_{(j)}}{N^*_{(j)}}$, $j = 1, 2, \ldots, i-1$. The Kaplan-Meier estimate is

$$\hat{S}(t) = \left(1 - \frac{d_{(1)}}{N^*_{(1)}}\right)\left(1 - \frac{d_{(2)}}{N^*_{(2)}}\right)\ldots\left(1 - \frac{d_{(i-1)}}{N^*_{(i-1)}}\right)$$

$$= \boxed{\prod_{y_{(j)} < t}\left(1 - \frac{d_{(j)}}{N^*_{(j)}}\right)}$$

Zero correlation between $\log \hat{q}_{(j)}$ and $\log \hat{q}_{(k)}$

Define $V_{(j)} = \log \hat{q}_{(j)} - \mathrm{E}[\log \hat{q}_{(j)}]$. Let $\mathcal{H}_{(j)}$ represent the data history (including all the censoring and failure times and the corresponding covariates, and censoring indicator information) up to $y^{-}_{(j)}$, $j = 1, ..., k$. Then, for $j < j'$,

$$\begin{aligned}
\mathrm{Cov}[\log \hat{q}_{(j)}, \log \hat{q}_{(j')}] &= \mathrm{E}[V_{(j)} \times V_{(j')}] \\
&= \mathrm{E}[\mathrm{E}[V_{(j)} \times V_{(j')} \mid \mathcal{H}_{(j')}] \,] \\
&= \mathrm{E}[V_{(j)} \times \mathrm{E}[V_{(j')} \mid \mathcal{H}_{(j')}] \,] \\
&= \mathrm{E}[V_{(j)} \times \mathrm{E}[V_{(j')} \mid R_{(j')}] \,] \\
&= \mathrm{E}[V_{(j)} \times 0 \,] \\
&= 0
\end{aligned}$$

Thus, $\log \hat{q}_{(j)}$ and $\log \hat{q}_{(j')}$ are uncorrelated if $j \neq j'$.