BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Nonparametric Estimator of Distribution Function

1. Observed survival data: (X_i, δ_i), $i = 1, \cdots, n$. We can sort the data based on X_i, $i = 1, \cdots, n$. Let $t_1, \cdots, t_K$ denote all distinctive uncensored event times. Let d_k denote the number of subjects having event time t_k. Let n_k denote the number of subjects having event time equal to or greater than t_k.
2. If all $\delta = 1$, then a nonparametric estimation of distribution is

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}}$$

When n is large, $\hat{F}$ tends to the true distribution. When censoring is present, consistentency does not hold.

Estimate the conditional probability

Time	d_k	n_k	$\hat{p}$	$\mathrm{Var}(\hat{p})$
1	3	48	$3/48$	$3 \times 45/48^3$
4	2	44	$2/44$	$2 \times 42/44^3$
5	4	42	$4/42$	$4 \times 38/42^3$
6	2	38	$2/38$	$2 \times 36/38^3$
8	1	35	$1/35$	$1 \times 34/35^3$

The estimated quantity is

$$P(T < t + \Delta | T \geq t) \approx h(t)\Delta.$$

Estimate the conditional probability

1. The survival function

$$
\begin{aligned}
S(t) &= P(T > t) \\
&= P(T \geq t_1 | T \geq t_0 = 0) \times P(T \geq t_2 | T \geq t_1) \\
&\quad \times \cdots \times P(T \geq t | T \geq t_k) \\
&= \{1 - P(T < t_1 | T \geq t_0)\} \times \{1 - P(T < t_2 | T \geq t_1) \\
&\quad \times \cdots \times \{1 - P(T < t | T \geq t_k)\}.
\end{aligned}
$$

Kaplan-Meier Estimator of Survival Function

1. Kaplan-Meier (also called product-limit) estimator of the survival function

$$\hat{S}(t) = \prod_{\{k | t_k \leq t\}} \left(1 - \frac{d_k}{n_k}\right).$$

   This estimator is consistent.

2. Variance estimate of $\hat{S}(t)$

$$\hat{V}\left\{\hat{S}(t)\right\} = \hat{S}^2(t) \sum_{\{k | t_k \leq t\}} \frac{d_k}{n_k(n_k - d_k)}.$$

   This is called Greenwood variance formula.

Example: Kaplan-Meier Estimator for Myeloma Data

Time	d_k	n_k	KME	$\sqrt{V\{log(KME)\}}$	$\sqrt{V(KME)}$
1	3	48	0.9375	0.0373	0.0349
4	2	44	0.8949	0.0497	0.0445
5	4	42	0.8097	0.0706	0.0571
6	2	38	0.7670	0.0802	0.0616
8	1	35	0.7451	0.0853	0.0636
10	4	34	0.6575	0.1058	0.0696
12	1	28	0.6340	0.1119	0.0710
13	1	26	0.6096	0.1186	0.0723
14	1	25	0.5852	0.1254	0.0734
15	1	24	0.5608	0.1324	0.0743
16	2	22	0.5098	0.1486	0.0758
17	1	20	0.4844	0.1572	0.0762

Example: Kaplan-Meier Estimator for Myeloma Data(Continuing)

Time	d_k	n_k	KME	$\sqrt{V\{log(KME)\}}$	$\sqrt{V(KME)}$
18	2	19	0.4334	0.1758	0.0762
23	1	15	0.4045	0.1889	0.0764
24	1	14	0.3756	0.2029	0.0762
36	1	13	0.3467	0.2181	0.0756
40	2	12	0.2889	0.2535	0.0732
50	1	9	0.2568	0.2795	0.0718
51	1	8	0.2247	0.3098	0.0696
65	1	5	0.1798	0.3821	0.0687
66	1	4	0.1348	0.4789	0.0646
88	1	2	0.0674	0.8540	0.0576
91	1	1	0	Inf	NaN

Confidence intervals for Kaplan-Meier Curve

1. Confidence interval for $\hat{S}(t)$ can be obtained directly by the approximation

$$\hat{S}(t) - S(t) \sim N\left(0, V\{\hat{S}(t)\}\right).$$

2. Confidence interval for $\hat{S}(t)$ can be obtained indirectly by the approximation

$$\log \hat{S}(t) - \log S(t) \sim N\left(0, V\{\log \hat{S}(t)\}\right).$$

Confidence intervals for Kaplan-Meier Curve (Continuing)

1. The variance for $\log\{-\log \hat{S}(t)\}$,

$$V\left[\log\{-\log \hat{S}(t)\}\right] = V\left\{\log \hat{S}(t)\right\} \Big/ \left[\log \hat{S}(t)\right]^2 .$$

2. Confidence interval for $\hat{S}(t)$ can also be obtained indirectly by the approximation

$$\log[-\log \hat{S}(t)] - \log[-\log S(t)] \sim N\left(0, V\{\log[-\log \hat{S}(t)]\}\right) .$$

Inference on the Survival Function

1. 95% confidence interval for $S(t)$

$$\hat{S}(t) \pm 1.96\sqrt{\hat{V}\left\{\hat{S}(t)\right\}}.$$

2. 95% confidence interval for $S(t)$ based on inversion of $\log S(t)$

$$\hat{S}(t)\exp\left[\pm 1.96\sqrt{\hat{V}\left\{\log \hat{S}(t)\right\}}\right].$$

3. 95% confidence interval for $S(t)$ based on inversion of $\log\{-\log S(t)\}$

$$\left\{\hat{S}(t)\right\}^{\exp\left(\mp 1.96\sqrt{\hat{V}\left[\log\{-\log \hat{S}(t)\}\right]}\right)}.$$

Comparison of the Confidence Intervals

1. Which one is better? The one with a better normal approximation.
2. The first can have bounds below 0 or above 1. The second can have upper bound above 1 but the low bound is always greater than 0. The third have bounds always in $[0, 1]$.

95% confidence intervals for Kaplan-Meier curve

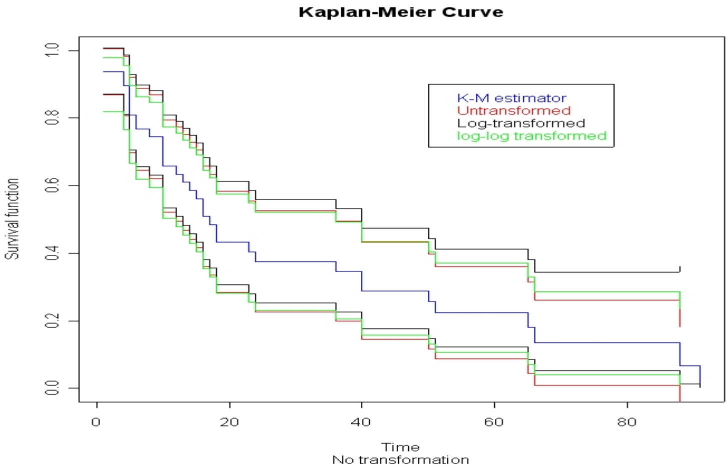

Figure: Three different ways to construct confidence intervals

Other estimators for Survival Function and Hazard Functions

1. Lifetable estimates

$$\tilde{S}(t) = \prod_{\{k|t_k \le t\}} \left(1 - \frac{d_k}{n_k - m_k/2}\right).$$

2. Nelson-Aalen estimator of the cumulative hazard

$$\hat{H}(t) = \sum_{\{k|t_k \le t\}} \frac{d_k}{n_k}.$$

3. The variance of the Nelson-Aalen estimator can be estimated by

$$\hat{V}\{\hat{H}(t)\} = \sum_{\{k|t_k \le t\}} \frac{d_k(n_k - d_k)}{n_k^3}.$$

4. A 95% confidence interval for $H(t)$

$$\hat{H}(t) \pm 1.96\sqrt{\hat{V}\{\hat{H}(t)\}}.$$

Example: Cumulative Hazard Function Estimates (Nelson-Aalen Estimator)

T_k	d_k	n_k	d_k/n_k	CH	V(CH)
1	3	48	0.0625	0.0625	0.0349
4	2	44	0.0455	0.1080	0.0470
5	4	42	0.0952	0.2032	0.0653
6	2	38	0.0526	0.2558	0.0746
8	1	35	0.0286	0.2844	0.0798
10	4	34	0.1176	0.4020	0.0970
12	1	28	0.0357	0.4378	0.1032
13	1	26	0.0385	0.4762	0.1099
14	1	25	0.0400	0.5162	0.1166
15	1	24	0.0417	0.5579	0.1236
16	2	22	0.0909	0.6488	0.1379
17	1	20	0.0500	0.6988	0.1463

Example: Cumulative Hazard Function Estimates (Nelson-Aalen Estimator)

T_k	d_k	n_k	d_k/n_k	CH	V(CH)
18	2	19	0.1053	0.8041	0.1623
23	1	15	0.0667	0.8707	0.1747
24	1	14	0.0714	0.9422	0.1877
36	1	13	0.0769	1.0191	0.2018
40	2	12	0.1667	1.1857	0.2286
50	1	9	0.1111	1.2969	0.2515
51	1	8	0.1250	1.4219	0.2774
65	1	5	0.2000	1.6219	0.3300
66	1	4	0.2500	1.8719	0.3947
88	1	2	0.5000	2.3719	0.5299
91	1	1	1	3.3719	0.5299

95% confidence intervals for cumulative hazard(Nelson-Aalen estimator)

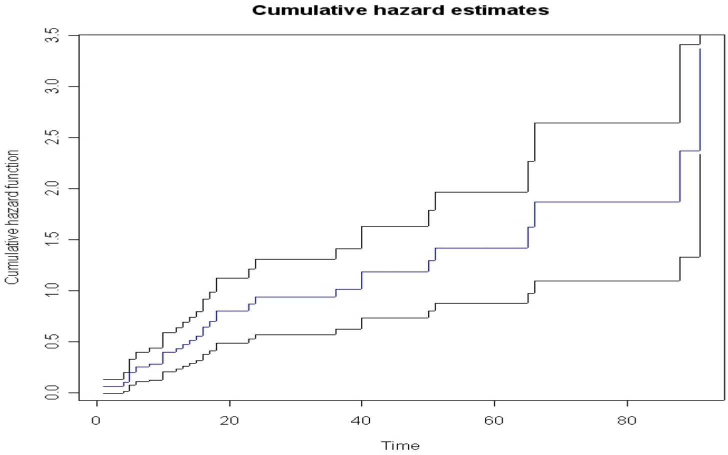

Figure: Cofidence interval for the cumulative hazard function

Median Survival Time

1. Mean survival time is not estimable because the survival distribution is estimable only up to the largest uncensored survival time.
2. Median survival time is often used in place of the mean survival time. Median survival time denoted by $t(0.5)$ is defined as

$$t(0.5) = S^{-1}(0.5) = \inf\{t | S(t) \leq 0.5\},$$

and is estimated by

$$\hat{t}(0.5) = \min\{t_k | \hat{S}(t_k) \leq 0.5\}$$

3. In general, let p be the probability and $100p$ be the percentile survival time denoted by $t(p)$ is defined as

$$t(p) = S^{-1}(1-p) = \inf\{t | S(t) \leq 1-p\}$$

and is estimated by

$$\hat{t}(p) = \min\{t_k | \hat{S}(t_k) \leq 1-p\}$$

Confidence Interval for Median Survival Time

1. The variance of the $100p$ percentile survival time estimator
$$V\{\hat{t}(p)\} \approx \frac{1}{f^2\{\hat{t}(p)\}} V[\hat{S}\{\hat{t}(p)\}],$$
where $V[\hat{S}\{\hat{t}(p)\}]$ can be estimated by the Greenwood variance formula.
2. $f(\hat{t}(p))$ can be approximated by
$$\hat{f}\{\hat{t}(p)\} = -\frac{\hat{S}\{t(p+\epsilon)\} - \hat{S}\{t(p-\epsilon)\}}{t(p+\epsilon) - t(p-\epsilon)},$$
for some small $\epsilon > 0$.
3. The confidence interval for the $100p$ percentile survival time estimator
$$\hat{t}(p) \pm 1.96\sqrt{V\{\hat{t}(p)\}}.$$
The confidence interval for the median survival can be obtained by setting $p = 0.5$.

Example: Median survival Time Estimates

1. The median survival time estimate
$$\hat{t}(0.5) = 17,$$
from the table for Kaplan -Meier estimates.
2. The variance
$$V[\hat{S}\{t(0.5)\}] = 0.0762^2$$
from the table.
3. Let $\epsilon = 0.1$. The estimate
$$\begin{aligned} \hat{f}\{\hat{t}(0.5)\} &= -\frac{\hat{S}\{t(0.6)\} - \hat{S}\{t(0.4)\}}{t(0.6) - t(0.4)} \\ &= -\frac{0.5852 - 0.3756}{14 - 24} = 0.021. \end{aligned}$$
4. The 95% confidence interval for the median survival time is
$$17 \pm 1.96 \times \frac{0.0762}{0.021} = 17 \pm 7.11 = (9.89, 24.11).$$

Another calculation of the Median Survival Time Estimates

1. Take the two neighboring values

$$t(1 - 0.4334) = 18, \qquad t(1 - 0.5098) = 16$$

2. Estimate

$$\begin{aligned} \hat{f}\{\hat{t}(0.5)\} &= -\frac{\hat{S}\{t(p_2)\} - \hat{S}\{t(p_1)\}}{t(p_2) - t(p_1)} \\ &= -\frac{0.5098 - 0.4334}{16 - 18} = 0.0382. \end{aligned}$$

3. The 95% confidence interval for the median survival time is

$$17 \pm 1.96 \times \frac{0.0762}{0.0382} = 17 \pm 3.91 = (13.09, 20.91).$$

Median Survival Time Estimates Based on Transformation

1. The confidence interval for the median survival time can potentially include negative values. Alternative normal approximations may be used.
2. The variance based on the log-transformation (the first approximation)

$$V\{\log t(0.5)\} = \frac{V\{t(0.5)\}}{t^2(0.5)} = \frac{0.0762^2}{0.021^2} \times \frac{1}{17^2} = 0.0456.$$

3. The 95% confidence interval for the median survival time based on the log-transformation

$$17 \times \exp(\pm 1.96 \times \sqrt{0.0456}) = (11.19, 25.84).$$

Median Survival Time Estimates Based on Transformation (Continuing)

1. The variance based on the log-transformation (the second approximation)

$$V\{\log t(0.5)\} = \frac{V\{t(0.5)\}}{t^2(0.5)} = \frac{0.0762^2}{0.0382^2} \times \frac{1}{17^2} = 0.01377.$$

2. The 95% confidence interval for the median survival time based on the log-transformation

$$17 \times \exp(\pm 1.96 \times \sqrt{0.01377}) = (13.5, 21.4).$$

Median Survival Time Estimate based on exponential model fit

1. The median survival time estimate from an exponential model can be derived from

$$\exp\{-\hat{\lambda}\hat{t}(0.5)\} = 0.5.$$

   This implies

$$\hat{t}(0.5) = -\frac{1}{\hat{\lambda}} log(0.5) = \frac{1}{\hat{\lambda}} log(2).$$

2. The variance of the median survival time is thus

$$V\{\hat{t}(0.5)\} = \frac{\{\log(2)\}^2}{\lambda^4} V(\hat{\lambda}).$$

3. A 95% confidence interval for $t(0.5)$ is

$$\hat{t}(0.5) \pm 1.96\sqrt{\hat{V}\{\hat{t}(0.5)\}} = \frac{log(2)}{\hat{\lambda}}\left(1 \pm \frac{1.96}{\hat{\lambda}}\sqrt{\hat{V}(\hat{\lambda})}\right\}.$$

*Derivation: Kaplan-Meier as the Nonparametric MLE

1. Let the observed data $(X_i, \delta_i), i = 1, \cdots, n$

$$\prod_{i=1}^{n} f^{\delta_i}(X_i) S^{1-\delta_i}(X_i).$$

2. Let the observed data be ordered from the smallest to the largest as

$$\begin{aligned} 0 \leq \quad & C_{11} \leq \cdots \leq C_{m_1 1} < T_1 \\ & \leq C_{12} \leq \cdots \leq C_{m_2 2} < T_2 \\ & \ldots \\ & \leq C_{1K} \leq \cdots \leq C_{m_K K} < T_K \\ & \leq C_{1(K+1)} \leq \cdots \leq C_{m_{(K+1)}(K+1)}. \end{aligned}$$

where T_j, $j = 1, \cdots, K$ are distinctive event times, and $C_{1j}, \cdots, C_{m_j j}$ are all the censoring times greater than or equal to T_{j-1} and less than T_j.

*Derivation: Kaplan-Meier as the Nonparametric MLE (continuing)

1. The nonparametric likelihood can be written as

$$\begin{gathered}\{\prod_{j=1}^{m_1} S(C_{j1})\}\{S(T_1-)-S(T_1)\}^{d_1}\\ \times\{\prod_{j=1}^{m_2} S(C_{j2})\}\{S(T_2-)-S(T_2)\}^{d_2}\\ \cdots\\ \times\{\prod_{j=1}^{m_K} S(C_{jK})\}\{S(T_K-)-S(T_K)\}^{d_K}\\ \times\{\prod_{j=1}^{m_{K+1}} S(C_{j(K+1)})\}.\end{gathered}$$

   where d_j is the number of observed failures at T_j.

2. The maximum has to satisfy

$$S(T_{k-1}) = S(C_{jk}) = S(T_k-),$$

   and $S(C_{j1}) = 1$, $j = 1, \cdots, m_1$.

*Derivation: Kaplan-Meier as the Nonparametric MLE (continuing)

1. The nonparametric likelihood reduces to

$$\prod_{k=1}^{K} \left[\{S(T_k-) - S(T_k)\}^{d_k} S^{m_k}(T_k) \right].$$

2. Let $\lambda_k = S(T_k-) - S(T_k)$. Then

$$\begin{aligned} S(T_1) &= 1 - \lambda_1 \\ S(T_2) &= 1 - \lambda_1 - \lambda_2 \\ &\dots \\ S(T_K) &= 1 - \lambda_1 - \cdots - \lambda_K. \end{aligned}$$

3. The likelihood can be re-expressed as

$$\prod_{k=1}^{K} \lambda_k^{d_k} (1 - \lambda_1 - \cdots - \lambda_k)^{m_k}.$$

*Derivation: Kaplan-Meier as the Nonparametric MLE (continuing)

1. The maximizers can be obtained recursively as

$$
\begin{aligned}
\hat{\lambda}_K &= (1 - \lambda_1 - \cdots - \lambda_{K-1}) \frac{d_K}{m_K + d_K}, \\
&\ldots \\
\hat{\lambda}_k &= (1 - \lambda_1 - \cdots - \lambda_{k-1}) \frac{d_k}{m_k + d_k + \cdots + m_K + d_K}, \\
&\ldots \\
\hat{\lambda}_1 &= \frac{d_1}{n}.
\end{aligned}
$$

*Derivation: Kaplan-Meier as the Nonparametric MLE (continuing)

1. The maximum likelihood estimator

$$\begin{aligned}
\hat{\lambda}_1 &= \frac{d_1}{n_1}, \\
\hat{\lambda}_2 &= (1 - \frac{d_1}{n_1})\frac{d_2}{n_2}, \\
& \cdots \\
\hat{\lambda}_k &= \prod_{j=1}^{k-1}(1 - \frac{d_j}{n_j})\frac{d_k}{n_k}, \\
& \cdots \\
\lambda_K &= \prod_{j=1}^{K-1}(1 - \frac{d_j}{n_j})\frac{d_K}{n_K}.
\end{aligned}$$

where $n_k = m_k + d_k + \cdots + m_K + d_K$.

*Derivation: Variance of the Kaplan-Meier estimator

1. The maximum likelihood estimator of the survival function

$$
\begin{aligned}
\hat{S}(T_1) &= 1 - \frac{d_1}{n_1}, \\
\hat{S}(T_2) &= (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2}), \\
&\cdots \\
\hat{S}(T_K) &= \prod_{j=1}^{K}(1 - \frac{d_j}{n_j}).
\end{aligned}
$$

2. The log survival function estimator

$$\log \hat{S}(t) = \sum_{\{k \mid T_k \leq t\}} \log\left(1 - \frac{d_k}{n_k}\right).$$

*Derivation: Variance of the Kaplan-Meier estimator (continuing)

1. The variance for $\log \hat{S}(t)$,

$$
\begin{aligned}
V\left\{\log \hat{S}(t)\right\} &= \sum_{\{k \mid T_k \le t\}} V\left\{\log\left(1 - \frac{d_k}{n_k}\right)\right\} \\
&= \sum_{\{k \mid T_k \le t\}} \left(1 - \frac{d_k}{n_k}\right)^{-2} V\left(\frac{d_k}{n_k}\right) \\
&= \sum_{\{k \mid T_k \le t\}} \left(1 - \frac{d_k}{n_k}\right)^{-2} \frac{1}{n_k}\frac{d_k}{n_k}\left(1 - \frac{d_k}{n_k}\right) \\
&= \sum_{\{k \mid T_k \le t\}} \frac{d_k}{n_k(n_k - d_k)}
\end{aligned}
$$

2. The variance for $\hat{S}(t)$,

$$
V\left\{\hat{S}(t)\right\} = \hat{S}^2(t) V\left\{\log \hat{S}(t)\right\}.
$$

*Derivation: Confidence Interval for Median Survival Time

1. Note that

$$
\begin{aligned}
0 &= \hat{S}\{\hat{S}^{-1}(1-p)\} - S\{S^{-1}(1-p)\} \\
&= \hat{S}\{S^{-1}(1-p)\} - S\{S^{-1}(1-p)\} \\
&\quad + S\{\hat{S}^{-1}(1-p)\} - S\{S^{-1}(1-p)\} \\
&\quad + \hat{S}\{\hat{S}^{-1}(1-p)\} - \hat{S}\{S^{-1}(1-p)\} \\
&\quad - S\{\hat{S}^{-1}(1-p)\} + S\{S^{-1}(1-p)\} \\
&= \hat{S}\{S^{-1}(1-p)\} - S\{S^{-1}(1-p)\} \\
&\quad - f\{S^{-1}(1-p)\}\{\hat{S}^{-1}(1-p) - S^{-1}(1-p)\} + o_p(1).
\end{aligned}
$$

*Derivation: Confidence Interval for Median Survival Time (Continuing)

1. It follows that

$$\hat{S}^{-1}(1-p) - S^{-1}(1-p) = f^{-1}\{S^{-1}(1-p)\}$$

$$\times \left[\hat{S}\{S^{-1}(1-p)\} - S\{S^{-1}(1-p)\} + o_p(1)\right]$$

2. This leads

$$\hat{t}(p) - t(p) \rightarrow N\left(0, f^{-2}\{S^{-1}(1-p)\} V\left[\hat{S}\{S^{-1}(1-p)\}\right]\right).$$