5.17 Multicensored Data

1 Warming-up: empirical distribution

Suppose that there are n observations, $x_1, x_2, \ldots, x_n$. We assume that there are k distinct observations, $y_1 < y_2 < \cdots < y_k$, with frequencies, $f_1, f_2, \ldots, f_k$, respectively. Then the likelihood function is given by

$$L(p_1, p_2, \ldots, p_k) \propto \prod_{j=1}^{k} {p_j}^{f_j},$$

with the constraint $\sum_{j=1}^{k} p_j = 1$. The log-likelihood with the constraint is

$$\ell(p_1, p_2, \ldots, p_k, \lambda) \propto \sum_{j=1}^{k} f_j \log p_j - \lambda\big(\sum_{j=1}^{k} p_j - 1\big),$$

where λ is the Lagrange multiplier. It is immediate from

$$\frac{\partial \ell}{\partial p_j} = \frac{f_j}{p_j} - \lambda = 0 \quad \text{and} \quad \frac{\partial \ell}{\partial \lambda} = -\big(\sum_{j=1}^{k} p_j - 1\big) = 0$$

that $\lambda = \sum_{j=1}^{k} f_j = n$ and $\hat{p}_j = f_j/\lambda = f_j/n$. Thus, the empirical distribution is obtained as

$$\hat{F}_n(t) = \sum_{j=1}^{k} \hat{p}_j \cdot \mathbb{I}(y_j \le t) = \frac{1}{n} \sum_{j=1}^{k} f_j \cdot \mathbb{I}(y_j \le t),$$

which is equivalent to

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \le t),$$

where $\mathbb{I}(\cdot)$ is an indicator function.

2 Nonparametric estimation method for the survival function

We consider the empirical survival function which can be obtained by the nonparametric maximum likelihood method.

Suppose that there are n observations $(x_1, x_2, \ldots, x_n)$ and that there are k distinct observations $(y_1 < y_2 < \cdots < y_k)$ at which *failures* occur. We set $y_0 = 0$ and $y_{k+1} = \infty$ by convention.

We assume that the values of survival function change only at the distinct failure points, that is, $S(y_j) = S(t)$ for $y_{j-1} < t \leq y_j$. Similarly, $F(y_j) = F(t)$ for $y_j \leq t < y_{j+1}$. Let d_j denote the number of observed failures at y_j and n_j denote the number of items on test just before time y_i.

Suppose that there are c_j *right-censored* observations within the interval $[y_j, y_{j+1})$. (Only for convenience, we summarize right-censored observations like this setup. Note that these are *not* interval-censored). Note that $n_1 = n$, $n_{k+1} = 0$, $n_{j+1} = n_j - c_j - d_j$, and $\sum_{j=i}^{k}(c_j + d_j) = n_i$.

Example 1 (6-MP data)**.** An experiment is conducted to determine the effect of a drug named 6-mercaptopurine (6-MP) on leukemia remission times.[1] The 6-MP experiment data set contains $n = 21$ patients on test whose failure observations are given by **6**, 6, 6, 6^+, **7**, 9^+, **10**, 10^+, 11^+, **13**, **16**, 17^+, 19^+, 20^+, **22**, **23**, 25^+, 32^+, 32^+, 34^+, 35^+.

j	1	2	3	4	5	6	7
y_j	**6**	**7**	**10**	**13**	**16**	**22**	**23**
$[y_j, y_{j+1})$	$[6, 7)$	$[7, 10)$	$[10, 13)$	$[13, 16)$	$[16, 22)$	$[22, 23)$	$[23, \infty)$
n_j	21	17	15	12	11	7	6
c_j	1	1	2	0	3	0	5
d_j	3	1	1	1	1	1	1

$\triangle$

Let $t_{j1}, t_{j2}, \ldots, t_{jc_j}$ be the censoring times within the interval $[y_j, y_{j+1})$. Then the likelihood function becomes

$$L = \prod_{j=1}^{k} f(y_j)^{d_j} \times \prod_{j=1}^{k}\prod_{i=1}^{c_j} \big\{1 - F(t_{ji})\big\}. \tag{1}$$

Since the values of survival function change only at the *distinct failure* points and the distribution $F(\cdot)$ is right-continuous, we have $F(y_j) = F(t_{ji})$ for $i = 1, 2, \ldots, c_j$. Thus, the above likelihood becomes

$$\boxed{L = \prod_{j=1}^{k} f(y_j)^{d_j} \times \prod_{j=1}^{k} \big\{1 - F(y_j)\big\}^{c_j}.} \tag{2}$$

[1]Example 10.2 in: Leemis, L. M. Reliability: Probabilistic Models and Statistical Methods. 2nd edition. Williamsburg, Virginia: Lawrence M. Leemis, 2009.

Theorem 1. *We have*

$$S(y_j) = \prod_{i=1}^{j-1} [1 - h(y_i)]$$

for $j \geq 2$ *and* $S(y_1) = 1$.

Proof. The survival function is a step function because it changes only at the distinct failure points. Since $S(y_j) = P(Y \geq y_j) = f(y_j) + f(y_{j+1}) + \cdots + f(y_k)$ and $S(y_{j+1}) = f(y_{j+1}) + f(y_{j+2}) + \cdots + f(y_k)$, we have $f(y_j) = S(y_j) - S(y_{j+1})$. Also, it is easily seen that $S(y_1) = f(y_1) + f(y_2) + \cdots + f(y_k) = 1$.

The hazard rate function is then given by

$$h(y_j) = \frac{f(y_j)}{S(y_j)} = \frac{S(y_j) - S(y_{j+1})}{S(y_j)} = 1 - \frac{S(y_{j+1})}{S(y_j)}.$$

It is immediate from solving the above for $S(y_{j+1})$ that we have

$$S(y_{j+1}) = S(y_j)\big[1 - h(y_j)\big].$$

Then using the mathematical induction with $S(y_1) = 1$, we have

$$S(y_{j+1}) = \big[1 - h(y_j)\big]\big[1 - h(y_{j-1})\big] \cdots \big[1 - h(y_1)\big] = \prod_{i=1}^{j} \big[1 - h(y_i)\big], \tag{3}$$

which completes the proof. □

For notational convenience, we let $h_j = h(y_j)$. We have $S(y_j) = \prod_{i=1}^{j-1}(1 - h_i)$ for $j \geq 2$. We can also rewrite as

$$S(y_j) = \prod_{i=1}^{j}(1 - h_i) \cdot (1 - h_j)^{-1} \tag{4}$$

for $j \geq 1$. It is easily seen from (4) that

$$\boxed{f(y_j) = h(y_j)S(y_j) = h_j S(y_j) = h_j \prod_{i=1}^{j}(1 - h_i) \cdot (1 - h_j)^{-1}.} \tag{5}$$

Since $F(y_j) = f(y_1) + f(y_2) + \cdots + f(y_j)$ and $S(y_{j+1}) = f(y_{j+1}) + f(y_{j+2}) + \cdots$, we have $F(y_j) + S(y_{j+1}) = 1$. Using this and (3), we have

$$\boxed{1 - F(y_j) = S(y_{j+1}) = \prod_{i=1}^{j}(1 - h_i).} \tag{6}$$

It should be noted that $1 - F(x) \neq S(x)$ for this discrete case since $1 - F(x) = 1 - P(X \leq x) = P(X > x)$ and $S(x) = P(X \geq x)$. Also, $F(x) = P(X \leq x)$ is right-continuous, but

$S(x) = P(X \geq x)$ is left-continuous. Note that $R(x) = P(X > x)$ is right-continuous due to $R(x) = 1 - F(x)$.

Substituting (5) and (6) into (2), we have

$$
\begin{aligned}
L &= \prod_{j=1}^{k} f(y_j)^{d_j} \times \prod_{j=1}^{k} \{1 - F(y_j)\}^{c_j} \\
&= \prod_{j=1}^{k} \left[h_j \prod_{i=1}^{j} (1 - h_i) \cdot (1 - h_j)^{-1} \right]^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{c_j}.
\end{aligned}
$$

Using the above results with tedious algebra, the likelihood is given by Substituting (5) and (6) into (2), we have

$$
\begin{aligned}
L &= \prod_{j=1}^{k} h_j{}^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \cdot (1 - h_j)^{-1} \right]^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{c_j} \\
&= \prod_{j=1}^{k} h_j{}^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{c_j} \times \prod_{j=1}^{k} (1 - h_j)^{-d_j} \\
&= \prod_{j=1}^{k} h_j{}^{d_j} \times \prod_{j=1}^{k} \left\{ \left[\prod_{i=1}^{j} (1 - h_i) \right]^{d_j} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{c_j} \right\} \times \prod_{j=1}^{k} (1 - h_j)^{-d_j} \\
&= \prod_{j=1}^{k} h_j^{d_j} \times \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - h_i) \right]^{d_j + c_j} \times \prod_{j=1}^{k} (1 - h_j)^{-d_j} \\
&= \prod_{j=1}^{k} h_j^{d_j} \times \prod_{j=1}^{k} \prod_{i=1}^{j} (1 - h_i)^{c_j + d_j} \times \prod_{j=1}^{k} (1 - h_j)^{-d_j} \qquad (7)
\end{aligned}
$$

Since $\prod_{j=1}^{k} \prod_{i=1}^{j} a_{ij} = \prod_{i=1}^{k} \prod_{j=i}^{k} a_{ij}$ and $\sum_{j=i}^{k} (c_j + d_j) = n_i$, we have

$$
\begin{aligned}
\prod_{j=1}^{k} \prod_{i=1}^{j} (1 - h_i)^{c_j + d_j} &= \prod_{i=1}^{k} \prod_{j=i}^{k} (1 - h_i)^{c_j + d_j} \\
&= \prod_{i=1}^{k} (1 - h_i)^{\sum_{j=i}^{k} (c_j + d_j)} = \prod_{i=1}^{k} (1 - h_i)^{n_i} = \prod_{j=1}^{k} (1 - h_j)^{n_j}.
\end{aligned}
$$

Substituting the above into (7), we have

$$
\begin{aligned}
L &= \prod_{j=1}^{k} h_j^{d_j} \times \prod_{j=1}^{k} (1 - h_j)^{n_j} \times \prod_{j=1}^{k} (1 - h_j)^{-d_j} \\
&= \prod_{j=1}^{k} h_j^{d_j} \times \prod_{j=1}^{k} (1 - h_j)^{n_j - d_j} \\
&= \prod_{j=1}^{k} h_j^{d_j} (1 - h_j)^{n_j - d_j}
\end{aligned}
$$

and

$$\ell = \log L = \sum_{j=1}^{k} \Big\{ d_j \log h_j + (n_j - d_j) \log(1 - h_j) \Big\}.$$

It is immediate from $\partial \ell / \partial h_i = 0$ that

$$\hat{h}_i = \frac{d_i}{n_i}.$$

Since $S(y_j) = \prod_{i=1}^{j-1}(1 - h_i)$ for $j \geq 2$ from Theorem 1, we can estimate the survival function as

$$\hat{S}(t) = \prod_{j: y_j < t} (1 - \hat{h}_j) = \prod_{j: y_j < t} \left(1 - \frac{d_j}{n_j}\right) = \prod_{j: y_j < t} \frac{n_j - d_j}{n_j}, \tag{8}$$

where $t > y_1$. This estimate is commonly known as the *Kaplan–Meier* or *product limit* estimate of the survival function. Note that $\hat{S}(t) = 1$ for $t \leq y_1$. The right-continuous version of the survival function estimate, denoted by $\hat{R}(t)$, is given by

$$\hat{R}(t) = \prod_{j: y_j \leq t} \frac{n_j - d_j}{n_j}.$$

It should be noted that we have $\hat{R}(t) = \hat{S}(t^{+0})$.

Example 2 (6-MP data)**.** For the previous example, we have following the survival function estimates.

j	1	2	3	4	5	6	7
y_j	**6**	**7**	**10**	**13**	**16**	**22**	**23**
$[y_j, y_{j+1})$	$[6, 7)$	$[7, 10)$	$[10, 13)$	$[13, 16)$	$[16, 22)$	$[22, 23)$	$[23, \infty)$
n_j	21	17	15	12	11	7	6
d_j	3	1	1	1	1	1	1
$\dfrac{n_j - d_j}{n_j}$	$\frac{18}{21}$	$\frac{16}{17}$	$\frac{14}{15}$	$\frac{11}{12}$	$\frac{10}{11}$	$\frac{6}{7}$	$\frac{5}{6}$
$\hat{R}(y_j)$	$\frac{18}{21}$	$\frac{18}{21} \cdot \frac{16}{17}$	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15}$	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{11}{12}$	$\frac{32}{51}$	$\frac{64}{119}$	$\frac{160}{357}$

$\triangle$

3 Variance of the Kaplan–Meier estimate of the survival function

From (8), the logarithm of the Kaplan–Meier estimator of the survival function is given by

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{\pi}_j,$$

where $k = \max\{j : y_j < t\}$ and $\hat{\pi}_j = (n_j - d_j)/n_j$. Then its variance becomes

$$\operatorname{Var}\big(\log \hat{S}(t)\big) = \sum_{j=1}^{k} \operatorname{Var}\big(\log \hat{\pi}_j\big). \tag{9}$$

We assume that the number of items which survive in the interval $[y_j, y_{j+1})$ has a binomial distribution with parameters n_j and π_j, that is,

$$n_j - d_j \sim \operatorname{Bin}(n_j, \pi_j).$$

Theorem 2 (Delta Method)**.** *Suppose that we are interested in the variance of $g(Y)$. Then we can estimate it approximately with*

$$\operatorname{Var}\big(g(Y)\big) \approx \big\{g'(\mu)\big\}^2 \operatorname{Var}(Y),$$

where $\mu = E(Y)$.

Proof. The first-order Taylor series expansion of $g(Y)$ around μ is

$$g(Y) - g(\mu) \approx g'(\mu)(Y - \mu),$$

where $E(Y) = \mu$. The result is easily obtained by taking the variance of the above. □

Using the above theorem, we have

$$\operatorname{Var}\big(\log \hat{\pi}_j\big) \approx \left(\frac{1}{\pi_j}\right)^2 \operatorname{Var}(\hat{\pi}_j).$$

Since $n_j - d_j \sim \operatorname{Bin}(n_j, \pi_j)$, we have $\operatorname{Var}(\hat{\pi}_j) = \pi_j(1 - \pi_j)/n_j$. Then we have

$$\operatorname{Var}\big(\log \hat{\pi}_j\big) \approx \frac{1 - \pi_j}{\pi_j n_j},$$

and it can thus be estimated by $\widehat{\operatorname{Var}}\big(\log \hat{\pi}_j\big) \approx (1 - \hat{\pi}_j)/(\hat{\pi}_j n_j)$. Again, note that $\hat{\pi}_j = (n_j - d_j)/n_j$. Thus, we can estimate $\operatorname{Var}\big(\log \hat{\pi}_j\big)$ with

$$\widehat{\operatorname{Var}}\big(\log \hat{\pi}_j\big) \approx \frac{d_j}{n_j(n_j - d_j)}. \tag{10}$$

It is immediate upon using (9) and (10) that we have

$$\widehat{\mathrm{Var}}\big(\log \hat{S}(t)\big) \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}, \tag{11}$$

where $y_j < t \leq y_{j+1}$ and $k = \max\{j : y_j < t\}$ again.

Applying the delta method again to $\mathrm{Var}\big(\log \hat{S}(t)\big)$ with $E\big[\hat{S}(t)\big] \approx S(t)$, we have

$$\mathrm{Var}\big(\log \hat{S}(t)\big) \approx \frac{1}{S(t)^2} \cdot \mathrm{Var}\big(\hat{S}(t)\big), \tag{12}$$

so that we also have

$$\mathrm{Var}\big(\hat{S}(t)\big) \approx S(t)^2 \cdot \mathrm{Var}\big(\log \hat{S}(t)\big).$$

Using (11) and $\hat{S}(t)$ as an estimate of $S(t)$, we can estimate the above with

$$\widehat{\mathrm{Var}}\big(\hat{S}(t)\big) \approx \hat{S}(t)^2 \cdot \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$

This result is known as *Greenwood's formula*. The approximate standard error is also obtained by

$$\mathrm{SE}\big(\hat{S}(t)\big) = \hat{S}(t) \cdot \Big[\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\Big]^{1/2}. \tag{13}$$

Example 3. Twenty five units of dry bearings are subjected to a creep test and their failure times are given by[2] **70**, **180**, 190^+, **200**, **210**, **230**, **275**, **295**, **310**, 370^+, **395**, **420**, **480**, **495**, **560**, 600^+, 620^+, **680**, **750**, **780**, **800**, **900**, 980^+, 1010^+, 1020^+. For more details, see the textbook.

```r
# Example 5.31 on Page 348 of Elsayed.
times = c( 70, 180, 190, 200, 210, 230, 275, 295, 310, 370,
           395, 420, 480, 495, 560, 600, 620, 680, 750, 780,
           800, 900, 980,1010,1020 )
      d= c(1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0 )
cbind(times, d)

# Estimates of Survivals Using library
library(survival)
Sn = survfit(Surv(times,d)~1)
summary(Sn)

```

[2] Example 5.31 in: Elsayed, E. A. Reliability Engineering. 2nd edition. Wiley, 2012.

```
# Compare the above with the following:
nj = c(25,24,22,21,20,19,18,17,15,14,13,12,11,8,7,6,5,4)
dj = rep(1,18)
cumprod( (nj-dj)/nj )

# Table 5.17 on Page 348
h.hat = dj / nj
H.hat = cumsum(h.hat)
R.ch  = exp( -cumsum(dj/nj) )
R.pl  = cumprod( (nj-dj)/nj )

out = cbind( times[d==1], nj, dj, h.hat, H.hat, R.ch, R.pl)
round(out,3)

# Plots of Survivals
Sn = survfit(Surv(times,d)~1)
plot(Sn)
plot(Sn, conf.int=FALSE) # without CI
```

△

4 Point-wise confidence interval for the survival function

A $100(1-\alpha)\%$ confidence interval for $S(t)$ for a given value of t can be obtained by using the approximate standard error in (13), which is called the standard (or linear plain) confidence interval. The confidence limits for this interval have

$$\hat{S}(t) \pm z_{\alpha/2} \cdot \mathrm{SE}\big(\hat{S}(t)\big),$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution. Note that this confidence interval is symmetric at $\hat{S}(t)$ so that it can be outside of the range of the survival function, $[0,1]$, when the survival estimate, $\hat{S}(t)$, is very close to zero or one. One easy simple solution for this problems is to replace a limit greater than one with one and any value less than zero with zero.

We can obtain better confidence intervals by transforming the survival function. Possible choices can be $\log S(t)$ (log-transformation), $\log\big\{-\log S(t)\big\}$ (log-log-transformation), $\log\big[S(t)/(1-S(t))\big]$ (logistic-transformation), arcsin-square root transformation, etc. For more details, one can refer to Borgan and Liestøl[3].

For the log-transformation alternative, the variance of $\log S(t)$ can be estimated by using the approximation in (11). Thus, the approximate standard error is given by

$$\mathrm{SE}\big(\log \hat{S}(t)\big) = \Big[\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}\Big]^{1/2}. \qquad (14)$$

Then a $100(1-\alpha)\%$ confidence limits for $\log S(t)$ is $\log \hat{S}(t) \pm z_{\alpha/2}\mathrm{SE}\big(\log \hat{S}(t)\big)$, which leads to a $100(1-\alpha)\%$ confidence limits for $S(t)$ of the form

$$\hat{S}(t)\ \exp\big[\pm z_{\alpha/2} \cdot \mathrm{SE}\big(\log \hat{S}(t)\big)\big].$$

This log-transformation guarantees that the confidence limits are always greater than zero. However, it can have a value greater than one.

For the log-log-transformation, we need to estimate the variance of $\log\big[-\log S(t)\big]$. Analogous with the approach used in (12) to obtain the variance of $\log \hat{S}(t)$, the variance of $\log\big\{-\log \hat{S}(t)\big\}$ is given by

$$\mathrm{Var}\Big[\log\big\{-\log \hat{S}(t)\big\}\Big] \approx \frac{1}{\{\log S(t)\}^2} \cdot \mathrm{Var}\big(\log \hat{S}(t)\big). \qquad (15)$$

[3] Borgan, Ørnulf/Liestøl, Knut A Note on Confidence Intervals and Bands for the Survival Function Based on Transformations. Scandinavian Journal of Statistics, 17 1990.

The approximate standard error of $\log\big\{-\log \hat{S}(t)\big\}$ is easily obtained using (11)

$$\mathrm{SE}\Big[\log\big\{-\log \hat{S}(t)\big\}\Big] = \Big[\frac{1}{\{\log \hat{S}(t)\}^2}\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}\Big]^{1/2}, \tag{16}$$

which results in the $100(1-\alpha)\%$ confidence limits of the form

$$\begin{aligned}\log\big\{-\log \hat{S}(t)\big\} \pm z_{\alpha/2}\cdot \mathrm{SE} &= \log\big\{-\log \hat{S}(t)\cdot \exp(\pm z_{\alpha/2}\mathrm{SE})\big\}\\ &= \log\big\{-\log \hat{S}(t)^{\exp(\pm z_{\alpha/2}\mathrm{SE})}\big\},\end{aligned}$$

where we denote the approximate standard error formula in (16) by SE for brevity. Note that the inverse of the function, $y = \log(-\log x)$, is given by $x = \exp\big(-\exp(y)\big)$. Using this inverse, we can obtain the $100(1-\alpha)\%$ confidence limits for $S(t)$

$$\hat{S}(t)^{\exp(\pm z_{\alpha/2}\mathrm{SE})}.$$

It should be noted that these limits are always in $[0,1]$.