§3.3. Confidence bands for the survival function

A 100(1-\(\alpha\))% confidence band for the survival curve on the interval \([t_L, t_U]\) is given by two functions \(L(t)\) and \(U(t)\) such that

\[
P[L(t) \leq S(t) \leq U(t); t_L \leq t \leq t_U] = 1-\alpha.
\]

The construction of a confidence band requires the following two quantities:

\[
a_L = \frac{n\sigma^2_H(t_L)}{1 + n\sigma^2_H(t_L)},
\]

\[
a_U = \frac{n\sigma^2_H(t_U)}{1 + n\sigma^2_H(t_U)}.
\]


In the construction of pointwise confidence interval, replacing $z_{1-\alpha/2}$ with $c_\alpha(a_L, a_U)$, which is given in Klein and Moeschberger (Appendix C, Table C.3), yields the confidence bands:

**Linear:**

$$\hat{S}(t) \pm c_\alpha(a_L, a_U)\sigma_H(t)\hat{S}(t).$$

**Log-log transformed:**

$$[\hat{S}(t)^{1/b(t)}, \hat{S}(t)^{b(t)}]$$

where

$$b(t) = \exp\left\{\frac{c_\alpha(a_L, a_U)\sigma_H(t)}{\ln \hat{S}(t)}\right\}.$$
**Arcsine-square root transformed:**

\[ [\sin^2(\text{ASR}(t)), \sin^2(\overline{\text{ASR}}(t))] \]

where

\[
\text{ASR}(t) = \max\left\{0, \arcsin(\sqrt{\hat{S}(t)}) - 0.5c_\alpha(a_L, a_U)\sigma_H(t)\sqrt{\frac{\hat{S}(t)}{1-S(t)}} \right\},
\]

\[
\overline{\text{ASR}}(t) = \min\left\{\frac{\pi}{2}, \arcsin(\sqrt{\hat{S}(t)}) + 0.5c_\alpha(a_L, a_U)\sigma_H(t)\sqrt{\frac{\hat{S}(t)}{1-S(t)}} \right\}.
\]

**Hall and Wellner (HW) bands.**

(Hall and Wellner (1980): *Biometrika* 67: 133-143.)

The HW bands is obtained by replacing \( z_{1-\alpha/2} \sigma_H(t) \) with

\[
\frac{k_\alpha(a_L, a_U)[1 + n\sigma_H^2(t)]}{\sqrt{n}},
\]

in the construction of pointwise confidence interval. The coefficients \( k_\alpha(a_L, a_U) \) are given in Klein and Moeschberger (Table C.4).
For example, the linear confidence band is given by

$$\hat{S}(t) \pm \frac{k_\alpha(a_L, a_U)[1 + n\sigma_H^2(t)]}{\sqrt{n}} \hat{S}(t).$$

**R function for computing confidence bands**

```r
conf.bands=function(critical.value, survfit.obj, band.type="EP", conf.type="plain", plot=T) {
    # critical.value = c.alpha(al,au) if band.type = "EP"
    # = k.alpha(al,au) if band.type ="HW"
    # band.type: "EP" equal probability bands,
    # "HW" Hall-Wellner bands.
    # conf.type: "plain", linear,
    # "log-log", log-log transformed,
    # "log", log transformed.
    time=summary(survfit.obj)$time
    survival=summary(survfit.obj)$surv
    std.err=summary(survfit.obj)$std.err
    n=summary(survfit.obj)$n.risk[1]
    sigma.s=sqrt(std.err^2/survival^2)
    if (band.type=="EP") {
        if (conf.type=="plain") {
            lower=survival*(1-critical.value*sigma.s)
            upper=survival*(1+critical.value*sigma.s)
        } else if (conf.type=="log-log") {

        } else if (conf.type=="log") {

    }
```
theta=exp(c.value*sigma.s/log(survival))
lower=survival^(1/theta)
upper=survival^(theta)
} else if (conf.type=="log") {
    lower=survival*exp(-critical.value*sigma.s)
    upper=survival*exp(critical.value*sigma.s)
}
} else if (band.type=="HW") {
    k.value=c.value*(1+n*sigma.s^2)/sqrt(n)
    if (conf.type=="plain") {
        lower=survival*(1-critical.value)
        upper=survival*(1+critical.value)
    } else if (conf.type=="log-log") {
        theta=exp(critical.value/log(survival))
        lower=survival^(1/theta)
        upper=survival^(theta)
    } else if (conf.type=="log") {
        lower=survival*exp(-critical.value)
        upper=survival*exp(critical.value)
    }
}
if (plot==T) {
    matplot(time,cbind(survival,lower,upper),type="l")
}
data.frame(time,survival,lower, upper)
Example 1.2 (cont.): Computation of confidence bands for 6-MP survival curve.

\[ t_L = 6, t_U = 23 \quad a_L = 0.44, a_U = 0.90 \]
\[ c_\alpha(a_L, a_U) = c_{0.05}(0.44, 0.90) = 2.8618 \]

```
bands.limits=conf.bands(2.8618, leuk.fit, 
    band.type="EP", 
    conf.type="log-log")
```

```
bands.limits
```

<table>
<thead>
<tr>
<th>time</th>
<th>survival</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.857</td>
<td>0.447</td>
<td>0.971</td>
</tr>
<tr>
<td>7</td>
<td>0.807</td>
<td>0.405</td>
<td>0.950</td>
</tr>
<tr>
<td>10</td>
<td>0.753</td>
<td>0.357</td>
<td>0.925</td>
</tr>
<tr>
<td>13</td>
<td>0.690</td>
<td>0.294</td>
<td>0.894</td>
</tr>
<tr>
<td>16</td>
<td>0.627</td>
<td>0.241</td>
<td>0.858</td>
</tr>
<tr>
<td>22</td>
<td>0.538</td>
<td>0.155</td>
<td>0.813</td>
</tr>
<tr>
<td>23</td>
<td>0.448</td>
<td>0.096</td>
<td>0.760</td>
</tr>
</tbody>
</table>
§3.4. Point and interval estimates of the mean and median survival time

- Estimation of the mean life

**Point Estimate:**

\[
\hat{\mu}_{\tau} = \int_0^{\tau} \hat{S}(t) dt,
\]

where \(\hat{S}(t)\) is the PL estimator, and \(\tau\) is either a prespecified the largest possible survival time (in the case that the largest observation is a censoring time) or the largest observation (in the case that it is a survival time).

**Variance and Confidence Interval**

\[
\hat{V}(\hat{\mu}_{\tau}) = \sum_{i=1}^{D} \left[ \int_{t_i}^{\tau} \hat{S}(t) dt \right]^2 \frac{d_i}{Y_i(Y_i - d_i)},
\]

\[
\hat{\mu}_{\tau} \pm z_{1-\alpha/2} / \sqrt{\hat{V}(\hat{\mu}_{\tau})}.
\]
Example 1.2 (cont.): Estimation of mean life for 6-MP group.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$Y_i$</th>
<th>$d_i$</th>
<th>$\hat{S}(t)$</th>
<th>$\int_t^\tau \hat{S}(t)dt$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>−</td>
<td>−</td>
<td>1</td>
<td>$17.286 + 1(6 - 0) = 23.286$</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>3</td>
<td>.857</td>
<td>$16.429 + .857(7 - 6) = 17.286$</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>1</td>
<td>.807</td>
<td>$14.011 + .807(10 - 7) = 16.429$</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>1</td>
<td>.753</td>
<td>$11.752 + .753(13 - 10) = 14.011$</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>1</td>
<td>.690</td>
<td>$9.682 + .690(16 - 13) = 11.752$</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
<td>.627</td>
<td>$5.914 + .628(22 - 16) = 9.682$</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>1</td>
<td>.538</td>
<td>$5.376 + .538(23 - 22) = 5.914$</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>1</td>
<td>.448</td>
<td>$5.376 + .538(23 - 22) = 5.376$</td>
</tr>
</tbody>
</table>

$\tau = 35$.

$\hat{\mu}_\tau = 23.286$. 
\[ \hat{V}(\hat{\mu}_\tau) = \frac{3(17.286^2)}{21(18)} + \frac{16.429^2}{17(16)} + \frac{14.011^2}{15(14)} + \frac{11.752^2}{12(11)} + \frac{9.682^2}{11(10)} + \frac{5.914^2}{7(6)} + \frac{5.376^2}{6(5)} = 7.993. \]

A 95% confidence interval for \( \mu_\tau \):

\[
\hat{\mu}_\tau \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{\mu}_\tau)} = 23.286 \pm 1.96\sqrt{7.993}.
\]

\[ [ 17.745, \ 28.827 ] \]
The R function `survfit` can be used to compute the mean, median and their confidence intervals. For example,

```R
leuk = read.table("leukemia.txt")
leuk.surv = Surv(leuk$time, leuk$status)
leuk.fit = survfit(leuk.surv ~ 1, se.fit = T,
                  type = "kaplan-meier",
                  error = "greenwood",
                  conf.type = "plain")

leuk.fit
```

```
 n  event  mean  se(mean)  median  LCL  UCL
21  9  23.3  2.83  23    13  NA
```
• Estimation of Quantiles

Point Estimate

\[ \hat{t}_p = \inf \{ t : \hat{S}(t) \leq 1 - p \} \]

Confidence Intervals

Let \( X \) be the totality of the observations and \( C(X) \) a random set such that

\[ P_\theta(\theta \in C(X)) = 1 - \alpha. \]

Then \( C(X) \) is called a 100(1 − \( \alpha \))% confidence set of \( \theta \).

Let \( \delta(X, \theta) \) be an indicator function of a random set in the space \( X \times \Theta \). If for any \( \theta \),

\[ P_\theta(\delta(X, \theta) = 1) = 1 - \alpha, \]

then the set

\[ C(X) = \{ \tilde{\theta} : \delta(X, \tilde{\theta}) = 1 \} \]

is a 100(1 − \( \alpha \))% confidence set of \( \theta \).
Confidence sets $C(X)$ of $t_p$ based on CIs of $S(t)$

(i) Based on linear CI of $S(t)$:

$$
\left\{ t : \left| \frac{\hat{S}(t)-(1-p)}{\hat{V}^{1/2}(\hat{S}(t))} \right| \leq z_{1-\alpha/2} \right\}
$$

(ii) Based on log-log transformed CI of $S(t)$:

$$
\left\{ t : \left| \frac{\ln(-\ln \hat{S}(t))-\ln(-\ln(1-p))}{\hat{V}^{1/2}(\hat{S}(t))/[\hat{S}(t) \ln \hat{S}(t)]} \right| \leq z_{1-\alpha/2} \right\}
$$

(iii) Based on arcsin square root transformed CI of $S(t)$:

$$
\left\{ t : \left| \frac{2[\arcsin(\sqrt{\hat{S}(t)})-\arcsin(\sqrt{1-p})]}{\hat{V}^{1/2}(\hat{S}(t))/[\hat{S}(t)(1-\hat{S}(t)]^{1/2}} \right| \leq z_{1-\alpha/2} \right\}
$$
R function for computing quantile confidence intervals

```r
quantile.ci=function(p,alpha,survfit.obj, conf.type=c("plain","log-log","arcsin","log")){
time=summary(survfit.obj)$time
surv=summary(survfit.obj)$surv
std.err=summary(survfit.obj)$std.err
z.alpha=qnorm(1-alpha/2)
    n.t=length(time)
q=1-p
if(conf.type=="plain") {
pivot=(surv-q)/std.err
} else if (conf.type=="log-log") {
    log.surv=log(surv)
pivot=( log(-log.surv) - log(-log(q)))*surv*log.surv/std.err
} else if (conf.type=="arcsin") {
    pivot=2*( asin(sqrt(surv)) - asin(sqrt(q)) )*sqrt(surv*(1-surv))/std.err
} else {
    pivot=(log(surv)-log(q))*surv/std.err
}
    ci.l_0
if (pivot[1]>z.alpha) {
    for (i in 2:n.t){

```
if (pivot[i]<=z.alpha) {
    ci.l_time[i]
    break
} else {
    next
}
}
} else {
    ci.l=ci.l
}
}
}
ci.u=10^10
if (pivot[n.t] < -z.alpha) {
    for (i in 2:n.t){
        if (pivot[n.t-i+1]>=-z.alpha) {
            ci.u=time[n.t-i+1]
            break
        } else {
            next
        }
    }
} else {
    ci.u=ci.u
}
list(conf.int=c(ci.l,ci.u),
     summary=data.frame(time,surv, std.err, pivot))
}
§3.5. Estimators of the survival function for left-truncated and right-censored data

• Data structure

In the left-truncated and right censored data, each individual is associated with a pair of observations:

$L_i$: left-truncation time, usually the age at entry of study.

$T_i$: lifetime or censoring time.

• Estimation

Still let $t_i$ denote distinct lifetimes:

$0 < t_1 < \cdots < t_D < \infty$.

Intuitive estimation of $h(t_i)$:

$$
\frac{\text{# of deaths at } t_i}{\text{# of individuals at risk at } t_i}.
$$
How to count the number of individuals at risk at $t_i$?

In order for a sample from a sub-population to be useful for estimating $h(t_i)$, its individuals must be subjected to risk at $t_i$. When $L_j > t_i$, none individual from the truncated population can be subjected to risk at $t_i$. Therefore an individual with $L_j > t_i$ does not provide information on $h(t_i)$. The individuals who provide information on $h(t_i)$ are those with $L_j < t_i \leq T_j$. Thus, in the presence of left truncation, $Y_i$ must be modified as:

$$Y_i = \# \text{ of individuals with } L_j < t_i \leq T_j.$$

Then all the estimators are obtained the same as in the right censored case with the modified definition of $Y_i$. 
Estimates and interpretation

\[ \hat{S}(t) = \prod_{t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right], \text{ PL} \]

\[ \hat{H}(t) = \prod_{t_i \leq t} \frac{d_i}{Y_i}, \text{ Nelson-Aalen.} \]

The estimates are interpreted as truncated quantities. Let \( L \) be the smallest left-truncation time. Then \( \hat{S}(t) \) is an estimate of \( S(t)/S(L) \) and \( \hat{H}(t) \) is an estimate of \( \int_L^t h(u)du \).

There could be the case that, at some early time, \( Y_i = d_i \), which causes that \( \hat{S}(t) = 0 \) beyond that time. The remedy for this is given as follows.
Find a constant $a$ such that, at all $t_i \geq a$, $Y_i > d_i$ except at $t_D$, then estimate $S(t)/S(a)$ by

$$\hat{S}(t) = \prod_{a \leq t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right].$$

§3.6. Summary curves for competing risks

Let $t_i$ be distinct time at which any of the competing risks occurs:

$$0 < t_1 < \cdots < t_D < \infty.$$ 

$Y_i$: # of individuals at risk at $t_i$.

$r_i$: # of events due to a particular competing risk.

$d_i$: # of events due to all other competing risks.

- Estimation of Net probabilities
For the estimation of net probabilities, treat all the competing risks but the particular one as censoring variables and then compute the PL estimate $\hat{S}_r(t)$.

$$\hat{S}_r(t) = \prod_{t_i \leq t} \left[ 1 - \frac{r_i}{Y_i} \right].$$

The net cumulative distribution function is then estimated by $1 - \hat{S}_r(t)$.

- **Estimation of cumulative incidence function**

Cumulative incidence function:

$$F_r(t) = P(T \leq t, \delta = r) = \int_0^t h_i(u)e^{-HT(u)}du.$$  

The estimate:

If $t < t_1$,  
$$\hat{F}_r(t) = 0.$$
if \( t \geq t_1 \),

\[
\hat{F}_r(t) = \sum_{t_i \leq t} \left[ \prod_{j=1}^{i-1} \left( 1 - \frac{r_j + d_j}{Y_j} \right) \right] \frac{r_i}{Y_i}
\]

\[
= \sum_{t_i \leq t} \hat{S}(t_i^-) \frac{r_i}{Y_i}
\]

\[
= \frac{r_1}{Y_1} + \left[ 1 - \frac{r_1 + d_1}{Y_1} \right] \frac{r_2}{Y_2} + \cdots
\]

A heuristic motivation of the estimate:

Let \( t_j \) be the largest \( t_i \) less than or equal to \( t \). Treat \( t_i \) as if they are fixed time points. Then

\[
P(T \leq t, \delta = r) = \sum_{i=1}^{J} P(t_{i-1} < T \leq t_i, \delta = r)
\]

\[
= \sum_{i=1}^{J} P(t_{i-1} < T \leq t_i, \delta = r | T > t_{i-1}) P(T > t_{i-1}).
\]

It is intuitively reasonable to estimate the first term by \( r_i/Y_i \) and the second term by \( \hat{S}(t_i^-) \).
Let $\tilde{d}_j = r_j + d_j$. It is easy to check that

$$1 - \sum_r \hat{F}_r(t)$$

$$= 1 - \frac{\tilde{d}_1}{Y_1} - (1 - \frac{\tilde{d}_1}{Y_1}) \frac{\tilde{d}_2}{Y_2} - \cdots - (1 - \frac{\tilde{d}_1}{Y_1}) \cdots (1 - \frac{\tilde{d}_{J-1}}{Y_{J-1}}) \frac{\tilde{d}_J}{Y_J}$$

$$= (1 - \frac{\tilde{d}_1}{Y_1}) \cdots (1 - \frac{\tilde{d}_{J-1}}{Y_{J-1}})(1 - \frac{\tilde{d}_J}{Y_J})$$

$$= \hat{S}_T(t).$$

The variance and CI:

$$\hat{V}(\hat{F}_r(t))$$

$$= \sum_{t_i \leq t} \hat{S}^2(t_i) \left\{ \left[ \hat{F}_r(t) - \hat{F}_r(t_i) \right]^2 \frac{r_i + d_i}{Y_i^2} \right. + \left. \left[ 1 - 2(\hat{F}_r(t) - \hat{F}_r(t_i)) \right] \frac{r_i}{Y_i^2} \right\}$$

$$\hat{F}_r(t) \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{F}_r(t))}$$

21
- Conditional probability:

$F_r(t)$: Cumulative incidence function for risk $r$,
$F_{rc}(t)$: Cumulative incidence function for all other risks lumped together.

The conditional probability of risk $r$ occurring by $t$ given that none of the other causes have occurred by $t$ is given by

$$CP_r(t) = \frac{F_r(t)}{1 - F_{rc}(t)}.$$

The estimate:

$$\hat{CP}_r(t) = \frac{\hat{F}_r(t)}{1 - \hat{F}_{rc}(t)}.$$

The estimated variance:

$$\hat{V}(\hat{CP}_r(t)) = \frac{\hat{S}^2(t-)}{[1 - \hat{F}_{rc}(t)]^4} \times \sum_{t_i \leq t} \frac{[1 - \hat{F}_{rc}(t_i)]^2r_i + \hat{F}_r^2(t_i)d_i}{Y_i^2}.$$
§3.7. Estimators of the survival function with left, double and interval censored data

• Some ideas behind estimators

Left-censoring only

If all the censored times are of the left censoring type, the data can be analyzed by the method for right censoring data. Suppose $T_1, \ldots, T_n$ are the observations, each $T_i$ is either a lifetime or a left-censored time. Let $a$ be a large constant. Consider the transformed data $\tilde{T}_i = a - T_i, i = 1, \ldots, n$. The transformed data become right censored data. After estimates are obtained with the transformed data, they can be transformed back to the original data.
Double and interval censored case.

The case refers to the situation that among observed data, all right-, left- and interval-censored times present.

The left censored can be considered as interval censored and can be treated the same way as interval censored data.

The idea behind the treatment of interval censored data is to estimate the number of deaths and number of individuals at risk at observed times and then used the estimated numbers to construct the PL or other estimators. This requires an iterative procedure. Starting from an initial estimate of the survival function, the numbers can be estimated and a new estimate of the survival function is obtained. The new estimated survival function is then used as the initial survival function, and the process goes
on until the estimated survival function converges.

- **R function for the analysis of left, double and interval censored data.**

Interval censored (including left and right censored) data object can be created by the R function `Surv` as follows:

```
Surv(time, time2, event, type=)
```

**time**: for right censored data, this is the follow up time. For interval data, the first argument is the starting time for the interval.

**event**: The status indicator, normally 0=alive, 1=dead. Other choices are T/F (TRUE = death) or 1/2 (2=death). For interval censored data, the status indicator is 0=right censored, 1= event at 'time', 2=left censored, 3=interval censored. Although unusual, the event indicator can be omitted, in which case all subjects are assumed to have an event.

**time2**: ending time of the interval for interval censored or counting process data only. Intervals are assumed to be open on the left and closed on the right, ’(start, end]’. For counting process
data, ‘event’ indicates whether an event occurred at the end of the interval.

**type:** character string specifying the type of censoring. Possible values are ”right”, ”left”, ”counting”, ”interval”, or ”interval2”. The default is ”right” or ”counting” depending on whether the ’time2’ argument is absent or present, respectively.