Stat 486 / 886 Survival Analysis

Chapter 2. Nonparametric Estimation and Graphical Methods

Professor: Wenyu Jiang

Department of Mathematics and Statistics
Queen's University

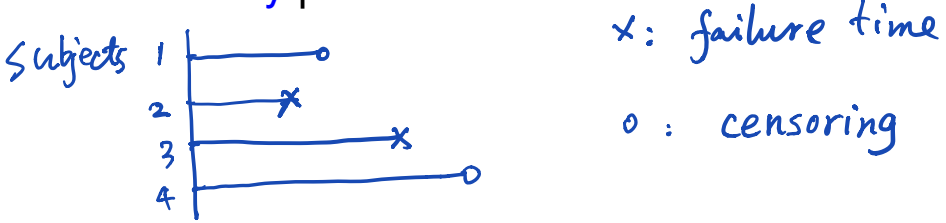
Acknowledgement: Profs. Jerry Lawless, David Matthews (Waterloo),
Profs. Paul Peng, Dongsheng Tu (Queen's)

Sections of Chapter 2

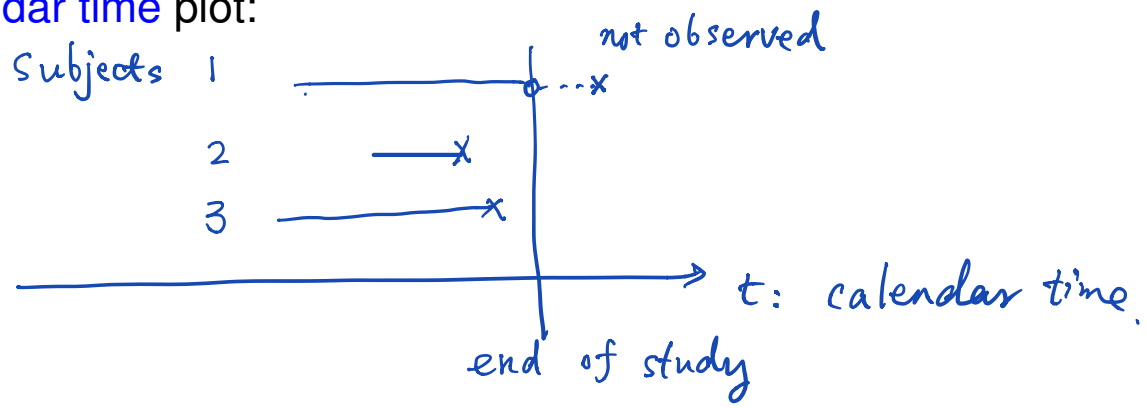
2.1 Plots of Univariate Survival Data

- Common plots such as histogram, boxplots may not be directly useful for survival data, because of censoring.
- Simple display of survival data

Time on study plot:



Calendar time plot:

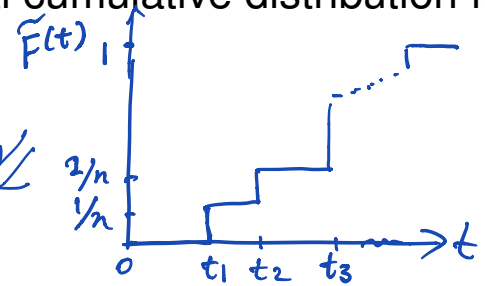


2.2 Kaplan-Meier Estimate of Survival Function

By Kaplan and Meier (Journal of American Statistical Association, 1958); also called the product-limit estimate.

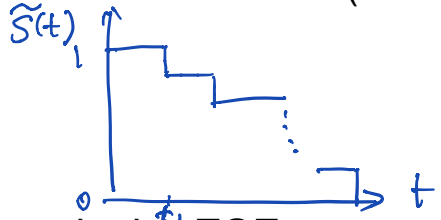
- A nonparametric estimate of $S(t)$ or $H(t)$.
- If there are no censoring, the empirical cumulative distribution function (ECDF) has the form

$$\tilde{F}(t) = \frac{\#\text{ of } t_i\text{'s} \le t}{n}$$



for failure times $t_1, \ldots, t_n$; the empirical survival function (ESF) is

$$\tilde{S}(t) = \frac{\#\text{ of } t_i\text{'s} \ge t}{n}.$$



- Kaplan-Meier (KM) estimate of $S(t)$ extends the ESF to censored survival data.

KM estimate for SF: a nonparametric method

Consider right censored data only, and
assume failure time T has a discrete distribution.

Assume that T take values $a_1, \ldots, a_K$.

Review 1.3.3

$f(t) = P(T=t)$

$h(t) = \frac{f(t)}{S(t)}$,

$S(t) = \prod_{j: a_j < t} [1 - h(a_j)]$

For subjects $i = 1, \ldots, n$, denote the observed data by (x_i, δ_i).
Denote $h(a_k)$, the hazard function at time a_k, by h_k.
The KM estimate is derived by taking $\boldsymbol{h} = (h_1, \ldots, h_K)^\mathsf{T}$ as the parameters.

The (nonparametric) likelihood function can be expressed by

$$L(\boldsymbol{h}) = \prod_{j=1}^{K} h_j^{d_j} (1 - h_j)^{n_j - d_j}, \tag{1}$$

where $d_j = \sum_{i=1}^n I(x_i = a_j, \delta_i = 1)$, : # of failures at time a_j.

$n_j = \sum_{i=1}^n I(x_i \geq a_j)$. # of subjects at risk of failure at a_j.

alive

Explanation of (1):

$a_1 < a_2 < \cdots < a_k$

are the distinct failure time. $(x_1, \delta_1 = 0)$

$\mathcal{D} = \{(x_i, \delta_i), i = 1, \cdots, n\}$

index of subjects

Likelihood fct'n:

$$L(\underset{\sim}{h}) = \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[h(x_i) S(x_i) \right]^{\delta_i} S(x_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} h(x_i)^{\delta_i} S(x_i).$$

$$= \prod_{i=1}^{n} \left\{ h(x_i)^{\delta_i} \prod_{j: a_j < x_i} \left[1 - h(a_j) \right] \right\}.$$

(i : index of subjects
j : index of time.)

$$= \prod_{j=1}^{K} \left[h(a_j) \right]^{d_j} \left[1 - h(a_j) \right]^{n_j - d_j}$$

$$= \prod_{j=1}^{K} h_j^{d_j} \left[1 - h_j \right]^{n_j - d_j}$$

About risk set

- Risk set at a_j:
  collection of subjects that are at risk to fail at time a_j.
  $n_j = \sum_{i=1}^{n} I(x_i \geq a_j)$: # of subjects in the risk set.
- Conventions for risk set specification at censoring/failure times:

Example: $(x_i = 10, \delta_i = 0)$ $(x_\ell = 10, \delta_\ell = 1)$.

i). The risk set at time 10 includes both subjects i and ℓ.

ii) At tied failure & censoring time: assume censoring occurs (infitesimally) after failure (at $10+$).

From the likelihood (1), we can derive the mle's of h_j,

$$\hat{h}_j = \frac{d_j}{n_j}, \text{ for } j = 1, \ldots, K.$$

Recall

$h_j = h(a_j)$

$= P(T = a_j \mid T \geq a_j)$

Log-likelihood:

$$\ell(\underset{\sim}{h}) = \sum_{j=1}^{K} \left[d_j \log h_j + (n_j - d_j) \log(1 - h_j) \right]$$

For each $j = 1, \cdots, K$.

$$\frac{\partial \ell}{\partial h_j} = \frac{d_j}{h_j} - \frac{n_j - d_j}{1 - h_j}$$

Solve $\frac{\partial \ell}{\partial h_j} = 0$. $\quad \frac{d_j}{h_j} = \frac{n_j - d_j}{1 - h_j}$

$\Rightarrow$ The mle of h_j is $\hat{h}_j = \frac{d_j}{n_j}$.

The KM estimate of the survival function is then obtained by

$$\hat{S}(t) = \prod_{j:a_j<t} \left(1 - \frac{d_j}{n_j}\right), \qquad (2)$$

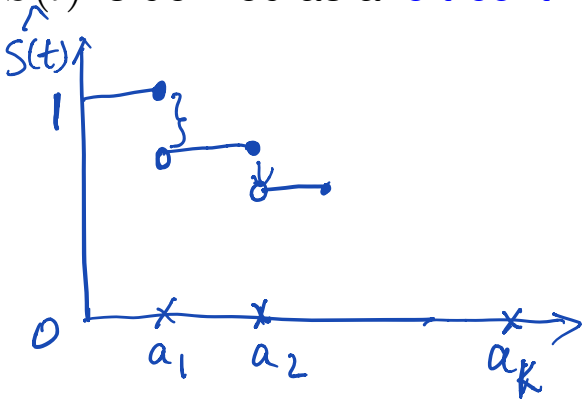
$\hat{h}_j$

or equivalently, by the iterative formula

$$\hat{S}(a_j) = \hat{S}(a_{j-1}) \left(1 - \frac{d_{j-1}}{n_{j-1}}\right).$$

Convention:

$\hat{S}(t)$ is defined as a left-continuous function.



$\hat{S}(0) = 1.$ $\hat{S}(a_1) = 1;$

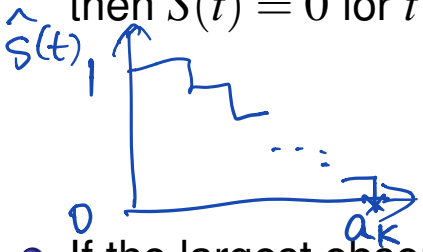
$\hat{S}(a_1+) = 1 - \frac{d_1}{n_1}$, $\hat{S}(a_2) = 1 - \frac{d_1}{n_1}$

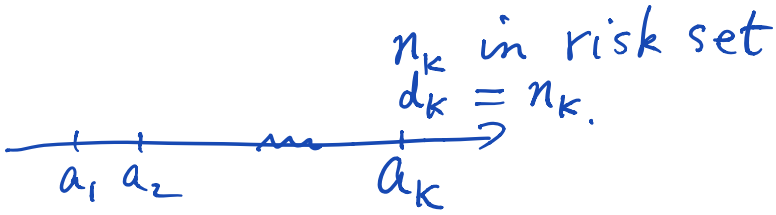
$\hat{S}(a_2+) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)$

. . . . . .

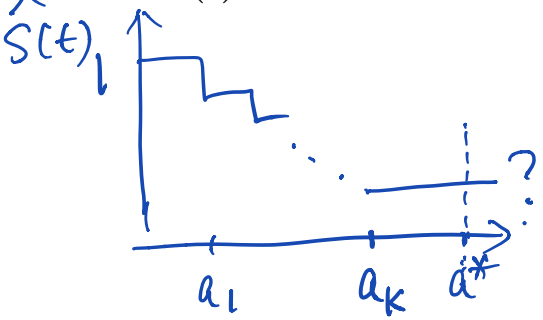
Remarks:

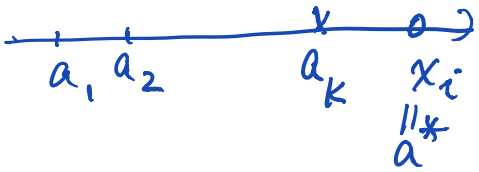
- If the largest observed time x_i is a failure time (it is a_K), then $\hat{S}(t) = 0$ for $t > a_K$.





- If the largest observed time x_i is a censoring time (call it a^*), then $\hat{S}(a^*) > 0$ since $\hat{S}(a_K+) > 0$, and $\hat{S}(t)$ is undefined for $t > a^*$.





Example 2.2.1 Equipment Field Failure

See Example 1.2.1 for details.

The data below are the number of days until the first malfunction for 10 equipment units. Find the KM estimate for the survival function.

2, 4, 14, 21*, 24, 27, 33, 51, 60*, 72*

Calculation based on the formula, (2):

Failure Times.	j	a_j	n_j	d_j
	1	2	10	1
	2	4	9	1
1	3	14	8	1
censoring →	4	24	6	1

$\hat{S}(t)$ at $t = a_j +$

$1 - \frac{d_1}{n_1} = 1 - \frac{1}{10} = 0.9$

$(1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2}) = 0.9(1 - \frac{1}{9}) = 0.8$

$0.8(1 - \frac{1}{8}) = 0.7$

$0.7(1 - \frac{1}{6}) = 0.583.$

5	27	5	1
6	33	4	1
7	51	3	1

$0.583\left(1-\frac{1}{5}\right)=0.466$

$0.466\left(1-\frac{1}{4}\right)=0.350$

$0.350\left(1-\frac{1}{3}\right)=0.233$

KM estimation can be obtained in R by the "survfit" function, or in SAS by the "lifetest" procedure.

Variance estimation for the KM Estimate of $S(t)$

The variance estimate, denoted by $\widehat{\text{Var}}[\hat{S}(t)]$, is needed for statistical inference about $S(t)$ at given time t.

The variance of the KM estimate $\hat{S}(t)$ of the form (2) can be estimated by the Greenwood's formula,

$$\widehat{\text{Var}}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)}. \tag{3}$$

The following theorem is needed for deriving the Greenwood's formula.

Theorem: Δ - Method

For a random variable Z (in $\mathbb{R}$) with finite variance, and a continuous function $g(z)$ (from $\mathbb{R} \to \mathbb{R}$), the variance of $g(Z)$ is approximately expressed as follows,

$$\mathrm{Var}[g(Z)] \approx [g'(\mu_Z)]^2 \mathrm{Var}(Z),$$

where $g'(z) = \frac{dg}{dz}$ and $\mu_Z = \mathrm{E}(Z)$.

Taylor Series:

$$g(z) = g(\mu_z) + g'(\mu_z)(z - \mu_z) + g''(\mu_z)\frac{(z-\mu_z)^2}{2} + \cdots$$

$g(\mu_z)$: constant; $g'(\mu_z)(z-\mu_z)$: constant; $g''(\mu_z)\frac{(z-\mu_z)^2}{2} + \cdots$: Ignore higher order terms.

$$\mathrm{Var}[g(z)] \approx [g'(\mu_z)]^2 \mathrm{Var}(z).$$

Derivations of the Greenwood's formula (3) in 3 steps.

Step 1. Variance estimation for mle's of $\boldsymbol{h} = (h_1, \ldots, h_K)$

For the mle's $\hat{\boldsymbol{h}} = (\hat{h}_1, \ldots, \hat{h}_K)$, show that

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{h}}) = [I(\hat{\boldsymbol{h}})]^{-1} = \mathrm{diag}\left[\frac{d_j(n_j - d_j)}{n_j^3}\right].$$

Recall $L(\underset{\sim}{h}) = \prod_{j=1}^{K} h_j^{d_j}(1-h_j)^{n_j-d_j}$

Log-likelihood: $\ell(\underset{\sim}{h}) = \sum_{j=1}^{K}\left[d_j \log h_j + (n_j - d_j)\log(1-h_j)\right]$

Information matrix: $I(\underset{\sim}{h}) = -\frac{\partial^2 \ell(\underset{\sim}{h})}{\partial \underset{\sim}{h}\, \partial \underset{\sim}{h}^T}$. $K \times K$ matrix.

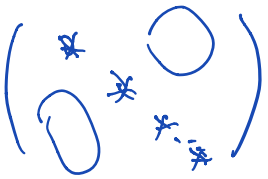
For mle, $\widehat{Var}(\hat{\underset{\sim}{h}}) = \left[I(\hat{\underset{\sim}{h}})\right]^{-1}$;

where $I(\hat{\underset{\sim}{h}})$: observed information matrix.

$$\frac{\partial \ell}{\partial h_j} = \frac{d_j}{h_j} - \frac{n_j - d_j}{1 - h_j}$$

$$\frac{\partial^2 \ell}{\partial h_j^2} = -\frac{d_j}{h_j^2} - \frac{n_j - d_j}{(1-h_j)^2}, \quad j = 1, \cdots, K.$$

$$\frac{\partial^2 \ell}{\partial h_j \partial h_\ell} = 0. \qquad j \neq \ell.$$



$I(\underset{\sim}{h})$: diagonal matrix

$$I(\underset{\sim}{h}) = \operatorname{diag}\left\{ \frac{d_j}{h_j^2} + \frac{n_j - d_j}{(1-h_j)^2} \right\}, \quad j = 1, \cdots, K.$$

with $\hat{h}_j = \frac{d_j}{n_j}$,

$$\frac{d_j}{\hat{h}_j^2} = \frac{n_j^2}{d_j}, \quad \frac{n_j - d_j}{(1-\hat{h}_j)^2} = \frac{n_j^2}{(n_j - d_j)}$$

$$I(\hat{\underset{\sim}{h}}) = \operatorname{diag}\left\{ \frac{n_j^2}{d_j} + \frac{n_j^2}{n_j - d_j} \right\}$$

$$= \operatorname{diag}\left\{ \frac{n_j^3}{d_j (n_j - d_j)} \right\}$$

$$\widehat{\operatorname{var}}(\hat{\underset{\sim}{h}}) = I^{-1}(\hat{\underset{\sim}{h}}) = \operatorname{diag}\left\{ \frac{d_j (n_j - d_j)}{n_j^3} \right\}$$

$j = 1, 2, \cdots, K.$

Step 2. Δ-method to estimate $\mathrm{Var}[\log \hat{S}(t)]$

Show that

$$\widehat{\mathrm{Var}}[\log \hat{S}(t)] = \sum_{j:a_j<t} \frac{d_j}{n_j(n_j-d_j)}.$$

$$\log \hat{S}(t) = \sum_{j:a_j<t} \log(1-\hat{h}_j).$$

[Know $\mathrm{Var}(Z)$. What's $\mathrm{Var}[\underbrace{\log Z}_{g(Z)}]$?

$g'(z) = \frac{1}{z}$.

Δ-method: $\mathrm{Var}[\log Z] \approx \left[\frac{1}{\underbrace{\mu_Z}_{\text{If unknown, just use } Z.}}\right]^2 \mathrm{Var}(Z) \approx \frac{1}{Z^2}\mathrm{Var}(Z)$.]

$$\widehat{\mathrm{Var}}\left[\log(1-\hat{h}_j)\right] = \frac{1}{(1-\hat{h}_j)^2}\,\underbrace{\widehat{\mathrm{Var}}(1-\hat{h}_j)}_{\widehat{\mathrm{Var}}(\hat{h}_j)}.$$

$\left(\hat{h}_j = \frac{d_j}{n_j}\right)$.

$$= \left(\frac{n_j}{n_j - d_j}\right)^2 \cdot \frac{d_j(n_j-d_j)}{n_j^3} = \frac{d_j}{n_j(n_j-d_j)}$$

$$\widehat{\mathrm{Var}}\left[\log \hat{S}(t)\right] = \sum_{j:a_j<t} \widehat{\mathrm{Var}}\left[\log(1-\hat{h}_j)\right]$$

$$= \sum_{j:a_j<t} \frac{d_j}{n_j(n_j-d_j)}$$

Step 3. Δ-method, relate $\mathrm{Var}[\log \hat{S}(t)]$ to $\mathrm{Var}[\hat{S}(t)]$

Obtain the Greenwoods formula (3):

$$\mathrm{Var}[\log \hat{S}(t)] \approx \frac{1}{[\hat{S}(t)]^2} \mathrm{Var}[\hat{S}(t)], \quad \Delta\text{-method}.$$

$$\mathrm{Var}[\hat{S}(t)] = [\hat{S}(t)]^2 \mathrm{Var}[\log \hat{S}(t)].$$

$$\widehat{\mathrm{Var}}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{j: a_j < t} \frac{d_j}{n_j (n_j - d_j)}$$

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

Greenwood's formula.

Example 2.2.1 Equipment Field Failure, continued.

See Example 1.2.1 for details.

Find the variance estimates for the KM estimate $\hat{S}(t)$.

j	a_j	n_j	d_j	$\hat{S}(t)$ at $t = a_j+$	$\widehat{Var}[\hat{S}(t)]$
1	2	10	1	$1-\frac{d_1}{n_1} = 0.9$.	$0.9^2\left[\frac{1}{10(10-1)}\right] = 0.0009$
2	4	9	1	$0.9\left(1-\frac{1}{9}\right) = 0.8$;	$0.8^2\left[\frac{1}{10(10-1)}+\frac{1}{9(9-1)}\right] =$
3	14	8	1		
⋮	⋮	⋮	⋮		⋮
7	51	⋮	⋮		

Confidence intervals (C.I.) for $S(t)$

Method 1. Plain C.I. from the Greenwood's formula.

Since KM estimates are based on the maximum likelihood method, by properties of mle's,

$$\hat{S}(t) \approx \mathrm{N}\left(S(t), \mathrm{Var}[\hat{S}(t)]\right).$$

At given time t.

The approximate $100(1-\alpha)\%$ plain C.I. for $S(t)$ is given by

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \left\{\widehat{\mathrm{Var}}[\hat{S}(t)]\right\}^{\frac{1}{2}},$$

s.e. of $\hat{S}(t)$.

where $\hat{S}(t)$ is the KM estimate of $S(t)$, $\widehat{\mathrm{Var}}[\hat{S}(t)]$ is from the Greenwood's formula, and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ (lower) quantile of $\mathrm{N}(0,1)$.

Note: Plain C.I. of $S(t)$ may fall outside $[0,1]$.

Confidence intervals (C.I.) for $S(t)$

Method 2. Construct C.I. based on a transformation $\psi = g[S(t)]$.
Idea: Choose a transformation function $g()$, find C.I. for $\psi = g[S(t)]$, transform back to C.I. for $S(t)$.

Common choices of $g()$: $\quad g(u) = \log u, \quad g(u) = \log(-\log u)$.
These lead to log based C.I. and log(-log) based C.I., for $S(t)$ respectively.
$g(u)$: an increasing, or decreasing fct'n, differentiable in u.
Sketches of building log(-log) based C.I., for $S(t)$

Take $\psi = g[S(t)] = \log[-\log S(t)]$, then $\hat{\psi} = \log[-\log \hat{S}(t)]$.

i) Estimate $\mathrm{Var}(\hat{\psi})$.
It can be shown that
From $\widehat{\mathrm{Var}}(\hat{S}(t))$, and Δ-method.

$$\widehat{\mathrm{Var}}(\hat{\psi}) = \widehat{\mathrm{Var}}\{\log[-\log \hat{S}(t)]\} = \frac{\sum_{j:a_j<t} \frac{d_j}{n_j(n_j-d_j)}}{\left[\sum_{j:a_j<t} \log\left(1-\frac{d_j}{n_j}\right)\right]^2}.$$

ii) Based on the result that $Z = \frac{\hat{\psi}-\psi}{[\widehat{\text{Var}}(\hat{\psi})]^{\frac{1}{2}}} \approx \text{N}(0,1)$, build a $100(1-\alpha)\%$ C.I. for ψ of the form

$$\hat{\psi} \pm z_{1-\frac{\alpha}{2}} [\widehat{\text{Var}}(\hat{\psi})]^{\frac{1}{2}}. = (\hat{\psi}_L, \hat{\psi}_u) \qquad (4)$$

iii) Apply $g^{-1}()$, the inverse function of $g()$, to the interval (4), and find the C.I. for $S(t)$ of the form $\hat{S}_L(t) < S(t) < \hat{S}_U(t)$.

$$p\left(\hat{\psi}_L \le \underbrace{\log(-\log S(t))}_{\psi} \le \hat{\psi}_u\right) = 1-\alpha.$$

$g^{-1}()$:

$$p\left(\underbrace{\hat{S}_L(t) \le S(t) \le \hat{S}_u(t)}_{\text{Log-Log based CI for } S(t).}\right) = 1-\alpha.$$

Remarks

$\psi = \log(-\log u)$

$u = e^{-e^{\psi}}$

- The plain C.I.,log(-log) based C.I. and the log based C.I. are the three most commonly used types of C.I. for $S(t)$ based on the KM estimation.
- The log(-log) based C.I. takes values in $[0,1]$ only. Why?

  $\log(-\log u) \in (-\infty, \infty)$, $e^{-e^{\psi}} \in$?
- The default C.I. type in "survfit" function in R is the log based.

- All the C.I. in this section are pointwise C.I., that is, they are for $S(t)$ at a given time t.

  At given time t, $P\left(\hat{S}_L(t) < S(t) < \hat{S}_U(t)\right) = 1-\alpha$.

  A more advanced inference procedure is to build a simultaneous C.I. for $S(t)$. Very challenging research topic.

  $P\left(\hat{S}_L(t) < S(t) < \hat{S}_U(t), \text{ for all } t \in [0, \tau]\right) = 1-\alpha$.

Kaplan-Meir estimation using R

Example 2.2.1 Equipment Field Failure Continued.

Data for days until the first malfunction for 10 equipment units:

2, 4, 14, 21*, 24, 27, 33, 51, 60*, 72*

Find the KM estimate for $S(t)$ using R.

Create a data file "eg121.txt" in the following format.

(x_i, δ_i)

```
time status
2 1
4 1
14 1
21 0
...
72 0
```

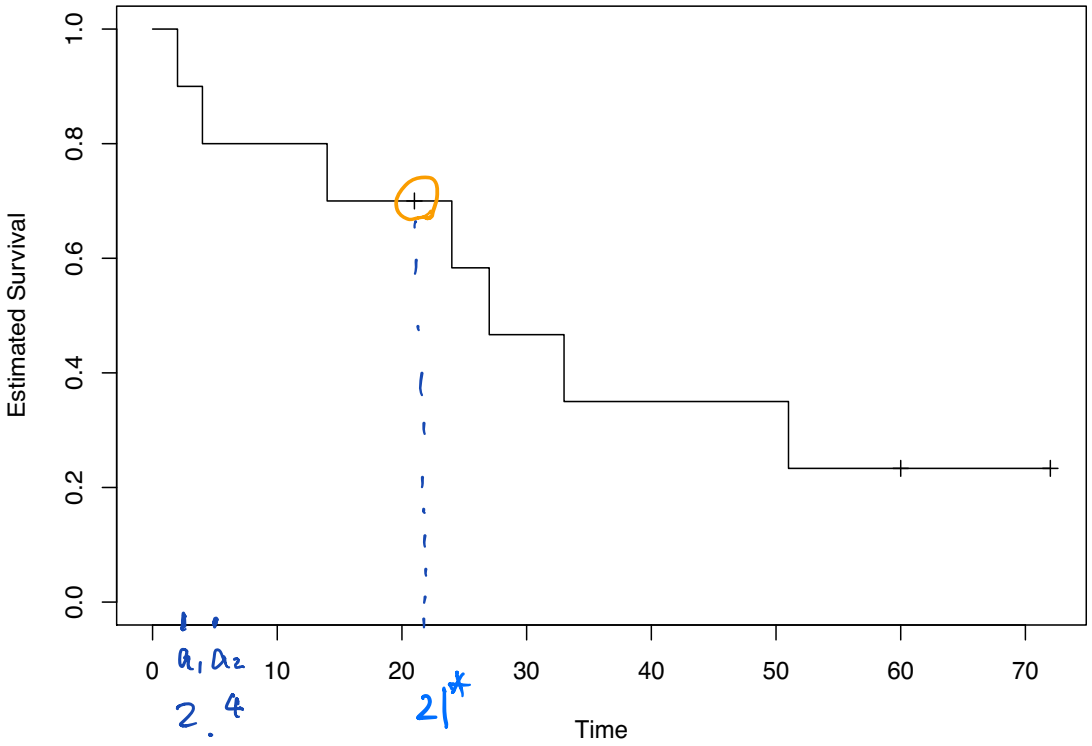


Figure: Kaplan-Meier (Product-Limit) estimate of the survival function for the time to equipment failure.

Example 2.2.1 Continued. R Listing:

KM estimates and plots

```
> library(survival)
> data<-read.table("eg121.txt", header=T)
> x<-data$time
> delta<-data$status
> x
 [1]  2  4 14 21 24 27 33 51 60 72
> delta
 [1] 1 1 1 0 1 1 1 1 0 0
> fit<-survfit(Surv(x, delta)~1)
> plot(fit, xlab = "Time", ylab = "Estimated
Survival", mark.time=T, conf.int=F)
```

= T for CI in the plot

Output: KM estimates... standard errors, confidence intervals...

default CI = log-based CI.

```
> summary(fit)
Call: survfit(formula = Surv(x, delta) ~ 1)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     10       1    0.900  0.0949        0.7320        1.000
    4      9       1    0.800  0.1265        0.5868        1.000
   14      8       1    0.700  0.1449        0.4665        1.000
   24      6       1    0.583  0.1610        0.3396        1.000
   27      5       1    0.467  0.1658        0.2326        0.936
   33      4       1    0.350  0.1602        0.1427        0.858
   51      3       1    0.233  0.1431        0.0701        0.776
>
> # "std.err": square root of variance estimate (Greenwood's formula).
>
> fit$conf.type  # The default type of confidence intervals.
[1] "log"
```

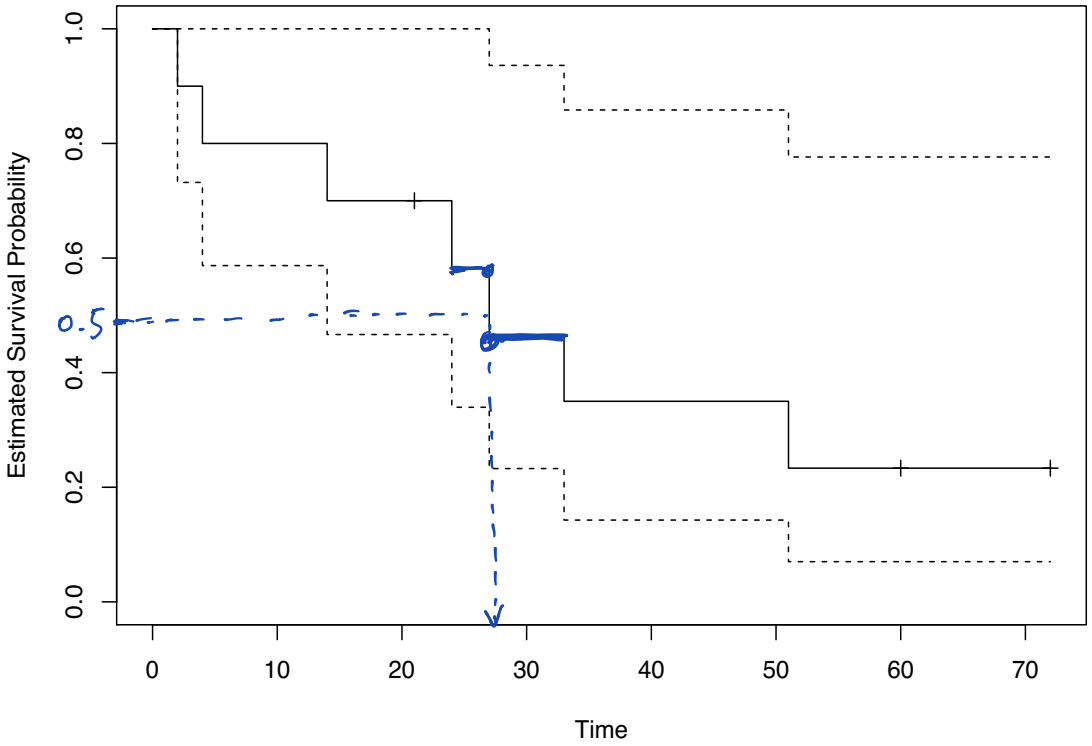


Figure: Kaplan-Meier estimate of the survival function for the time to equipment failure with pointwise 95% log-based (the default) confidence intervals.

Different types of confidence interval for $S(t)$

"log" for log-based CI

```
> fit.plain<-survfit(Surv(x, delta)~1, conf.type="plain")
> summary(fit.plain)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     10       1    0.900  0.0949        0.714        1.000
    4      9       1    0.800  0.1265        0.552        1.000
    ... ...
   33      4       1    0.350  0.1602        0.036        0.664
   51      3       1    0.233  0.1431        0.000        0.514
> fit.loglog<-survfit(Surv(x, delta)~1, conf.type="log-log")
> summary(fit.loglog)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     10       1    0.900  0.0949       0.4730        0.985
    4      9       1    0.800  0.1265       0.4087        0.946
    ... ...
   33      4       1    0.350  0.1602       0.0848        0.640
   51      3       1    0.233  0.1431       0.0360        0.529
```

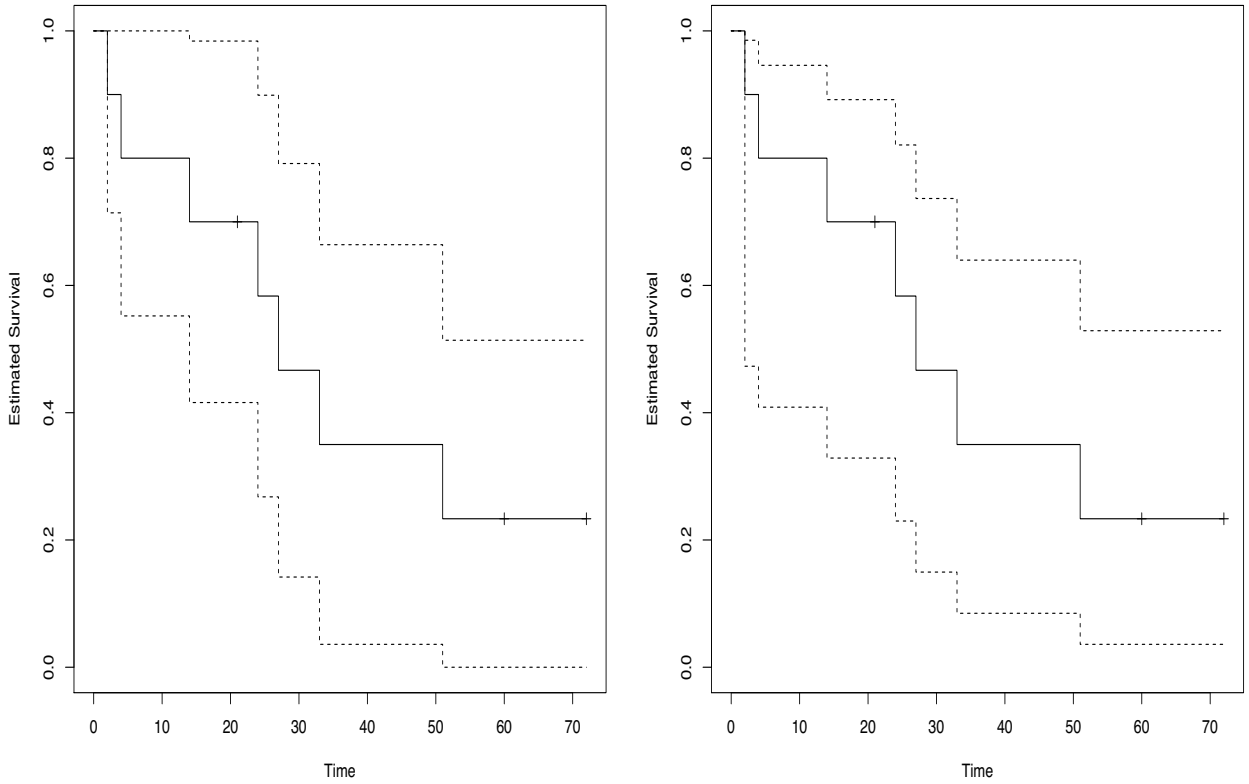


Figure: Pointwise 95% confidence intervals for $S(t)$ for the equipment field failure data. Plain (left panel) and log(-log)-based (right) confidence intervals.

```
> par(mfrow=c(1,2))
> plot(fit.plain, xlab = "Time", mark.time=T,
ylab = "Estimated Survival", conf.int=T)
> plot(fit.loglog, xlab = "Time", mark.time=T,
ylab = "Estimated Survival", conf.int=T)
```

2.3 Nelson-Aalen Estimate of the Cumulative Hazard Function $H(t)$

Now assume that T has a continuous distribution, recall that

$$H(t) = \int_0^t h(u)du = -\log S(t), \quad S(t) = e^{-H(t)}.$$

Nelson-Aalen (NA) estimate of $H(t)$ is obtained by approximating the integral by a summation,

$$\hat{H}_{NA}(t) = \sum_{j:a_j<t} \hat{h}_j = \sum_{j:a_j<t} \frac{d_j}{n_j},$$

where $a_1 < a_2 < \ldots < a_K$ are the distinct failure times.

NA estimate of the survival function,

$$\hat{S}_{NA}(t) = e^{-\hat{H}_{NA}(t)}.$$

Example 2.3.1 Advanced non-Hodgkin's Lymphoma

The data below give times (in months) from diagnosis to death for 31 individuals with advanced non-Hodgkin's lymphoma, and presenting with clinical symptoms. The 11 censored observations correspond to patients who were still alive at last follow-up.

2.5, 4.1, 4.6, 6.4, 6.7, 7.4, 7.6, 7.7, 7.8, 8.8, 13.3, 13.4, 18.3, 19.7, 21.9, 24.7, 27.5, 29.7, 30.1*, 32.9, 33.5, 35.4*, 37.7*, 40.9*, 42.6*, 45.4*, 48.5*, 48.9*, 60.4*, 64.4*, 66.4*

Find NA estimate of $H(t)$, and compare KM and NA estimates for $S(t)$.

```
> library(survival)
> data<-read.table("eg231.txt", header=T)
> data
   time status
1   2.5      1
2   4.1      1
3   4.6      1
  ...    ...
```

```
> x<-data$time
> delta<-data$status
>
> # Check > ?survfit.formula for details.
> # NA estimates for H(t) and S(t).
> fit.na<-survfit(Surv(x, delta)~1,ctype=1,stype=2)
> # Default: KM estimte for S(t), NA estimate for H(t).
> fit<-survfit(Surv(x, delta)~1,ctype=1,stype=1)
>
> # Both give the NA estimate for H(t).
> fit.na$cumhaz
 [1] 0.03225806 0.06559140 0.10007416 0.13578844 0.17282548 0.21128702
... ...
[25] 1.02135386 1.02135386 1.02135386 1.02135386 1.02135386 1.02135386
[31] 1.02135386
> fit$cumhaz
 [1] 0.03225806 0.06559140 0.10007416 0.13578844 0.17282548 0.21128702
... ...
[25] 1.02135386 1.02135386 1.02135386 1.02135386 1.02135386 1.02135386
[31] 1.02135386
```

```
> # Compare the estimated S(t) using NA and KM methods.
> print(cbind(fit.na$time,fit.na$surv,fit$surv))
      [,1]      [,2]      [,3]
 [1,]  2.5 0.9682567 0.9677419
 [2,]  4.1 0.9365134 0.9354839
 [3,]  4.6 0.9047703 0.9032258
... ...
[30,] 64.4 0.3601071 0.3494624
[31,] 66.4 0.3601071 0.3494624
>
> # Plot NA estimate of H(t).
> plot(fit.na, xlab="Time (months)", ylab="Estimated H(t)",
ylim=c(0, 1.3),conf.int=F, cumhaz=T,mark.time=T)
>
> # Plot NA and KM estimates of S(t).
> plot(fit.na, xlab="Time (months)", ylab="Estimated S(t)",
mark.time=T,conf.int=F)
> lines(fit, conf.int=F,lty=2)
> legend(40,0.95,c("Nelson-Aalen","Kaplan-Meier"),lty=1:2)
```

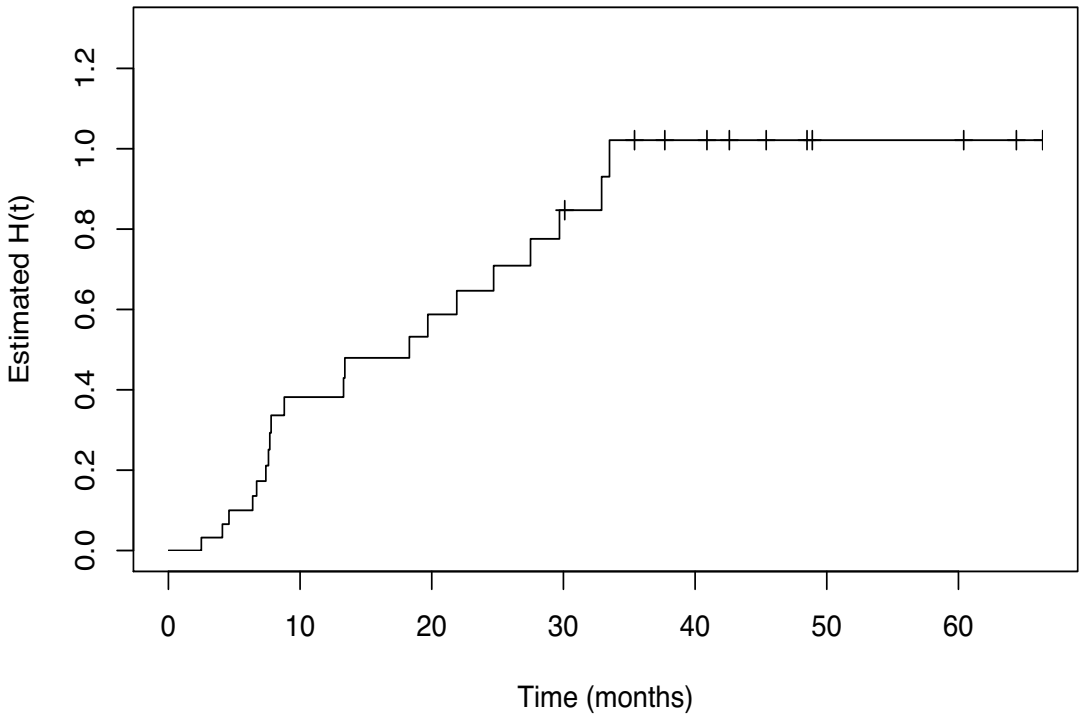


Figure: Nelson-Aalen estimate of the cumulative hazard function for the advanced non-Hodgkin's lymphoma data.

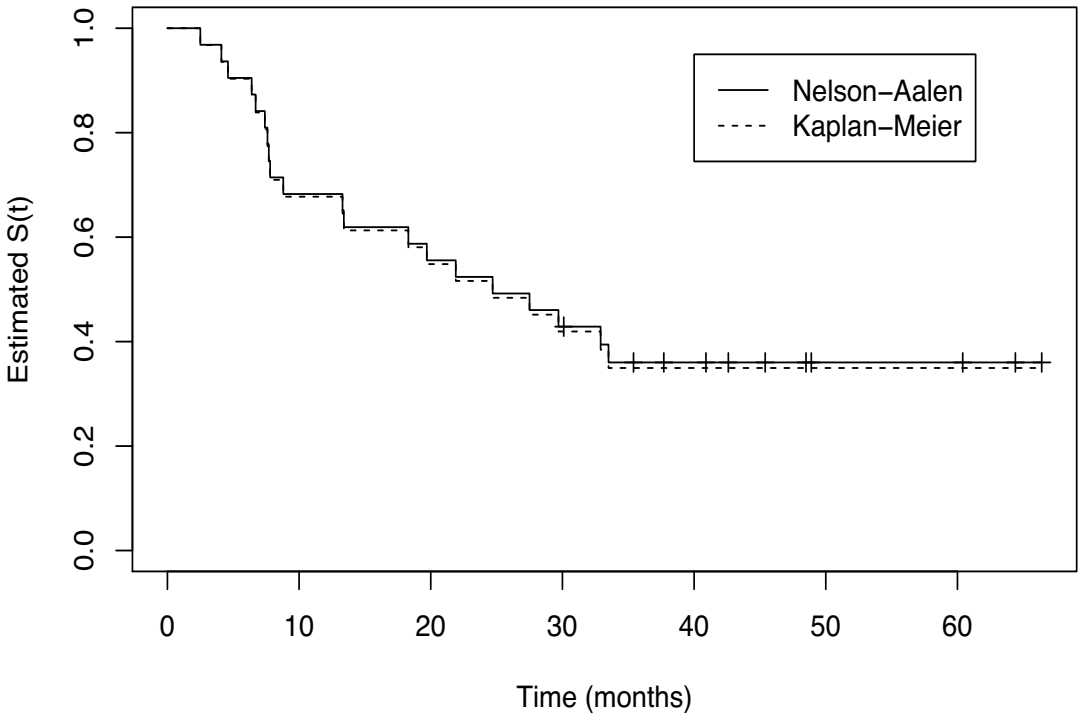


Figure: Nelson-Aalen and Kaplan-Meier estimates of the survival function for the advanced non-Hodgkin's lymphoma data.

Discussion

- The NA estimate and the KM estimate for $S(t)$ are very similar.
- Alternatively, we can also find KM estimate of $S(t)$ first, then estimate $H(t)$ by

$$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t).$$

Not surprisingly, $\hat{H}_{NA}(t)$ and $\hat{H}_{KM}(t)$ are also very similar.

2.4 Nonparametric Estimation of Other Characteristics

The following characteristics describing the distribution of T can also be estimated based on the nonparametric estimation for $S(t)$.

- Mean survival time, or mean lifetime

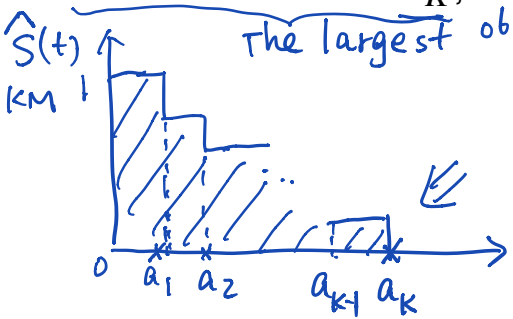
$$\mu = \mathrm{E}(T) = \int_0^\infty t f(t)dt = \int_0^\infty S(t)dt.$$

Its point estimate: $\hat{\mu} = \int_0^\infty \hat{S}(t)dt$

Distinct failure time $a_1 < a_2 < \cdots < a_K$.

largest failure time.

If a failure occurs at a_K, then $\hat{\mu}$ = area under $\hat{S}(t)$.



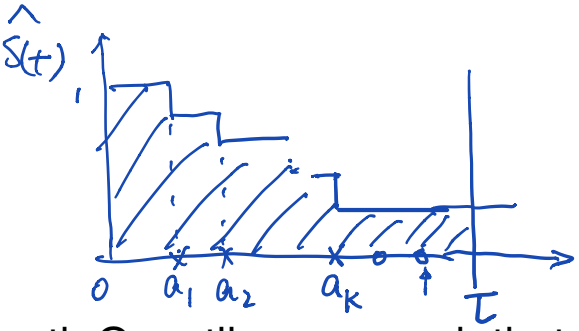
The largest observed time is a failure, denote it by a_K.

$$\hat{\mu} = a_1 + \hat{S}(a_1)(a_2 - a_1) + \cdots + \hat{S}(a_{K-1})(a_K - a_{K-1})$$

Question: Can we use the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n t_i$ to estimate μ?

If a censoring occurs at the largest observed time a_K, we usually calculate instead the restricted mean survival time, the mean survival time for $T \in [0, \tau]$, where τ is pre-determined according to duration of study,

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$$



- pth Quantile: t_p such that $P(T \le t_p) = p$, or equivalently, $S(t_p) = 1 - p$.
  Its point estimate: $\hat{t}_p = \inf\{t : \hat{S}(t) \le 1 - p\}$
  Median survival time: $t_{0.5}$
  Find $\hat{t}_{0.5}$ for Example 2.2.1.

$\hat{t}_{0.5} = 27$ days.

Further reading: C.I. for $t_{0.5}$, in the book by Klein & Moeschberger.

Find C.I. for t_p.

1) Find C.I. for $S(\hat{t}_p)$.

$S'(t) = -f(t)$.

Δ – method ...

$$Var\left[S(\hat{t}_p)\right] \approx \underbrace{\left[S'(t_p)\right]^2}_{\left[f(t_p)\right]^2} Var(\hat{t}_p).$$

2). Need estimate $\widehat{f(t_p)}$.

pdf

Kernel density estimation for $f(t)$ ... //

2.5. Applications of Nonparametric Estimation

The non-parametric (Kaplan-Meier, Nelson-Aalen) estimates of $S(t)$ and/or $H(t)$ are often used for

- initial exploration of data:
  for example, look for patterns in $S(t)$ and $H(t)$;
- compare survival functions for different groups in the population,
  e.g. treatment versus control group;
- assess the fit of a parametric model,
  for example, plot and compare the nonparametric estimate of $S(t)$ (or $H(t)$) with the corresponding parametric estimate.

Question: If the Nelson-Aalen estimate, $\hat{H}(t)$, shows a linear increasing pattern against t, what is a suitable parametric model for the failure time?

$\approx$ straight line $\approx \lambda t$?

$h(t) = \frac{d\hat{H}(t)}{dt} \approx \lambda$ ..... Maybe $T \sim Exp(\lambda)$?

rate

Example 2.5.1. Leukemia Remission Comparison

The data (given in Example 1.2.2) are remission times (in weeks) for patients with acute leukemia who were randomized in a clinical trial to receive either a drug (6-MP) or a placebo.

The intended purpose of the drug was to maintain remission in an individual.

Compare survival distributions of 6-MP and placebo groups through Kaplan-Meier estimates.

Create data file “eg122.txt” (grp: 0–“placebo”; 1–“6-MP”).

```
time status grp
1 1 0
1 1 0
2 1 0
...
34 0 1
35 0 1
```

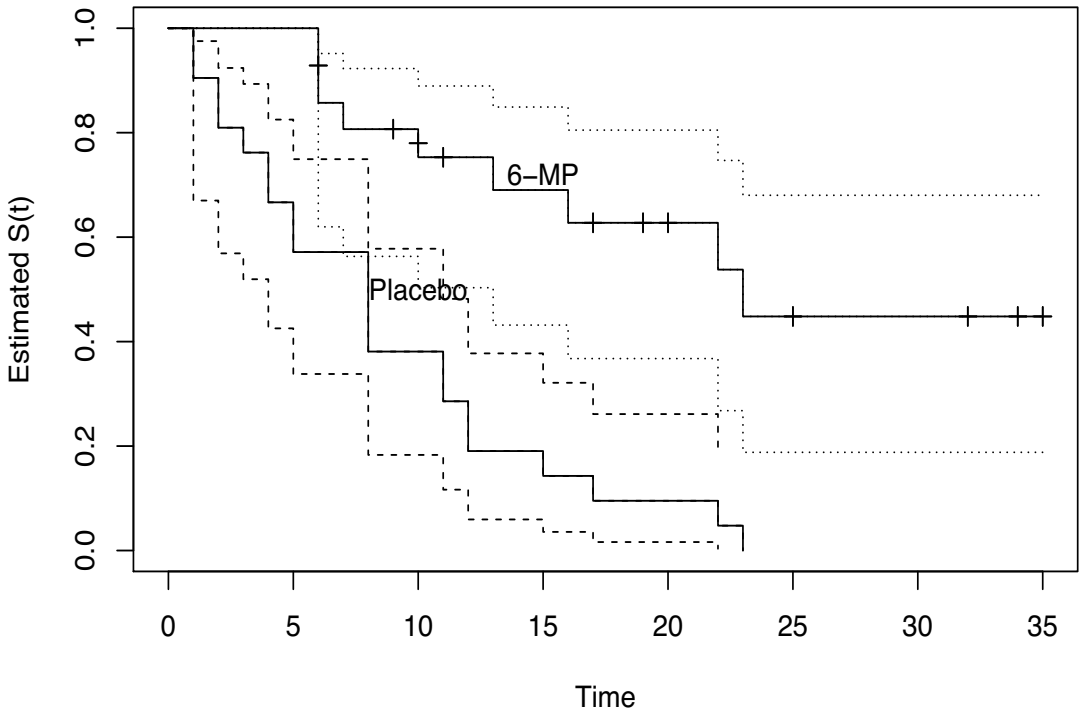


Figure: Leukemia Remission: Kaplan-Meier estimates of survival functions for patients given 6-MP versus placebo, with log-log based 95% confidence intervals.

Description: KM estimates by groups; respective log-log base CIs.

Comments on the survival distributions of the 6-MP and placebo groups:

censoring by groups: { No censoring in placebo grp. 12/21 > 50% censoring in 6-MP.

$\hat{S}(t)$ different for the 2 grp. More discussions about difference in survival prob. between groups, and CIs...

Put code and lengthy output in the Appendix.

```
> library(survival)
> leukemia<-read.table("eg122.txt", header=T)
> fit.km = survfit(Surv(time, status)~grp, data=leukemia,
conf.type="log-log")
> plot(fit.km,xlab="Time", ylab="Estimated S(t)",
mark.time=T)
> lines(fit.km,conf.int=T,lty=2:3,mark.time=T)
> text(10,0.5,"Placebo")
> text(15,0.72,"6-MP")
```