Chapter 2

Estimating the Survival Function

2.1 One-sample nonparametric methods

We will consider three methods for estimating a survivorship function

$$S(t) = P(T > t)$$

without resorting to parametric methods:

(1) Kaplan-Meier (most popular)

(2) Life-table (Actuarial Estimator, for grouped data)

(3) via the Nelson-Aalen cumulative hazard estimator

2.2 Estimating the survivorship function - without censoring

To motivate the derivation of the non-parametric *Kaplan-Meier estimator,* we first consider a set of survival times where there is NO censoring.

The following are times to relapse (weeks) for **21** leukemia patients receiving control treatment (Table **1.1** of Cox & Oakes, **1984**):

$$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$$

How would we estimate $\tilde{S}(10)$, the probability that an individual survives to time 10 or later?

What about $\tilde{S}(8)$, the probability that an individual survives more than 8 weeks?

$$\tilde{S}(8) \; = \; \hat{P}(T > 8) = \frac{8}{21} \; = \; 0.381$$

(at $t = 8$, we count the 4 events at time=8 as already having failed)

Empirical Survival Function:

When there is no censoring, the general formula is:

$$\tilde{S}(t) = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size}$$

And the standard error, how would one find that???

This is easy, since when there is no censoring, our estimated survival is just a proportion p:

$$se[\tilde{S}(8)] = \sqrt{p(1-p)/n}$$

$$= \sqrt{(0.38)(0.62)/21} = 0.106$$
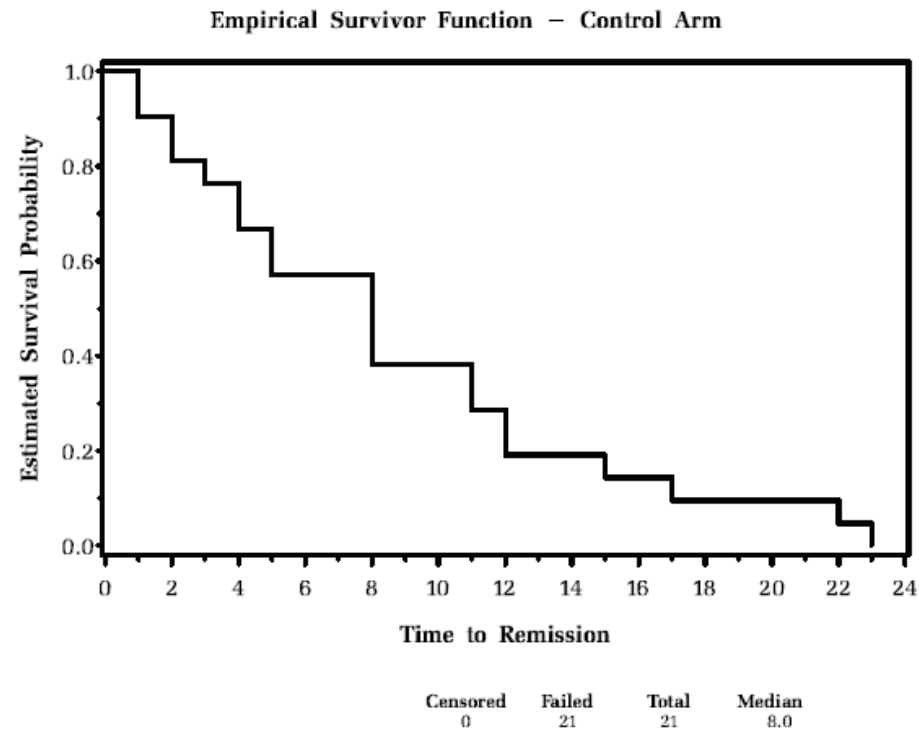
It is helpful to construct a table of $\tilde{S}(t)$:

Values of t	$\tilde{S}(t)$
$t < 1$	**21/21=1.000**
$1 \leq t < 2$	**19/21=0.905**
$2 \leq t < 3$	**17/21=0.809**
$3 \leq t < 4$	
$4 \leq t < 5$	
$5 \leq t < 8$	
$8 \leq t < 11$	
$11 \leq t < 12$	
$12 \leq t < 15$	
$15 \leq t < 17$	
$17 \leq t < 22$	
$22 \leq t < 23$	

Remission Times	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23
Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Figure 2.1 Example for leukemia data (control arm):

Figure 1.5: Example for leukemia data (control arm)

Empirical Survivor Function − Control Arm

Censored	Failed	Total	Median
0	21	21	8.0

The LIFETEST Procedure

Product-Limit Survival Estimates

t	Survival	Survival Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
1.0000	.	.	.	1	20
1.0000	0.9048	0.0952	0.0641	2	19
2.0000	.	.	.	3	18
2.0000	0.8095	0.1905	0.0857	4	17
3.0000	0.7619	0.2381	0.0929	5	16
4.0000	.	.	.	6	15
4.0000	0.6667	0.3333	0.1029	7	14
5.0000	.	.	.	8	13
5.0000	0.5714	0.4286	0.1080	9	12
8.0000	.	.	.	10	11
8.0000	.	.	.	11	10
8.0000	.	.	.	12	9
8.0000	0.3810	0.6190	0.1060	13	8
11.0000	.	.	.	14	7
11.0000	0.2857	0.7143	0.0986	15	6
12.0000	.	.	.	16	5
12.0000	0.1905	0.8095	0.0857	17	4
15.0000	0.1429	0.8571	0.0764	18	3
17.0000	0.0952	0.9048	0.0641	19	2
22.0000	0.0476	0.9524	0.0465	20	1
23.0000	0	1.0000	0	21	0

R Commands for Survival Estimation

library(survival)

```
> t<-c(1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15,17, 22, 23)
> km<-survfit(Surv(t, rep(1,21))~1)
> summary(km)
Call: survfit(formula = Surv(t, rep(1, 21)) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000      NA           NA           NA

>
```

2.4 (1) The Kaplan-Meier Estimator

We've developed an estimator for $S(t)$ when there is no censoring, but when we do have censoring we need to extend this approach using the Kaplan-Meier estimator.

The Kaplan-Meier (KM) estimator is one of the most widely-used measures in survival analysis and in medical research as a whole.

It can be justified from more than one perspective:

- product limit estimator
- likelihood justification

We will start with an intuitive motivation based on conditional probabilites, then review some other justifications.

Remember the Empirical Survival Function:

$$\text{(without censoring:)} \quad \tilde{S}(t) = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size}$$

What if there is censoring?

Consider the treated group of Leukemia patients from Table 1.1 of Cox and Oakes [Note: times with $^+$ are right-censored]:

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

$$19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$$

- Clearly $\hat{S}(5) = 21/21$, because everyone survived at least until time 6 or greater.

- But, we can't necessarily say $\hat{S}(6) = 17/21$, because we don't know the status of the person who was censored at time 6.

In a landmark 1958 paper in the *Journal of the American Statistical Association*, Kaplan and Meier proposed a way to nonparametrically estimate S(t), even in the presence of censoring. The method is based on the ideas of conditional probability.

This is one of the most cited papers in mathematics, statistics and computer science.

We **estimate $S(t)$ in the discrete case** first, heuristically:

Suppose we want to estimate $S(t)$ at at time t which is between two observed survival times, a_k and a_{k+1}, eg., $a_k \le t < a_{k+1}$. Then

$$S(t) = P(T > t) = P(T > a_k)$$

$$= P(T > a_1, T > a_2, \ldots, T > a_k)$$

$$= P(T > a_1) \times \prod_{j=2}^{k} P(T > a_j | T > a_{j-1})$$

$$\overset{(*)}{=} \prod_{j=1}^{k} [1 - P(T = a_j | T > a_{j-1})] = \prod_{j=1}^{k} [1 - \lambda_j]$$

$$\textbf{so} \quad \hat{S}(t) \cong \prod_{j=1}^{k} \left(1 - \frac{d_j}{r_j}\right) = \prod_{j:a_j \le t} \left(1 - \frac{d_j}{r_j}\right)$$

d_j **is the number of deaths at a_j**
r_j **is the number at risk at a_j**
(*) Initial assumptions: $a_0 = 0$, $P(T > a_0) = 1$.

Remember the rules of conditional probability:

Conditional Probability: Suppose A and B are two events. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication law of probability: can be obtained from the above relationship, by multiplying both sides by $P(B)$:

$$P(A \cap B) = P(A|B)\,P(B)$$

Suppose $A_1, A_2...A_k$ are k different events. Then, the probability of all k events happening together can be written as a product of conditional probabilities:

$$P(A_1 \cap A_2... \cap A_k) =$$

$$P(A_k|A_{k-1} \cap ... \cap A_1) \times P(A_{k-1}|A_{k-2} \cap ... \cap A_1) \times ... \times P(A_2|A_1) \times P(A_1).$$

The Kaplan-Meier Estimator in the *continuous* case

Think of dividing the observed timespan of the study into a series of fine intervals so that there is a separate interval for each time of death or censoring (with possible ties):

Using the law of conditional probability,

$$P(T > t) = \prod_{j} P(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j)$$

where the product is taken over all the intervals including or preceding time t.

Four possibilities for each interval:

(1) No events (death or censoring) - conditional probability of surviving the interval is 1

(2) Censoring - assume they survive to the end of the interval, so that the conditional probability of surviving the interval is 1

(3) Death, but no censoring - conditional probability of *not* surviving the interval is # deaths (d) divided by # 'at risk' (r) at the beginning of the interval. So the conditional probability of surviving the interval is $1 - (d/r)$.

(4) Tied deaths and censoring - assume censorings last to the end of the interval, so that conditional probability of surviving the interval is still $1 - (d/r)$

General Formula for jth interval:

It turns out we can write a general formula for the conditional probability of surviving the j-th interval that holds for all 4 cases:

$$\left(1 - \frac{d_j}{r_j}\right)$$

We could use the same approach by grouping the event times into intervals (say, one interval for each month), and then counting up the number of deaths (events) in each to estimate the probability of surviving the interval (this is called the *lifetable estimate*).

However, the assumption that those censored last until the end of the interval wouldn't be quite accurate, so we would end up with a cruder approximation.

As the intervals get finer and finer, the approximations made in estimating the probabilities of getting through each interval become smaller and smaller, so that the estimator converges to the true $S(t)$ as the sample size increases.

This intuition clarifies why an alternative name for the KM is the product limit estimator.

The Kaplan-Meier estimator of the survivorship function (or survival probability) $S(t) = P(T > t)$ is:

$$\hat{S}(t) = \prod_{j:\tau_j \leq t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

where

- $\tau_1, ... \tau_K$ **is the set of K distinct death times observed in the sample**

- d_j **is the number of deaths at** τ_j

- r_j **is the number of individuals "at risk" right before the** j**-th death time (everyone dead or censored <u>at or after</u> that time).**

- c_j **is the number of censored observations between the** j**-th and** $(j{+}1)$**-st death times. Censorings tied at** τ_j **are included in** c_j

- **Note that the risk set** r_j **can be calculated as either** $r_{j-1} - d_{j-1} - c_{j-1}$ **or alternatively as** $\sum_{l \geq j}(c_l + d_l)$

Calculating the KM - for treated leukemia patients, Cox and Oakes

Make a table with a row for every death or censoring time:

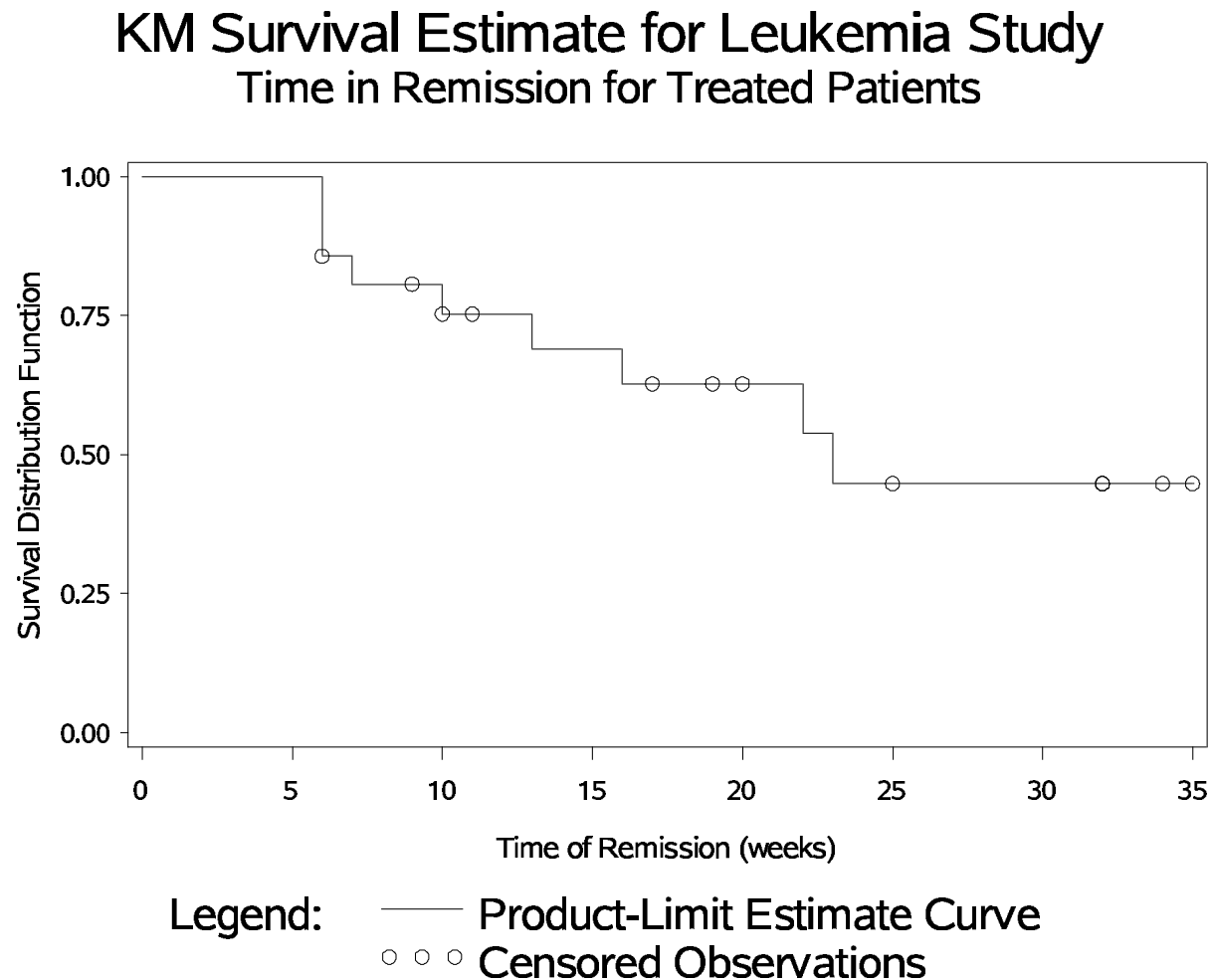| τ_j	d_j	c_j	r_j	$1 - (d_j/r_j)$	$\hat{S}(\tau_j)$
6	3	1	21	$\frac{18}{21} = 0.857$	
7	1	0	17		
9	0	1	16		
10					
11					
13					
16					
...					
23					

Note that:

- $\hat{S}(t)$ only changes at death (failure) times

- $\hat{S}(t)$ is 1 up to the first death time

- $\hat{S}(t)$ only goes to 0 if the last event is a death

Figure 2.2 Example for leukemia data (treated arm):

KM Survival Estimate for Leukemia Study
Time in Remission for Treated Patients

2.5 Likelihood-based derivation of the KM Estimator

For a discrete failure time variable, define:

$$
\begin{array}{ll}
d_j & \text{number of failures at } a_j \\
r_j & \text{number of individuals at risk at } a_j \\
& \text{(including those censored at } a_j). \\
\lambda_j & \text{P('death') in } j\text{-th interval} \\
& \text{(conditional on survival to start of interval)}
\end{array}
$$

A likelihood of g independent binomials can be constructed:

$$
L(\lambda_1, .., \lambda_J) \;=\; \prod_{j=1}^{J} \lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j}
$$

Therefore, the maximum likelihood estimator of λ_j is:

$$
\hat{\lambda}_j = d_j / r_j
$$

Let $d_{ij} = 1$ if subject i had event at time a_j and 0 otherwise; $d_j = \sum_{i=1}^{n} d_{ij}$

Let $c_{ij} = 1$ if subject i was censored at time a_j and 0 otherwise; $c_j = \sum_{i=1}^{n} c_{ij}$

$$
\begin{aligned}
L(\lambda_1, .., \lambda_J) &= \prod_{i=1}^{n} \prod_{j=1}^{J} P(T = a_j)^{d_{ij}} P(T > a_j)^{c_{ij}} \\
&= \prod_{j=1}^{J} P(T = a_j)^{d_j} P(T > a_j)^{c_j} \\
&= \prod_{j=1}^{J} f_j^{d_j} S(a_j)^{c_j} \\
&= \prod_{j=1}^{J} [\lambda_j S(a_{j-1})]^{d_j} S(a_j)^{c_j} \\
&= \prod_{j=1}^{J} [\lambda_j S(a_{j-1})]^{d_j} S(a_{j-1})^{c_j} (1 - \lambda_j)^{c_j} \\
&= \prod_{j=1}^{J} \lambda_j^{d_j} S(a_{j-1})^{d_j + c_j} (1 - \lambda_j)^{c_j} \\
&= \prod_{j=1}^{J} \lambda_j^{d_j} \prod_{k=1}^{j-1} (1 - \lambda_k)^{d_j + c_j} (1 - \lambda_j)^{c_j}
\end{aligned}
$$

$$= \prod_{j=1}^{J} \lambda_j^{d_j} (1 - \lambda_j)^{c_j + \sum_{k=j}^{J}(d_{k+1} + c_{k+1})}$$

$$= \prod_{j=1}^{J} \lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j}$$

Since we can write the estimated survival function as:

$$\hat{S}(t) \;=\; \prod_{j:a_j \leq t} (1 - \hat{\lambda}_j)$$

We can then use the fact that the maximum likelihood estimator of λ_j is:

$$\hat{\lambda}_j = d_j / r_j,$$

and plug in the MLE's of λ_j to estimate $S(t)$:

$$\hat{S}(t) \;=\; \prod_{j:a_j \leq t} \left(1 - \frac{d_j}{r_j} \right)$$

2.6 Properties of the KM estimator

In the case of no censoring (with $n=\#$ on study):

$$\hat{S}(t) = \tilde{S}(t) = \frac{\# \ deaths \ greater \ than \ t}{n}$$

Just as for $\hat{p}$ from the binomial distribution, we know that for large n:

$$\hat{S}(t) \overset{d}{\simeq} \mathcal{N}\left(S(t), \ S(t)[1-S(t)]/n\right)$$

How does censoring affect this?

- $\hat{S}(t)$ is still approximately normal

- The mean of $\hat{S}(t)$ converges to the true S(t)

- The variance is a bit more complicated (since the denominator n includes some censored observations).

Once we get the variance (or standard error), then we can construct (pointwise) $(1-\alpha)$ 100% confidence bands about $\hat{S}(t)$:

$$\hat{S}(t) \pm z_{1-\alpha/2} \, se[\hat{S}(t)]$$

2.7 Greenwood's formula

(Also see Section 2.2 of Collett)

We can think of the KM estimator as

$$\hat{S}(t) = \prod_{j:\tau_j \leq t} (1 - \hat{\lambda}_j)$$

where $\hat{\lambda}_j = d_j/r_j$. Since the $\hat{\lambda}_j$'s are (conditionally) seen as binomial proportions, we can apply standard likelihood theory to show that each $\hat{\lambda}_j$ is approximately normal, with mean the true λ_j, and estimated variance:

$$\widehat{\text{Var}}(\hat{\lambda}_j) = \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{r_j}$$

<u>Note:</u> the true variance is $\text{Var}(\hat{\lambda}_j) = \frac{\lambda_j(1-\lambda_j)}{r_j}$ (just take the "hats" off). Also, the $\hat{\lambda}_j$'s are independent in large enough samples.

Since $\hat{S}(t)$ is a function of the λ_j's, we can estimate its variance using the 'delta method' - an approach for calculating the variance of *non-linear* functions:

> <u>Delta method</u>: If Y is normal with mean μ and variance σ^2, then $g(Y)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.

(also see the review notes)

Two specific examples of the delta method:

(A) $Z = g(Y) = log(Y)$, then $Z \sim N\left[log(\mu), \left(\frac{1}{\mu}\right)^2\sigma^2\right]$

(B) $Z = g(Y) = exp(Y)$ then $Z \sim N\left[e^\mu, [e^\mu]^2\sigma^2\right]$

[**Note:** If $g(y) = log(y)$ then $g'(y) = (1/y)$, and if $g(y) = e^y$ then $g'(y) = e^y$]

Greenwood's formula (continued)

Instead of dealing with $\hat{S}(t)$ directly, we will look at its log (Why? because calculating the variance of a sum is easier than calculating the variance of a product!):

$$\log[\hat{S}(t)] = \sum_{j:\tau_j \leq t} \log(1 - \hat{\lambda}_j)$$

Thus, by approximate independence of the $\hat{\lambda}_j$'s,

$$\widehat{\textbf{Var}}(\log[\hat{S}(t)]) = \sum_{j:\tau_j \leq t} \hat{\textbf{Var}}[\log(1 - \hat{\lambda}_j)]$$

$$\textbf{by (A)} \quad = \sum_{j:\tau_j \leq t} \left(\frac{1}{1 - \hat{\lambda}_j}\right)^2 var(\hat{\lambda}_j)$$

$$= \sum_{j:\tau_j \leq t} \left(\frac{1}{1 - \hat{\lambda}_j}\right)^2 \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{r_j}$$

$$= \sum_{j:\tau_j \leq t} \frac{\hat{\lambda}_j}{(1 - \hat{\lambda}_j)r_j} \quad = \sum_{j:\tau_j \textbf{E} t} \frac{d_j}{(r_j - d_j)r_j}$$

Now, $\hat{S}(t) = \exp[\log[\hat{S}(t)]]$. **Thus (by B), the "true" variance of the estimated survival function is:**

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 \, \mathbf{Var} \left[\log[\hat{S}(t)] \right]$$

and its estimated value is:

$$\widehat{\mathbf{Var}}(\hat{S}(t)) = [\hat{S}(t)]^2 \, \widehat{\mathbf{Var}} \left[\log[\hat{S}(t)] \right]$$

to obtain:

Greenwood's Formula:

$$\widehat{\mathbf{Var}}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j}$$

2.8 Confidence Intervals for KM Estimator

For a 95% confidence interval, we could use

$$\hat{S}(t) \pm z_{1-\alpha/2} \, se[\hat{S}(t)]$$

and estimate $se[\hat{S}(t)]$ using Greenwood's formula.

Problem: This approach can yield values > 1 or < 0.

Better approach: Get a 95% confidence interval for

$$L(t) = \log[-\log(S(t))]$$

Since this quantity is unrestricted, the confidence interval will be in the proper range when we transform back.

Note: this is similar to the justification for calculating a CI for the $\log(OR)$ rather than the OR itself.

To see why this works, note the following:

- Since $\hat{S}(t)$ is an estimated probability

$$0 \leq \hat{S}(t) \leq 1$$

- Taking the log of $\hat{S}(t)$ has bounds:

$$-\infty \leq \log[\hat{S}(t)] \leq 0$$

- Taking the opposite:

$$0 \leq -\log[\hat{S}(t)] \leq \infty$$

- Taking the log again:

$$-\infty \leq \log\left[-\log[\hat{S}(t)]\right] \leq \infty$$

To transform back, reverse steps with $S(t) = \exp[-\exp(L(t))]$

Log-log Approach for Confidence Intervals:

(1) Define $L(t) = \log[-\log(S(t))]$

(2) Form a 95% confidence interval for $L(t)$ **based on** $\hat{L}(t)$**,**
yielding $[\hat{L}(t) - A, \hat{L}(t) + A]$**, with** $A = 1.96 \times se(\hat{L}(t))$**.**

(3) Since $S(t) = \exp[-\exp(L(t))]$**, the confidence bounds for the 95% CI on** $S(t)$ **are:**

$$[\exp\{-e^{(\hat{L}(t)+A)}\}, \quad \exp\{-e^{(\hat{L}(t)-A)}\}]$$

(note that the upper and lower bounds switch)

(4) Substituting $\hat{L}(t) = \log[-\log(\hat{S}(t))]$ **back into the above bounds, we get confidence**
bounds of

$$([\hat{S}(t)]^{e^A}, [\hat{S}(t)]^{e^{-A}})$$

Software for Kaplan-Meier Curves

- **SAS -** PROC LIFETEST

- **Splus/R -** `surv.fit(Surv(time,status)~1)`

- **Stata - stset and sts commands**

Defaults for Confidence Interval Calculations

- **SAS and Stata - "log-log"** $\Rightarrow \hat{L}(t) \pm 1.96 \, \widehat{se}[\hat{L}(t)]$
  where $L(t) = \log[-\log(S(t))]$

- **R - "log"** $\Rightarrow \log(\hat{S}(t)) \pm 1.96 \, \widehat{se}[\log(\hat{S}(t))]$

Output from SAS Proc Lifetest

Leukemia data - trt arm - from Table 1.1 of Cox and Oakes

The LIFETEST Procedure
Product-Limit Survival Estimates

WEEKS	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
6.0000	.	.	.	1	20
6.0000	.	.	.	2	19
6.0000	0.8571	0.1429	0.0764	3	18
6.0000*	.	.	.	3	17
7.0000	0.8067	0.1933	0.0869	4	16
9.0000*	.	.	.	4	15
10.0000	0.7529	0.2471	0.0963	5	14
10.0000*	.	.	.	5	13
11.0000*	.	.	.	5	12
13.0000	0.6902	0.3098	0.1068	6	11
16.0000	0.6275	0.3725	0.1141	7	10
17.0000*	.	.	.	7	9
19.0000*	.	.	.	7	8
20.0000*	.	.	.	7	7
22.0000	0.5378	0.4622	0.1282	8	6
23.0000	0.4482	0.5518	0.1346	9	5
25.0000*	.	.	.	9	4
32.0000*	.	.	.	9	3
32.0000*	.	.	.	9	2
34.0000*	.	.	.	9	1
35.0000*	.	.	.	9	0

* Censored Observation
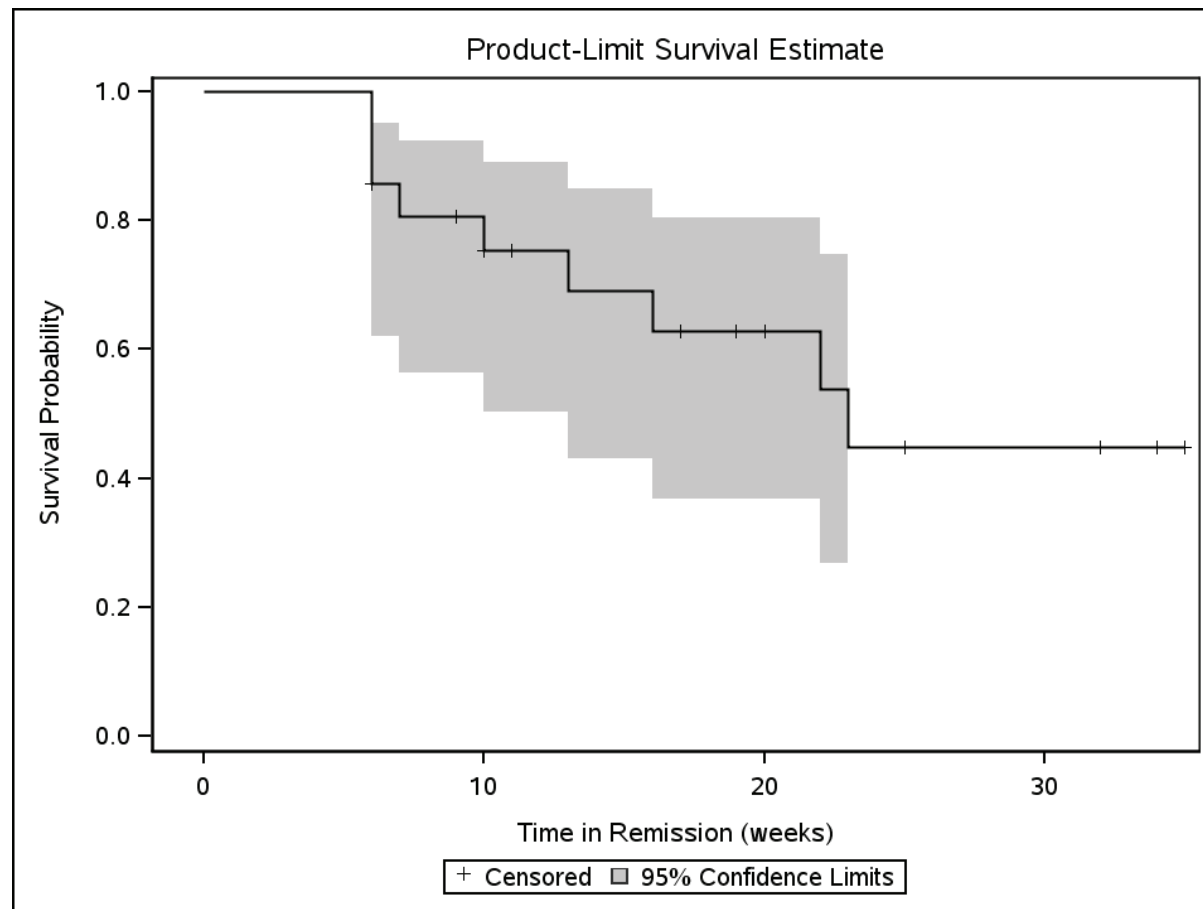
Output from printing the CONFINT file - SAS V9

```
95% Confidence Intervals for Estimated Survival
    Obs     weeks    _CENSOR_     SURVIVAL     SDF_LCL     SDF_UCL

     1        0          .        1.00000     1.00000     1.00000
     2        6          0        0.85714     0.61972     0.95155
     3        6          1        0.85714        .           .
     4        7          0        0.80672     0.56315     0.92281
     5        9          1        0.80672        .           .
     6       10          0        0.75294     0.50320     0.88936
     7       10          1        0.75294        .           .
     8       11          1        0.75294        .           .
     9       13          0        0.69020     0.43161     0.84907
    10       16          0        0.62745     0.36751     0.80491
    11       17          1        0.62745        .           .
    12       19          1        0.62745        .           .
    13       20          1        0.62745        .           .
    14       22          0        0.53782     0.26778     0.74679
    15       23          0        0.44818     0.18805     0.68014
    16       25          1           .           .           .
    17       32          1           .           .           .
    18       32          1           .           .           .
    19       34          1           .           .           .
    20       35          1           .           .           .
```

The output dataset will have one observation for each unique combination of WEEKS and _CENSOR_. It will also add an observation for failure time equal to 0.

Pointwise Confidence Intervals for KM Estimate

Output from printing the CI_LINEAR file

```
         95% Confidence Intervals for Estimated Survival
  Obs     weeks     _CENSOR_     SURVIVAL     CONFTYPE     SDF_LCL     SDF_UCL

    1       0          .          1.00000                 1.00000     1.00000
    2       6          0          0.85714      LINEAR      0.70748     1.00000
    3       6          1          0.85714                    .           .
    4       7          0          0.80672      LINEAR      0.63633     0.97711
    5       9          1          0.80672                    .           .
    6      10          0          0.75294      LINEAR      0.56410     0.94178
    7      10          1          0.75294                    .           .
    8      11          1          0.75294                    .           .
    9      13          0          0.69020      LINEAR      0.48084     0.89955
   10      16          0          0.62745      LINEAR      0.40391     0.85099
   11      17          1          0.62745                    .           .
   12      19          1          0.62745                    .           .
   13      20          1          0.62745                    .           .
   14      22          0          0.53782      LINEAR      0.28648     0.78915
   15      23          0          0.44818      LINEAR      0.18439     0.71197
   16      25          1            .                        .           .
   17      32          1            .                        .           .
   18      32          1            .                        .           .
   19      34          1            .                        .           .
   20      35          1            .                        .           .
```

Output options in R:

```
>surv.fit(y$t,y$c)
 95 percent confidence interval is of type "log"
 time n.risk n.event  survival    std.dev lower 95% CI upper 95% CI
    6     21       3 0.8571429 0.07636035    0.7198171    1.0000000
    7     17       1 0.8067227 0.08693529    0.6531242    0.9964437
   10     15       1 0.7529412 0.09634965    0.5859190    0.9675748
   13     12       1 0.6901961 0.10681471    0.5096131    0.9347692
   16     11       1 0.6274510 0.11405387    0.4393939    0.8959949
   22      7       1 0.5378151 0.12823375    0.3370366    0.8582008
   23      6       1 0.4481793 0.13459146    0.2487882    0.8073720


> surv.fit(y$t,y$c,conf.type="log-log")
 95 percent confidence interval is of type "log-log"
 time n.risk n.event  survival    std.dev lower 95% CI upper 95% CI
    6     21       3 0.8571429 0.07636035    0.6197180    0.9515517
    7     17       1 0.8067227 0.08693529    0.5631466    0.9228090
   10     15       1 0.7529412 0.09634965    0.5031995    0.8893618
   13     12       1 0.6901961 0.10681471    0.4316102    0.8490660
   16     11       1 0.6274510 0.11405387    0.3675109    0.8049122
   22      7       1 0.5378151 0.12823375    0.2677789    0.7467907
   23      6       1 0.4481793 0.13459146    0.1880520    0.6801426


> surv.fit(y$t,y$c,conf.type="plain")
 95 percent confidence interval is of type "plain"
 time n.risk n.event  survival    std.dev lower 95% CI upper 95% CI
    6     21       3 0.8571429 0.07636035    0.7074793    1.0000000
    7     17       1 0.8067227 0.08693529    0.6363327    0.9771127
   10     15       1 0.7529412 0.09634965    0.5640993    0.9417830
   13     12       1 0.6901961 0.10681471    0.4808431    0.8995491
   16     11       1 0.6274510 0.11405387    0.4039095    0.8509924
   22      7       1 0.5378151 0.12823375    0.2864816    0.7891487
   23      6       1 0.4481793 0.13459146    0.1843849    0.7119737
```

2.9 Confidence Bands for the Survival Function

In addition to the pointwise confidence intervals for $S(t)$ at each specific failure time t, we can also create a confidence band for the entire survival distribution.

This was proposed by Hall and Wellner in 1980.

Just as for linear regression models, the confidence bands are wider than the pointwise confidence intervals.

They are based on the same $log[-log(S(t))]$ transformation in order for the bounds to be in the correct (0,1) parameter space once they are back transformed (since the confidence bands are wider, this is even more of a possibility without transformation).

The Hall-Wellner Confidence Bands are formed as:

$$log[-log(S(t))] \ \pm \ H_{\hat{a},\alpha} \ \frac{\left(1 + n\hat{\sigma}^2\right)}{\sqrt{n} \ [ln\hat{S}(t)]}$$

where σ^2 is the variance of $log(\hat{S}(t))$ developed previously, and

$H_{\hat{a},\alpha}$ is a percentile developed by Hall and Wellner.

It had been difficult to obtain these bands through standard software because of the need to calculate the H value, but now these can be obtained in SAS Version 9.

They are provided in an Appendix of the Klein and Moeschboerger textbook.

KM Survival Estimate with Hall−Wellner Confidence Band **2**
Time in Remission for Treated Leukemia Patients
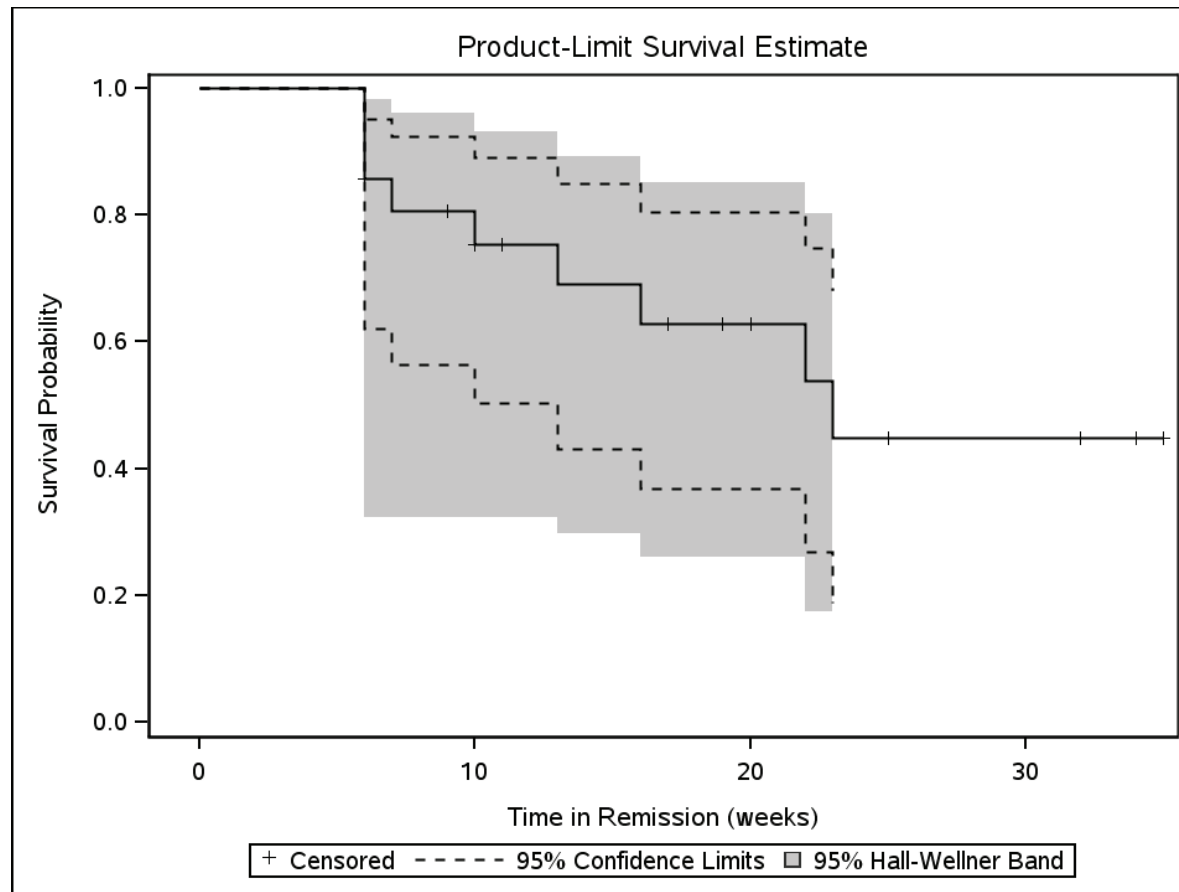
The LIFETEST Procedure

Product-Limit Survival Estimate

Summary of the Number of
Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
21	9	12	57.14

2.10   Means, Medians, and Quantiles based on the KM

First, can we calculate these when there is no censoring?

Consider the untreated leukemia patients (N=21) who all failed prior to week 23.

- Mean: we can calculate as $\sum_{j=1}^{J} \tau_j P(T = \tau_j)$, which is just adding up the event times and dividing by the number of events.

  For untreated leukemia patients, $\sum t_j = 182$, and so mean=182/21=8.667.

- Median: this is the smallest observed event time τ such that $\hat{S}(\tau) \leq 0.50$.

  If we look back at our survival estimates on p.7, $\hat{S}(5) = 0.571$ and $\hat{S}(8) = 0.381$. There are no event times between 5 and 8.

  So $\tau = 8$ is the median survival time.

Now let's consider the more general case with censored data:

- **Mean:** This is estimated as $\sum_{j=1}^{J} \hat{S}(\tau_{j-1})(\tau_j - \tau_{j-1})$

$$
\begin{aligned}
\text{mean } T \;&=\; \sum_{j=1}^{J} \tau_j Pr(T = \tau_j) \\[2mm]
&=\; \sum_{j=1}^{J} \tau_j f_j \\[2mm]
&=\; \sum_{j=1}^{J} \tau_j [S(\tau_{j-1}) - S(\tau_j)] \\[2mm]
&=\; \sum_{j=1}^{J} \tau_j S(\tau_{j-1}) - \sum_{j=1}^{J} \tau_j S(\tau_j) \\[2mm]
&=\; \sum_{j=1}^{J} \tau_j S(\tau_{j-1}) - \sum_{j=2}^{J+1} \tau_{j-1} S(\tau_{j-1}) \\[2mm]
&=\; \sum_{j=2}^{J} (\tau_j - \tau_{j-1}) S(\tau_{j-1}) + \tau_1 S(\tau_0) - \tau_J S(\tau_J) \\[2mm]
&=\; \sum_{j=1}^{J} (\tau_j - \tau_{j-1}) S(\tau_{j-1})
\end{aligned}
$$

However, if the very last observation is censored, then this will underestimate the true mean. (SAS actually tells you this!)

- Median - in practice, this is defined as the smallest time such that $\hat{S}(\tau) \leq 0.5$. The median is more appropriate for censored survival data than the mean.

  For the treated leukemia patients, we find:

$$\hat{S}(22) = 0.5378$$
$$\hat{S}(23) = 0.4482$$

  The median is thus **23**. This can also be seen on Figure **2.2 (p.23)**.

- 75th percentile of $S(t)$:
  the smallest time t such that $\hat{S}(t) \leq 0.25$

- 25th percentile of $S(t)$:
  the smallest time t such that $\hat{S}(t) \leq 0.75$

- **General formula for the 100*p^{th} of** $S(t)$**: smallest time** t **such that** $\hat{S}(t) \leq (1 - p)$

2.11   (2) Lifetable Estimator of Survival

- also called the "actuarial estimator"

- one of the oldest techniques around

- used by actuaries, demographers, etc.

- applies when the data are grouped

Our goal is still to estimate the survival function, hazard, and density function, but this is complicated by the fact that we don't know exactly when during each time interval an event occurs.

The book by Lee (Statistical Methods for Survival Data Analysis; 4th edition, section 4.3) provides a good description of lifetable methods, and distinguishes several types according to the data sources (population life tables vs clinical life tables).

Population Life Tables

- <u>cohort life table</u> - describes the mortality experience from birth to death for a particular cohort of people born at about the same time. People at risk at the start of the interval are those who survived the previous interval.

- <u>current life table</u> - constructed from (1) census information on the number of individuals alive at each age, for a given year and (2) vital statistics on the number of deaths or failures in a given year, by age. This type of lifetable is often reported in terms of a hypothetical cohort of 100,000 people.

Generally, censoring is not an issue for Population Cohort Life Tables.
For the current life tables we tend to have records of all failures, but one must think hard about whether the derived distribution could be relevant for any given cohort.

Clinical Life tables - applies to grouped survival data from studies in patients with specific diseases. Because patients can enter the study at different times, or be lost to follow-up, censoring must be allowed.

Life Expectancy

Life expectancy continues to increase, and the gap in life expectancy between males and females has been narrowing.

Life expectancy is a measure often used to gauge the overall health of a population. As a summary measure of mortality, life expectancy represents the average number of years of life that could be expected if current death rates were to remain constant. Shifts in life expectancy are often used to describe trends in mortality. Life expectancy at birth is strongly influenced by infant and child mortality. Life expectancy later in life reflects death rates at or above a given age and is independent of the effect of mortality at younger ages (1).

From 1900 through 2003, life expectancy at birth increased from 48 to 75 years for men and from 51 to 80 years for women (Figure 24). Life expectancy at age 65 has also increased since the beginning of the 20th century. Among men, life expectancy at age 65 rose from 12 to 17 years and among women from 12 to 20 years. In contrast to life expectancy at birth, which increased sharply early in the 20th century, life expectancy at age 65 improved primarily after mid-century. Improved access to health care, advances in medicine, healthier lifestyles, and better health before age 65 are factors underlying decreased death rates among older Americans (2).

Athough the overall trend in life expectancy for the United States was upward throughout the 20th century, the gain in years of life expectancy for women generally exceeded that for men until the 1970s, widening the gap in life expectancy between men and women. After the 1970s, the gain in life expectancy for men exceeded that for women and the gender gap in life expectancy began to narrow. Between 1990 and 2003, the total gain in life expectancy for women was 1.3 years compared with 3.0 years for men, reflecting proportionately greater decreases in heart disease and cancer mortality for men than for women and proportionately larger increases in chronic lower respiratory diseases mortality among women (3).

Chartbook on Trends in the Health of Americans | Health, United States, 2006

Life expectancy
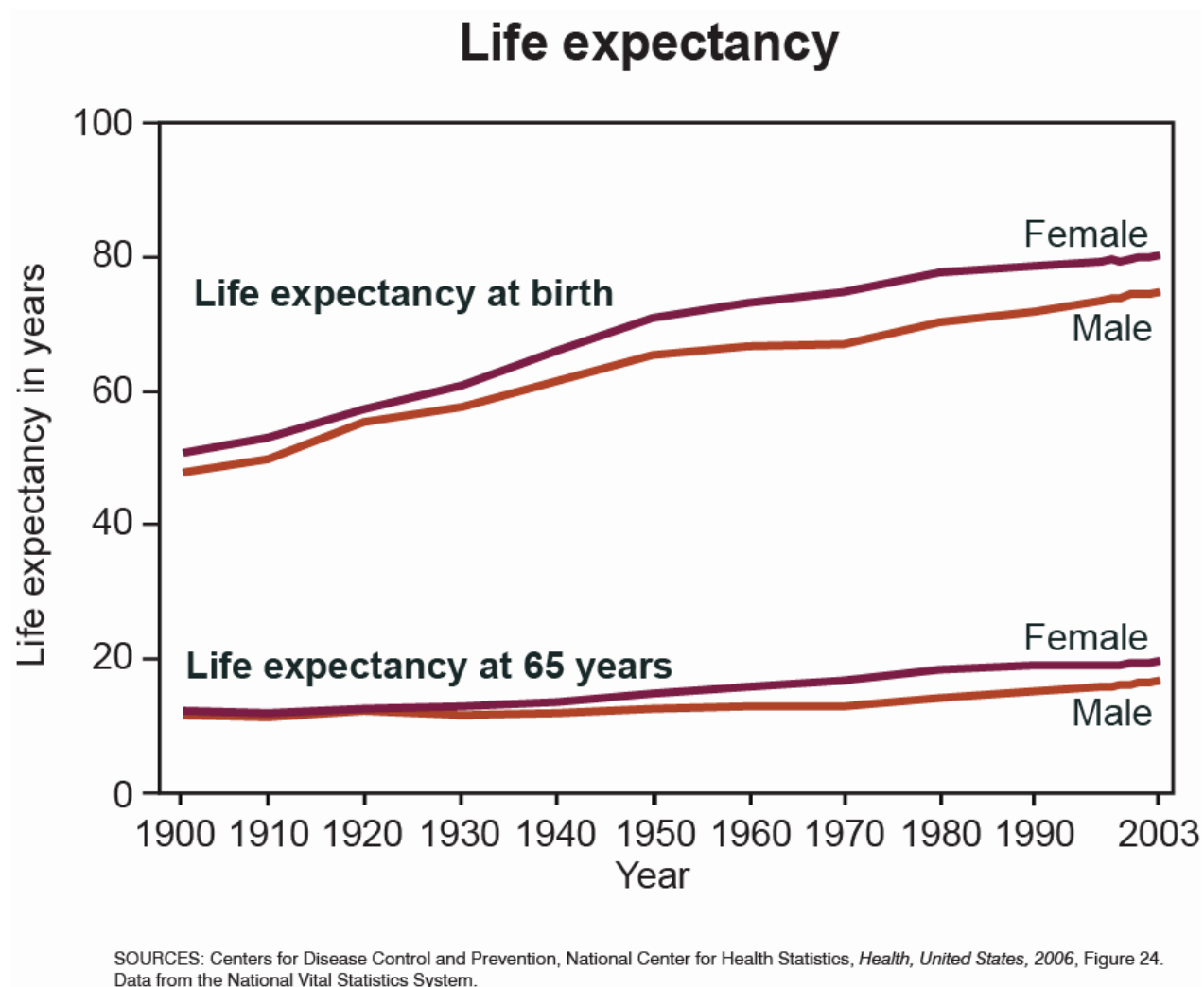
Notation

- the j-th time interval is $[t_{j-1}, t_j)$ (the way of grouping in SAS). For continuous data, there is probability 0 of falling exactly on the boundary.

- c_j - the number of censorings in the j-th interval

- d_j - the number of failures in the j-th interval

- r_j is the number <u>entering</u> the interval
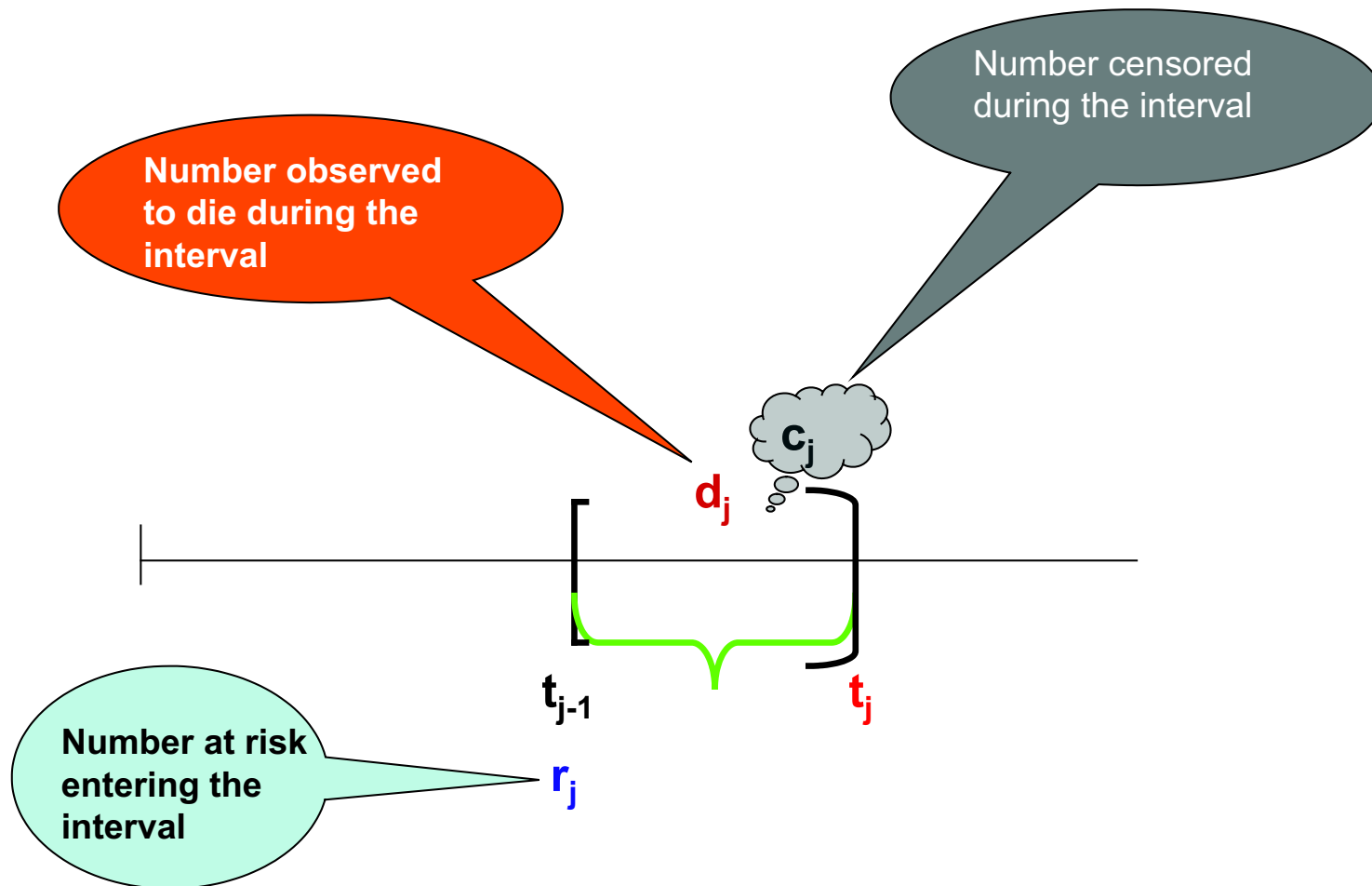
Example: 2418 Males with Angina Pectoris (Lee, p.95)

Year after Diagnosis	j	d_j	c_j	r_j
$[0, 1)$	1	456	0	2418
$[1, 2)$	2	226	39	1962
$[2, 3)$	3	152	22	1697
$[3, 4)$	4	171	23	1523
$[4, 5)$	5	135	24	1329
$[5, 6)$	6	125	107	1170
$[6, 7)$	7	83	133	938
etc..				

Estimating the survivorship function

We could apply the K-M formula directly to the numbers in the table on the previous page, estimating S(t) as

$$\hat{S}(t) \;=\; \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

However, this approach is unsatisfactory for grouped data.... it treats the problem as though it were in discrete time, with events happening only at 1 yr, 2 yr, etc. In fact, what we are trying to calculate here is the conditional probability of dying <u>within the interval</u>, given survival to the beginning of it.

Options for treating the censored data...

We can assume that censorings occur:

- at the <u>beginning</u> of each interval: $r'_j = r_j - c_j$

- at the <u>end</u> of each interval: $r'_j = r_j$

- <u>on average halfway through</u> the interval:

$$r'_j = r_j - c_j/2$$

The last assumption yields the Actuarial Estimator.

It is appropriate if censorings occur uniformly throughout the interval.

Example: 2418 Males with Angina Pectoris (Lee, p.95)

Year after Diagnosis	j	d_j	c_j	r_j	$r'_j = r_j - c_j/2$	
$[0,1)$	1	456	0	2418	2418.0	
$[1,2)$	2	226	39	1962	1942.5	$(1962 - \frac{39}{2})$
$[2,3)$	3	152	22	1697	1686.0	
$[3,4)$	4	171	23	1523	1511.5	
$[4,5)$	5	135	24	1329	1317.0	
$[5,6)$	6	125	107	1170	1116.5	
$[6,7)$	7	83	133	938	871.5	
etc..						

Constructing the lifetable

First, some additional notation for the j-th interval, $[t_{j-1}, t_j)$:

- Midpoint ($t_{mj} = \frac{t_j + t_{j-1}}{2}$) - useful for plotting the density and the hazard function

- Width ($b_j = t_j - t_{j-1}$) needed for calculating the hazard in the j-th interval

Quantities estimated:

- Conditional probability of dying

$$\hat{q}_j = d_j / r'_j$$

- Conditional probability of surviving

$$\hat{p}_j = 1 - \hat{q}_j$$

- Cumulative probability of surviving at t_j:

$$\hat{S}(t_j) \;=\; \prod_{\ell \le j} \hat{p}_\ell = \prod_{\ell \le j} \left(1 - \frac{d_\ell}{r'_\ell} \right)$$

Other estimated quantities:

Most other quantities are estimated at the midpoint of the j-th interval:

- **Hazard** in the j-th interval:

$$\hat{\lambda}(t_{mj}) = \frac{d_j}{b_j(r'_j - d_j/2)}$$

$$= \frac{\hat{q}_j}{b_j(1 - \hat{q}_j/2)}$$

  (the # of deaths in interval divided by the average # of survivors at the midpoint)

- **Density** at the midpoint of the j-th interval:

$$\hat{f}(t_{mj}) = \frac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{b_j} = \frac{\hat{S}(t_{j-1})\,\hat{q}_j}{b_j}$$

Actuarial Estimator for Angina Pectoris Example

The LIFETEST Procedure

Life Table Survival Estimates

Interval [Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure
0	1	456	0	2418.0	0.1886	0.00796	1.0000	0
1	2	226	39	1942.5	0.1163	0.00728	0.8114	0.1886
2	3	152	22	1686.0	0.0902	0.00698	0.7170	0.2830
3	4	171	23	1511.5	0.1131	0.00815	0.6524	0.3476
4	5	135	24	1317.0	0.1025	0.00836	0.5786	0.4214
5	6	125	107	1116.5	0.1120	0.00944	0.5193	0.4807
6	7	83	133	871.5	0.0952	0.00994	0.4611	0.5389
7	8	74	102	671.0	0.1103	0.0121	0.4172	0.5828
8	9	51	68	512.0	0.0996	0.0132	0.3712	0.6288
9	10	42	64	395.0	0.1063	0.0155	0.3342	0.6658
10	11	43	45	298.5	0.1441	0.0203	0.2987	0.7013
11	12	34	53	206.5	0.1646	0.0258	0.2557	0.7443
12	13	18	33	129.5	0.1390	0.0304	0.2136	0.7864
13	14	9	27	81.5	0.1104	0.0347	0.1839	0.8161
14	15	6	23	47.5	0.1263	0.0482	0.1636	0.8364
15	.	0	30	15.0	0	0	0.1429	0.8571

Evaluated at the Midpoint of the Interval

Interval [Lower,	Upper)	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	1	0	5.3313	0.1749	0.1886	0.00796	0.208219	0.009698
1	2	0.00796	6.2499	0.2001	0.0944	0.00598	0.123531	0.008201
2	3	0.00918	6.3432	0.2361	0.0646	0.00507	0.09441	0.007649
3	4	0.00973	6.2262	0.2361	0.0738	0.00543	0.119916	0.009154
4	5	0.0101	6.2185	0.1853	0.0593	0.00495	0.108043	0.009285
5	6	0.0103	5.9077	0.1806	0.0581	0.00503	0.118596	0.010589
6	7	0.0104	5.5962	0.1855	0.0439	0.00469	0.1	0.010963
7	8	0.0105	5.1671	0.2713	0.0460	0.00518	0.116719	0.013545
8	9	0.0106	4.9421	0.2763	0.0370	0.00502	0.10483	0.014659
9	10	0.0107	4.8258	0.4141	0.0355	0.00531	0.112299	0.017301
10	11	0.0109	4.6888	0.4183	0.0430	0.00627	0.155235	0.023602
11	12	0.0111	.	.	0.0421	0.00685	0.17942	0.030646
12	13	0.0114	.	.	0.0297	0.00668	0.149378	0.03511
13	14	0.0118	.	.	0.0203	0.00651	0.116883	0.038894
14	15	0.0123	.	.	0.0207	0.00804	0.134831	0.054919
15	.	0.0133	.	.	.	.	.	.

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
2418	1625	793	32.80

Actuarial method for non-grouped data

Suppose we wish to use the actuarial method, but the data are not already grouped.

<u>Example:</u> Consider the treated nursing home patients, with length of stay (LOS) grouped into 100 day intervals:

SAS Commands for lifetable analysis - grouping data

```
Title 'Actuarial Estimator for nursing home data'; data morris ;
   infile 'ch12.dat' ;
   input los age trt gender marstat hltstat cens ;

data morristr;
   set morris;
   if trt=1;

proc lifetest data=morristr outsurv=survres
              intervals=0 to 1100 by 100 method=act;
   time los*cens(1);
run;

proc print data=survres; run;
```

Actuarial estimator for treated nursing home patients

Actuarial Estimator for Nursing Home Patients

The LIFETEST Procedure

Life Table Survival Estimates

| Interval [Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure
0	100	330	0	712.0	0.4635	0.0187	1.0000	0
100	200	86	0	382.0	0.2251	0.0214	0.5365	0.4635
200	300	65	0	296.0	0.2196	0.0241	0.4157	0.5843
300	400	38	0	231.0	0.1645	0.0244	0.3244	0.6756
400	500	32	1	192.5	0.1662	0.0268	0.2711	0.7289
500	600	13	0	160.0	0.0813	0.0216	0.2260	0.7740
600	700	13	0	147.0	0.0884	0.0234	0.2076	0.7924
700	800	10	30	119.0	0.0840	0.0254	0.1893	0.8107
800	900	4	29	79.5	0.0503	0.0245	0.1734	0.8266
900	1000	4	30	46.0	0.0870	0.0415	0.1647	0.8353
1000	1100	0	27	13.5	0	0	0.1503	0.8497

Evaluated at the Midpoint of the Interval

Interval [Lower,	Upper)	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	100	0	130.2	15.5136	0.00463	0.000187	0.006033	0.000317
100	200	0.0187	306.2	30.4597	0.00121	0.000122	0.002537	0.000271
200	300	0.0185	398.8	65.7947	0.000913	0.000108	0.002467	0.000304
300	400	0.0175	617.0	74.5466	0.000534	0.000084	0.001792	0.00029
400	500	0.0167	.	.	0.000451	0.000078	0.001813	0.000319
500	600	0.0157	.	.	0.000184	0.000050	0.000847	0.000235
600	700	0.0152	.	.	0.000184	0.000050	0.000925	0.000256
700	800	0.0147	.	.	0.000159	0.000050	0.000877	0.000277
800	900	0.0143	.	.	0.000087	0.000043	0.000516	0.000258
900	1000	0.0142	.	.	0.000143	0.000070	0.000909	0.000454
1000	1100	0.0147	.	.	0	.	0	.

2.12 (3) Estimating $S(t)$ via the cumulative hazard

Nelson-Aalen estimator

Suppose we want to estimate $\Lambda(t) = \int_0^t \lambda(u)du$, the cumulative hazard at time t, in the continuous case. Just as we did for the KM, think of dividing the observed time span of the study into a series of fine intervals so that there is only one event per interval:

$$\boxed{|\quad|\quad|\quad|D\quad|\quad|C\quad|\quad|C\quad|D\quad|D\quad|D\quad|}$$
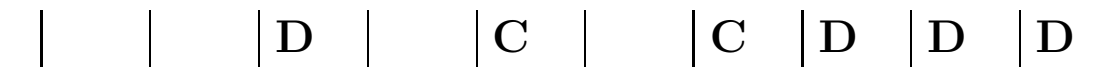
$\Lambda(t)$ can then be approximated by a sum:

$$\Lambda(t) \approx \sum_j \lambda_j \Delta$$

where the sum is over intervals, λ_j is the value of the hazard in the j-th interval and Δ is the width of each interval. Since $\hat{\lambda}\Delta$ is approximately the probability of dying in the interval, we can further approximate by

$$\hat{\Lambda}(t) = \sum_{t_j \leq t} d_j/r_j = \hat{\Lambda}_{NA}(t)$$

which is called the Nelson-Aalen estimator.

			D		C		C	D	D	D
r_j	n	n	n	n-1	n-1	n-2	n-2	n-3	n-4	
d_j	0	0	1	0	0	0	0	1	1	
c_j	0	0	0	0	1	0	1	0	0	
$\hat{\lambda}(t_j)$	0	0	1/n	0	0	0	0	$\frac{1}{n-3}$	$\frac{1}{n-4}$	
$\hat{\Lambda}(t_j)$	0	0	1/n	1/n	1/n	1/n	1/n	$+\frac{1}{n-3}$	$+\frac{1}{n-4}$	

Once we have $\hat{\Lambda}_{NA}(t)$, we can also find another estimator of $S(t)$ (**Fleming-Harrington**):

$$\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t))$$

In general, this estimator of the survival function will be close to the Kaplan-Meier estimator, $\hat{S}_{KM}(t)$

We can also go the other way ... we can take the Kaplan-Meier estimate of $S(t)$, and use it to calculate an alternative estimate of the cumulative hazard function:

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

Nelson-Aalen Cumulative Hazard Estimator

Cumulative Hazard Function

Time of remission (in weeks)
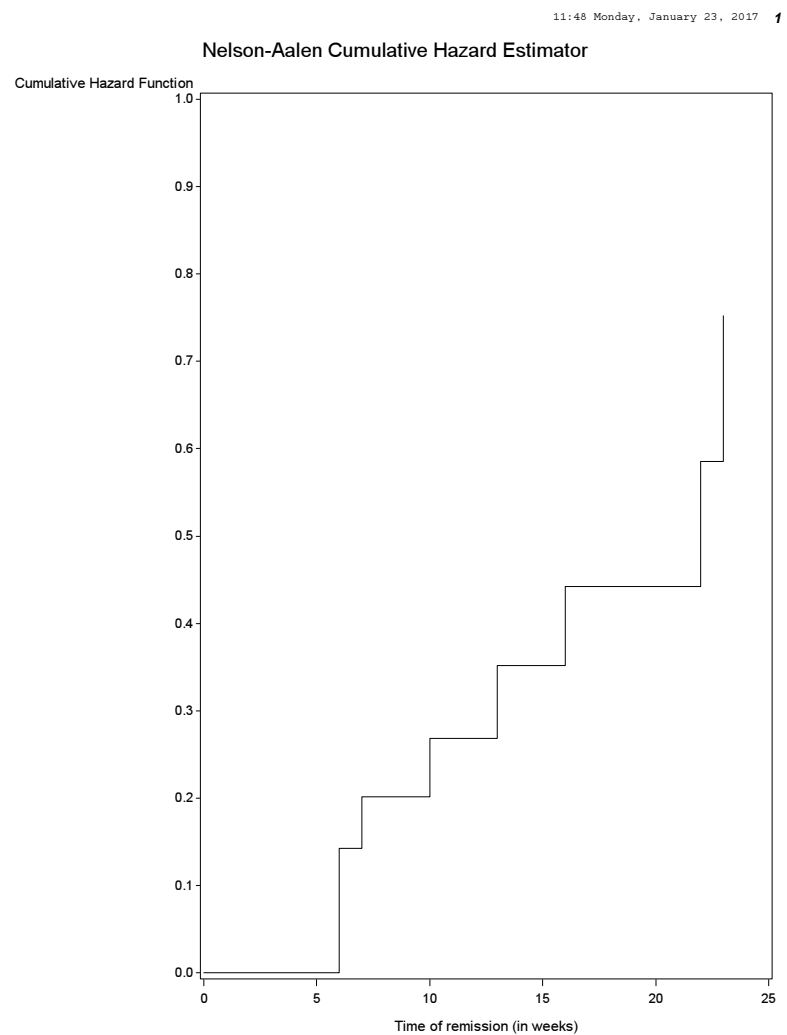
R Commands for Fleming-Harrington Estimator:
(Nursing home data: females, untreated, married, healthy)

Fleming-Harrington:

```
>fh<-survfit(Surv(los,cens)~1,type="f",conf.type="log-log")
>fh
```

```
 95 percent confidence interval is of type "log-log"
```

time	n.risk	n.event	survival	std.dev	lower 95% CI	upper 95% CI
14	12	1	0.9200444	0.08007959	0.5244209125	0.9892988
24	11	1	0.8400932	0.10845557	0.4750041174	0.9600371
25	10	1	0.7601478	0.12669130	0.4055610500	0.9200425
38	9	1	0.6802101	0.13884731	0.3367907188	0.8724502
64	8	1	0.6002833	0.14645413	0.2718422278	0.8187596
89	7	1	0.5203723	0.15021856	0.2115701242	0.7597900
113	6	1	0.4404857	0.15045450	0.1564397006	0.6960354
123	5	1	0.3606392	0.14723033	0.1069925657	0.6278888
149	4	1	0.2808661	0.14043303	0.0640979523	0.5560134
168	3	1	0.2012493	0.12990589	0.0293208029	0.4827590
185	2	1	0.1220639	0.11686728	0.0058990525	0.4224087
234	1	1	0.0449048	0.06216787	0.0005874321	0.2740658

Kaplan-Meier:

```
>km<-survfit(Surv(los,cens)~1,conf.type="log-log")
>km
```

```
 95 percent confidence interval is of type "log-log"
 time n.risk n.event  survival    std.dev lower 95% CI upper 95% CI
   14     12       1 0.91666667 0.07978559  0.538977181    0.9878256
   24     11       1 0.83333333 0.10758287  0.481714942    0.9555094
   25     10       1 0.75000000 0.12500000  0.408415913    0.9117204
   38      9       1 0.66666667 0.13608276  0.337018933    0.8597118
   64      8       1 0.58333333 0.14231876  0.270138924    0.8009402
   89      7       1 0.50000000 0.14433757  0.208477143    0.7360731
  113      6       1 0.41666667 0.14231876  0.152471264    0.6653015
  123      5       1 0.33333333 0.13608276  0.102703980    0.5884189
  149      4       1 0.25000000 0.12500000  0.060144556    0.5047588
  168      3       1 0.16666667 0.10758287  0.026510427    0.4129803
  185      2       1 0.08333333 0.07978559  0.005052835    0.3110704
  234      1       1 0.00000000         NA           NA           NA
```

Contents