Estimating the Survival Function

One-sample nonparametric methods:

We will consider three methods for estimating a survivorship function

$$S(t) = Pr(T \geq t)$$

without resorting to parametric methods:

(1) **Kaplan-Meier**

(2) **Life-table** (Actuarial Estimator)

(3) **Cumulative hazard estimator**

The Kaplan-Meier Estimator

The Kaplan-Meier (or KM) estimator is probably the most popular approach. It can be justified from several perspectives:

- product limit estimator
- likelihood justification
- redistribute to the right estimator

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

Motivation:

First, consider an example where there is no censoring.

The following are times of remission (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes):

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 S(10), the probability that an individual survives to time 10 or later?

What about $\tilde{S}(8)$? Is it $\frac{12}{21}$ or $\frac{8}{21}$?

Let's construct a table of $\tilde{S}(t)$:

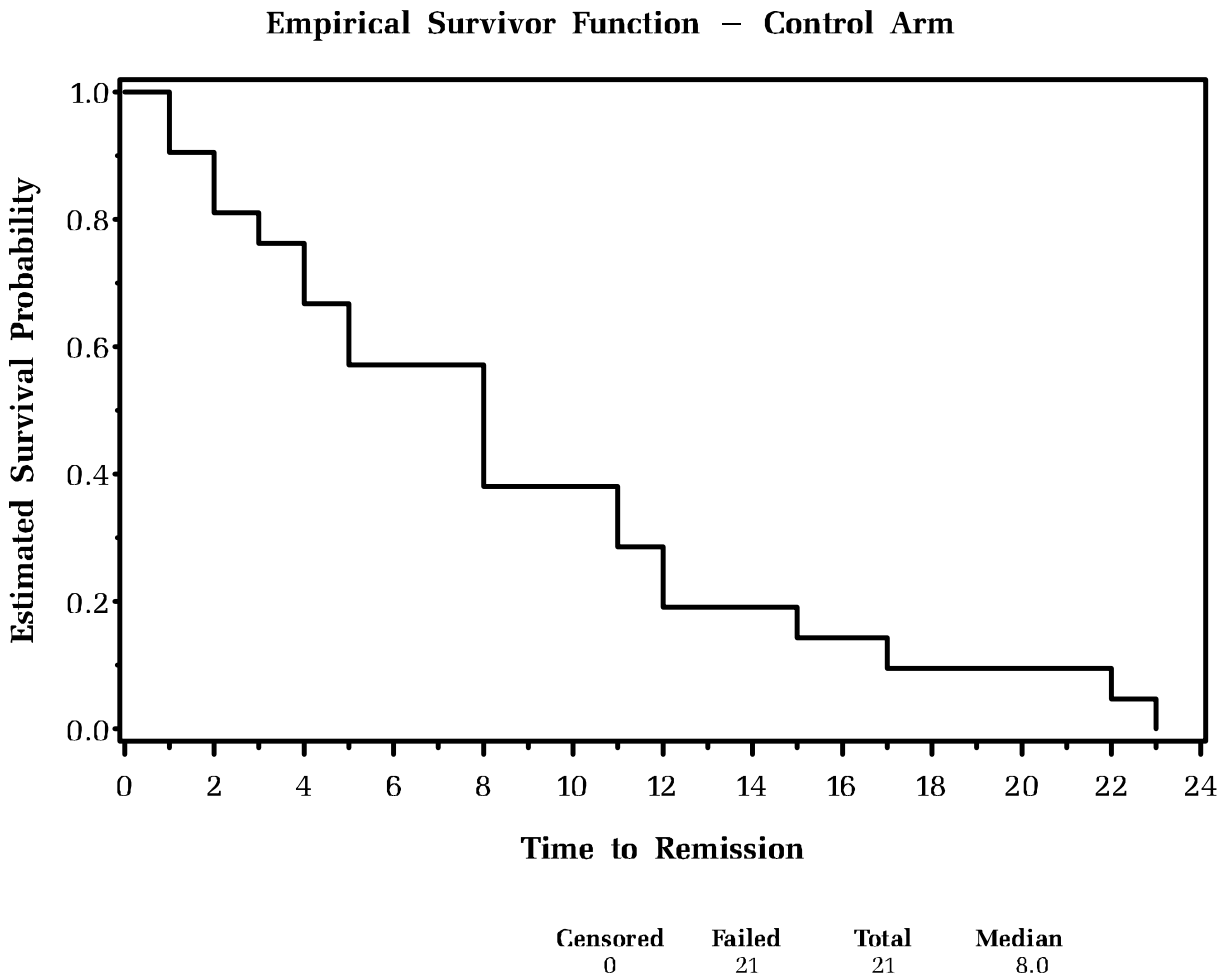
Values of t	$\hat{S}(t)$
$t \leq 1$	21/21=1.000
$1 < t \leq 2$	19/21=0.905
$2 < t \leq 3$	17/21=0.809
$3 < t \leq 4$	
$4 < t \leq 5$	
$5 < t \leq 8$	
$8 < t \leq 11$	
$11 < t \leq 12$	
$12 < t \leq 15$	
$15 < t \leq 17$	
$17 < t \leq 22$	
$22 < t \leq 23$	

Empirical Survival Function:

When there is no censoring, the general formula is:

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

Example for leukemia data (control arm):



What if there is censoring?

Consider the treated group from Table 1.1 of Cox and Oakes:

$$6^+, 6, 6, 6, 7, 9^+, 10^+, 10, 11^+, 13, 16, 17^+$$
$$19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$$

[Note: times with $^+$ are right censored]

We know S(6)= 21/21, because everyone survived at least until time 6 or greater. But, we can't say S(7) = 17/21, because we don't know the status of the person who was censored at time 6.

In a 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**.

A quick review 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 $P(B)$:

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

Extension to more than 2 events:

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:

$$\begin{aligned} P(A_1 \cap A_2 ... \cap A_k) &= P(A_k | A_{k-1} \cap ... \cap A_1) \times \\ &\quad \times P(A_{k-1} | A_{k-2} \cap ... \cap A_1) \\ &\quad ... \\ &\quad \times P(A_2 | A_1) \\ &\quad \times P(A_1) \end{aligned}$$

Now, let's apply these ideas to estimate $S(t)$:

Suppose $a_k < t \leq a_{k+1}$. Then

$$\begin{aligned}
S(t) &= P(T \geq a_{k+1}) \\
&= P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{k+1}) \\
&= P(T \geq a_1) \times \prod_{j=1}^{k} P(T \geq a_{j+1} | T \geq a_j) \\
&= \prod_{j=1}^{k} [1 - P(T = a_j | T \geq a_j)] \\
&= \prod_{j=1}^{k} [1 - \lambda_j]
\end{aligned}$$

So,

$$\begin{aligned}\hat{S}(t) &\cong \prod_{j=1}^{k}\left(1-\frac{d_j}{r_j}\right)\\ &= \prod_{j:a_j<t}\left(1-\frac{d_j}{r_j}\right)\end{aligned}$$

d_j is the number of deaths at a_j

r_j is the number at risk at a_j

Intuition behind the Kaplan-Meier Estimator

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:

Using the law of conditional probability,

$$Pr(T \geq t) \quad = \quad \prod_j Pr(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j)$$

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

There are 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:

$$1 - \frac{d_j}{r_j}$$

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.

The Kaplan-Meier - product-limit - estimator

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 $S(t)$.

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) $S(t) = Pr(T \geq t)$ is:

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

where,

- $\tau_1, ...\tau_K$ are the 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 at or after 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: two useful formulas are:

$$(1) \quad r_j = r_{j-1} - d_{j-1} - c_{j-1}$$

$$(2) \quad r_j = \sum_{l \geq j} (c_l + d_l)$$

Calculating the KM - Cox and Oakes example

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

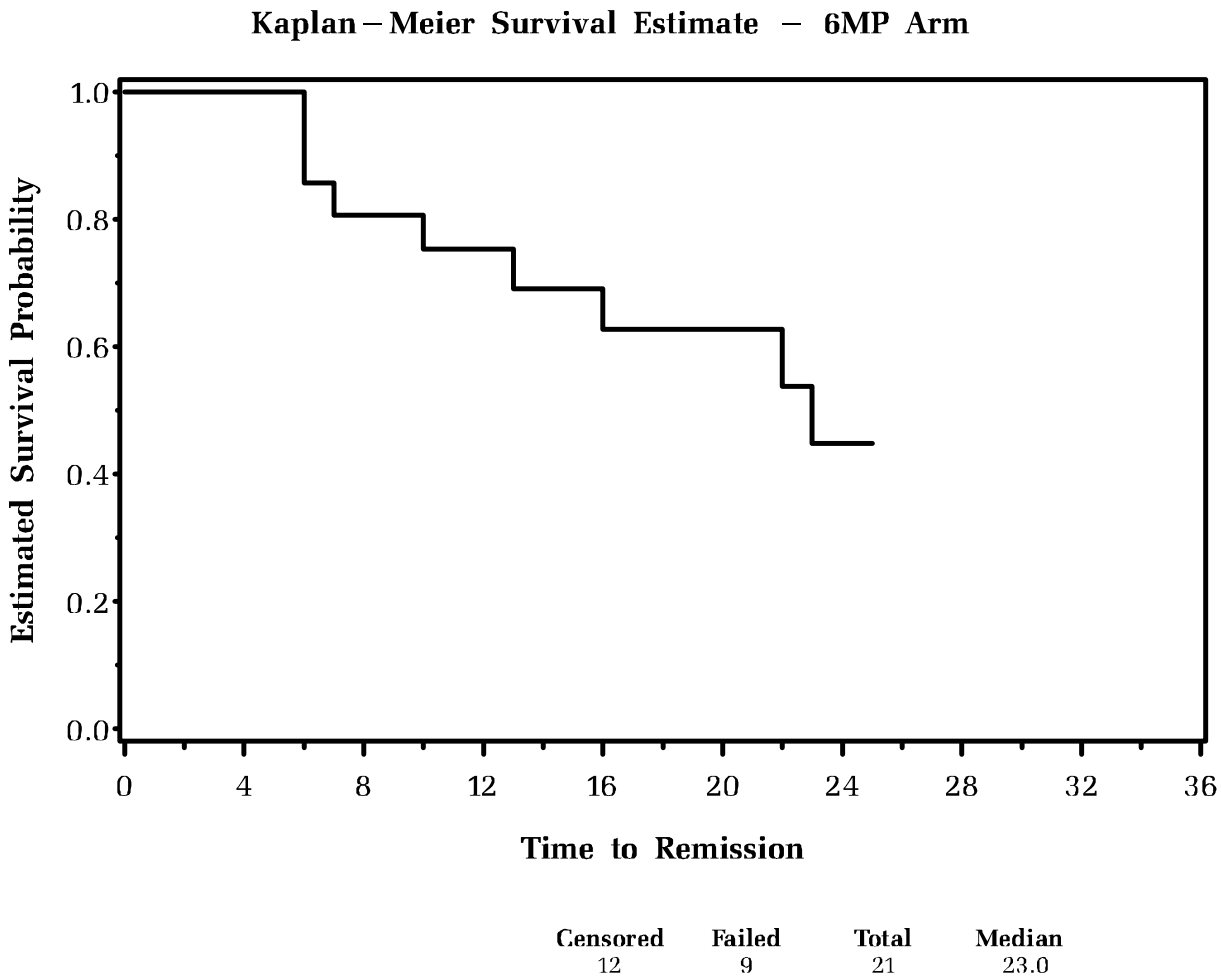
τ_j	d_j	c_j	r_j	$1-(d_j/r_j)$	$\hat{S}(\tau_j^+)$
6	3	1	21	$\frac{18}{21}=0.857$	
7	1	0	17		
9	0	1	16		
10					
11					
13					
16					
17					
19					
20					
22					
23					

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

KM plot for treated leukemia patients



Censored	Failed	Total	Median
12	9	21	23.0

Note: most statistical software packages summarize the KM survival function at τ_j^+, i.e., *just after* the time of the j-th failure.

In other words, they provide $\hat{S}(\tau_j^+)$.

When there is no censoring, the empirical survival estimate would then be:

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

Output from STATA KM Estimator:

```
    failure time:  weeks
  failure/censor:  remiss
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
6	21	3	1	0.8571	0.0764	0.6197	0.9516
7	17	1	0	0.8067	0.0869	0.5631	0.9228
9	16	0	1	0.8067	0.0869	0.5631	0.9228
10	15	1	1	0.7529	0.0963	0.5032	0.8894
11	13	0	1	0.7529	0.0963	0.5032	0.8894
13	12	1	0	0.6902	0.1068	0.4316	0.8491
16	11	1	0	0.6275	0.1141	0.3675	0.8049
17	10	0	1	0.6275	0.1141	0.3675	0.8049
19	9	0	1	0.6275	0.1141	0.3675	0.8049
20	8	0	1	0.6275	0.1141	0.3675	0.8049
22	7	1	0	0.5378	0.1282	0.2678	0.7468
23	6	1	0	0.4482	0.1346	0.1881	0.6801
25	5	0	1	0.4482	0.1346	0.1881	0.6801
32	4	0	2	0.4482	0.1346	0.1881	0.6801
34	2	0	1	0.4482	0.1346	0.1881	0.6801
35	1	0	1	0.4482	0.1346	0.1881	0.6801

Two Other Justifications for KM Estimator

I. Likelihood-based derivation (Cox and Oakes)

For a discrete failure time variable, define:

d_j	number of failures at a_j
r_j	number of individuals at risk at a_j (including those censored at a_j).
λ_j	Pr(death) in j-th interval (conditional on survival to start of interval)

The likelihood is that of g independent binomials:

$$L(\lambda) = \prod_{j=1}^{g} \lambda_j^{d_j} (1-\lambda_j)^{r_j - d_j}$$

Therefore, the **maximum likelihood estimator** of λ_j is:

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

Now we plug in the MLE's of λ to estimate S(t)::

$$\begin{aligned}\hat{S}(t) &= \prod_{j:a_j<t} (1-\hat{\lambda}_j) \\ &= \prod_{j:a_j<t} \left(1-\frac{d_j}{r_j}\right)\end{aligned}$$

II. Redistribute to the right justification (Efron, 1967)

In the absence of censoring, $\hat{S}(t)$ is just the proportion of individuals with $T \geq t$. The idea behind Efron's approach is to spread the contributions of censored observations out over all the possible times to their right.

Algorithm:

- Step (1): arrange the n observed times (deaths or censorings) in increasing order. If there are ties, put censored after deaths.
- Step (2): Assign weight $(1/n)$ to each time.
- Step (3): Moving from left to right, each time you encounter a censored observation, distribute its mass to all times to its right.
- Step (4): Calculate $\hat{S}_j$ by subtracting the final weight for time j from $\hat{S}_{j-1}$

Example of "redistribute to the right" algorithm

Consider the following event times:

$$2,\ \ 2.5+,\ \ 3,\ \ 3,\ \ 4,\ \ 4.5+,\ \ 5,\ \ 6,\ \ 7$$

The algorithm goes as follows:

(Step 1) Times	Step 2	Step 3a	Step 3b	(Step 4) $\hat{S}(\tau_j)$
2	1/9=0.11			0.889
2.5^+	1/9=0.11	0		0.889
3	2/9=0.22	0.25		0.635
4	1/9=0.11	0.13		0.508
4.5^+	1/9=0.11	0.13	0	0.508
5	1/9=0.11	0.13	0.17	0.339
6	1/9=0.11	0.13	0.17	0.169
7	1/9=0.11	0.13	0.17	0.000

This comes out the same as the product limit approach.

Properties of the KM estimator

In the case of no censoring:

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

where n is the number of individuals in the study.

This is just like an estimated probability from a binomial distribution, so we have:

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

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, then we can construct (pointwise) $(1-\alpha)\%$ confidence bands about $\hat{S}(t)$:

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

Greenwood's formula (Collett 2.1.3)

We can think of the KM estimator as

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

where $\hat{\lambda}_j = d_j/r_j$. Since the $\hat{\lambda}_j$'s are just 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

$$var(\hat{\lambda}_j) \approx \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{r_j}$$

The $\hat{\lambda}_j$'s are independent in large samples. Since $\hat{S}(t)$ is a function of the λ_j's, we can estimate its variance using the **delta method**:

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

Two specific examples of the delta method:

(A) $Z = log(Y)$

$$\text{then} \quad Z \sim N\left[\log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2\right]$$

(B) $Z = exp(Y)$

$$\text{then} \quad Z \sim N\left[e^{\mu}, [e^{\mu}]^2 \sigma^2\right]$$

The examples above use the following results from calculus:

$$\frac{d}{dx} \log u = \frac{1}{u} \left(\frac{du}{dx}\right)$$

$$\frac{d}{dx} e^u = e^u \left(\frac{du}{dx}\right)$$

Greenwood's formula (continued)

Instead of dealing with $\hat{S}(t)$ directly, we will look at its log:

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

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

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

By (A)

$$\begin{aligned} var(\log[\hat{S}(t)]) &= \sum_{j:\tau_j<t} \left(\frac{1}{1-\hat{\lambda}_j}\right)^2 var(\hat{\lambda}_j) \\ &= \sum_{j:\tau_j<t} \left(\frac{1}{1-\hat{\lambda}_j}\right)^2 \hat{\lambda}_j(1-\hat{\lambda}_j)/r_j \\ &= \sum_{j:\tau_j<t} \frac{\hat{\lambda}_j}{(1-\hat{\lambda}_j)r_j} = \sum_{j:\tau_j<t} \frac{d_j}{(r_j-d_j)r_j} \end{aligned}$$

Since $\hat{S}(t) = \exp[\log[\hat{S}(t)]]$, (by B),

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

Greenwood's Formula:

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:\tau_j<t} \frac{d_j}{(r_j-d_j)r_j}$$

Back to confidence intervals

For a 95% confidence interval, we could use

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

where $se[\hat{S}(t)]$ is calculated 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 right range when we transform back.

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 $\hat{S}(t)$ and bounds:

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

- Taking the opposite:

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

- Taking the log again:

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

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 $[\hat{L}(t) - A, \hat{L}(t) + A]$

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

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

(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

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

What is A?

- A is $1.96\, se(\hat{L}(t))$
- To calculate this, we need to calculate

$$var(\hat{L}(t)) \quad = \quad var\left[\log(-\log(\hat{S}(t)))\right]$$

- From our previous calculations, we know

$$var(\log[\hat{S}(t)]) \quad = \quad \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

- Applying the delta method as in example (A), we get:

$$\begin{aligned} var(\hat{L}(t)) &= var(\log(-\log[\hat{S}(t)])) \\ &= \frac{1}{[\log\ \hat{S}(t)]^2} \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j} \end{aligned}$$

- We take the square root of the above to get $se(\hat{L}(t))$, and then form the confidence intervals as:

$$\hat{S}(t)^{e^{\pm 1.96\ se(\hat{L}(t))}}$$

- This is the approach that Stata uses. Splus also gives an option to calculate these bounds.

Summary of Confidence Intervals on $S(t)$

- Calculate $\hat{S}(t) \pm 1.96\, se[\hat{S}(t)]$ where $se[\hat{S}(t)]$ is calculated using Greenwood's formula, and replace negative lower bounds by 0 and upper bounds greater than 1 by 1 (not very satisfactory).
  - Recommended by Collett
  - This is the default using SAS
- Use a log transformation to stabilize the variance and allow for non-symmetric confidence intervals. This is what is normally done for the confidence interval of an estimated odds ratio.
  - Use $var[\log(\hat{S}(t))] = \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$ already calculated as part of Greenwood's formula
  - This is the default in Splus
- Use the log-log transformation just described
  - Somewhat complicated, but always yields proper bounds
  - This is the default in Stata!

Software for Kaplan-Meier Curves

- Stata - stset and sts commands
- SAS - PROC LIFETEST
- Splus - surv.fit(time,censor)

Defaults for Confidence Interval Calculations

- Stata - "log-log" $\Rightarrow \hat{L}(t) \pm 1.96\, se[\hat{L}(t)]$
  where $L(t) = \log[-\log(S(t))]$
- SAS - "plain" $\Rightarrow \hat{S}(t) \pm 1.96\, se[\hat{S}(t)]$
- Splus - "log" $\Rightarrow \log S(t) \pm 1.96\, se[\log(\hat{S}(t))]$
  but Splus will also give either of the other two options if you request them.

Stata Commands

Create a file called "leukemia.dat" with the raw data, with a column for treatment, weeks to relapse (i.e., duration of remission), and relapse status:

```
.infile trt remiss status using leukemia.dat

.stset remiss status          (sets up a failure time dataset,
                               with failtime status in that order,
                               type help stset to get details)

.sts list                    (estimated S(t), se[S(t)], and 95% CI)

.sts graph, saving(kmtrt)    (creates a Kaplan-Meier plot, and
                               saves the plot in file kmtrt.gph,
                               type ‘‘help gphdot’’ to get some
                               printing instructions)

.graph using kmtrt           (redisplays the graph at any later time)
```

If the dataset has already been created and loaded into Stata, then you can substitute the following commands for initializing the data:

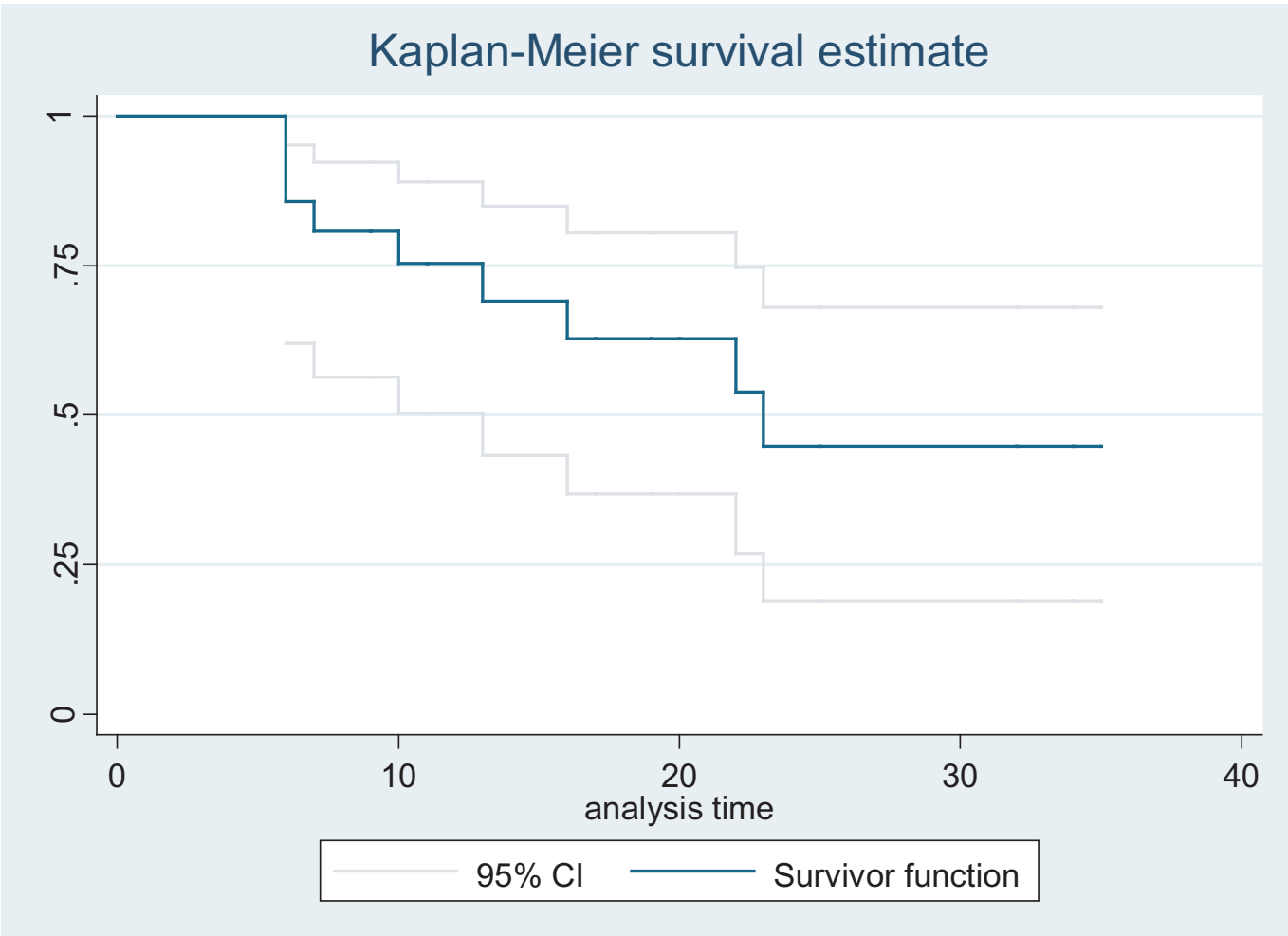
```
.use leukem                    (finds Stata dataset leukem.dta)

.describe                      (provides a description of the dataset)

.stset remiss status           (declares data to be failure type)

.stdes                         (gives a description of the survival dataset)
```

STATA Output for Treated Leukemia Patients:

```
.use leukem
.stset remiss status if trt==1
.sts list
    failure time:  remiss
  failure/censor:  status
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
6	21	3	1	0.8571	0.0764	0.6197	0.9516
7	17	1	0	0.8067	0.0869	0.5631	0.9228
9	16	0	1	0.8067	0.0869	0.5631	0.9228
10	15	1	1	0.7529	0.0963	0.5032	0.8894
11	13	0	1	0.7529	0.0963	0.5032	0.8894
13	12	1	0	0.6902	0.1068	0.4316	0.8491
16	11	1	0	0.6275	0.1141	0.3675	0.8049
17	10	0	1	0.6275	0.1141	0.3675	0.8049
19	9	0	1	0.6275	0.1141	0.3675	0.8049
20	8	0	1	0.6275	0.1141	0.3675	0.8049
22	7	1	0	0.5378	0.1282	0.2678	0.7468
23	6	1	0	0.4482	0.1346	0.1881	0.6801
25	5	0	1	0.4482	0.1346	0.1881	0.6801
32	4	0	2	0.4482	0.1346	0.1881	0.6801
34	2	0	1	0.4482	0.1346	0.1881	0.6801
35	1	0	1	0.4482	0.1346	0.1881	0.6801

KM Survival Estimate and Confidence intervals



```
. stsum

         failure _d:  status
   analysis time _t:  remiss

      |                incidence    no. of  |------ Survival time -----|
      | time at risk      rate     subjects      25%       50%       75%
------+-------------------------------------------------------------------
total |          359   .0250696        21       13        23         .
```

Means, Medians, Quantiles based on the KM

- **Mean:** $\sum_{j=1}^{k} \tau_j \, Pr(T = \tau_j)$
- **Median** - by definition, this is the time, τ, such that $S(\tau) = 0.5$. However, in practice, it 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 visually on the graph to the left.

- **Lower quartile (25^{th} percentile):**
  the smallest time (LQ) such that $\hat{S}(LQ) \leq 0.75$
- **Upper quartile (75^{th} percentile):**
  the smallest time (UQ) such that $\hat{S}(UQ) \leq 0.25$

Stata command for median and quartiles: `stsum`

(2) The Lifetable Estimator of Survival:

We said that we would consider the following three methods for estimating a survivorship function

$$S(t) = Pr(T \geq t)$$

without resorting to parametric methods:

(1) $\surd$ **Kaplan-Meier**

(2) $\Longrightarrow$ **Life-table** (Actuarial Estimator)

(3) $\Longrightarrow$ **Cumulative hazard estimator**

(2) The Lifetable or 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. Lee (section 4.2) provides a good description of lifetable methods, and distinguishes several types according to the data sources:

Population Life Tables

- **cohort life table -** 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.
- **current life table -** 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 Life Tables.

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.

Notation

- the j-th time interval is $[t_{j-1}, t_j)$
- 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 entering the interval

Example: 2418 Males with Angina Pectoris (Lee, p.91)

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..						

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

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

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 within the interval, given survival to the beginning of it.

What should we do with the censored people?

We can assume that censorings occur:

- at the beginning of each interval: $r'_j = r_j - c_j$
- at the end of each interval: $r'_j = r_j$
- **on average halfway through** 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.

Constructing the lifetable

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

- **Midpoint** (t_{mj}) - 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 is $\hat{q}_j = d_j / r'_j$
- Conditional probability of surviving is $\hat{p}_j = 1 - \hat{q}_j$

- Cumulative probability of surviving at t_j:

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

Some important points to note:

- Because the intervals are defined as $[t_{j-1}, t_j)$, the first interval typically starts with $t_0 = 0$.

- Stata estimates the survival function at the right-hand endpoint of each interval, i.e., $S(t_j)$

- However, SAS estimates the survival function at the left-hand endpoint, $S(t_{j-1})$.

- The implication in SAS is that $\hat{S}(t_0) = 1$ and $\hat{S}(t_1) = p_1$

Other quantities 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 number of deaths in the interval divided by the average number of survivors at the midpoint

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

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

  Note: Another way to get this is:

$$\hat{f}(t_{mj}) = \hat{\lambda}(t_{mj})\hat{S}(t_{mj}) = \hat{\lambda}(t_{mj})[\hat{S}(t_j) + \hat{S}(t_{j-1})]/2$$

Constructing the Lifetable using Stata

Uses the `ltable` command.

If the raw data are already grouped, then the `freq` statement must be used when reading the data.

```
. infile years status count using angina.dat
(32 observations read)
. ltable years status [freq=count]
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	1	2418	456	0	0.8114	0.0080	0.7952	0.8264
1	2	1962	226	39	0.7170	0.0092	0.6986	0.7346
2	3	1697	152	22	0.6524	0.0097	0.6329	0.6711
3	4	1523	171	23	0.5786	0.0101	0.5584	0.5981
4	5	1329	135	24	0.5193	0.0103	0.4989	0.5392
5	6	1170	125	107	0.4611	0.0104	0.4407	0.4813
6	7	938	83	133	0.4172	0.0105	0.3967	0.4376
7	8	722	74	102	0.3712	0.0106	0.3505	0.3919
8	9	546	51	68	0.3342	0.0107	0.3133	0.3553
9	10	427	42	64	0.2987	0.0109	0.2775	0.3201
10	11	321	43	45	0.2557	0.0111	0.2341	0.2777
11	12	233	34	53	0.2136	0.0114	0.1917	0.2363
12	13	146	18	33	0.1839	0.0118	0.1614	0.2075
13	14	95	9	27	0.1636	0.0123	0.1404	0.1884
14	15	59	6	23	0.1429	0.0133	0.1180	0.1701
15	16	30	0	30	0.1429	0.0133	0.1180	0.1701

It is also possible to get estimates of the hazard function, $\hat{\lambda}_j$, and its standard error using the "`hazard`" option:

```
. ltable years status [freq=count], hazard
```

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf	Int]
0	1	2418	0.1886	0.0080	0.2082	0.0097	0.1892	0.2272
1	2	1962	0.2830	0.0092	0.1235	0.0082	0.1075	0.1396
2	3	1697	0.3476	0.0097	0.0944	0.0076	0.0794	0.1094
3	4	1523	0.4214	0.0101	0.1199	0.0092	0.1020	0.1379
4	5	1329	0.4807	0.0103	0.1080	0.0093	0.0898	0.1262
5	6	1170	0.5389	0.0104	0.1186	0.0106	0.0978	0.1393
6	7	938	0.5828	0.0105	0.1000	0.0110	0.0785	0.1215
7	8	722	0.6288	0.0106	0.1167	0.0135	0.0902	0.1433
8	9	546	0.6658	0.0107	0.1048	0.0147	0.0761	0.1336
9	10	427	0.7013	0.0109	0.1123	0.0173	0.0784	0.1462
10	11	321	0.7443	0.0111	0.1552	0.0236	0.1090	0.2015
11	12	233	0.7864	0.0114	0.1794	0.0306	0.1194	0.2395
12	13	146	0.8161	0.0118	0.1494	0.0351	0.0806	0.2182
13	14	95	0.8364	0.0123	0.1169	0.0389	0.0407	0.1931
14	15	59	0.8571	0.0133	0.1348	0.0549	0.0272	0.2425
15	16	30	0.8571	0.0133	0.0000	.	.	.

There is also a "`failure`" option which gives the number of failures (like the default), and also provides a 95% confidence interval on the cumulative failure probability.

Suppose we wish to use the actuarial method, but the data do not come grouped.

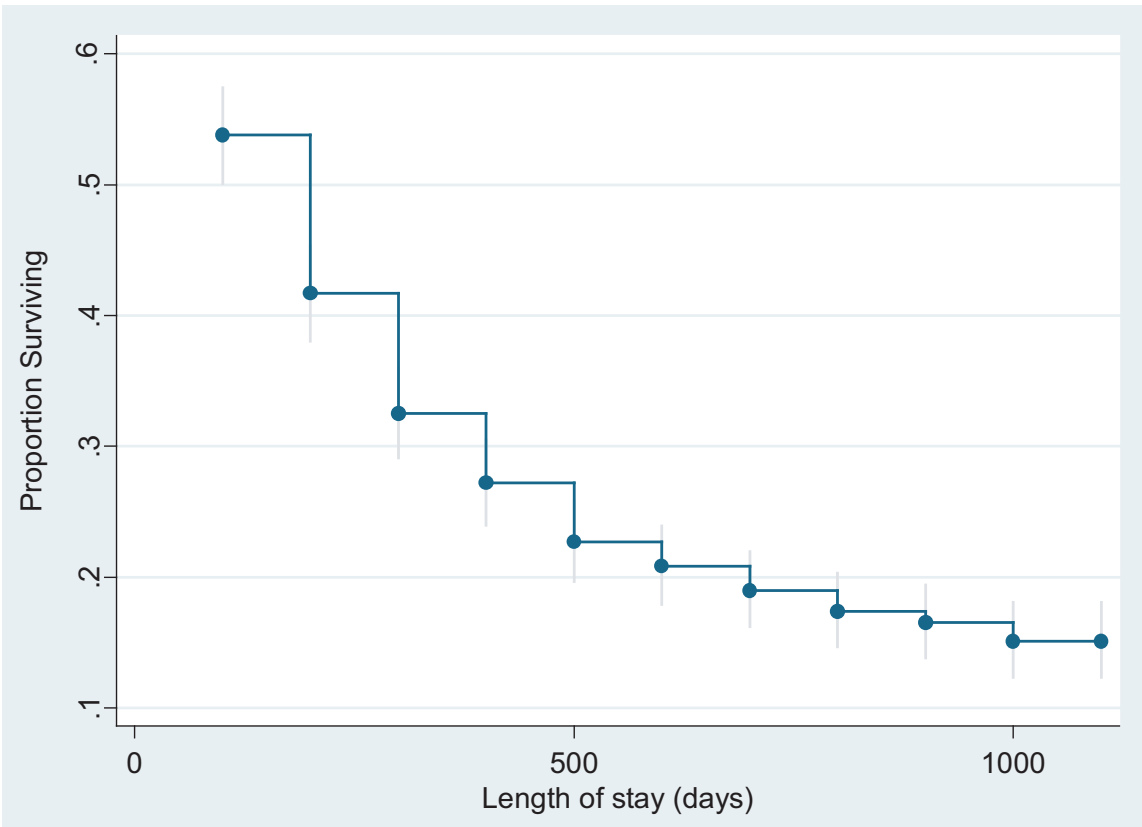
Consider the treated nursing home patients, with length of stay (los) grouped into 100 day intervals:

```
.use nurshome

.drop if rx==0                    (keep only the treated patients)
 (881 observations deleted)

.stset los fail

.ltable los fail, intervals(100)

              Beg.                                  Std.
   Interval   Total Deaths   Lost   Survival   Error     [95% Conf. Int.]
-------------------------------------------------------------------------
     0   100     710    328      0    0.5380   0.0187    0.5006    0.5739
   100   200     382     86      0    0.4169   0.0185    0.3805    0.4529
   200   300     296     65      0    0.3254   0.0176    0.2911    0.3600
   300   400     231     38      0    0.2718   0.0167    0.2396    0.3050
   400   500     193     32      1    0.2266   0.0157    0.1966    0.2581
   500   600     160     13      0    0.2082   0.0152    0.1792    0.2388
   600   700     147     13      0    0.1898   0.0147    0.1619    0.2195
   700   800     134     10     30    0.1739   0.0143    0.1468    0.2029
   800   900      94      4     29    0.1651   0.0143    0.1383    0.1941
   900  1000      61      4     30    0.1508   0.0147    0.1233    0.1808
  1000  1100      27      0     27    0.1508   0.0147    0.1233    0.1808
-------------------------------------------------------------------------
```

Examples for Nursing home data:

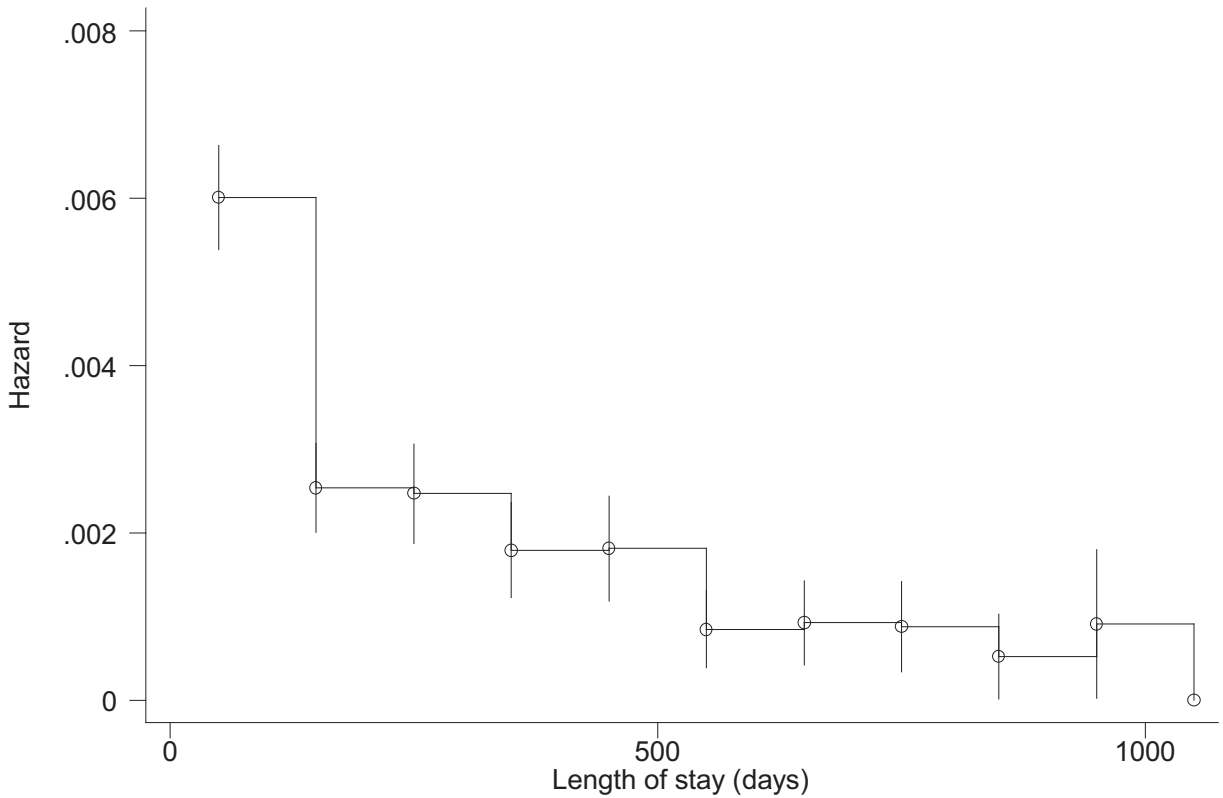
Estimated Survival:

```
ltable los fail,  intervals(100) graph connect(J)
```



Estimated hazard:

```
version 7
ltable los fail, hazard intervals(100) graph connect(J)
```



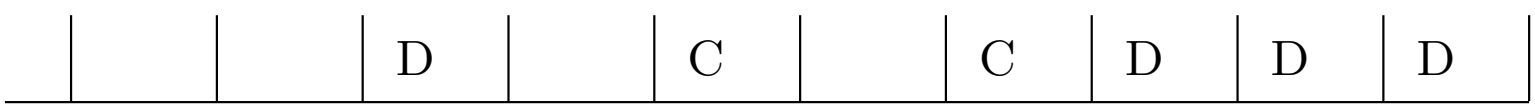
Note: This command is not supported by version 9.0 in Stata.

(3) Estimating 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.

Just as we did for the KM, think of dividing the observed timespan of the study into a series of fine intervals so that there is only one event per interval:



$\Lambda(t)$ can then be approximated by a sum:

$$\hat{\Lambda}(t) = \sum_j \lambda_j \Delta$$

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

$$\hat{\Lambda}(t) = \sum_j d_j / r_j$$

It follows that $\Lambda(t)$ will change only at death times, and hence we write the Nelson-Aalen estimator as:

$$\hat{\Lambda}_{NA}(t) = \sum_{j:\tau_j < t} d_j / r_j$$

The Fleming-Harrington (FH) 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			

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:

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

Stata commands for FH Survival Estimate

Say we want to obtain the Fleming-Harrington estimate of the survival function for married females, in the healthiest initial subgroup, who are randomized to the untreated group of the nursing home study.

First, we use the following commands to calculate the Nelson-Aalen cumulative hazard estimator:

```
. use nurshome

. keep if rx==0 & gender==0 & health==2 & married==1
 (1579 observations deleted)
```

```
. sts list, na

        failure _d:  fail
  analysis time _t:  los
```

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
14	12	1	0	0.0833	0.0833	0.0117	0.5916
24	11	1	0	0.1742	0.1233	0.0435	0.6976
25	10	1	0	0.2742	0.1588	0.0882	0.8530
38	9	1	0	0.3854	0.1938	0.1438	1.0326
64	8	1	0	0.5104	0.2306	0.2105	1.2374
89	7	1	0	0.6532	0.2713	0.2894	1.4742
113	6	1	0	0.8199	0.3184	0.3830	1.7551
123	5	1	0	1.0199	0.3760	0.4952	2.1006
149	4	1	0	1.2699	0.4515	0.6326	2.5493
168	3	1	0	1.6032	0.5612	0.8073	3.1840
185	2	1	0	2.1032	0.7516	1.0439	4.2373
234	1	1	0	3.1032	1.2510	1.4082	6.8384

After generating the Nelson-Aalen estimator, we manually have to create a variable for the survival estimate:

```
. sts gen nelson=na
. gen sfh=exp(-nelson)
. list sfh

          sfh
  1.  .9200444
  2.  .8400932
  3.  .7601478
  4.  .6802101
  5.  .6002833
  6.  .5203723
  7.  .4404857
  8.  .3606392
  9.  .2808661
 10.  .2012493
 11.  .1220639
 12.  .0449048
```

Additional built-in functions can be used to generate 95% confidence intervals on the FH survival estimate (to be covered in lab session).

We can compare the Fleming-Harrington survival estimate to the KM estimate by rerunning the `sts list` command:

```
. sts list
. sts gen skm=s
. list skm sfh
             skm         sfh
  1.  .91666667   .9200444
  2.  .83333333   .8400932
  3.         .75   .7601478
  4.  .66666667   .6802101
  5.  .58333333   .6002833
  6.          .5   .5203723
  7.  .41666667   .4404857
  8.  .33333333   .3606392
  9.         .25   .2808661
 10.  .16666667   .2012493
 11.  .08333333   .1220639
 12.           0   .0449048
```

In this example, it looks like the Fleming-Harrington estimator is slightly higher than the KM at every time point, but with larger datasets the two will typically be much closer.