Lecture 2
ESTIMATING THE SURVIVAL FUNCTION

— One-sample nonparametric methods

There are commonly three methods for estimating a survivorship function

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

without resorting to parametric models:

(1) **Kaplan-Meier**

(2) **Nelson-Aalen** or **Fleming-Harrington** (via estimating the cumulative hazard)

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

We will mainly consider the first two.

(1) The Kaplan-Meier Estimator

The Kaplan-Meier (or KM) estimator is probably the most popular approach.

Motivation (no censoring):

Remission times (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes; Freireich *et al.*, 1963):

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

We estimate S(10), the probability that an individual survives to week 10 or later, by $\frac{8}{21}$.

How would you calculate the standard error of the estimated survival?

$$\hat{S}(10) = \hat{P}(T > 10) = \frac{8}{21} \; = \; 0.381$$

(Answer: $se[\hat{S}(10)] = 0.106$)

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

A table of $\hat{S}(t)$:

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

In most software packages, the survival function is evaluated just after time t, i.e., at t^+. In this case, we only count the individuals with $T > t$.

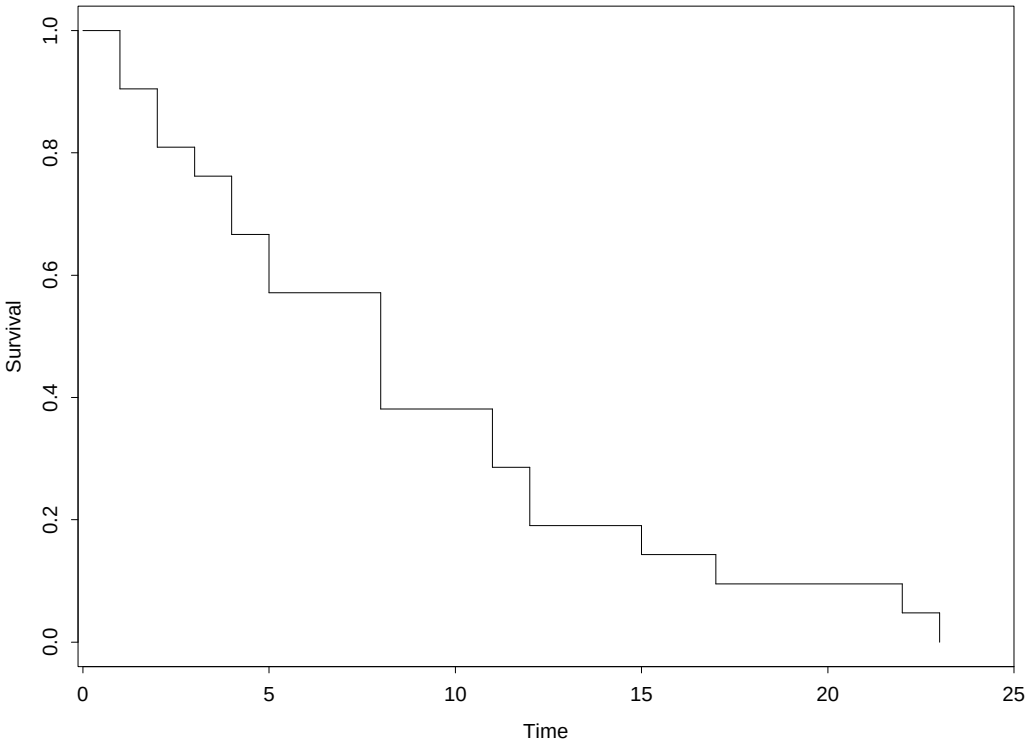


Figure 1: Example for leukemia data (control arm)

Empirical Survival Function:

When there is no censoring, the general formula is:

$$S_n(t) = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size} = \frac{\sum_{i=1}^{n} I(T_i > t)}{n}$$

Note that $F_n(t) = 1 - S_n(t)$ is the **empirical CDF**.

Also $I(T_i > t) \sim$ Bernoulli$(S(t))$, so that

1. $S_n(t)$ converges in probability to $S(t)$ (consistency);
2. $\sqrt{n}\{S_n(t) - S(t)\} \to N(0, S(t)[1 - S(t)])$ in distribution.

[**Make sure that you know these.**]

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(5)= 21/21, because everyone survived at least until week 5 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 estimate S(t) nonparametrically, even in the presence of censoring. The method is based on the ideas of **conditional probability**.

[**Reading:**]
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 $P(B)$:

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

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:

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

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

– 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:

			D		C		C	D	D	D

Using the law of conditional probability,

$$\begin{aligned} P(T > t) &= \prod_j P(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j) \\ &= \prod_j \lambda_j \end{aligned}$$

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

4 possibilities for each interval:

(1) **No death or censoring** - conditional probability of surviving the interval is estimated to be 1;

(2) **Censoring** - assume they survive to the end of the interval (the intervals are very small), so that the conditional probability of surviving the interval is again estimated to be 1;

(3) **Death, but no censoring** - conditional probability of *not* surviving the interval is estimated by # deaths (d) divided by # '**at risk**' (r) at the beginning of the interval. So the estimated 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 the estimated 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 estimated conditional probability of surviving the j-th interval that holds for all 4 cases:

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

Here as the intervals get finer and finer, the approximations made in estimating the probabilities of getting through each interval become more and more accurate, at the end the estimator converges to the true $S(t)$ in probability (proof not shown here).

This intuition explains 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:**

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

where

- $\tau_1, ...\tau_K$ is the set of K distinct uncensored failure times observed in the sample
- d_j is the number of failures at τ_j
- r_j is the number of individuals "at risk" right before the j-th failure time (everyone who died or censored at or after that time).

Furthermore, let c_j be the number of censored observations between the j-th and $(j+1)$-st failure times. Any censoring tied at τ_j are included in c_j, but not censorings tied at τ_{j+1}.

Note: two useful formulas are:

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

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

Calculating the KM - leukemia treated group

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 observation is uncensored;
- When there is no censoring, the KM estimator equals the empirical survival estimate

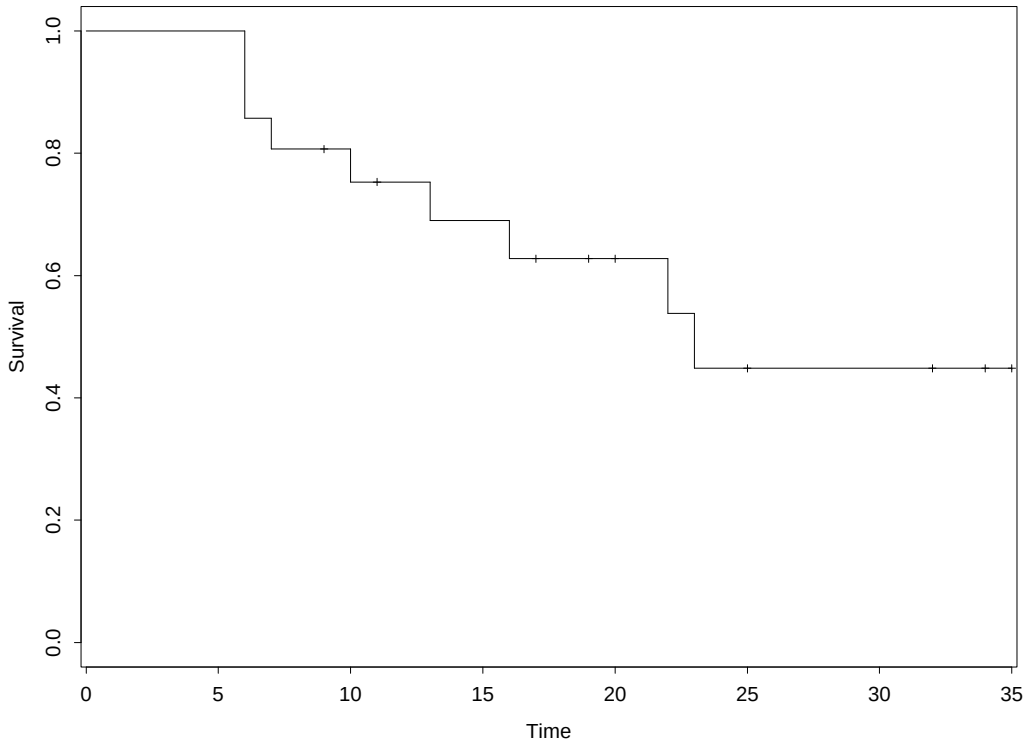


Figure 2: KM plot for treated leukemia patients

Output from a software 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

(Note: the above is from Stata, but most software output similar KM tables.)

[Reading:] Redistribution to the right algorithm (Efron, 1967)

There is another way to compute the KM estimator.

In the absence of censoring, $\hat{S}(t)$ is just the proportion of individuals with $T > t$. 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 observation times (failures or censorings) in increasing order. **If there are ties, put censored after failures.**
- 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

When there is no censoring, KM estimator is the same as the empirical estimator:

$$\hat{S}(t) = \frac{\#\text{ deaths after time } t}{n}$$

where n is the number of individuals in the study.

As said before

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

How does censoring affect this?

- $\hat{S}(t)$ is still consistent for the true S(t);
- $\hat{S}(t)$ is still asymptotically normal;
- The variance is more complicated.

The proofs can be done using the usual method (by writing as sum of i.i.d. terms plus $o_p(1)$) but it is laborious, or it can be done using counting processes which was considered more elegant, or by *empirical processes* method which is very powerful for semiparametric inferences.

The KM estimator is also an MLE

You can read in Cox and Oakes book Section 4.2.

Here we need to think of the distribution function $F(t)$ as an (infinite dimensional) parameter, and we try to find the $\hat{F}$ (or $\hat{S} = 1 - \hat{F}$) that maximizes a **nonparametric** likelihood. Such a MLE is called a **NPMLE**.

As it turns out, such a $\hat{F}(t)$ has to be discrete in order to for the likelihood to be bounded (otherwise MLE does not exist), with masses only at the uncensored failure times $\{a_j\}_j$. ($a_j = \tau_j$ in our previous notation)

Ex. Show the following for an absolutely continuous distribution function F(t): a) for a random sample of size n from F(t), if we allow an estimator of F(t) to be absolutely continuous (i.e. having a positive density) on any open interval covering the observed data point(s), then the likelihood function can be made arbitrarily large; b) if we restrict to those estimates of F(t) such that the likelihood function is bounded, the maximum of the likelihood is achieved when we assign point mass of 1/n to each observed data point. This shows that the empirical CDF is the NPMLE.

Cox and Oakes book Section 4.2 (please read if you have time) shows that the right-censored data likelihood for such a discrete distribution can be written as

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

where λ_j is the discrete hazard (i.e. conditional probability) at a_j. Note that the likelihood is the same as that of g independent binomials.

Therefore, the maximum likelihood estimator of λ_j is (why):

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

For a discrete survival distribution

$$S(t) = \prod_{j:a_j \le t} (1-\lambda_j)$$

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

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

This is the NPMLE of S.

One can often show that an NPMLE behaves like a classic MLE:

- consistent for the true parameter (function);
- asymptotically normal (converges in distribution to a Gaussian process).

For a semparametric model (that we'll talk about later),

it is often **semiparametrically efficient**.

Greenwood's formula for variance

Note that the KM estimator is

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

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

Since the $\hat{\lambda}_j$'s are just binomial proportions given r_j's, then

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

Also, the $\hat{\lambda}_j$'s are asymptotically independent.

Since $\hat{S}(t)$ is a function of the λ_j's, we can estimate its variance using the **Delta method**.

> **Delta method**: If Y_n is (asymptotically) normal with mean μ and variance σ^2, g is differentiable and $g'(\mu) \neq 0$, then $g(Y_n)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.

Greenwood's formula (continued)

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

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

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

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

Now, $\hat{S}(t) = \exp[\log[\hat{S}(t)]]$. Using Delta method once again,

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

Greenwood's Formula:

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

Confidence intervals

For a 95% confidence interval, we could use

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

where $\mathrm{se}[\hat{S}(t)]$ is calculated using Greenwood's formula. Here t is fixed, and this is refered to as pointwise confidence interval.

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:

$$S(t) = \exp(-\exp(L(t)).$$

[To see why this works:

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

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

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

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

]

[Read] 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 of $S(t)$ are:

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

(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

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

What is A?

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

$$\text{Var}(\hat{L}(t)) \;=\; \text{Var}\left[\log(-\log(\hat{S}(t)))\right]$$

- From our previous calculations, we know

$$\widehat{\text{Var}}(\log[\hat{S}(t)]) \;=\; \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j}$$

- Applying the delta method again, we get:

$$\begin{aligned} \widehat{\text{Var}}(\hat{L}(t)) \;&=\; \widehat{\text{Var}}(\log(-\log[\hat{S}(t)])) \\ &=\; \frac{1}{[\log\, \hat{S}(t)]^2} \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j} \end{aligned}$$

- We take the square root of the above to get $\text{se}(\hat{L}(t))$,

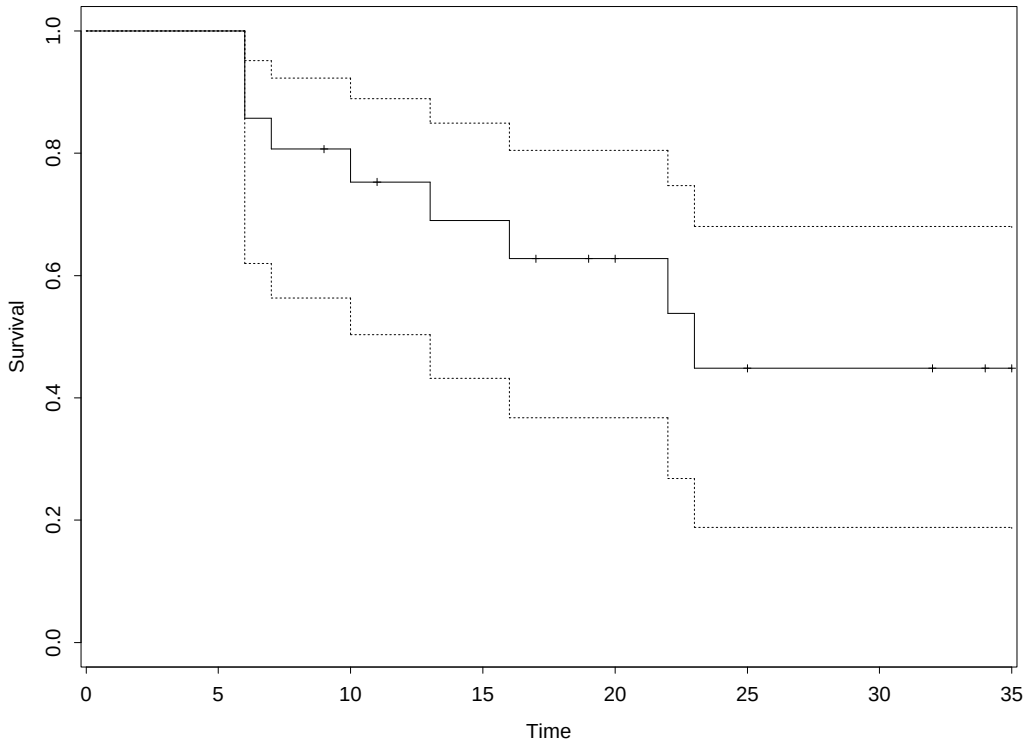


Figure 3: KM Survival Estimate and Confidence intervals (type 'log-log')

- Different software might use different approaches to calculate the CI.

Software for Kaplan-Meier Curves

- Stata - stset and sts commands
- SAS - PROC LIFETEST
- R - survfit() in 'survival' package

R allows different types of CI (some are truncated to be between 0 and 1):

```
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


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


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
```

Mean, Median, Quantiles based on the KM

- Mean is not well estimated with censored data, since we often don't observe the right tail.

- **Median** - by definition, this is the time, τ, such that $S(\tau) = 0.5$. In practice, it is often 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:

$$\begin{aligned} \hat{S}(22) &= 0.5378 \\ \hat{S}(23) &= 0.4482 \end{aligned}$$

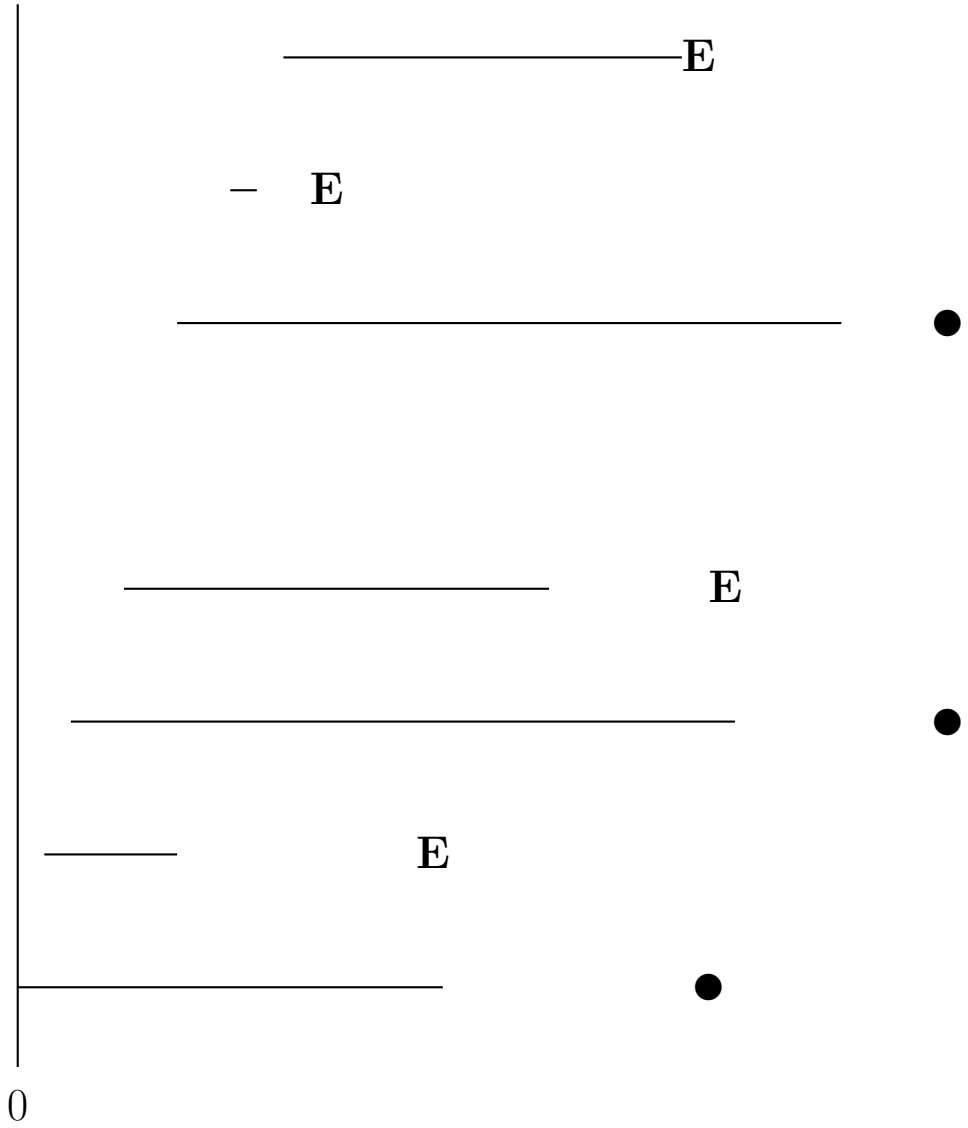
  The median is thus 23.

- **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 $\hat{S}(UQ) \leq 0.25$

From class discussion: 1) how do we bootstrap survival data? 2) how do we estimate the distribution of C? Recall that the observed data is (X_i, δ_i), where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, $i = 1, ..., n$.

Left truncated KM estimate



● = censored observation

E = event

When there is left truncation, the observed data is (Q_i, X_i, δ_i), $i = 1, ..., n$.

Now the 'risk set' at any time t consists of subjects who have entered the study, and have not failed or been censored by that time, i.e. $\{i : Q_i < t \leq X_i\}$.

So $r_j = \sum_{i=1}^{n} I(Q_i < \tau_j \leq X_i)$.

We still have

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

where $\tau_1, ...\tau_K$ is the set of K distinct uncensored failure times observed in the sample, d_j is the number of failures at τ_j, and r_j is the number of individuals "at risk" right before the j-th failure time (everyone **who had entered and** who died or censored at or after that time).

Read [required] Tsai, Jewell and Wang (1987).

- When $\min_{i=1}^{n} Q_i = t_0 > 0$, then the KM estimates $P(T > t | T > t_0)$.
- The left truncated KM is still an NPMLE (Wang, 1991).
- The Greenwood's formula for variance still applies.
- In R (and most other softwares) it is handled by something like 'Surv(time=Q, time2=X, event=δ)'.
- This approach is referred to as conditional inference, i.e. conditional on the Q's. This is implemented in most survival software.
- Another approach is the unconditional inference, where assumptions are made about the distribution of the Q's; eg. uniform, in which case this is referred to as 'length biased'.

(2) Nelson-Aalen (Fleming-Harrington) estimator – Estimating the cumulative hazard

If we can estimate $\Lambda(t) = \int_0^t \lambda(u)du$, the cumulative hazard at time t, then we can estimate $S(t) = e^{-\Lambda(t)}$.

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:

			D		C		C	D	D	D

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

$$\Lambda(t) \approx \sum_{j:\tau_j \leq t} \tilde{\lambda}_j \cdot \Delta_j$$

where the sum is over intervals up to t, $\tilde{\lambda}_j$ is the value of the hazard in the j-th interval and Δ_j is the width of that interval.

Since $\tilde{\lambda}\Delta$ is approximately the conditional probability of dying in the interval, we can further estimate $\tilde{\lambda}_j \cdot \Delta_j$ by d_j/r_j.

This gives the **Nelson-Aalen estimator**:

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

It follows that $\hat{\Lambda}(t)$, like the KM, changes only at the observed death (event) times.

Example:

			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	
$\hat{\lambda}(t_j)\Delta$	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 obtain the **Fleming-Harrington estimator** of $S(t)$:

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

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

We can compare the Fleming-Harrington survival estimate to the KM estimate using a subgroup of the nursing home data:

	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.

Question: do you think that the two estimators are *asymptotically equivalent*?

Note: 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:

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

[Reading:]
(3) 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 sometimes complicated by the fact that we don't know exactly when during a time interval an event occurs.

There are several types of lifetable methods 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 <u>entering</u> the interval

Example: 2418 Males with Angina Pectoris (chest pain, from book by 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

If we apply the KM formula directly to the numbers in the table on the previous page, estimating S(t) as

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

this approach is unsatisfactory for grouped data because it treats the problem as though it were in discrete time, with events happening only at 1 yr, 2 yr, etc. In fact, we should try to calculate the conditional probability of dying within the interval, given survival to the beginning of it.

What should we do with the censored subjects?

Let r'_j denote the 'effective' number of subjects at risk. If we 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 (event)

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

- Conditional probability of surviving

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

- Cumulative probability of surviving at t_j:

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

Other quantities estimated at the midpoint of the j-th interval:

- **Hazard** in the j-th interval (why)

$$\begin{aligned}\hat{\lambda}(t_{mj}) &= \frac{d_j}{b_j(r'_j - d_j/2)} \\ &= \frac{\hat{q}_j}{b_j(1 - \hat{q}_j/2)}\end{aligned}$$

- **density** at the midpoint of the j-th interval (why)

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

Note: Another way to get this is:

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

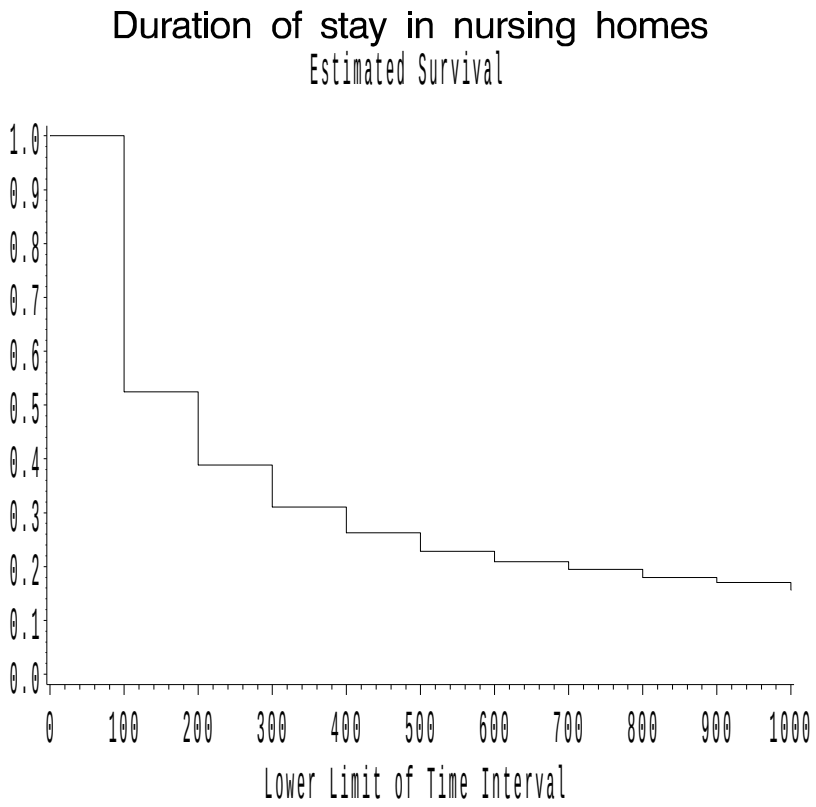


Figure 4: Life table estimate of survival

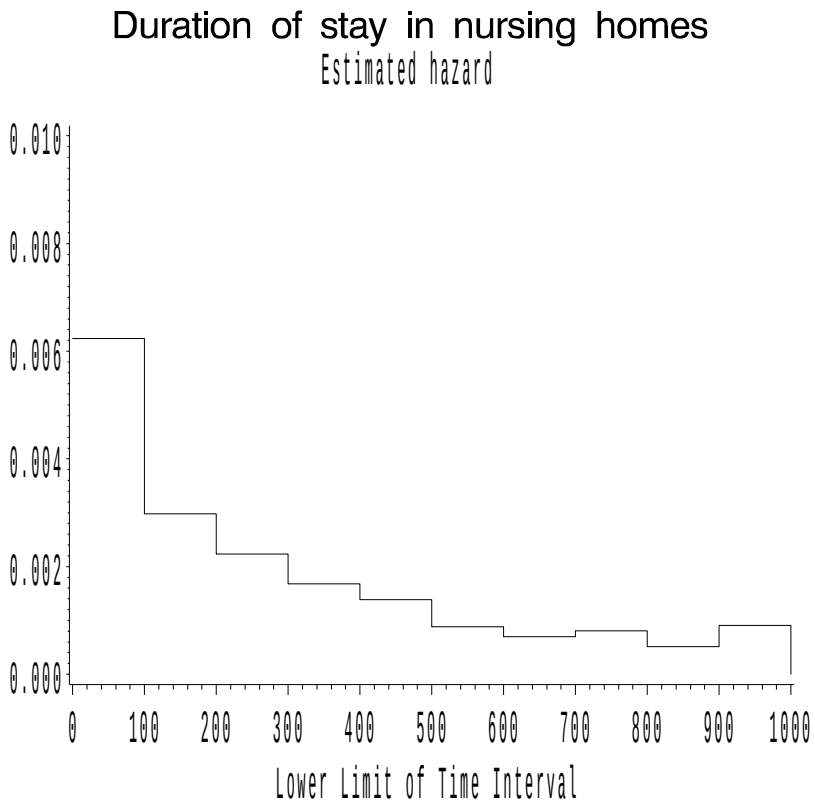


Figure 5: Estimated discrete hazard