STK4080/9080 SURVIVAL AND EVENT HISTORY ANALYSIS

Slides 8: The Kaplan-Meier estimator

Bo Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim

https://www.ntnu.edu/employees/bo.lindqvist
bo.lindqvist@ntnu.no
boli@math.uio.no

University of Oslo, Spring 2021

Towards a general formula for hazard rate

First, let T be a lifetime with an absolutely continuous distribution with density named $f(t)$. Then the hazard rate and cumulative hazard rate are

$$\alpha(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}, \quad A(t) = \int_0^t \alpha(s)ds$$

Also

$$S(t) = e^{-A(t)} \qquad\qquad (*)$$

Questions:

- How can we define $\alpha(t)$ and $A(t)$ if T does not have a continuous distribution?
- How can we in that case generalize $(*)$?

General definition of hazard rate

The basic function defining the distribution of a lifetime T is $S(t) = P(T > t)$ (continuous from the right). Given this, let

$$\begin{aligned} dA(t) &= P(t \leq T < t + dt | T \geq t) \\ &= \frac{P(t \leq T < t + dt)}{P(T \geq t)} \\ &= \frac{P(T \geq t) - P(T \geq t + dt)}{S(t-)} \\ &= \frac{S(t-) - S((t + dt)-)}{S(t-)} \\ &= -\frac{dS(t)}{S(t-)} \end{aligned}$$

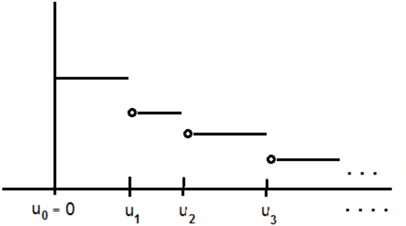
From this we get the generally valid definition of **cumulative hazard**,

$$A(t) = -\int_0^t \frac{dS(u)}{S(u-)}$$

Note: $P(T \geq t) = 1 - P(T < t) = 1 - F(t-) = S(t-)$

The discrete case

Let $u_0 = 0, u_1, u_2, \ldots$ be the values of T. Then $S(t) = P(T > t)$ is:



Recall: $dA(t) = \frac{S(t-)-S((t+dt)-)}{S(t-)}$

Now

$$\begin{aligned} dA(u_j) &= \frac{S(u_{j-1}) - S(u_j)}{S(u_{j-1})} = \frac{P(T > u_{j-1}) - P(T > u_j)}{P(T > u_{j-1})} \\ &= \frac{P(T = u_j)}{P(T \geq u_j)} \equiv \alpha_{u_j} \end{aligned}$$

a discrete hazard rate

The discrete survival function

Let $t = u_m$. Then $S(t) = P(T > u_m)$ and

$$\begin{aligned}
P(T > u_m) &= P(T > u_m, T > u_{m-1}, \cdots, T > u_1, T > u_0) \\
&= P(T > u_0) P(T > u_1 | T > u_0) \cdots P(T > u_m | T > u_{m-1}) \\
&= \prod_{j=1}^{m} P(T > u_j | T > u_{j-1}) \\
&= \prod_{j=1}^{m} (1 - P(T = u_j | T > u_{j-1})) \\
&= \prod_{j=1}^{m} (1 - \alpha_{u_j}) \equiv \prod_{j=1}^{m} (1 - dA(u_j))
\end{aligned}$$

This is a special case of the so called **product-integral** expression for $S(t)$ (next slide),

$$S(t) = \prod\nolimits_{0 \le u \le t} (1 - dA(u))$$

The product-integral

Recall the general expressions

$$dA(u) = -\frac{dS(u)}{S(u-)}, \quad A(t) = -\int_0^t \frac{dS(u)}{S(u-)}$$

Solving $S(t)$ from these equations in the general case, leads to the so-called **product-integral** ,

$$S(t) = \prod\nolimits_{0 \le u \le t} (1 - dA(u))$$

which can be defined as a limit as follows:

Let $t > 0$ be fixed, and let $0 = u_0 < u_1 < u_2 < \ldots < u_m \equiv t$ define a partition of $[0, t]$. Let $m \to \infty$ in a way such that the spacings $u_j - u_{j-1}$ tend to 0. Then

$$S(t) = \prod\nolimits_{0 \le u \le t} (1 - dA(u)) =_{def} \lim_{m \to \infty} \prod_{j=1}^{m} [1 - (A(u_j) - A(u_{j-1}))]$$

A quick check for the continuous case

Suppose $A(t) = \int_0^t \alpha(u)du$ so that $A'(t) = \alpha(t)$.

Then for a given partition with large m and fixed $t = u_m$,

$$\begin{aligned} & \prod_{j=1}^{m}[1-(A(u_j)-A(u_{j-1}))] \approx \prod_{j=1}^{m}(1-\alpha(u_j)(u_j-u_{j-1})) \\ \approx \quad & \prod_{j=1}^{m} e^{-\alpha(u_j)(u_j-u_{j-1})} = e^{-\sum_{j=1}^{m}\alpha(u_j)(u_j-u_{j-1})} \\ \approx \quad & e^{-\int_0^t \alpha(u)du} \equiv S(t) \end{aligned}$$

Thus, the product-integral gives the correct result for continuous distribution, as we saw that it did for the discrete case.

The Kaplan-Meier estimator

Let's go back to the discrete survival function, with $t = u_m$:

$$\begin{aligned} S(t) = P(T > u_m) &= \prod_{j=1}^{m} (1 - P(T = u_j | T > u_{j-1})) \\ &= \prod_{j=1}^{m} (1 - \alpha_{u_j}) = \prod_{j=1}^{m} (1 - dA(u_j)) \end{aligned}$$

An estimator of $S(t)$ can hence be obtained by putting an estimate for the function $A(t)$ in the above. This may be done by the Nelson-Aalen estimator and turns out to lead to the Kaplan-Meier estimator.

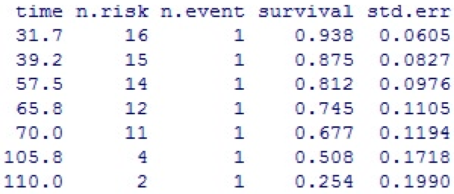
Recall the NA-estimator $\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)}$. This is a discrete function with jumps $\frac{dN(s)}{Y(s)}$, i.e., it jumps at the failure times T_i. Thus we can assume that the u_j above are these jump times. This leads to:

$$\hat{S}(t) = \prod_{0 \le u \le t} (1 - d\hat{A}(u)) = \prod_{0 \le u \le t} \left(1 - \frac{dN(s)}{Y(s)}\right) = \prod_{T_i \le t} \left(1 - \frac{1}{Y(T_i)}\right)$$

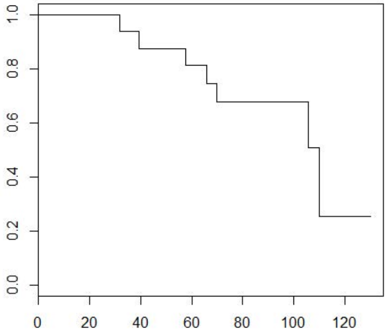
Kaplan-Meier estimator in R

```
library(survival)
testdata=read.table("https://folk.ntnu.no/bo/STK4080/
nelson-aalen-hand.txt",header=T)
fitB=survfit(Surv(Time,Status)~1, data=testdata,
conf.type="none")
summary(fitB)
plot(fitB, mark.time=F)
```

R-Output with testdata



```
 time n.risk n.event survival std.err
 31.7     16       1    0.938  0.0605
 39.2     15       1    0.875  0.0827
 57.5     14       1    0.812  0.0976
 65.8     12       1    0.745  0.1105
 70.0     11       1    0.677  0.1194
105.8      4       1    0.508  0.1718
110.0      2       1    0.254  0.1990
```



Example 3.8: Second births

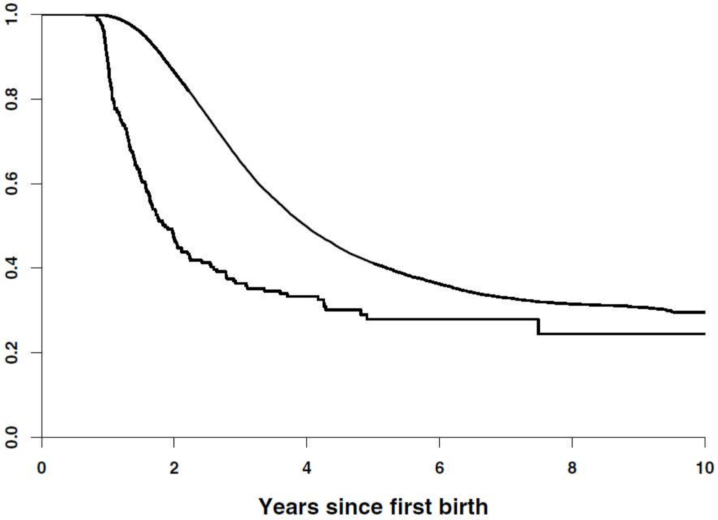


Fig. 3.11 *Kaplan-Meier estimates for the time between first and second birth. Upper curve: first child survived one year; lower curve: first child died within one year.*

Properties of the Kaplan-Meier estimator

Recall first for the NA-estimator,

$$\hat{A}(t) = \int_0^t J(s)\alpha(s)ds + \int_0^t \frac{J(s)}{Y(s)} dM(s) \equiv A^*(t) + I(t)$$

so that $I(t)$ is a mean zero martingale. Recall also the KM-estimator,

$$\hat{S}(t) = \prod_{0 \le s \le t} (1 - d\hat{A}(s)) = \prod_{T_i \le t} \left(1 - \frac{1}{Y(T_i)}\right)$$

Now let $S^*(t) = \prod_{0 \le s \le t}(1 - dA^*(s)) = e^{-A^*(t)} \approx S(t)$

Duhamel's equation (see book p. 461, eq. (A12)) gives

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = -\int_0^t \frac{\hat{S}(s-)}{S^*(s)} d(\hat{A} - A^*)(s)$$

Properties of the Kaplan-Meier estimator (cont.)

...Duhamel's equation gives

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = -\int_0^t \frac{\hat{S}(s-)}{S^*(s)} d(\hat{A} - A^*)(s)$$

so that $\frac{\hat{S}(t)}{S^*(t)} - 1$ is a martingale since $\hat{A} - A^*$ is a martingale.

Hence $E\left(\frac{\hat{S}(t)}{S^*(t)}\right) = 1.$ Thus, for large n we will have $E(\hat{S}(t)) \approx S(t)$.

Further, for large n we will have $\frac{\hat{S}(s-)}{S^*(s)} \approx 1.$ Hence from the expression on top of the page,

$$\begin{aligned} \frac{\hat{S}(t)}{S(t)} - 1 &\approx -(\hat{A}(t) - A(t)), \quad \text{so} \\ \hat{S}(t) - S(t) &\approx -S(t)(\hat{A}(t) - A(t)) \end{aligned}$$

Variance estimator for the KM estimator

From

$$\hat{S}(t) - S(t) \approx -S(t)(\hat{A}(t) - A(t))$$

we get that $\hat{S}(t)$ is asymptotically normal with

$$Var(\hat{S}(t)) \approx S(t)^2 Var(\hat{A}(t))$$

which can be estimated by

$$\hat{\tau}^2(t) = \hat{S}(t)^2 \sum_{T_i \le t} \frac{1}{Y(T_i)^2}$$

Note the classical *Greenwood's formula*, which gives the estimator

$$\tilde{\tau}^2(t) = \hat{S}(t)^2 \sum_{T_i \le t} \frac{1}{Y(T_i)(Y(T_i) - 1)}$$

$100(1-\alpha)\%$ confidence limits for $S(t)$

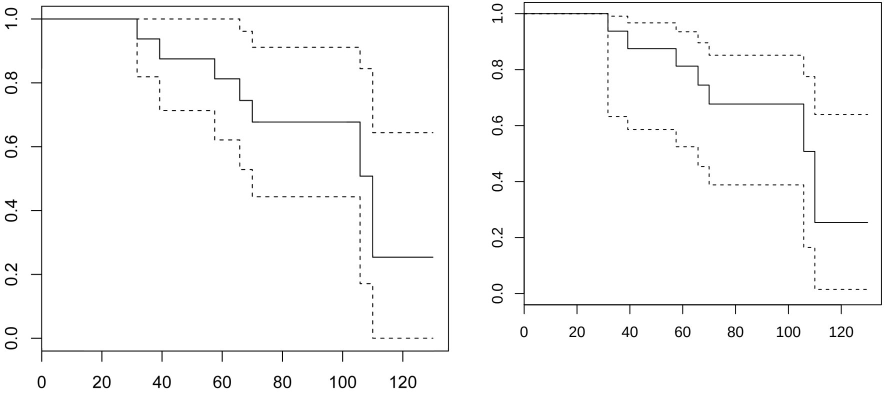
Let

$$\hat{\sigma}^2 = \sum_{T_i \leq t} \frac{1}{Y(T_i)^2}$$

Standard interval: $\hat{S}(t) \pm z_{1-\alpha/2} \hat{S}(t) \hat{\sigma}(t)$

Log-log transformed interval: $\hat{S}(t) \exp\{\pm z_{1-\alpha/2} \hat{\sigma}(t) / \log \hat{S}(t)\}$

Example with test-data



Left: Standard 95% confidence limits. *Right:* 95% log-log transformed limits.

3.1.3 Handling of ties for the NA estimator

Suppose that at T_j there are observed $d_j > 1$ events. We may deal with these by two conceptually different approaches:

(i) Assume that **the events actually happen in continuous time**, and that **in reality** no two event times coincide.

Then in the NA estimator one may replace $1/Y(T_j)$ by

$$\Delta\hat{A}(T_j) = \sum_{k=0}^{d_j-1} \frac{1}{Y(T_j) - k}$$

and in the corresponding variance estimator one may replace $1/Y(T_j^2)$ by

$$\Delta\hat{\sigma}^2(T_j) = \sum_{k=0}^{d_j-1} \frac{1}{(Y(T_j) - k)^2}$$

thus obtaining the estimators

$$\hat{A}(t) = \sum_{T_j \le t} \Delta\hat{A}(T_j) \quad \text{and} \quad \hat{\sigma}^2(t) = \sum_{T_j \le t} \Delta\hat{\sigma}^2(T_j)$$

3.1.3 Handling of ties for the NA estimator (cont.)

(ii) Assume that the data are **genuinely discrete**, so that **tied event times are real** and not due to grouping and rounding.

Then one uses instead

$$\begin{aligned} \Delta\hat{A}(T_j) &= \frac{d_j}{Y(T_j)} \\ \Delta\hat{\sigma}^2(T_j) &= \frac{(Y(T_j) - d_j)d_j}{(Y(T_j)^3} \end{aligned}$$

obtaining again the estimators

$$\hat{A}(t) = \sum_{T_j \le t} \Delta\hat{A}(T_j) \quad \text{and} \quad \hat{\sigma}^2(t) = \sum_{T_j \le t} \Delta\hat{\sigma}^2(T_j)$$

3.2.2 Handling of ties for the KM estimator

We may write the KM-estimator as

$$\hat{S}(t) = \prod_{T_j \le t} \left\{ 1 - \Delta \hat{A}(T_j) \right\}$$

Consider again the two cases: (i) Ties are due to **rounding**, (ii) The data are genuinely **discrete**.

For both cases, one uses $\Delta \hat{A}(T_j) = \frac{d_j}{Y(T_j)}$

For estimation of variance, in case (i) one uses

$$\hat{\tau}^2(t) = \hat{S}(t)^2 \sum_{T_j \le t} \sum_{k=0}^{d_j - 1} \frac{1}{(Y(T_j) - k)^2}$$

In case (ii), one uses Greenwood's formula,

$$\tilde{\tau}^2(t) = \hat{S}(t)^2 \sum_{T_j \le t} \frac{d_j}{Y(T_j)(Y(T_j) - d_j)}$$

Estimation of median survival time and other fractiles

The pth fractile ξ_p of the survival distribution is given by (Exercise 1.2)

$$F(\xi_p) = p \quad \text{or, equivalently,} \quad S(\xi_p) = 1 - p$$

Median corresponds to $p = 0.5$;

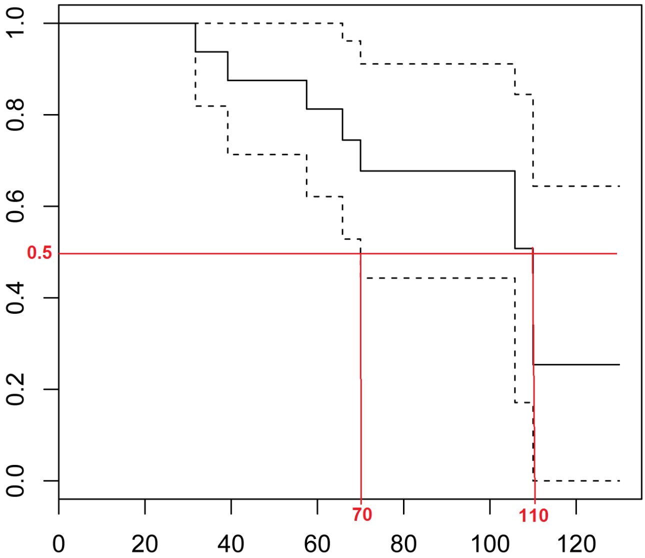
ξ_p is estimated by

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

(draw a horizontal line in the KM plot at height $1 - p$ until it crosses the KM curve - see next slide!)

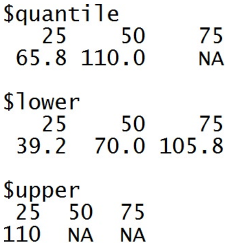
Confidence intervals are obtained by "inverting" the confidence limits for the survival function (to be considered in Exercise 3.8).

Test-example: Graphical estimation of median lifetime



Test-example: R-estimation of quartiles and median

```
library(survival)
testdata=read.table("https://folk.ntnu.no/bo/STK4080/
nelson-aalen-hand.txt",header=T)
fitC=survfit(Surv(Time,Status)~1,data=testdata,conf.type="plain")
summary(fitC)
quantile(fitC,probs=c(.25,.5,.75))
```



```
$quantile
   25    50    75
 65.8 110.0    NA

$lower
   25    50    75
 39.2  70.0 105.8

$upper
 25  50  75
110  NA  NA
```

Example 3.10: Median time between first and second birth

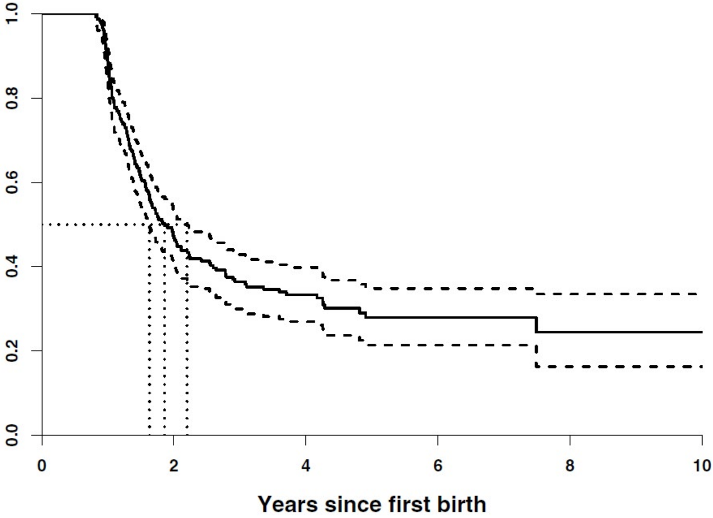


Fig. 3.13 *Kaplan-Meier estimate with 95% log-log-transformed confidence intervals for the time between first and second birth for women who lost the first child within one year after birth. It is indicated at the figure how one may obtain the estimated median time with confidence limits.*

Can we estimate mean survival time, $E(T)$, from KM?

Recall (Exercise 1.3) that

$$E(T) = \int_0^\infty S(u)du$$

So, can we estimate $E(T)$ by

$$\widehat{E(T)} = \int_0^\infty \hat{S}(u)du \ \ ?$$

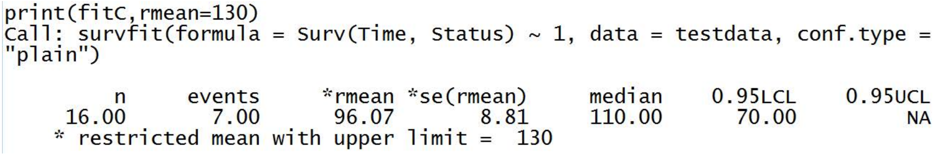
This is, however, problematic due to censoring, and the fact that the right tail is poorly estimated (and $\hat{S}(t)$ may even be constant and positive for all large t.)

But we can instead estimate the *restricted mean*, i.e., the expected survival in $[0, t]$, $\mu_t = \int_0^t S(u)du$. This may be estimated by

$$\hat{\mu}_t = \int_0^t \hat{S}(u)du$$

Test-example: R-estimation of restricted mean

```
print(fitC,rmean=130)
```



```
print(fitC,rmean=130)
Call: survfit(formula = Surv(Time, Status) ~ 1, data = testdata, conf.type =
"plain")

        n    events    *rmean *se(rmean)    median   0.95LCL   0.95UCL
    16.00      7.00     96.07       8.81    110.00     70.00        NA
    * restricted mean with upper limit =  130
```