STK4080/9080
SURVIVAL AND EVENT HISTORY ANALYSIS
Slides 4: Introduction to nonparametric estimation

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

NONPARAMETRIC ESTIMATION OF $S(t) = P(T > t)$

We are interested in estimating the distribution of the lifetime T of some equipment or the time to some given event in a medical context.

We have indicated how parametric models like exponential and Weibull can be fitted to data.

Now we shall instead see how in particular $S(t)$ can be estimated without making parametric assumptions.

Thus, instead of having to restrict to estimation of one or two parameters, we now have an infinite number of possible functions $S(t)$ to choose from. (Essentially, the only restriction is that it is decreasing, starts in 1 and converges to 0 as $t \to \infty$.)

NONPARAMETRIC ESTIMATION FOR NON-CENSORED DATA

In this case our observations are the exact failure times $T_1, \ldots, T_n$, assumed to be i.i.d. observations of a lifetime T.

Hence we can estimate $S(t) = P(T > t)$ for a given $t > 0$ by the relative proportion of lifetimes that exceed t:

$$\hat{S}(t) = \frac{\text{number of } T_i > t}{n}$$

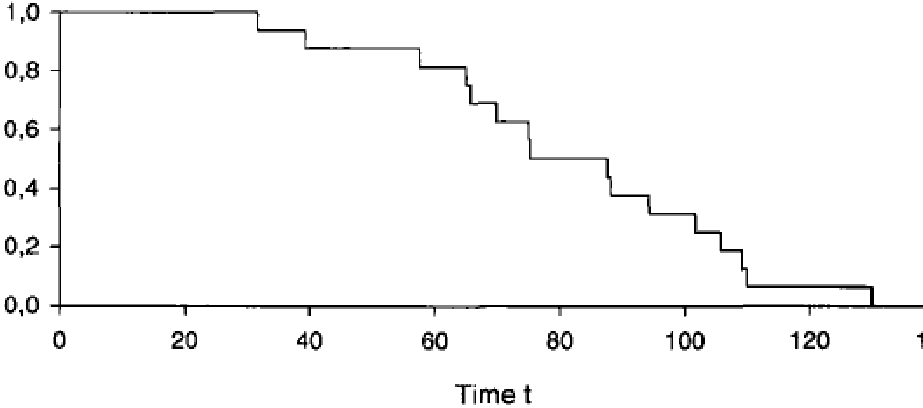
This is called the *empirical survivial function*.

If we order the observations as $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$, then $\hat{S}(t)$ starts at 1 for $t = 0$ and makes a downward jump of $1/n$ at $T_{(1)}$, a new downward jump of $1/n$ at $T_{(2)}$, and so on until it jumps from $1/n$ to 0 at $T_{(n)}$.

EXAMPLE OF EMPIRICAL SURVIVAL PLOT, $\hat{S}(t)$

$n = 16$ observed lifetimes:

31.7, 39.2, 57.5, 65.0, 65.8, 70.0, 75.0, 75.2, 87.7, 88.3, 94.2, 101.7, 105.8, 109.2, 110.0, 130.0



CENSORED DATA: KAPLAN-MEIER ESTIMATOR FOR $S(t)$

Consider n individuals, where the ith individual has potential lifetime T_i and potential censoring time C_i. We *observe* the pair $(\tilde{T}_i, D_i)$, where

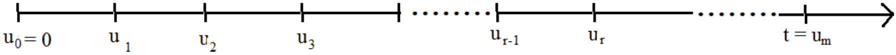
$$\begin{aligned} \tilde{T}_i &= \min(T_i, C_i) \\ D_i &= \begin{cases} 1 & \text{if} \quad \tilde{T}_i = T_i \\ 0 & \text{if} \quad \tilde{T}_i = C_i \end{cases} \end{aligned}$$

Assume:

- $T_1, T_2, \cdots, T_n$ are *independent and identically distributed* with common reliability function $S(t)$.
- The censoring mechanism satisfies the property of *independent censoring*.

The estimator is constructed in the following.

MAIN IDEA OF CONSTRUCTION

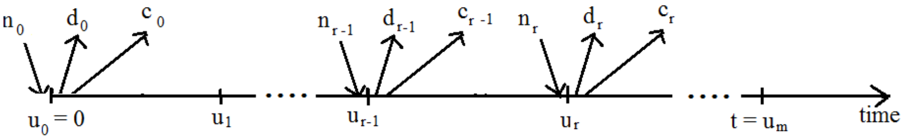


Assume first that time is measured on a discrete scale with values $u_0 = 0 \leq u_1 \leq u_2 \leq \cdots$, so that all T_i, C_i and hence $\tilde{T}_i$ are among these. Let $t = u_m$. Then

$$
\begin{aligned}
S(t) &= P(T > t) = P(T > u_m) \\
&= P(T > u_m \cap T > u_{m-1} \cap \cdots \cap T > u_2 \cap T > u_1 \cap T > u_0) \\
&= P(T > u_0) \cdot P(T > u_1 \mid T > u_0) \cdot P(T > u_2 \mid T > u_1 \cap T > u_0) \\
&\cdots P(T > u_m \mid T > u_{m-1} \cap \cdots \cap T > u_0) \\
&= P(T > u_0) \cdot P(T > u_1 \mid T > u_0) \cdot P(T > u_2 \mid T > u_1) \\
&\cdots P(T > u_r \mid T > u_{r-1}) \cdots P(T > u_m \mid T > u_{m-1})
\end{aligned}
$$

Idea: Estimate each factor $P(T > u_r \mid T > u_{r-1})$, from data (Y_i, δ_i); $i = 1, \cdots, n$.

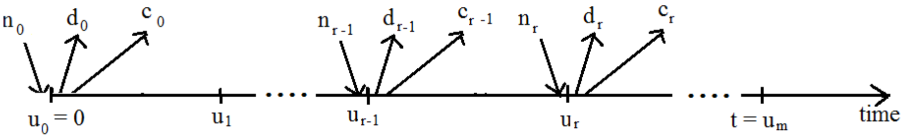
CONSTRUCTION OF ESTIMATOR



Define:

- n_r = number at risk at time u_r; i.e. number that can fail at u_r; counted immediately before u_r.
- d_r = number failing at u_r (those with $Y = u_r$, $\delta = 1$)
- c_r = number censored at u_r (those with $Y = u_r$, $\delta = 0$); assumed to be censored right after u_r, and by convention after all failures at u_r (in practice in the interval following u_r)

CONSTRUCTION OF ESTIMATOR (CONT.)



Note: The d_i, c_i are found directly from the data, while the n_i are found recursively as:

$$
\begin{aligned}
n_0 &= n \\
n_1 &= n_0 - d_0 - c_0 \\
\dots \\
n_r &= n_{r-1} - d_{r-1} - c_{r-1}
\end{aligned}
$$

Then estimate,

$$P(T > u_r \mid T > u_{r-1}) = 1 - P(T = u_r \mid T > u_{r-1}) \approx 1 - \frac{d_r}{n_r} = \frac{n_r - d_r}{n_r}$$

$$\& \quad P(T > u_0) = 1 - P(T = u_0) \approx 1 - \frac{d_0}{n_0} = \frac{n_0 - d_0}{n_0}$$

THE FINAL KM-ESTIMATOR

It follows that $S(t) = P(T > t)$ can be estimated by

$$\hat{S}(t) = \frac{n_0 - d_0}{n_0} . \frac{n_1 - d_1}{n_1} \cdots \frac{n_r - d_r}{n_r} \cdots \frac{n_m - d_m}{n_m}$$

Note that these factors are 1, whenever $d_r = 0$. Thus

$$\hat{S}(t) = \prod_{\substack{\text{all } u_r \leq t \\ \text{with } d_r \geq 1}} \frac{n_r - d_r}{n_r}$$

In practice we have continous time. But this case can be approximated by making the grid $u_1 < u_2 < \cdots$ finer and finer.

Thus in general the KM-estimator is given by:

If $T_{(1)} < T_{(2)} < \cdots$, are the times with at least one failure, and n_i, d_i are, respectively, the number at risk and the number of failures at $T_{(i)}$, then

$$\hat{S}(t) = \prod_{i: T_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$

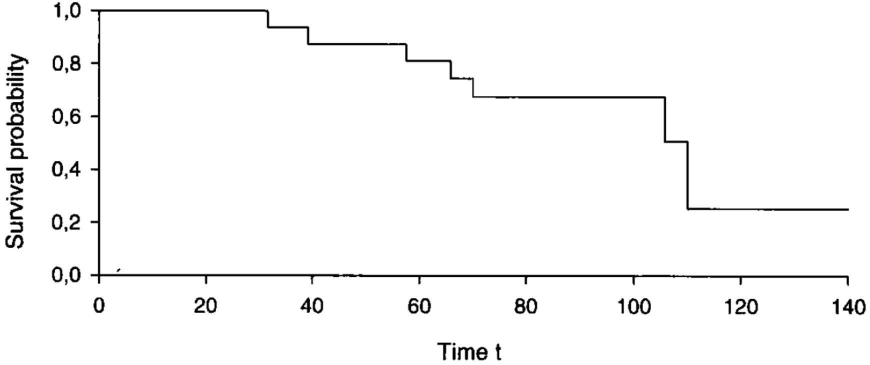
KM-ESTIMATOR FOR CENSORED DATA

Suppose data are (+ means right censored)

31.7	39.2	57.5	65.0+	65.8	70.0	75.0+	75.2+
87.5+	88.3+	94.2+	101.7+	105.8	109.2+	110.0	130.0+

Time	at risk	$\frac{n_i - d_i}{n_i}$	KM estimate	
31.7	16	$\frac{15}{16}$	$\frac{15}{16}$	$= 0.9375$
39.2	15	$\frac{14}{15}$	$\frac{15}{16} \cdot \frac{14}{15} = \frac{14}{16}$	$= 0.8750$
57.5	14	$\frac{13}{14}$	$\frac{14}{16} \cdot \frac{13}{14} = \frac{13}{16}$	$= 0.8125$
65.8	12	$\frac{11}{12}$	$\frac{13}{16} \cdot \frac{11}{12}$	$= 0.7448$
70.0	11	$\frac{10}{11}$	$\frac{13}{16} \cdot \frac{11}{12} \cdot \frac{10}{11} = \frac{13}{16} \cdot \frac{10}{12}$	$= 0.6771$
105.8	4	$\frac{3}{4}$	$\frac{13}{16} \cdot \frac{10}{12} \cdot \frac{3}{4}$	$= 0.5078$
110.0	2	$\frac{1}{2}$	$\frac{13}{16} \cdot \frac{10}{12} \cdot \frac{3}{4} \cdot \frac{1}{2}$	$= 0.2539$

KM-PLOT FOR CENSORED DATA



BREAST CANCER DATA (from Collett (book))

Table 1.2 *Survival times of women with tumours that were negatively or positively stained with HPA.*

Negative staining	Positive staining	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

BREAST CANCER DATA: KM PLOTS (Collett)

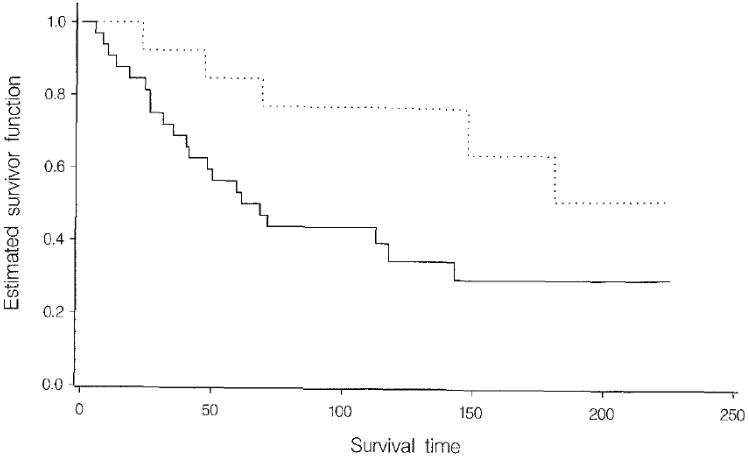


Figure 2.9 *Kaplan-Meier estimate of the survivor functions for women with tumours that were positively stained (—) and negatively stained (···).*

NONPARAMETRIC ESTIMATION OF THE CUMULATIVE HAZARD $A(t)$

Why is an estimate of $A(t)$ useful?

Note first that $A'(t) = \alpha(t)$. Thus,

- T is IFR $\Leftrightarrow \alpha(t)$ is *increasing* $\Leftrightarrow A(t)$ is *convex*
- T is DFR $\Leftrightarrow \alpha(t)$ is *decreasing* $\Leftrightarrow A(t)$ is *concave*

Thus a plot of an estimate $\hat{A}(t)$ can give us information on whether the distribution of T is IFR (*increasing failure rate*) or DFR (*decreasing failure rate*).

More generally, a plot of $\hat{A}(t)$ can give us information on the shape of the hazard rate.

ESTIMATING $A(t)$ BY THE KM-ESTIMATOR

Recall that $S(t) = e^{-A(t)}$, so

$$A(t) = -\ln S(t)$$

Thus, if $\hat{S}_{KM}(t)$ is the KM-estimator for $S(t)$, then we can define,

$$\begin{aligned}
\hat{A}_{KM}(t) &= -\ln \hat{S}_{KM}(t) \\
&= -\ln \prod_{T_{(i)} \leq t} \frac{n_i - d_i}{n_i} \\
&= -\sum_{T_{(i)} \leq t} \ln \left(1 - \frac{d_i}{n_i}\right) \\
&\approx \sum_{T_{(i)} \leq t} \frac{d_i}{n_i}
\end{aligned}$$

where we used that for small x is

$$-\ln(1-x) \approx x$$

THE NELSON-AALEN ESTIMATOR FOR $A(t)$

The Nelson-Aalen estimator (NA-estimator) is simply defined by

$$\hat{A}_{NA}(t) = \sum_{T_{(i)} \leq t} \frac{d_i}{n_i}$$

Note: *A more thorough treatment of the Nelson-Aalen and Kaplan-Meier estimators are at the main core of the course – using theory for counting processes and martingales.*

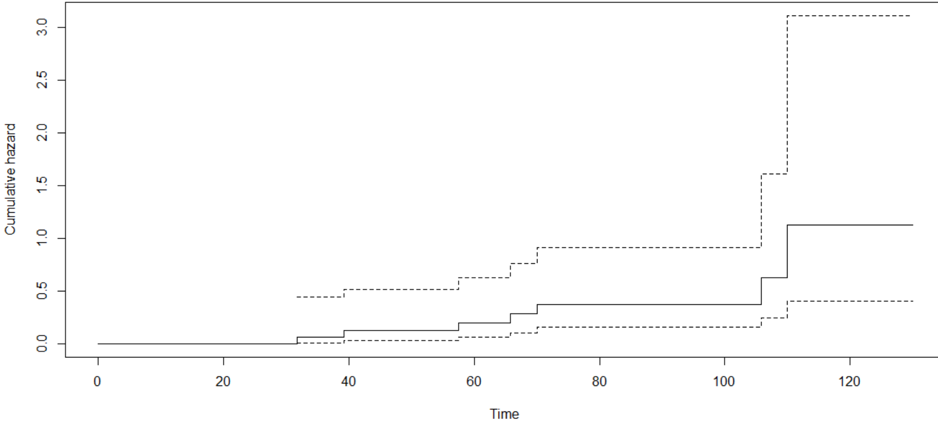
NELSON-AALEN ESTIMATOR

Suppose data are as for the KM example (+ means right censored)

31.7	39.2	57.5	65.0+	65.8	70.0	75.0+	75.2+
87.5+	88.3+	94.2+	101.7+	105.8	109.2+	110.0	130.0+

Time	at risk	$\frac{d_i}{n_i}$	Nelson-Aalen estimate	
31.7	16	$\frac{1}{16}$	$\frac{1}{16}$	$= 0.06250$
39.2	15	$\frac{1}{15}$	$\frac{1}{16}+\frac{1}{15}$	$= 0.12917$
57.5	14	$\frac{1}{14}$	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}$	$= 0.20060$
65.8	12	$\frac{1}{12}$	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}+\frac{1}{12}$	$= 0.28393$
70.0	11	$\frac{1}{11}$	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}+\frac{1}{12}+\frac{1}{11}$	$= 0.37484$
105.8	4	$\frac{1}{4}$	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}+\frac{1}{12}+\frac{1}{11}+\frac{1}{4}$	$= 0.62484$
110.0	2	$\frac{1}{2}$	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}+\frac{1}{12}+\frac{1}{11}+\frac{1}{4}+\frac{1}{2}$	$= 1.12484$

PLOT OF NELSON-AALEN ESTIMATOR



R-CODE

https://folk.ntnu.no/bo/STK4080/2021/R-code-KM-NA.txt

```
library(survival)
# READ DATA
testdata=read.table("https://folk.ntnu.no/bo/STK4080/nelson-aalen-hand.txt",header=T)
#
# KAPLAN-MEIER ESTIMATOR
#
fitKM=survfit(Surv(Time,Status)~1, data=testdata,conf.type="log-log")
summary(fitKM)
plot(fitKM)
#
#NELSON-AALEN ESTIMATOR
#
fitNA=survfit(coxph(Surv(Time,Status)~1, data=testdata))
plot(fitNA, fun="cumhaz",xlab="Time", ylab="Cumulative hazard")
```