TMA4275 LIFETIME ANALYSIS

Slides 5: Censoring and Kaplan-Meier estimator

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

http://www.math.ntnu.no/∼bo/
bo@math.ntnu.no

NTNU, Spring 2014

Lifetime data typically include *censored* data, meaning that:

- some lifetimes are known to have occurred only within certain intervals.
- The remaining lifetimes are known exactly.

Categories of censoring:

- right censoring
- left censoring
- interval censoring

Right censoring is the most common way of censoring. Different subtypes of right censoring can be considered. A common way of presenting right-censored data is as follows:

n units are observed, with potential i.i.d. lifetimes $T_1, T_2, \cdots, T_n$. For each i, we observe a time Y_i which is either the true lifetime T_i, or a censoring time $C_i < T_i$, in which case the true lifetime is "to the right" of the observed time C_i.

The observation from a unit is the pair (Y_i, δ_i) where the *censoring indicator* δ_i is defined by

$$\delta_i = \begin{cases} 1 & \text{if} \quad Y_i = T_i \\ 0 & \text{if} \quad Y_i = C_i, \text{ in which case it is known that } T_i > Y_i \end{cases}$$

TYPE I CENSORING

n units put on test at time $t = 0$. Experiment stopped at time $t = t_0$.

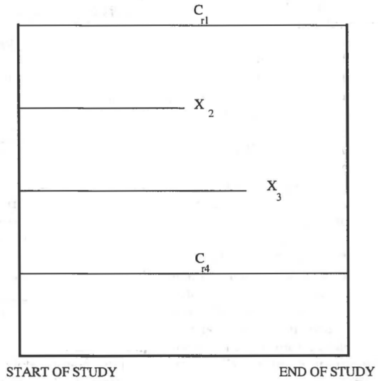


Figure 3.1 *Example of Type I censoring*

GENERALIZED TYPE I CENSORING

Individuals enter the study at different times, and the terminal point of the study is predetermined.

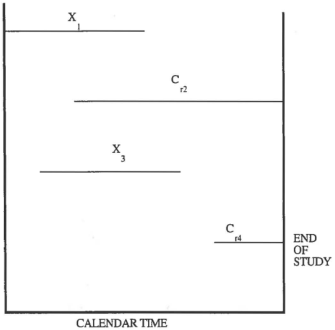


Figure 3.3 *Generalized Type I censoring when each individual has a different starting time*

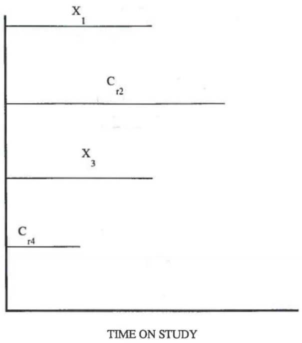


Figure 3.4 *Generalized Type I censoring for the four individuals in Figure 3.3 with each individual's starting time backed up to 0. $T_1 = X_1$ (death time for first individual) ($\delta_1 = 1$); $T_2 = C_{r2}$ (right censored time for second individual) ($\delta_2 = 0$); $T_3 = X_3$ (death time for third individual) ($\delta_3 = 1$); $T_4 = C_{r4}$ (right censored time for fourth individual) ($\delta_4 = 0$).*

In Type II (right) censoring, the study continues until the failure of the first r individuals, where r is some predetermined integer ($r < n$).

Usual application: Testing of equipment life, where all items are put on test at the same time, and the test is terminated when r of the n items have failed.

Advantage: It could take a very long time for all items to fail. Also, the statistical treatment of Type II censored data is simpler because the joint distribution of the order statistics is available.

Type III CENSORING

This is a mix of Type I and Type II censoring. Choose both an end time t_0 as for Type I censoring and an $r < n$ as for Type II censoring. Stop the experiment at time t_0 or at the rth failure, whatever comes first.

RANDOM CENSORING (TYPE IV CENSORING)

- For each unit we define
  - T_i to be the potential lifetime
  - C_i to be the potential censoring time

  where
- T_i, C_i are **independent random variables**.
- Then we *observe* the pair (Y_i, δ_i), where

$$
\begin{aligned}
Y_i &= \min(T_i, C_i) \\
\delta_i &= \begin{cases} 1 & \text{if} \quad T_i \leq C_i \\ 0 & \text{if} \quad T_i > C_i \end{cases}
\end{aligned}
$$

Example of use: Cancer treatment, with T_i being the time of death due to this cancer; while C_i is the time of death of another cause, or an accident, or migration, etc.

INDEPENDENT CENSORING

Consider a situation where n individuals are followed from time $t = 0$. The ith individual is followed until $Y_i = \min(T_i, C_i)$, i.e. until either failure (death) or censoring at time C_i.

The ith individual is said to be at risk at time t if $t < Y_i$.

A sensoring scheme is said to satisfy the property of **independent censoring** if, at any time t, the individuals that are *at risk* are representative for the distribution of T in the sense that their probaility of failing in a small time interval $(t, t+h)$ is (in the limit as h tends to 0) is $z(t)h$.

For example this would not be the case if individuals are censored because they are supposed to fail very soon. (By considering them as censored instead of failed could lead to a more optimistic lifetime estimate than the correct one).

The censoring types we have considered so far all satisfy this independent censoring property.

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 $R(t)$ 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 $R(t)$ to choose from. (Essentially, the only restriction is that it is decreasing, starts in 1 and converges to 0 as $t \to \infty$.)

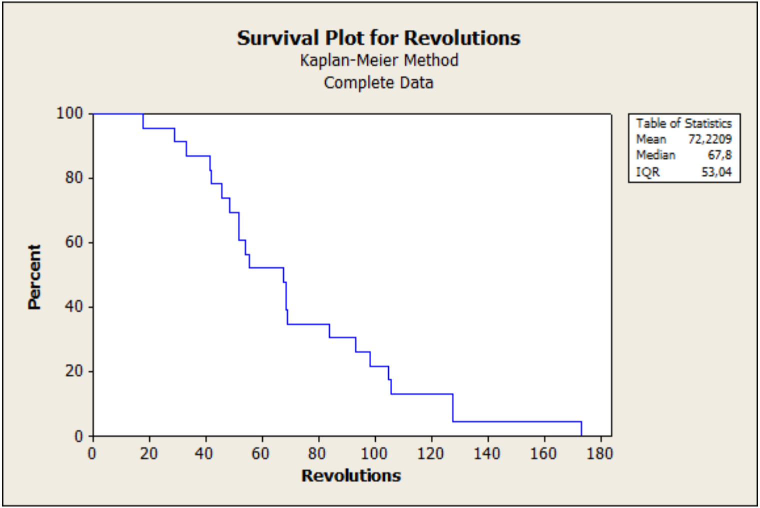
In this case our observations are the exact failure times $T_1, \ldots, T_n$, assumed to be i.i.d. observations of a lifetime T.

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

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

This is called the *empirical survivial function*.

If we order the observations as $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$, then $\hat{R}(t)$ 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 $T_{(n)}$.


Survival Plot for Revolutions
Kaplan-Meier Method
Complete Data
Percent
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
Revolutions
Table of Statistics
Mean 72,2209
Median 67,8
IQR 53,04

Consider n individuals, where the ith individual has potential lifetime T_i and potential censoring time C_i. We *observe* the pair (Y_i, δ_i), where

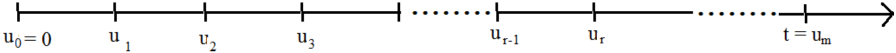
$$\begin{aligned} Y_i &= \min(T_i, C_i) \\ \delta_i &= \begin{cases} 1 & \text{if} \quad T_i \leq C_i \\ 0 & \text{if} \quad T_i > C_i \end{cases} \end{aligned}$$

Assume:

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

The estimator is constructed in the following.

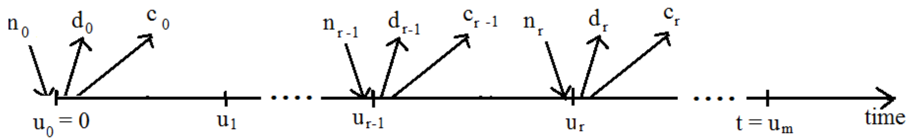
MAIN IDEA OF CONSTRUCTION



Assume first that time is measured on a discrete scale $u_0 = 0 \leq u_1 \leq u_2 \leq \cdots$, so that all T_i, C_i, Y_i are among these. Now suppose $t = u_m$ and we want to compute (estimate) $R(t)$.

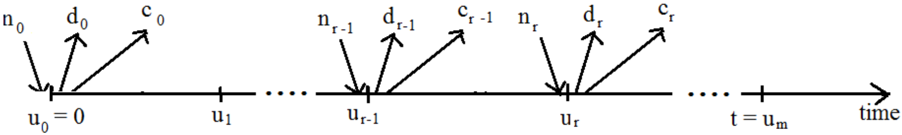
$$
\begin{aligned}
R(t) &= P(T > t) \\
&= P(T > u_m) \\
&= P(T > u_m \cap T > u_{m-1} \cap \cdots \cap T > u_2 \cap T > u_1 \cap T > u_0) \\
&= P(T > u_0) \cdot P(T > u_1 \mid T > u_0) \cdot P(T > u_2 \mid T > u_1) \\
&\cdots P(T > u_r \mid T > u_{r-1}) \cdots P(T > u_m \mid T > u_{m-1})
\end{aligned}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

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



Define:

- n_r = number at risk at time u_r = 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)



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

Then estimate,

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

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

THE FINAL KM-ESTIMATOR

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

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

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

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

In practice we have continous time. But this 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

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

GREENWOOD'S FORMULA FOR VARIANCE OF THE KM-ESTIMATOR

$$\widehat{Var(\hat{R}(t))} = \left(\hat{R}(t)\right)^2 \cdot \sum_{T_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

It can shown that for large n, $\hat{R}(t)$ is approximately normally distributed,

$$\hat{R}(t) \approx N(R(t), \widehat{SD(\hat{R}(t))})$$

Thus an approximate 95% confidence interval can be obtained for each t by

$$P\left(\hat{R}(t) - 1.96 \cdot \widehat{SD(\hat{R}(t))} \le R(t) \le \hat{R}(t) + 1.96 \cdot \widehat{SD(\hat{R}(t))}\right)$$

Recall thet MTTF $= \int_0^\infty R(t)dt$. Hence it seems natural to estimate MTTF by $\widehat{\text{MTTF}} = \int_0^\infty \hat{R}(t)dt$.

But - recall that

$$\hat{R}(t) = \prod_{T_{(i)} \le t} \frac{n_i - d_i}{n_i}$$

- If *largest observed time is a failure time: the last factor is 0, so* $\int_0^\infty \hat{R}(t)dt$ *is a finite number.*
- *If largest observed time is censored: the last factor is* $\frac{n_i - d_i}{n_i} > 0$. *So the estimate* $\hat{R}(t)$ *is constant and positive from this time on, making* $\int_0^\infty \hat{R}(t)dt = \infty$.

But - MINITAB uses the common convention:

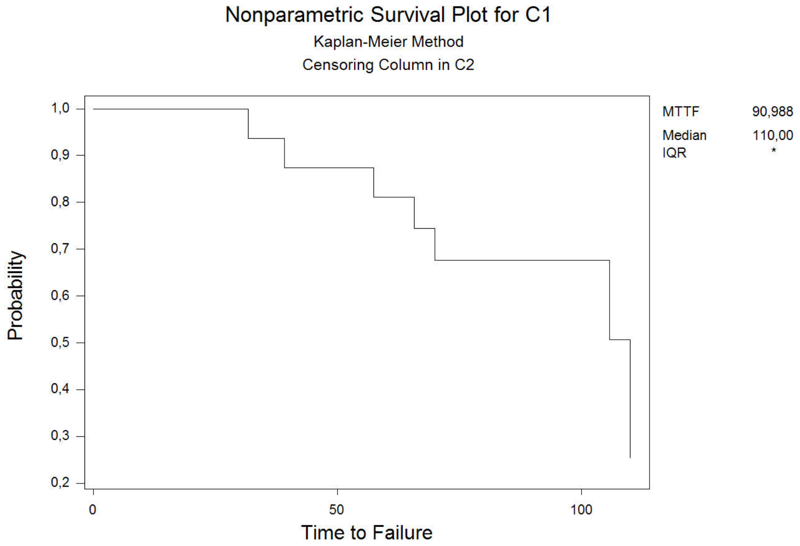
$$\widehat{MTTF} = \int_0^{\text{largest observed time}} \hat{R}(t)dt$$

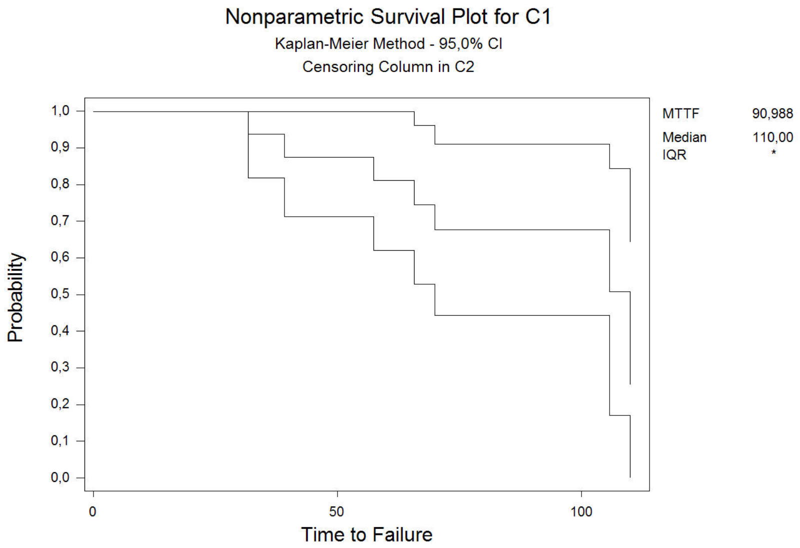
.

Row	C1	C2
1	31,7	1
2	39,2	1
3	57,5	1
4	65,0	0
5	65,8	1
6	70,0	1
7	75,0	0
8	75,2	0
9	87,5	0
10	88,3	0
11	94,2	0
12	101,7	0
13	105,8	1
14	109,2	0
15	110,0	1
16	130,0	0

	Number	Number	Survival	Standard	95,0% Normal CI	
Time	at Risk	Failed	Probability	Error	Lower	Upper
31,7000	16	1	0,9375	0,0605	0,8189	1,0000
39,2000	15	1	0,8750	0,0827	0,7130	1,0000
57,5000	14	1	0,8125	0,0976	0,6213	1,0000
65,8000	12	1	0,7448	0,1105	0,5283	0,9613
70,0000	11	1	0,6771	0,1194	0,4431	0,9111
105,8000	4	1	0,5078	0,1718	0,1711	0,8445
110,0000	2	1	0,2539	0,1990	0,0000	0,6440

KM-PLOT FOR THE CENSORED DATA




Nonparametric Survival Plot for C1
Kaplan-Meier Method - 95,0% CI
Censoring Column in C2
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Probability
0
50
100
Time to Failure
MTTF 90,988
Median 110,00
IQR *