3. Estimating the Lifetime Distribution – Censoring & the Kaplan Meier Estimate

CHRIS SUTTON, OCTOBER 2023

Our question in this topic:

How can we estimate $F_x(t)$?

Our wish list for a complete understanding of statistical models

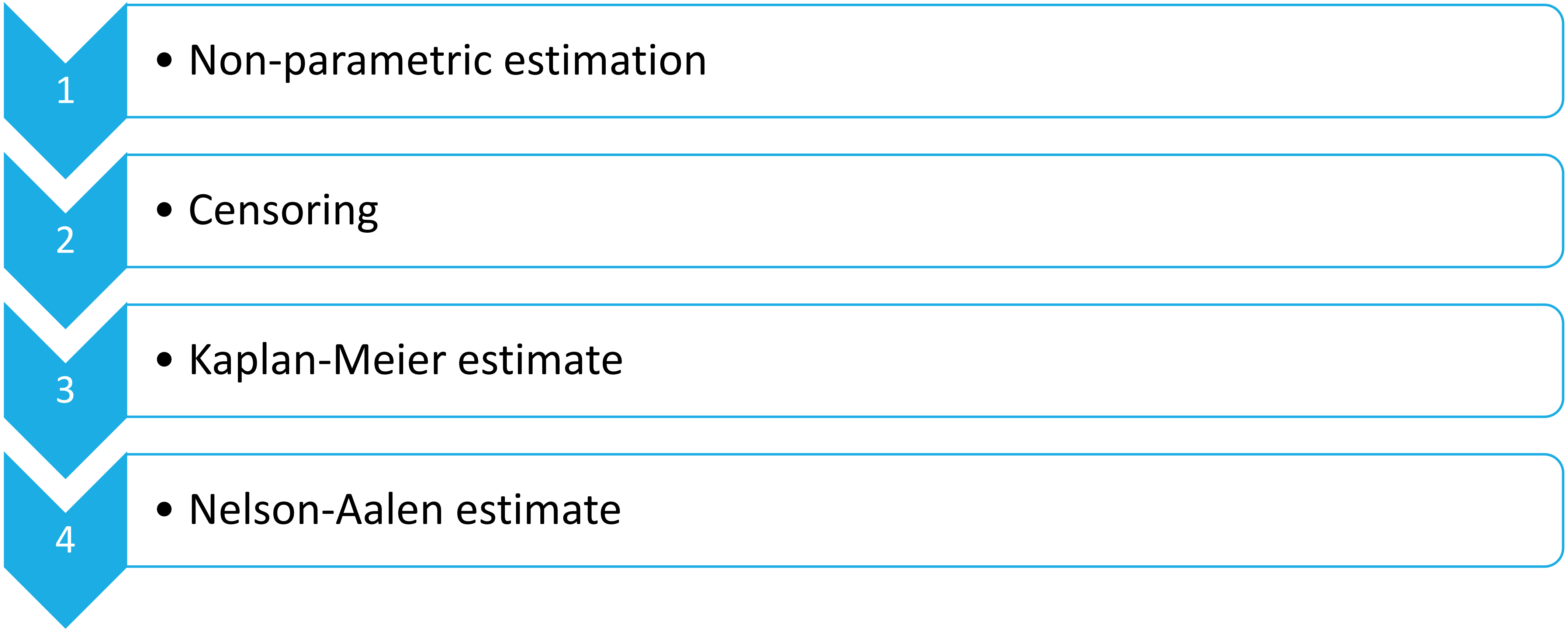
$F_x(t)$

$S_x(t)$

$f_x(t)$

μ_x

Topic outline



1
- Non-parametric estimation

2
- Censoring

3
- Kaplan-Meier estimate

4
- Nelson-Aalen estimate

Non-parametric estimation

Non-parametric

A model with no parameters

So, the [observed] data does all the work and completely defines the model

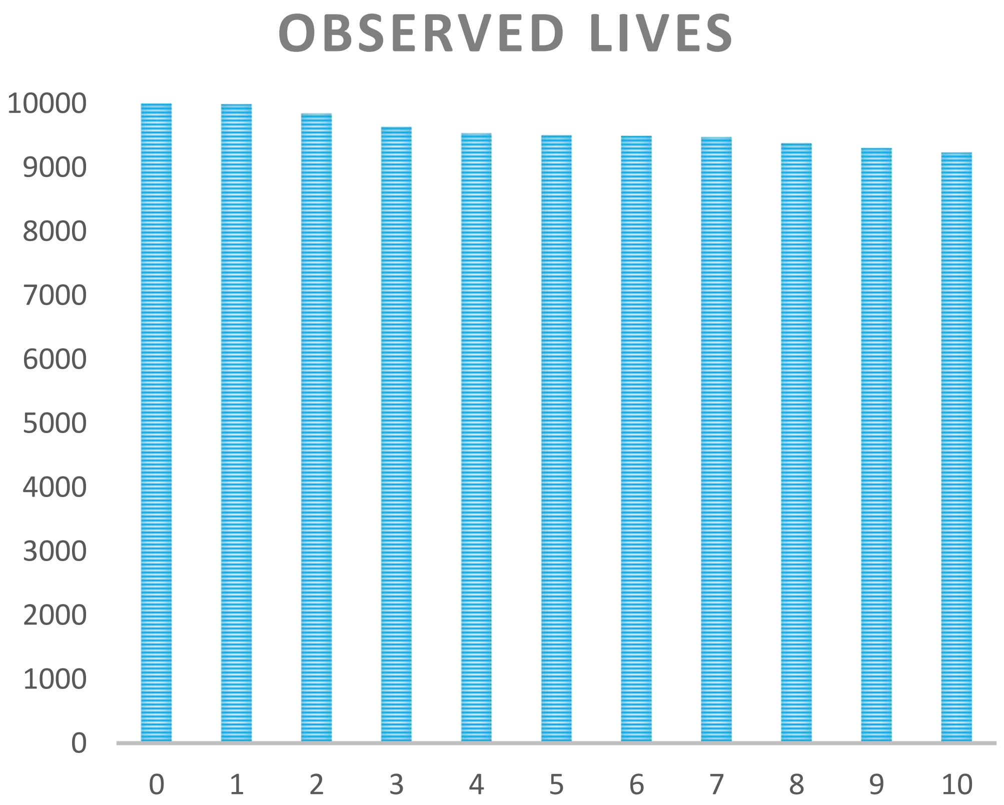
Attractive in medical statistics if we want observed results of a medical trial dominate rather than mathematical assumptions

But means that the nature and quality of our data is key

introducing non-parametric estimation

The idea here is to observe a large number of lives from t=0 onwards and use observed data to give S(t) and F(t)

- an empirical distribution function of T
- the data would give a step function
- this could be smoothed



Very simple example

We observe 1000 people and see how many are alive after t = 0, 1, 2, 3, 4, 5 years

Time t	Number observed n_t
0	1000
1	998
2	996
3	992
4	986
5	977

Non-parametric survival model

t	n_t	deaths	hazard	1 - hazard	Survival S(t)
0 – 1	1000	2	2/1000	0.998	0.998
1 – 2	998	2	2/998	0.997996	0.996
2 – 3	996	4	4/996	0.995984	0.992
3 – 4	992	6	6/992	0.993952	0.986
4 - 5	986	9	9/988	0.990872	0.977
5	977				

Practical problems with this approach

Would take > 100 years to complete a full study of human lives

We will lose track of some people

- this problem is called “censoring”
- just excluding these people will introduce bias
- e.g. if life assurance company collecting data we have the problem of lapsed policies

If we shorten the observation period to a small number of years and study people of ages simultaneously we introduce a new problem of sampling from cohorts with different distributions

Despite this, non-parametric estimation is important in medical statistics where lifetimes short

This week we will examine a non-parametric approach called the Kaplan-Meier estimator in some detail

Censoring

Censoring

This is where we do not observe the whole length of a lifetime but only an interval

Important concept in survival models as in practice we are nearly always relying on censored data

3 types of censoring to consider

Types of censoring

Right censoring

- observations stop before all lives have died [the most common type]
- we do not know the precise value of these lifetimes, only that they exceed the right-censored limit

Left censoring

- we do not know the precise time a life entered the state we are observing
- e.g. medical study for some condition where patients are only examined every 3 months

Interval censoring

- there is both left and right censoring
- e.g. a mortality investigation where we only given year of death

Censoring notation

let C_i = time at which the observation of the i^{th} life is censored

- a random variable

T_i = lifetime of that same i^{th} life

- also a random variable

then the observation is censored if $C_i < T_i$

- in this case the censoring is “random”
- we can also have cases of non-random or degenerately-random censoring

non-random censoring

type I censoring

- censoring times $\{C_i\}$ are known in advance

type II censoring

- observations continue until a pre-determined number of deaths observed

comments

In medical studies we need to be open to right-censoring – ending a medical trial early - dependant on the results observed

- unexpectedly positive results mean the treatment should be open to all
- unexpectedly negative results mean the treatment should be withdrawn

Censoring is “non-informative” if the set $\{C_i\}$ give no information about $\{T_i\}$

- random censoring is non-informative
- we must be very careful with which statistical methods are valid with informative censoring
- watch the wording in questions

Medical trials example (BMJ)

Survival (in months) of 49 patients with Duke’s C colorectal cancer (BMJ 1987) split into 2 groups

Lineolic acid treatment	Control treatment
1* 5* 6 6 9* 10 10 10* 12 12 12 12 12* 13* 15* 16* 20* 24 24* 27 32 34* 36* 36* 44*	3* 6 6 6 6 8 8 12 12 12* 15* 16* 18* 18* 20 22* 24 28* 28* 28* 30 30* 33* 42

* = censored observation

https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/12-survival-analysis

McIllmurray MB, Turkie W. Controlled trial of linoleic acid in Dukes' C colorectal cancer. BMJ 1987; 294 :1260, 295 :475.

Initial questions:

- what observations would you make from simply looking at this data set?
- what would you say about the nature of censoring in this trial?
- what challenges do we need to overcome in survival modelling here?

Kaplan-Meier estimate

K-M

The original 1958 paper

Kaplan E.L. & Meier P. (1958) ‘Nonparametric estimation from incomplete observations’ *Journal of the American Statistical Association* vol. 53 pp.457–481

A good example of its application today

Dudley, W.N., Wickham, R. & Coombs, N. (2016) ‘An introduction to survival statistics: Kaplan-Meier Analysis’ *Journal of the Advanced Practitioner in Oncology* vol.7(1) pp.91-100

Introduction to Kaplan-Meier

a [non-parametric] method for estimating the survival function $S_x(t)$ [*and hence also the lifetime distribution $F_x(t)$*] which allows for censoring

- in the last topic we introduced the **force of mortality** μ_x for a theoretical, continuous lifetime distribution $F_x(t)$
- here observed data will give us a discrete distribution from which we are trying to estimate $F_x(t)$ and we will use the **hazard** λ (which is analogous to μ_x)

Kaplan-Meier makes no reference to age x, only to duration t, the time from the beginning of observation

observe a population of n lives

non-informative, right censoring takes place

Setting up the Kaplan-Meier scenario

we observe m deaths at times $t_1, t_2, \ldots, t_k$

we order the times: $t_1 < t_2 < \ldots < t_k$

$k \leq m$

- k does not necessarily equal m as we could observe more than one death at a particular observation point

assume d_j deaths are observed at time t_j $(0 \leq j \leq k)$

so $d_1 + d_2 + \ldots + d_k = m$

remaining n-m lives are censored with c_j lives censored between times t_j and t_{j+1}

- we define $t_0 = 0$ and $t_{k+1} = \infty$

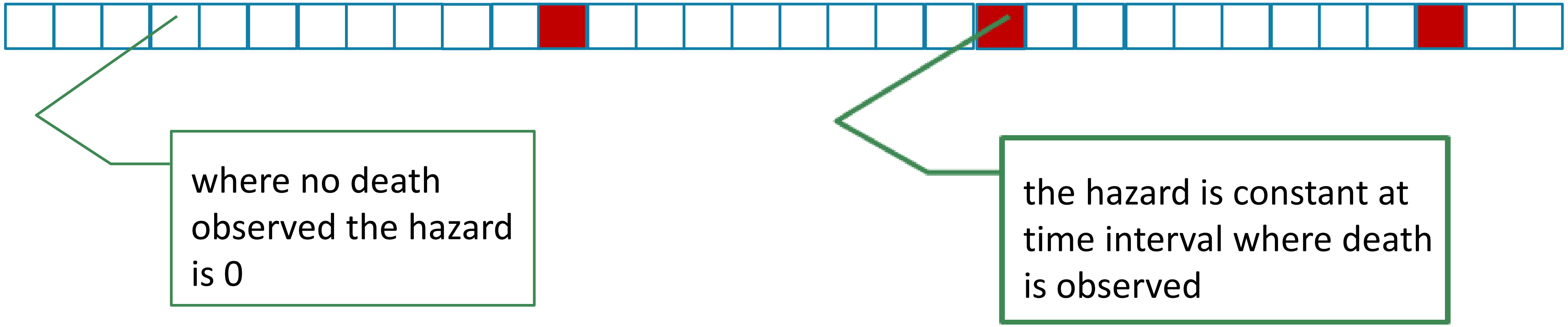
then $c_1 + c_2 + \ldots + c_k = n - m$

Kaplan-Meier assumptions

Kaplan-Meier estimation then assumes:

1. the hazard of experiencing the event [death] is zero at all times except where the event is actually observed in our sample
2. the hazard of experiencing the event at time t_j is d_j / n_j where n_j is the “risk set” or the number of lives still at risk of experiencing the event just prior to t_j
3. censored lives are removed just after the event (so lives censored at t_j are removed after those who die at t_j and therefore censored lives are still in the risk set at t_j for the hazard calculation)

$\hat{\lambda}_j$



where no death observed the hazard is 0

the hazard is constant at time interval where death is observed

$$\hat{\lambda}_j = \frac{d_j}{n_j} \qquad (1 \le j \le k)$$

This is actually a maximum likelihood estimate given our data set

$\hat{S}(t)$

$\lambda_j = P[\, T = t_j \mid T \geq t_j \,]$

- remembering that λ is the discreet distribution version of μ_x

then $S(t) = 1 - F(t) = \prod_{t_j \leq t} (1 - \lambda_j)$ and we can estimate the survival function with

Kaplan-Meier estimator

$$\boxed{\hat{S}(t) = \prod_{t_j \leq t} (1 - \hat{\lambda}_j)}$$

The Kaplan-Meier estimator

The Kaplan-Meier estimator for the survival function S(t) is $\hat{S}(t)$

- found by multiplying together survival probabilities in each interval up to and including t
- hence is sometimes called the “product limit estimate”

This estimator:

- is always specified in terms of duration t not age x
- is constant for durations after the last observed death
- is not defined for durations after the last censoring

Its main application is in medical statistics

- comparing lifetime distributions for 2 or more groups undergoing different treatments

Medical trials example (BMJ)

Survival (in months) of 49 patients with Duke’s C colorectal cancer (BMJ 1987) split into 2 groups

Lineolic acid treatment	Control treatment
1* 5* 6 6 9* 10 10 10* 12 12 12 12 12* 13* 15* 16* 20* 24 24* 27 32 34* 36* 36* 44*	3* 6 6 6 6 8 8 12 12 12* 15* 16* 18* 18* 20 22* 24 28* 28* 28* 30 30* 33* 42

* = censored observation

https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/12-survival-analysis

McIllmurray MB, Turkie W. Controlled trial of linoleic acid in Dukes' C colorectal cancer. BMJ 1987; 294 :1260, 295 :475.

Initial questions:

- what observations would you make from simply looking at this data set?
- what would you say about the nature of censoring in this trial?
- what challenges do we need to overcome in survival modelling here?

Nelson-Aalen estimate

μ_s with λ_j

the Nelson-Aalen estimate is an alternative to Kaplan-Meier, adding to it

- it combines continuous parts of the distribution (which have hazard μ_s) and discrete parts (with hazard λ_j)

We define the “integrated hazard” A_t to be

$$A_t = \int_0^t \mu_s \, ds + \sum_{t_j \le t} \lambda_j$$

and the Nelson-Aalen estimator of this integrated hazard is

$$\hat{A}_t = \sum_{t_j \le t} \frac{d_j}{n_j}$$