INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: SY BSc

Subject : Statistical and Risk Modelling - 1

Chapter: Unit 1 Chapter 2

Chapter Name: Nonparametric estimation of a lifetime distribution

Today's Agenda

1. Censoring in lifetime data
    1. Censoring - Introduction
    2. Types of Censoring
        a. Right Censoring
        b. Left Censoring
        c. Interval Censoring
        d. Type I & Type II Censoring
        e. Informative & Non-Informative Censoring

2. Non-Parametric Estimates
    1. Kaplan-Meier Estimate
    2. Nelson-Aalen Estimate
    3. Relationship between KM and NA

1 Censoring in lifetime data

1.1a Censoring – Introduction

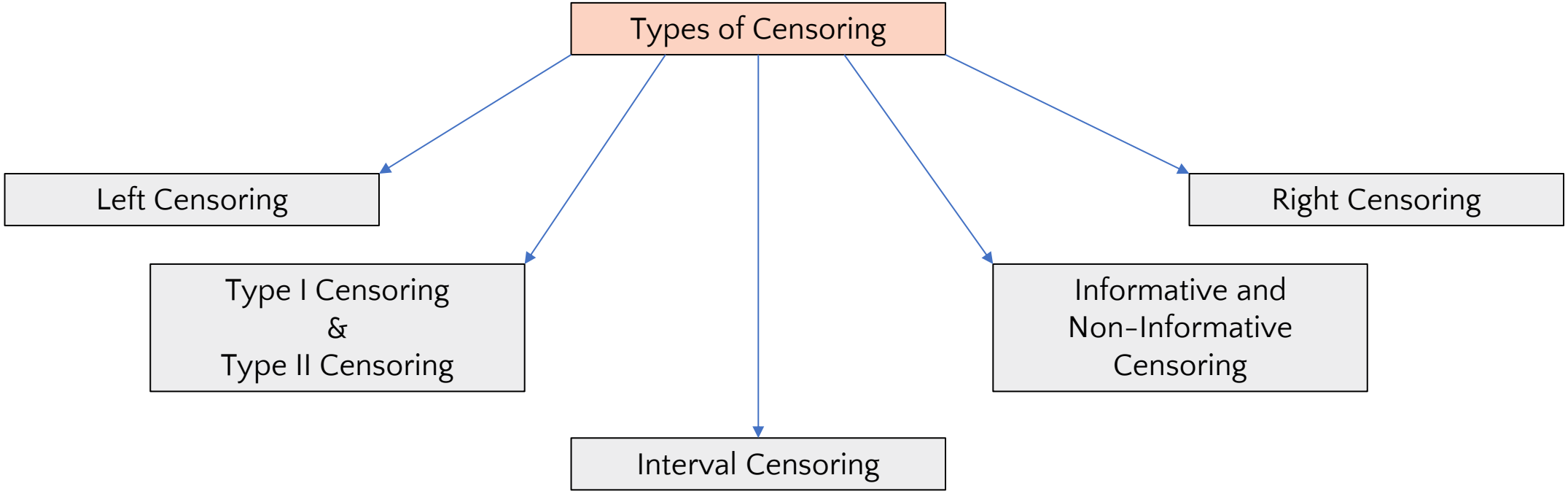
Censoring is a situation in which the value of a measurement or observation is only partially known. Censoring also takes place when a value occurs outside the range of a measuring instrument.
For example, suppose a study is conducted to measure the impact of a drug on mortality rate. In such a study, it may be known that an individual's age at death is *atleast*75 years (but may be more). Such a situation could occur if the individual withdrew from the study at age 75, or if the individual is currently alive at the age of 75.

Thus, subjects are said to be censored :-

- If they are lost to follow up.
- If the study ends before they die or have an outcome of interest.
- If the researcher doesn't know for how long before the study, the subject had been facing a particular condition/disease.

1.2 Types of Censoring

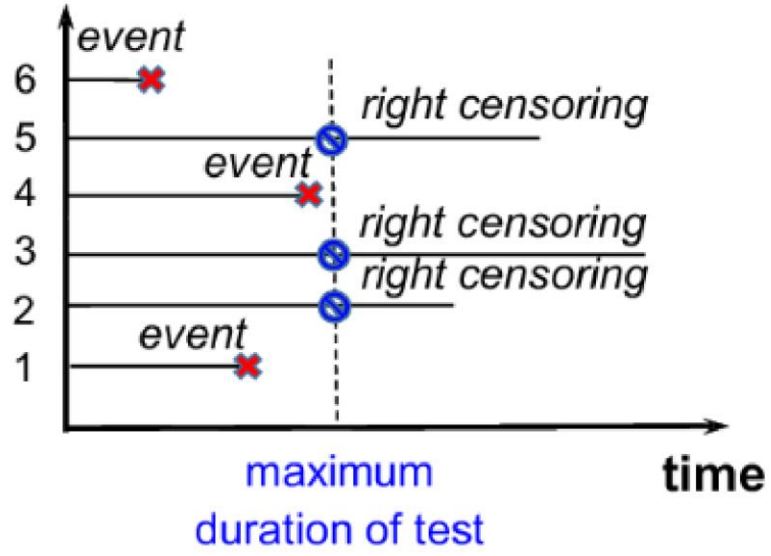
There are different types of censoring based on the kind or time of information missing. They are :-



1.2a Right Censoring

Right censoring–a data point is above a certain value but it is unknown by how much.

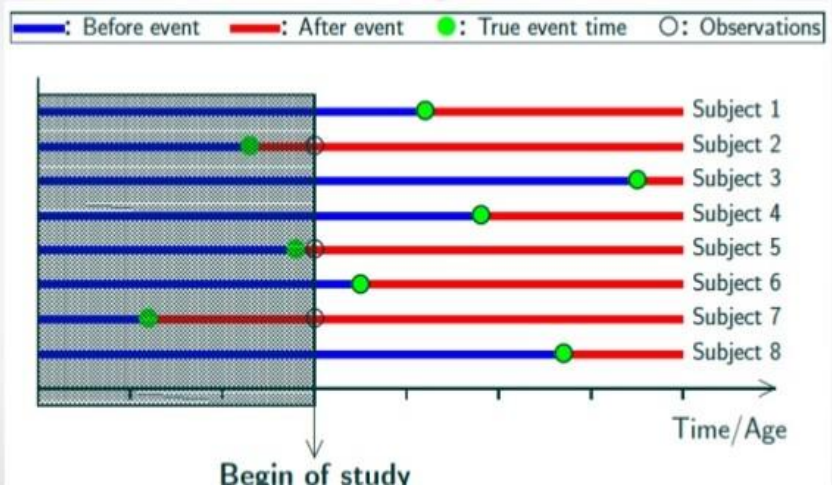
- Right censoring is the most common occurrence for lifetime data.
- It means that we are not certain what happened to people after some point in time.
- This happens when some people cannot be followed the entire time because they were lost to follow-up or withdrew from the study.
- For example, we consider patients in a clinical trial to study the effect of treatments on stroke occurrence. The study ends after 5 years. Those patients who have had no strokes by the end of the year are censored.



1.2b Left Censoring

Left censoring–a data point is below a certain value but it is unknown by how much.

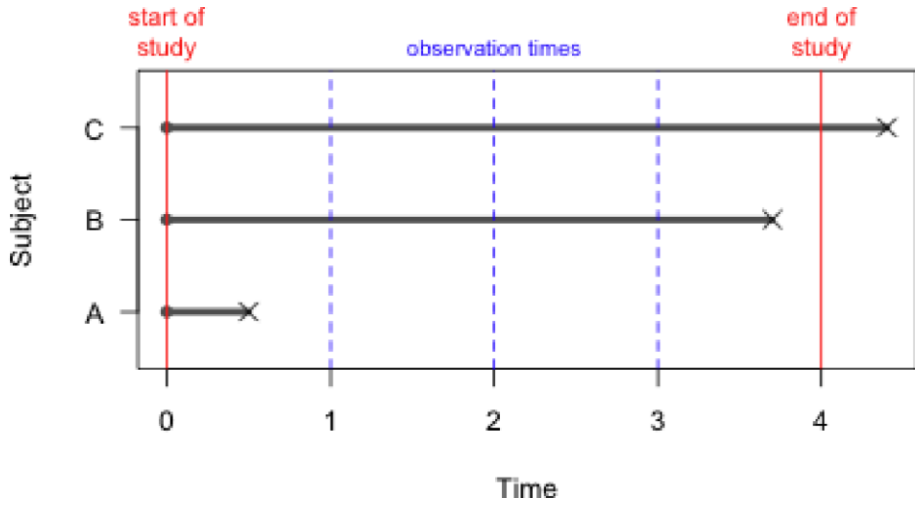
- This occurs less frequently.
- Left Censoring is when we are not certain what happened to people before some point of time.
- Commonest example is when people already have the disease of interest when the study starts. Here we do not know the exact date of onset of disease.
- Can you think of a recent example of this kind of a disease and when such censoring might have happened?
- For example as shown in the graph, subject 2,5 and 7 are left censored.



1.2c Interval Censoring

Interval censoring–a data point is somewhere on an interval between two values

- Interval Censoring is when we don't know when exactly the event happened, but are aware of the interval of time in which the event happened.
- For example, we know that the patient was well at the time of start of the study and was diagnosed with disease at time of end of study, so when did the disease actually begin? All we know is the interval.
- Interval censoring can also occur if the observations are studied at intervals rather than continuously.
- For Example, as shown in the graph, subject A is interval censored between times 0 and 1, subject B is censored in the interval of 3 & 4. Subject C is right censored.



1.2d Type I & Type II Censoring

Type I Censoring

- ***Type I censoring*** occurs if an experiment has a set number of subjects or items and stops the experiment at a predetermined time, at which point any subjects remaining are right-censored. Here, the time is fixed!

- Example: Suppose a clinical trial is conducted for a new lung cancer treatment. A Researcher is studying the effect of this treatment on lung cancer on 100 participants.
- She will end the study exactly one year from now. That's an example of a Type I censoring.

Type II Censoring

- ***Type II censoring*** occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number are observed to have failed; the remaining subjects are then right-censored. Here, the number of events is fixed!
- Example: Supposing in the same example, if it is decided to end the trial if a certain number of participants in the clinical trial face adverse reactions and side effects due to the treatment.
- If 10 participants facing adverse reactions is the benchmark and it's breached, then Type II censoring occurs as the study will end post that.

1.2e Informative & Non-Informative Censoring

- ***Informative censoring*** occurs when events are not counted in the analysis due to reasons related to the study design. It affects the interest of subjects in the study.
- Censoring is ***non-informative*** if it gives no information about the lifetimes. It does not affect the interest of study.
- Censoring in survival analysis should be "non-informative," i.e. participants who drop out of the study should do so due to reasons unrelated to the study.
- Informative censoring occurs when participants are lost to follow-up due to reasons related to the study.

- For example, in a study comparing disease-free survival after two treatments for cancer, the control arm may be ineffective, leading to more recurrences and patients becoming too sick to follow-up. On the other hand, patients on the intervention arm may be completely cured by an effective treatment and may no longer feel the need to follow-up.
- If these participants are routinely censored, the true treatment effect will not be picked up and the results of the study will be biased. Disease-free survival rates would be based on the patients who continued to be followed-up in the study, and would be overestimated for the control arm and underestimated for the treatment arm.

Question

- An investigation is carried out into the mortality rates of married male accountants. A group of 10,000 married male accountants is selected at random on 1 January 2016. Each member of the sample group supplies detailed personal information as at 1 January 2016 including name, address and date of birth. The same information is collected as at each 1 January in the years 2017, 2018, 2019 and 2020. The investigation closes in 2020. Describe the ways in which the available data for this investigation may be censored.

Solution

- There will be left censoring of all lives that change marital status from single (or divorced or widowed) to married during the investigation. We only know that the change of status occurred since the previous set of information was collected.
- There will be interval censoring if the exact date of death is unknown, eg if only the calendar year of death is known.
- There will be random censoring of all lives that change marital status from married to divorced or widowed, or give up accountancy, and consequently no longer qualify as participants in the mortality investigation. There will also be random censoring of all lives from whom data cannot be collected.
- There will be right censoring of all lives that survive until the end of the investigation in 2020.

Question

- You have been asked to investigate whether the rate of ill-health retirement of the employees of a large company varies with their duration of employment.
  - the date on which an employee was hired
  - the calendar year in which they retired, if an employee left employment as a result of
  - ill-health retirement
  - the date of retirement, if an employee reached the normal retirement age of 65
  - the date of leaving, if an employee left the company for any other reason.

- In the context of this investigation consider the following types of censoring and in each case:
  - describe the nature of the censoring
  - state whether or not that type of censoring is present in these data
  - if that particular type of censoring is present, explain how it arises.

  i. Left censoring
  ii. Right censoring
  iii. Interval censoring
  iv. Informative censoring

Solution

Left censoring

- Data in this study would be left censored if the censoring mechanism prevented us from knowing when an employee joined the company.
- This is not present because the exact date of joining is given.

Right censoring

- Data in this study would be right censored if the censoring mechanism cuts short observations in progress, so that we are not able to discover if and when an employee retired as a result of ill health.
- Here there is right censoring of those lives who leave employment before their normal retirement date for reasons other than ill health.

Interval censoring

- Data in this study would be interval censored if the observational plan only allows us to say that the duration of employment at the date of ill-health retirement fell within some interval of time (and does not allow us to find the exact duration of employment).
- Here we know the calendar year of ill-health retirement and the date of employment, so we will know that the duration of employment falls within a one-year interval. Interval censoring is present.

Solution

Informative censoring

- Censoring in this study would be informative if the censoring event divided individuals into two groups whose subsequent experience of ill-health retirement was thought to be different.
- Here the censoring event of leaving the company might be suspected to be informative. Those who leave are more likely to be in better health (less likely to have retired on ill-health grounds had they remained in employment) because they probably left to take another (perhaps better paid and more responsible) job for which they may have been required to pass a medical examination. Similarly, those not resigning their jobs are more likely to retire on ill-health grounds. Informative censoring is present if these groups have different subsequent experience.

2 Non-Parametric Estimates

- **Non-parametric** estimates do not rely on any underlying distributions. The distributions are developed empirically. Non-parametric methods are called distribution free.
- In this chapter, we focus on estimating the survival function of the failure time without any specified distribution assumption
- The two very important and widely used are :-
  1) ***Kaplan Meier Estimate***
  2) ***Nelson Aalen Estimate***

2.1 Kaplan-Meier Estimate (KM)

In this model, we assume there is non informative right censoring present. By assuming that the type of censoring present is non-informative, we are assuming that the mortality of those lives remaining in the group under observation is not systematically higher or lower than the mortality of the lives that have been censored.

We will consider lifetimes as a function of time *t* without mention of a starting age *x*.

•Let $t_1 < t_2 < \ldots . < t_m$ denote the distinct ordered times of death (not counting censoring times). We also assume that more than one deaths could occur at the same times.

•Let d_i be the number of deaths at t_i , and

•Let n_i be the number alive just before t_i . This is the number exposed to risk at time t_i .

•Let c_j be the lives that are censored between times t_j and t_{j+1}. Thus c_j represents the number of lives that are removed from the investigation between t_j and t_{j+1} for a reason other than the decrement we are investigating.

2.1 KM Important Points

- The hazard of experiencing the event is zero at all durations except those where an event actually happens in our sample.
- The hazard of experiencing the event at any particular duration, t_j, when an event takes place is estimated by d_j/n_j. i.e $\widehat{\lambda_j} = \frac{d_j}{n_j}$
- The conditional probability of surviving time t_i is the complement $1 - d_i/n_i$.
  i.e. $1 - \hat{\lambda}_j = \frac{n_j - d_j}{n_j} = \frac{number\ of\ survivors}{number\ at\ risk}$
- If any of the individuals are observed to be censored at the same time as one of the deaths, the convention is to treat the censoring as if it happened shortly afterwards, *ie* the deaths are assumed to have occurred first.
- The Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times.

2.1 The Kaplan-Meier Estimate

The overall probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t . Then the Kaplan Meier or product limit estimate of the survival function is:

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

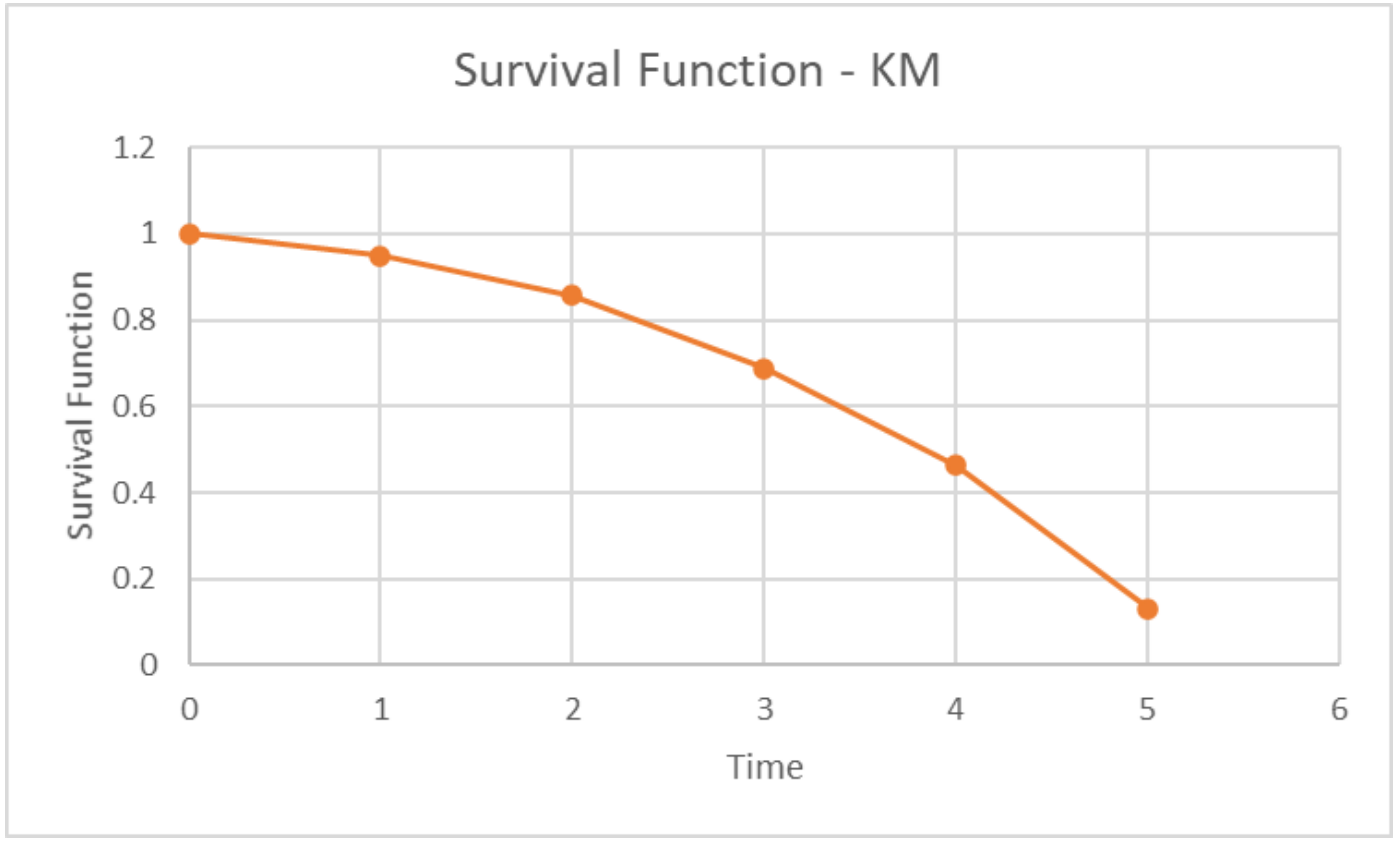
Because the Kaplan-Meier estimate involves multiplying up survival probabilities, it is sometimes called the *product limit estimate.*

Kaplan-Meier estimate of Survival Function

$$\hat{S}(\mathrm{t}) = \prod_{t_j \le t} (1 - \hat{\lambda}_j) = \prod_{t_j \le t} \frac{n_j - d_j}{n_j}$$

2.1 The Kaplan-Meier Estimate

A graph of a typical KM survival function:



Question

A clinical trial is being carried out to test the effectiveness of a new drug. Hundred patients were involved in the trial, which followed them for 5 years from the start of their treatment. The following data show the period in complete years from the start of treatment to the end of observation for those patients who died or withdrew from the trial before the end of the 5-year period. We assume that we get to know about the withdrawal at the end of the year after deaths.

Year	1	2	3	4	5
Deaths	5	9	15	19	25
Withdrawals	3	7	3	4	5

i. Calculate the Kaplan-Meier estimate of the survival function.

Solution

The solution for the question would be:

				Survived			
1	100	3	5	95	5/100 = 0.05	0.95	0.95
2	92	7	9	83	8/92 = 0.098726	0.902174	0.857065
3	76	3	15	61	15/76 = 0.197368	0.802632	0.687908
4	58	4	19	39	19/58 = 0.327586	0.672414	0.462559
5	35	5	25	10	25/35 = 0.714286	0.285714	0.13216

2.1 Variance of KM

Kaplan-Meier estimates have wide range of application. They are also used in comparison of lifetime distributions. Thus having the statistical properties is vital. We already learnt how to find the survival estimate. Now we look at the variance. An approximate **formulae for the variance of $\tilde{F}(t)$ are available.** We're using $\tilde{F}(t)$ to denote the estimator of the distribution function at time t and $\hat{F}(t)$ to represent our estimate.

The Variance is given by the Greenwood's formulae:

Greenwood's Formula

$$var[\tilde{F}(t)] \approx \left(1-\tilde{F}(t)\right)^2 \sum_{t_j \le t} \left(\frac{d_j}{n_j(n_j-d_j)}\right)$$

Can you find the variance of $\hat{S}(t)$?
Hint: $Var(1-x) = Var(x)$

2.2 Nelson–Aalen Estimate (NA)

An alternative non-parametric approach is to estimate the integrated hazard. The definitions of t_j, d_j , n_j and c_j remain same as in KM model.

This is denoted by Λ(capital lambda).

Nelson-Aalen estimate of the integrated hazard function

$$\widehat{\Lambda}_{t_i} = \sum_{j=1}^{i} \frac{d_j}{n_j}$$

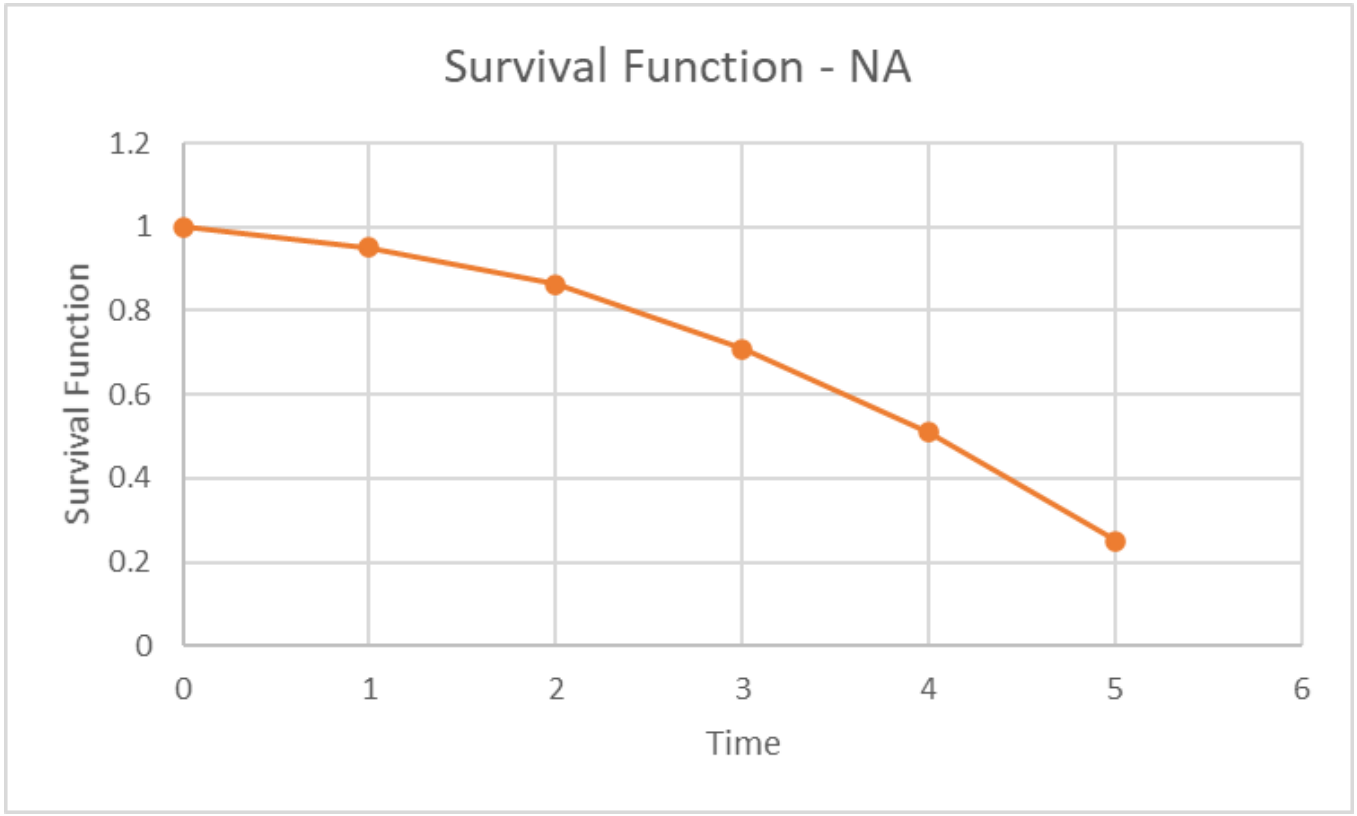
Now that we know our hazard function, we find the survival estimate.

Nelson-Aalen estimate of Survival function

$$\tilde{S}(t) = \exp[-\widehat{\Lambda}_t]$$

2.2 Nelson–Aalen Estimate (NA)

A graph of a typical NA survival function:



2.2 Solved Example

We will Take the same example as that in the KM estimate. This time we find the NA estimate:

				Survived			
1	100	3	5	95	5/100 = 0.05	0.05	0.951229
2	92	7	9	83	9/92 = 0.097826	0.147826	0.862581
3	76	3	15	61	15/76 = 0.197368	0.345195	0.708083
4	58	4	19	39	19/58 = 0.327586	0.672781	0.510288
5	35	5	25	10	25/35 = 0.714286	1.387066	0.249807

2.2 Variance of NA

There is a formula for the variance of the Nelson-Aalen estimator:

Variance of the Integrated hazard function

$$var[\widetilde{\Lambda}_t] \approx \sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}$$

This formula gives the variance of the integrated hazard estimator, not the variance of $\tilde{F}(t)$.

2.3 Relation between KM and NA

To be clear with estimates, here we will denote the Kaplan-Meier estimate of the distribution function by $\hat{F}_{\mathrm{KM}}(\mathrm{t})$ and the Nelson-Aalen estimate of the distribution function by $\hat{F}_{\mathrm{NA}}(\mathrm{t})$.
So we have:

$$\hat{F}_{\mathrm{KM}}(\mathrm{t}) = 1 - \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

Using the approximation $e^x \approx 1 + x$ for small x, and replacing x by $\frac{d_j}{n_j}$, we have:

$$\hat{F}_{KM}(t) \approx 1 - \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j}\right) = 1 - \exp(-\hat{\Lambda}_t) = \hat{F}_{NA}(t)$$

Note: Since $e^x \ge 1 + x$ for all values of x, the NA estimate will always be greater than KM estimate at any given time. The NA estimate has been shown to perform better than the KM estimate for small samples, but in many circumstances the estimates will be very similar.

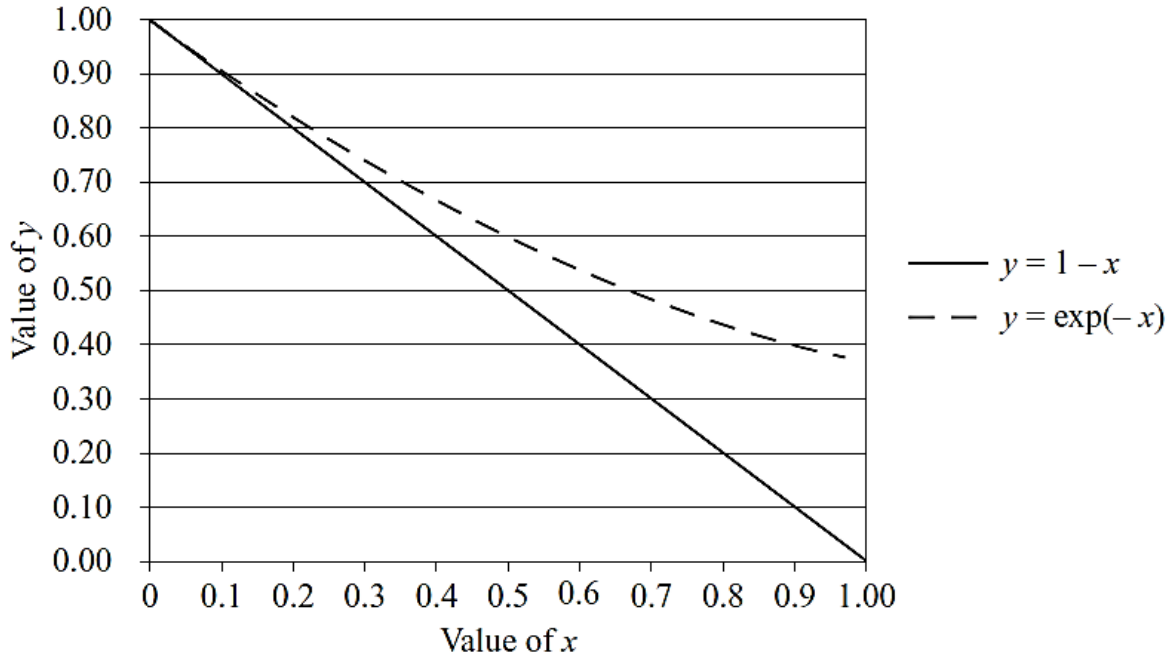
Question

CT4 September 2017 Q10

(i) Write down the formulae for the Kaplan-Meier estimator $\hat{S}(t)$ and Nelson-Aalen estimator $\tilde{S}(t)$ of survival in the presence of a stated hazard, defining all terms used. [2]

The following graph shows the functions:

$y = 1 - x$ and
$y = e^{-x}$ over the range $0 \le x \le 1$.



(ii) Demonstrate that the Nelson–Aalen estimator is never lower than the Kaplan–Meier estimator. [2]

Solution

(i) The Kaplan-Meier estimator is

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j} \right)$$

and the Nelson-Aalen estimator is

$$\tilde{S}(t) = \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j} \right),$$

where d_j represents the number of occurrences of the event of interest at duration t_j,

and n_j represents the number exposed to the hazard at duration t_j.

Solution

(ii) Expanding into the individual terms:

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right).\left(1 - \frac{d_2}{n_2}\right).....\left(1 - \frac{d_j}{n_j}\right), \qquad +½$$

and

$$\tilde{S}(t) = \exp\left(-\sum_{t_j <= t} \frac{d_j}{n_j}\right) = \exp\left(-\frac{d_1}{n_1}\right).\exp\left(-\frac{d_2}{n_2}\right).....\exp\left(-\frac{d_j}{n_j}\right). \qquad +½$$

As each d_j/n_j must be between 0 and 1 the chart shows each term in the Nelson-Aalen estimator is no lower than the parallel in the Kaplan-Meier. +½

Hence the Nelson-Aalen estimator is always no lower than the Kaplan-Meier estimator. +½

[2]

Question

A trial is conducted amongst 20 patients who have suffered from eczema but are in remission (that is, they are clear of the condition). The trial is to assess whether continuing with periodic doses of a certain steroid cream in remission reduces the rate at which eczema recurs. Patients are invited to tests every 3 months for a period of up to 5 years from when first declared to be in remission.

(iii) Describe THREE types of censoring present in the investigation. [3]

Solution

(iii)

- **Interval censoring** is present because the tests only take place every three months and recurrence of eczema could occur between tests. +1
- **Type 1** censoring is present because it is specified in advance that the study will end after 5 years. +1
- **Random** censoring is present as for patients who leave the study, the time of their censoring can be considered a random variable. +1
- **Right** censoring is present for patients still free of eczema after 5 years or patients who left the study, as we do not know when the reoccurrence of eczema happened, just that it happened after a certain date. +1
- **Non-informative** censoring could be said to be present as we have no reason to believe that those patients who left the study were any more or less likely to have the eczema recur than those who remained in the study. +1
- **Informative** censoring could be said to be present as we could argue that those who left the study may have done so because they considered themselves cured, and were therefore less likely to suffer a recurrence than those still in the study. +1

[max. 3]

Question

The data for the trial are subdivided into a group who continued to receive the steroid cream, and a control group who did not receive the steroid cream. The data for the patients in the trial showing the quarterly test at which eczema recurred, or censoring occurred, are as follows (an * indicates a patient who was censored):

For group receiving steroid cream: 3, 5, 6*, 7*, 10, 10, 12*, 14*, 18, 19*
For control group: 6, 8, 8, 10*, 11*, 12*, 14, 15*, 18, 18

(iv) Calculate the Kaplan-Meier estimates of the survival function for remaining clear of eczema for:
(a) the group who continued to receive the steroid cream; and
(b) the control group. [8]

(v) (a) Recommend, without performing any calculations, a method of establishing whether the hazard of eczema returning is statistically lower for those continuing to receive the steroid cream.
(b) Comment on the chance of being able to conclude from the trial data that continuing to receive the steroid cream reduces the risk of recurrence of eczema. [3]
[Total 18]

Solution

(iv) For the group continuing to receive steroid cream:

t_j	n_j	d_j	c_j	λ_j	$1-\lambda_j$
3	10	1	0	1/10	9/10
5	9	1	2	1/9	8/9
10	6	2	2	1/3	2/3
18	2	1	1	1/2	1/2

+2

The Kaplan-Meier estimate of the survival function, $S(t)_{KM}$, is

t	$S(t)_{KM}$
$0 \le t < 3$	1
$3 \le t < 5$	9/10
$5 \le t < 10$	4/5
$10 \le t < 18$	8/15
$18 \le t < 20$	4/15
+1	*+1*

+2

Solution

For the control group:

t_j	n_j	d_j	c_j	λ_j	$(1-\lambda_j)$
6	10	1	0	1/10	9/10
8	9	2	3	2/9	7/9
14	4	1	1	1/4	3/4
18	2	2	0	2/2	0

+2

The Kaplan-Meier estimate of the survival function, $S(t)_{KM}$, is:

t	$S(t)_{KM}$
$0 \leq t < 6$	1
$6 \leq t < 8$	9/10
$8 \leq t < 14$	7/10
$14 \leq t < 18$	21/40
$18 \leq t < 20$	0

+2

Solution

(v) (a) In order to assess whether the risk is statistically lower a simple and quick approach would be to calculate confidence intervals around each survival function. +½

If the confidence intervals do not overlap the survival rate is statistically higher or lower at the chosen confidence level. +½

For the Kaplan Meier estimate the variance can be estimated using Greenwood's formula, +½

which is:

$$\text{Var}[\hat{S}(t)] \approx (\hat{S}(t))^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}.$$ +½

Methods such as the log-rank test or Wilcoxon's test could be used. +½

Solution

(b) In this case it is unlikely it could be shown that continuing to receive steroid cream statistically reduces the risk of recurrence, +½

as the sample size is small +½

and the survival rates do not appear markedly better for the group receiving steroid cream. +½

[max. 3]

[Total 18]

Quick Recap

- Censoring is a situation in which the value of a measurement or observation is only partially known. The different types of censoring are right censoring, left censoring, interval censoring, Type I & II censoring, informative & non-informative censoring, etc...
- Right censoring occurs when a data point is above a certain value but it is unknown by how much.
- Left censoring occurs when a data point is below a certain value but it is unknown by how much.
- Interval censoring occurs when a data point is somewhere on an interval between two values but we don't know exactly where.
- Type I censoring occurs if an experiment has a set number of subjects or items and stops the experiment at a predetermined time, at which point any subjects remaining are right-censored. Here the time if fixed.
- Type II censoring occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number are observed to have failed; the remaining subjects are then right-censored. Here, the number of events is fixed.
- Informative censoring occurs when events are not counted in the analysis due to reasons related to the study design. It affects the interest of subjects in the study.
- Censoring is non-informative if it gives no information about the lifetimes. It does not affect the interest of study.

Continued

- Non-parametric estimates do not rely on any underlying distributions. The distributions are developed empirically. Non-parametric methods are called distribution free. We learnt about the Kaplan-Meier and Nelson-Aalen estimate, both of which do not rely on any underlying distribution.
- For KM estimate, we assume there is non informative right censoring present. The hazard of experiencing the event is zero at all durations except those where an event actually happens in our sample.
- If any of the individuals are observed to be censored at the same time as one of the deaths, the convention is to treat the censoring as if it happened shortly afterwards, *ie* the deaths are assumed to have occurred first.
- The Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times.
- Kaplan-Meier estimate of Survival Function : $\hat{S}(t) = \prod_{t_j \le t}(1 - \hat{\lambda}_j) = \prod_{t_j \le t} \frac{n_j - d_j}{n_j}$
- Greenwood's Formula : $var[\tilde{F}(t)] \approx \left(1 - \tilde{F}(t)\right)^2 \sum_{t_j \le t} \left(\frac{d_j}{n_j(n_j - d_j)}\right)$
- An alternative non-parametric approach, the NA estimate, uses the integrated hazard. The definitions of t_j, d_j , n_j and c_j remain same as in KM model. This is denoted by Λ(capital lambda).
- Nelson-Aalen estimate of the integrated hazard function : $\hat{\Lambda}_{t_i} = \sum_{j=1}^{i} \frac{d_j}{n_j}$
- Nelson-Aalen estimate of Survival function : $\tilde{S}(t) = \exp[-\hat{\Lambda}_t]$
- Variance of the Integrated hazard function : $var[\tilde{\Lambda}_t] \approx \sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}$