Survival Models

Rahul Telang

Survival

- **Survival analysis** is a set of statistical approaches used to find out the time it takes for an event of interest to occur. Survival analysis is used to study the **time** until some **event** of interest (often referred to as **death**) occurs. Time could be measured in years, months, weeks, days, etc. The event of interest could be anything of interest. It could be an actual death, a birth, a retirement, etc.

Where is it used?

- **Survival analysis** is used in a variety of field such as:
  - Cancer studies for patient survival time analyses.
  - Sociology for “event-history analysis”.
  - In Engineering for “failure-time analysis”.
  - Time until product failure.
  - Time until a warranty claim.
  - Time until a process reaches a critical level.
  - Time from initial sales contact to a sale.
  - Time from employee hire to either termination or quit.
  - Time from a salesperson hire to their first sale.

Method

- most of the survival analyses use following methods.

  Kaplan-Meier plots to visualize survival curves.

  Nelson-Aalen plots to visualize the cumulative hazard.

  Log-rank test to compare the survival curves of two or more groups

  Cox proportional hazards regression to find out the effect of different covariates like age, sex, weight on survival.

- Here, we start by defining fundamental terms of survival analysis, :
  - Survival time and event.
  - Censoring of data.
  - Survival function and hazard function.

What makes Survival unique?

- The data involves an interval till an event occurs. Since it is timing data, it is widely understood that data generation process is not normal distribution. This make regression analysis infeasible.
- Survival analysis focuses on the occurrence of an event of interest (e.g., birth, death, retirement). But there is still a possibility that the event ***may not be observed for various reasons***. Such observations are known as **censored observations**.
- Censoring may arise in the following ways :
  - A patient has not (yet) experienced the event of interest (death or relapse in our case) within the study period.
  - A patient is not followed anymore.
  - A employee may not quit if one is studying duration of employment.

Censoring

- There are three general types of censoring, right-censoring, left-censoring, and interval-censoring.
  - **Right Censoring**: The death of the person.
  - **Left Censoring**: The event can’t be observed for some reason. It includes events that occurred before the experiment started. (e.g., number of days from birth when the kid started walking.)
  - **Interval Censoring**: When we have data for some intervals only.

Left and right censoring

- Suppose you're conducting a study on pregnancy duration. You're ready to complete the study and run your analysis, but some women in the study are still pregnant, so you don't know exactly how long their pregnancies will last. These observations would be *right-censored*. The "failure," or birth in this case, will occur after the recorded time.

Right Censoring ***

↓	C1	C2	C3-T
	Patient ID	Days	Exact or Censored
1	1	286	Exact
2	2	279	Exact
3	3	269	Exact
4	4	256	Exact
5	5	293	Censored
6	6	262	Exact
7	7	285	Censored

Left Censoring ***

↓	C1	C2	C3
	Patient ID	Start	End
1	1	*	250
2	2	282	282
3	3	*	250
4	4	253	253
5	5	258	258
6	6	295	295
7	7	268	268

- Now suppose you survey some women in your study at the 250-day mark, but they already had their babies. You know they had their babies before 250 days, but don't know *exactly* when. These are therefore *left-censored* observations, where the "failure" occurred before a particular time.

Interval censoring

- If we don’t know exactly when some babies were born but we know it was within some interval of time, these observations would be *interval-censored*. We know the “failure” occurred within some given time period. For example, we might survey expectant mothers every 7 days and then count the number who had a baby within that given week.

Interval Censoring ***

↓	C1	C2	C3
	Start	End	No. of Births
1	245	251	1
2	252	258	7
3	259	265	5
4	266	272	27
5	273	279	44
6	280	286	86
7	287	293	87

Survival

- So how does one make sense of timing data?
  - How to account for censored observations?
- There are parametric models where one uses some distribution to model survival functions and then proceed to make predictions.
  - Geometric, Weibull are commonly used distribution
  - We will discuss one particular approach which is commonly used.
- Use non-parametric approach
  - In this approach, one uses data (without any assumptions) to make predictions.
  - A very commonly used method here is Kaplan-Meier method

KM estimator

- The **Kaplan–Meier estimator** is a non-parametric statistic used to estimate the survival function (probability of a person surviving) from lifetime data.
- In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. For example, calculating the amount of time certain patient lived after he/she was diagnosed with the cancer or when his treatment starts. The estimator is named after **Edward L. Kaplan** and **Paul Meier**.
- Probability of survival is how many subject (patients) survive (do not perish) out of the total events (patients) at that time.
- The probability of survival at time t_i, $S(t_i)$, is calculated as

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i} = \frac{survive}{total}$$

KM Estimator

The probability of survival at time t_i, $S(t_i)$, is calculated as

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

- We can write this as a recurve function, where

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

$S(t_{i-1})$= probability of being alive at t_{i-1}

N_i = number alive just before t_i

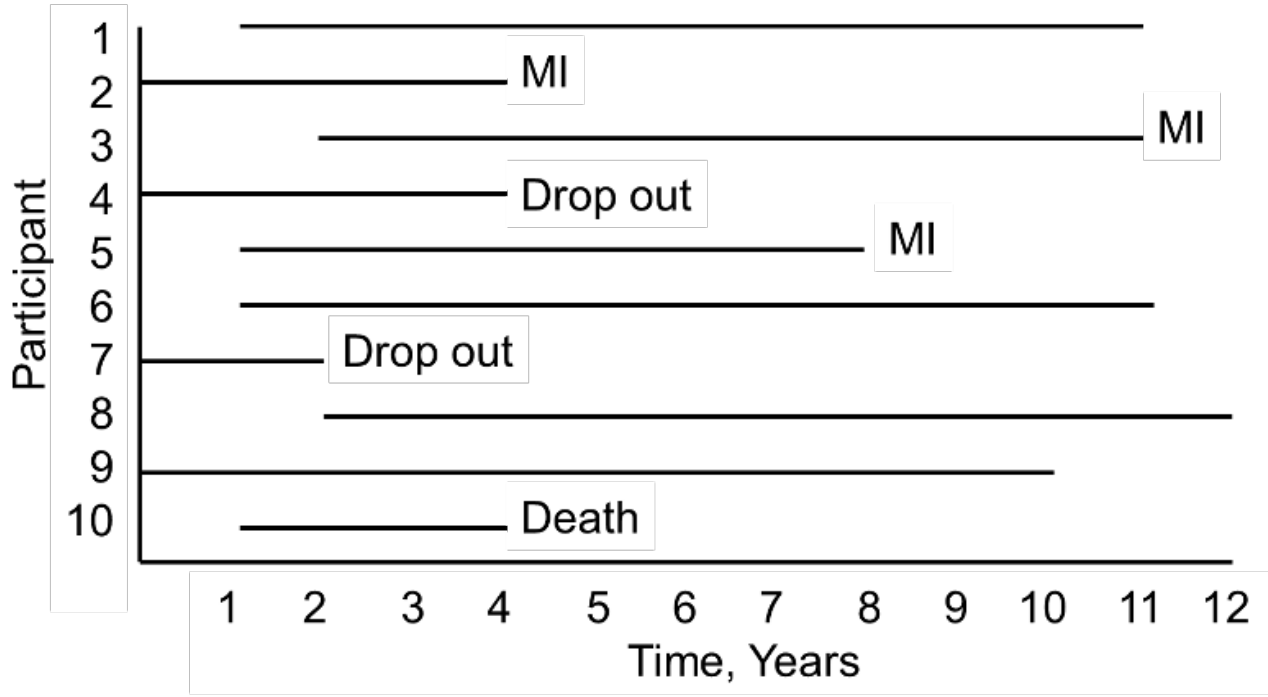
Di = number of events at t_i

This means,

$$S(1) = S(0) * \left(1 - \frac{d_1}{n_1}\right) = \left(1 - \frac{d_1}{n_1}\right) \qquad S(2) = S(1) * \left(1 - \frac{d_2}{n_2}\right)$$

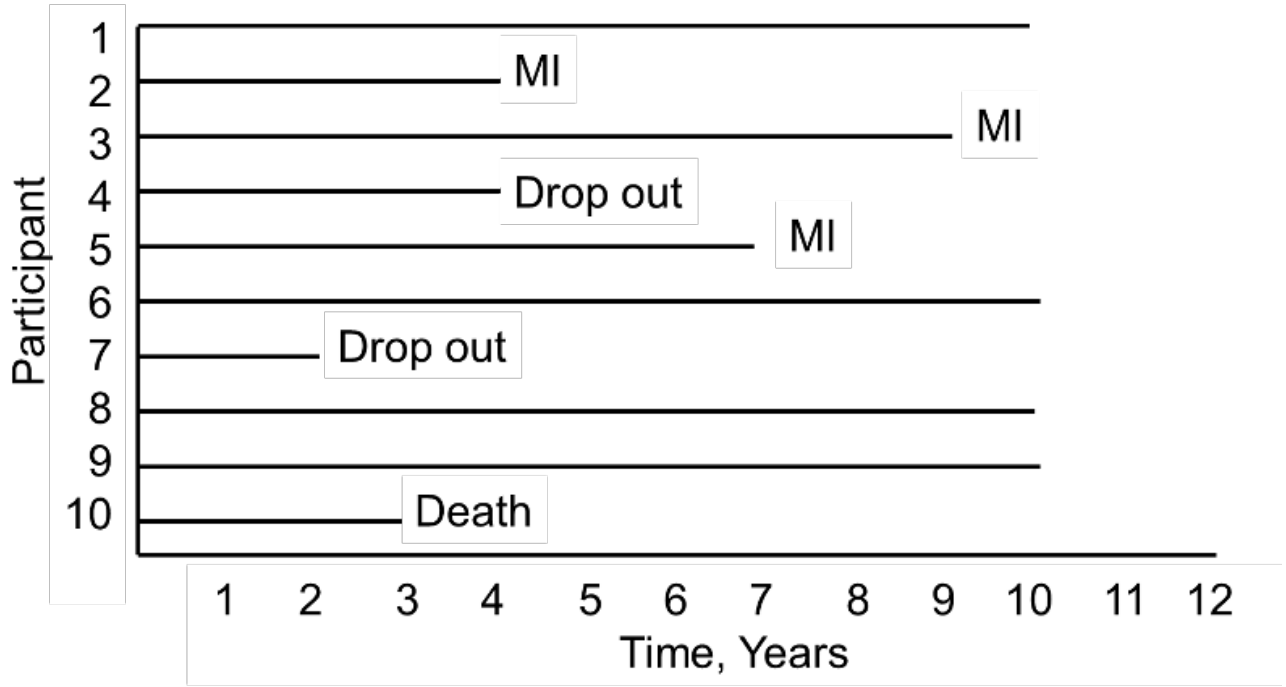
K-M example

- A study follows ten participants for the development of myocardial infarction (MI, or heart attack) over a period of 10 years. Participants are recruited into the study over a period of two years and are followed for up to 10 years. The plot here indicates when they enrolled and what subsequently happened to them during the observation period.



Example - 1

- Usually, we want to normalize the starting time, survival time starting at a common time zero (i.e., as if all participants enrolled in the study at the same time).
- As the data shows, during the study period, three participants suffered myocardial infarction (MI), one dies, two drop out of the study (for unknown reasons), and four complete the 10-year follow-up without suffering MI.



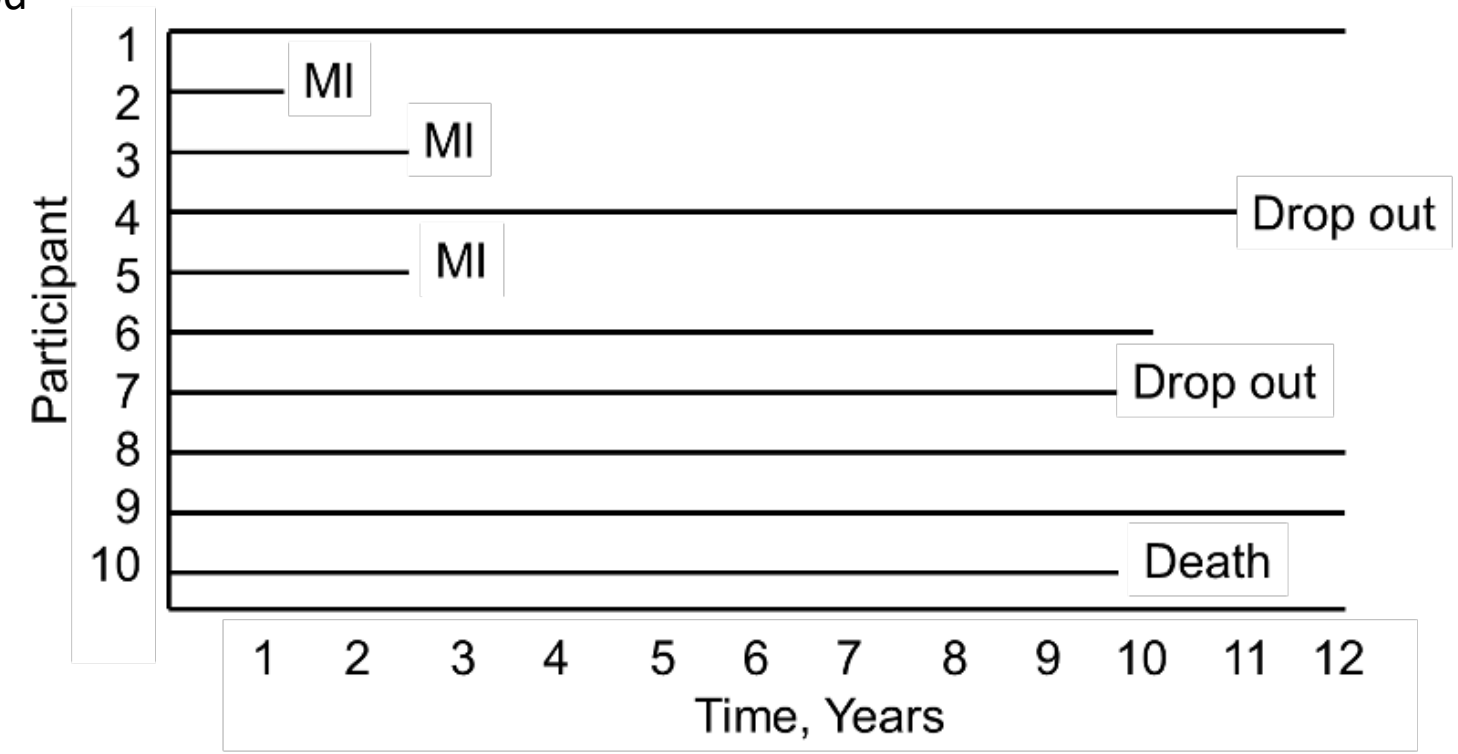
What is the likelihood that a participant will suffer an MI over 10 years?

Is it 30%?
What do you do with 2 drops (censored observations)
And 1 death during the study?

Example -2

Consider the same study with different data.
Here also, 3 people suffer MI but in early part but drop-outs and death occur in later period
Now, what are the odds of survival?

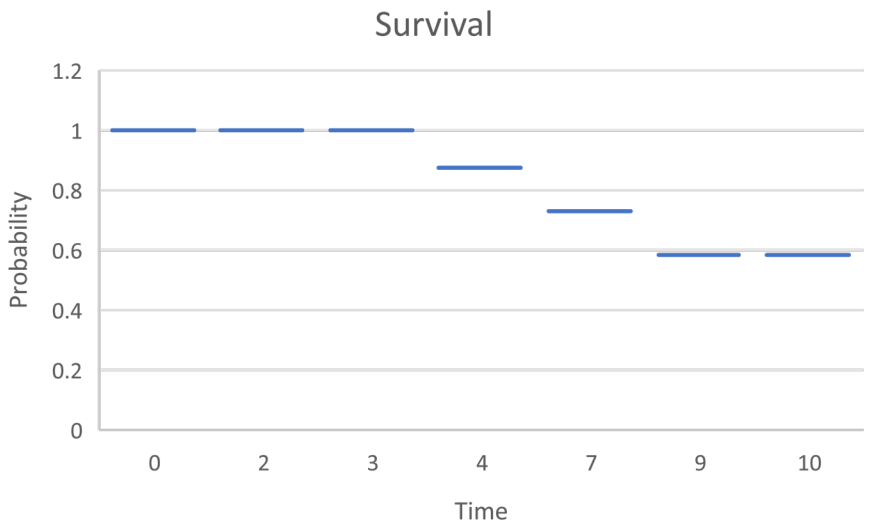
It is clear that
(i) Time when the event occurs
(ii) Number of subjects at a time.
Play a role in survival



Calculating the numbers (example – 1)

Year	Number at Risk (N)	Number of death (d)	Censor	Survival Probability $S_t = S_{(t-1)} *(1-d_t/N_t)$
0	10	0		1
2	10	0	1	1*(1-0/10)=1
3	9	0	1	1*(1-0/9)=1
4	8	1	1	1*(1-1/8)=0.875
7	6	1		0.875*(1-1/6)=0.73
9	5	1		0.73*(1-1/5)=0.584
10	4	0		0.584*(1-0/4)=0.584

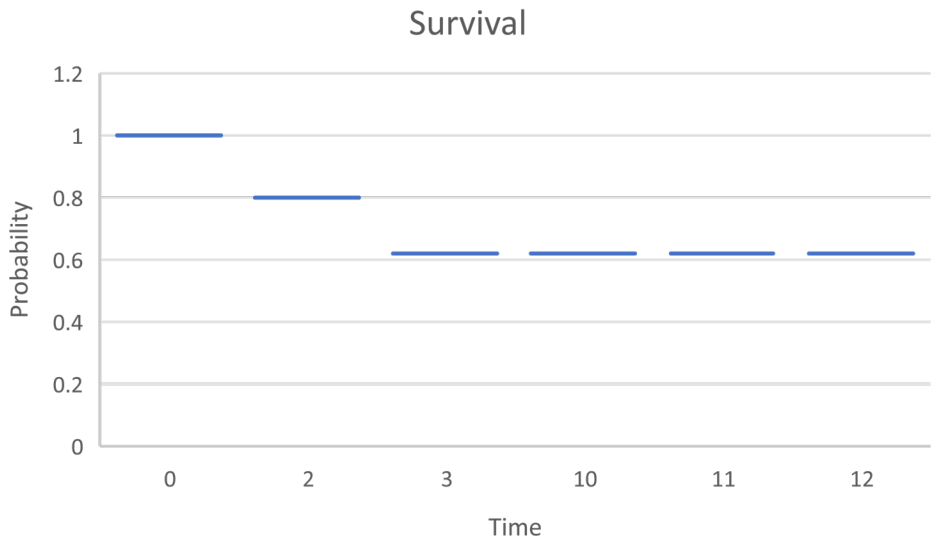
- We only consider the time when event occurs (death, censoring, drop out).
- censoring and dropouts reduce number of items at risk (N).
- Survival probability changes only when event occurs (But N may change sue to censoring)



Example 2

How do numbers look for example2?

	Number at Risk	Number of death	Censor	Survival Probability = S*(t-1)*(1-d/N)
0	10	0		1
2	10	1		1*(1-1/10)=0.8
3	9	2		0.8*(1-2/9)=0.62
10	7	0	2	0.62*(1-0/7)=0.62
11	5	0	1	0.62*(1-0/5)=0.62
12	4	0	0	0.62*(1-0/4)=0.62



One more example

- A study involves 20 participants who are 65 years of age and older; they are enrolled over a 5-year period and are followed for up to 24 years until they die, the study ends, or they drop out of the study (lost to follow-up). [Note that if a participant enrolls after the study start, their maximum follow up time is less than 24 years. e.g., if a participant enrolls two years after the study start, their maximum follow up time is 22 years.]
- The data are shown. In the study, there are 6 deaths and 3 participants with complete follow-up (i.e., 24 years). The remaining 11 have fewer than 24 years of follow-up due to enrolling late or loss to follow-up.

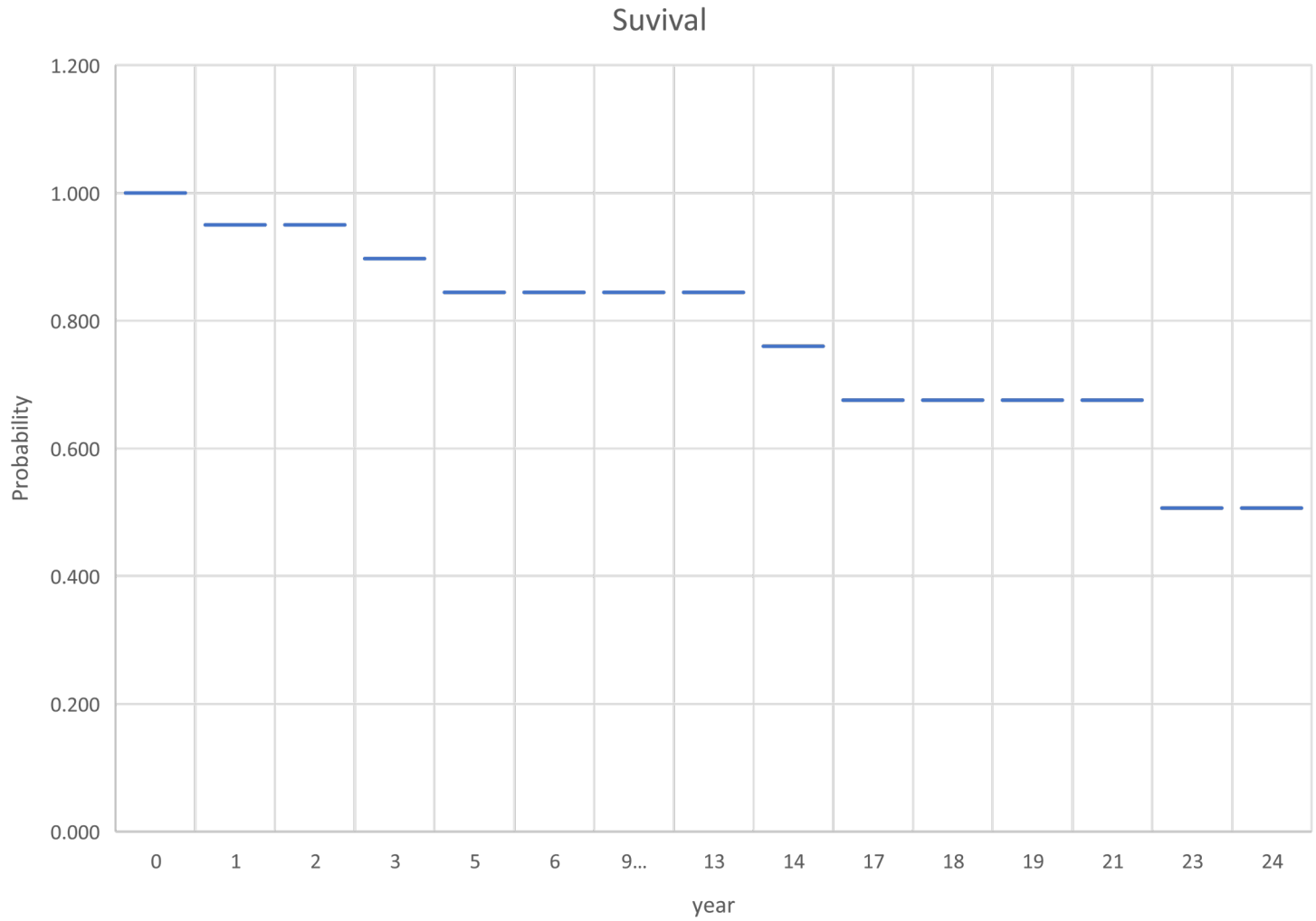
participant	Year of Death	Year of Last Contact
1		24
2	3	
3		11
4		19
5		24
6		13
7	14	
8		2
9		18
10		17
11		24
12		21
13		12
14	1	
15		10
16	23	
17		6
18	5	
19		9
20	17	

KM table

- We use the same approach.
- Key to note is that censoring does not change the survival probability.

Time, Year	No at Risk - Nt	No of Deaths-Dt	No Censored - Ct	Survival Probability $S_{t+1} = S_t*(1-D_{t+1}/N_{t+1})$
0	20			1
1	20	1		=1*(1-1/20)=0.95
2	19		1	=0.95*(1-0/19)=0.95
3	18	1		=0.95*(1-1/18)=0.897
5	17	1		=0.897*(1-1/17) = 0.844
6	16		1	=0.844
9…	15…		1…	=0.844
13	11		1	=0.844
14	10	1		=0.844*(1-1/10) = 0.760
17	9	1	1	=0.760*(1-1/9) = 0.676
18	7		1	=0.676
19	6		1	=0.676
21	5		1	=0.676
23	4	1		=0.507
24	3		3	=0.507

KM Curve

- 
Suvival
1.200
1.000
0.800
0.600
0.400
0.200
0.000
Probability
0
1
2
3
5
6
9...
13
14
17
18
19
21
23
24
year

Standard Errors and Confidence Interval

- There are many different ways to produce standard errors and confidence interval estimates of survival probabilities. A popular formula to estimate the standard error of the survival estimates is called Greenwoods formula

  - $SE(S_t) = S_t\sqrt{\sum \frac{D_t}{N_t(N_t - D_t)}}$

  where $S_t = N_{t+1}/N_t$

- It is cumulative. When N is higher compared to D, standard error tends to be lower.

Standard Error and CI

-

Time, Year	Num at Risk - Nt	Num of Deaths- D_t	Num Censored - C_t	Survival Probability S_t	$S_t\sqrt{\sum \frac{D_t}{N_t(N_t - D_t}}$	1.96*SE(St)
0	20			1		
1	20	1		0.95	0.049	0.096
2	19		1	0.95	0.069	0.096
3	18	1		0.897	0.083	0.135
5	17	1		0.844	0.083	0.162
6	16		1	0.844	0.083	0.162

KM in Python

- While these numbers can be calculated by hands, with larger dataset, this is not optimal.
- There are packages that will allow us the calculate these numbers more efficiently.
- Lets look at this in an example using Python.
- We use KaplanMeierFitter to implement KM on a dataset.
  - KaplanMeierFitter() will allows us to create the table on at risk, death, censor etc which is needed to calculate survival probability
- To be able to use this, we first have to import “lifelines” library.

KM in python

- In jupyter, use this to install lifelines package

  conda install -c conda-forge lifelines

- And, then import KaplanMeierFitter

  from lifelines import KaplanMeierFitter

- Lets do an example using Python, the data is lung.csv which is posted on canvas.
  - Survival in patients with advanced lung cancer (227 patients). Besides survival time, there are other covariates and a dummy (status) for whether the person is dead=2 or alive =1) at that time.

Data

- Status=2, means the person is dead, otherwise alive (hence censored).
- All non available (NaN) data is dropped in Kaplan-Meier
  - Inst – institution;
  - status – dead or alive,
  - ph.ecog - ECOG performance score (0-5),
  - ph.karno - Karnofsky performance score (0-100) rated by physician,
  - pat.karno – rated by patient

	Unnamed: 0	Inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
0	1	3.0	306	2	74	1	1.0	90.0	100.0	1175.0	NaN
1	2	3.0	455	2	68	1	0.0	90.0	90.0	1225.0	15.0
2	3	3.0	1010	1	56	1	0.0	90.0	90.0	NaN	15.0
3	4	5.0	210	2	57	1	1.0	90.0	60.0	1150.0	11.0
4	5	1.0	883	2	60	1	0.0	100.0	90.0	NaN	0.0

Use python KM

- Fit this data to create the table we want
  - kmf.fit(durations = data["time"], event_observed = data["dead"])
- Where “time” is the survival that we want to model and “dead” is the event of interest. One generates a table like

	removed	observed	censored	entrance	at_risk
event_at					
0.0	0	0	0	228	228
5.0	1	1	0	0	228
11.0	3	3	0	0	227
12.0	1	1	0	0	224
13.0	2	2	0	0	223
…	…	…	…	…	…

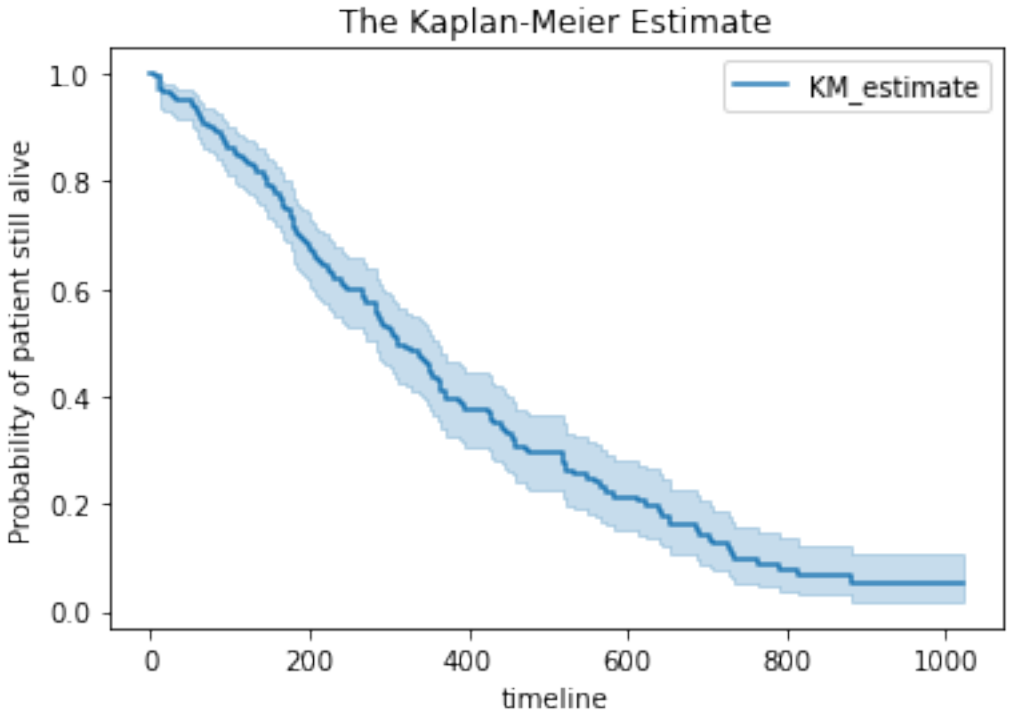
Survival time

- One can either generate statistics on survival probability at different times or generate the table for all times
  - kmf.survival_function_

timeline	KM_estimate
0.0	1.000000
5.0	0.995614
11.0	0.982456
12.0	0.978070
13.0	0.969298
...	...

Plot survival time

- Kmf.plot()



Generate confidence interval

- kmf.confidence_interval_

	KM_estimate_lower_0.95	KM_estimate_upper_0.95
0.0	1.000000	1.000000
5.0	0.969277	0.999381
11.0	0.953935	0.993379
12.0	0.948120	0.990813
13.0	0.936682	0.985244
…	…	…

Nelson-Aalen hazard

- KM is focused on survival (how many observations survive), but one can look at the hazard – (how many observations perish). Nelson-Aalen estimator is non-parametric hazard.
- Hazard is hot many objects perish out of total at that time.
- As the name indicate, it is calculated based on data. It is *cumulative summation*. So

$$\dot{H(t)} = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

- Note that KM survival is (*cumulative multiplication*)

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

- We import NelsonAalenFitter from lifelines and follow same steps as in Kaplan Meier

Male-Female Survival

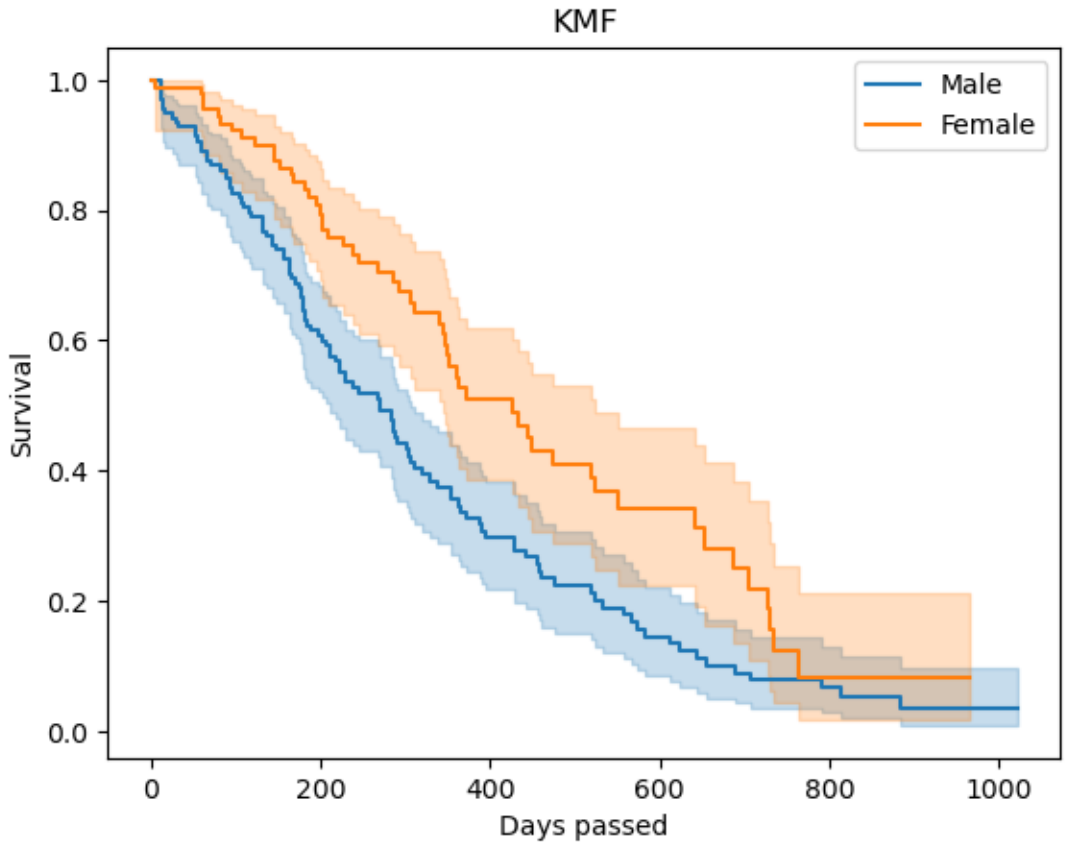
- KM curve can accommodate a discrete covariate.
- We want to plot survival for males and females differently.
  - kmf_m.fit(durations = Male["time"], event_observed = Male["dead"], label="Male")
  - kmf_f.fit(durations = Female["time"], event_observed = Female["dead"], label="Female")
- We can generate event tables for each

event_at	removed	observed	censored	entrance	at_risk
0.0	0	0	0	138	138
11.0	3	3	0	0	138
12.0	1	1	0	0	135
13.0	2	2	0	0	134
15.0	1	1	0	0	132
...	...	...	...	...	...

event_at	removed	observed	censored	entrance	at_risk
0.0	0	0	0	90	90
5.0	1	1	0	0	90
60.0	1	1	0	0	89
61.0	1	1	0	0	88
62.0	1	1	0	0	87
...	...	...	...	...	...

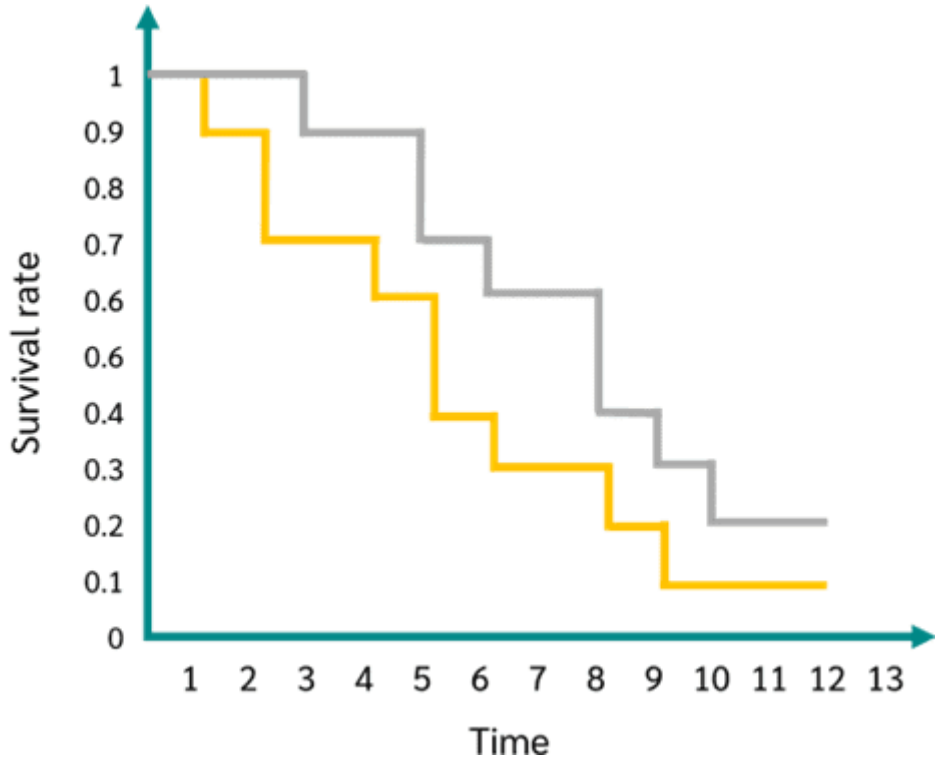
plot

-


KMF
Male
Female
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Days passed

Log Rank Test

- The Log Rank Test is used in survival time analysis and compares the distribution of time to event occurrence of two or more independent samples.

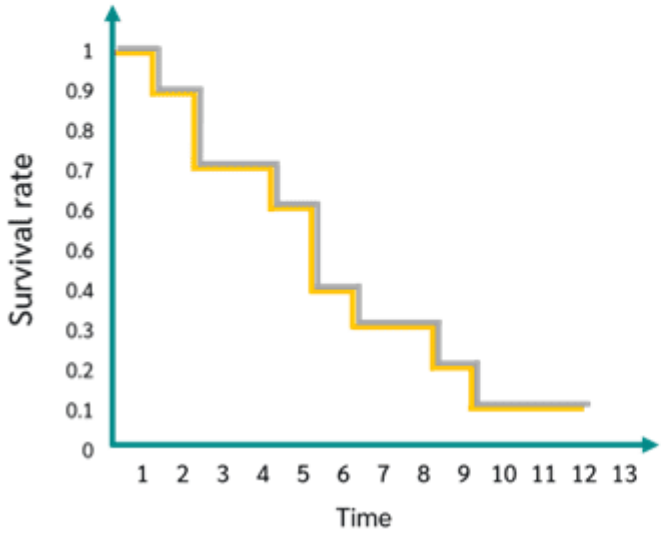


Log Rank Test

- One tests whether the two distributions (Kaplan Meier curves) are the same?

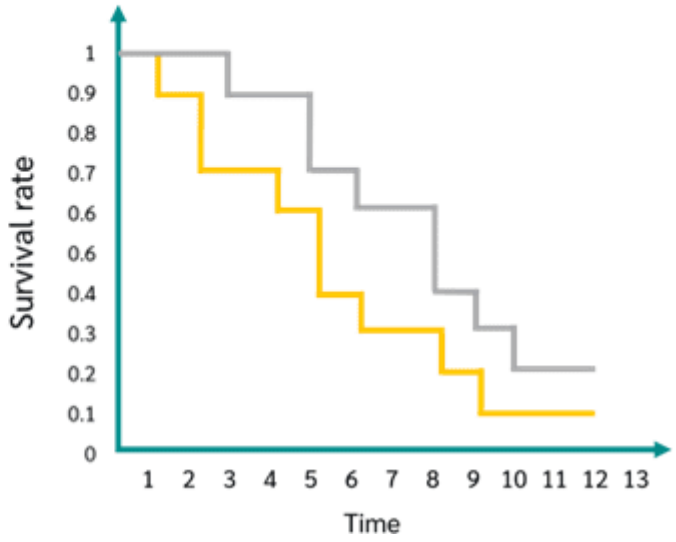
Null hypothesis:

The groups have identical distribution curves.



Alternative hypothesis:

The groups have different distribution curves.



Log Rank Test

- The intuition behind the log-rank test for comparison of two life tables is simple
  - if there were no difference between the groups, the total deaths occurring at any time should split between the two groups at that time.
  - If the numbers at risk in the first and second groups in (say) in a month were 70 and 30, respectively, and 10 deaths occurred in that month we would expect

    $10 \times 70/(70+30) = 7$
  - of these deaths to have occurred in the first group, and

    $10 \times 30/(70+30) = 3$

  to have occurred in the second group.
- Log Rank test compares actual numbers with expected numbers.

Statistics behind Log Rank test

- The log rank statistic is approximately distributed as a chi-square test statistic. There are several forms of the test statistic, and they vary in terms of how they are computed. One can calculate following

$$\chi^2 = \Sigma \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}}$$

where ΣO_{jt} represents the sum of the **observed number of events** in the j^{th} group over time (e.g., j=1,2) and ΣE_{jt} represents the sum of the **expected number of events** in the j^{th} group over time.

We compare observed and expected events for these groups. The outcome is a χ^2 statistics of degree 1. One can look at χ^2 table and see if the observed number is higher than noted.

Let's take an example to understand this clearly

Log Rank test

- A small clinical trial is run to compare two combination treatments in patients with advanced gastric cancer. Twenty participants consent to participate in the trial are randomly assigned to receive chemotherapy before surgery or chemotherapy after surgery. The primary outcome is death and participants are followed for up to 48 months (4 years) following enrollment into the trial. The experiences of participants in each arm of the trial are shown below.

Chemotherapy Before Surgery		Chemotherapy After Surgery	
Month of Death	Month of Last Contact	Month of Death	Month of Last Contact
8	8	33	48
12	32	28	48
26	20	41	25
14	40		37
21			48
27			25
			43

Log Rank Example

Life Table for Group Receiving Chemotherapy Before Surgery

Time, Months	Number at Risk	Number of Deaths	Number Censored	Survival Probability
	N_t	D_t	C_t	
0	10			1
8	10	1	1	0.9
12	8	1		0.788
14	7	1		0.675
20	6		1	0.675
21	5	1		0.54
26	4	1		0.405
27	3	1		0.27
32	2		1	0.27
40	1		1	0.27

Life Table for Group Receiving Chemotherapy After Surgery

Time, Months	Number at Risk	Number of Deaths	Number Censored	Survival Probability
	N_t	D_t	C_t	
0	10			1
25	10		2	1
28	8	1		0.875
33	7	1		0.75
37	6		1	0.75
41	5	1		0.6
43	4		1	0.6
48	3		3	0.6

Log Rank Test

Time, Months	Number at Risk in Group 1	Number at Risk in Group 2	Number of Events (Deaths) in Group 1	Number of Events (Deaths) in Group 2
	N_{1t}	N_{2t}	O_{1t}	O_{2t}
8	10	10	1	0
12	8	10	1	0
14	7	10	1	0
21	5	10	1	0
26	4	8	1	0
27	3	8	1	0
28	2	8	0	1
33	1	7	0	1
41	0	5	0	1

We take all the times from both cohort where incidences are positive. We ignore times where data is censored

Log Rank Test

Expected Numbers of Events in Each Group

Time, Months	Number at Risk in Group 1	Number at Risk in Group 2	Total Number at Risk	Number of Events in Group 1	Number of Events in Group 2	Total Number of Events	Expected Number of Events in Group1	Expected Number of Events in Group2
	N_{1t}	N_{2t}	N_t	O_{1t}	O_{2t}	O_t	$E_{1t} = N_{1t}*(O_t/N_t)$	$E_{2t} = N_{2t}*(O_t/N_t)$
8	10	10	20	1	0	1	0.5	0.5
12	8	10	18	1	0	1	0.444	0.556
14	7	10	17	1	0	1	0.412	0.588
21	5	10	15	1	0	1	0.333	0.667
26	4	8	12	1	0	1	0.333	0.667
27	3	8	11	1	0	1	0.273	0.727
28	2	8	10	0	1	1	0.2	0.8
33	1	7	8	0	1	1	0.125	0.875
41	0	5	5	0	1	1	0	1

Expected at any time is proportion to the size of each group. So total evens are split according to this proportion. Actual events are the data we observe in each group at every period.

Chi-Square test

We difference actual and expected value for each group and calculate

$$\chi^2 = \Sigma \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}} = \frac{(6-2.620)^2}{2.620} + \frac{(3-6.380)^2}{6.380} = 4.360 + 1.791 = 6.151$$

X^2 with degree 1 - at 95% confidence interval it is 3.84

Time, Months	Number at Risk in Group 1	Number at Risk in Group 2	Total Number at Risk	Number of Events in Group 1	Number of Events in Group 2	Total Number of Events	Expected Number of Events in Group 1	Expected Number of Events in Group 2
	N_{1t}	N_{2t}	N_t	O_{1t}	O_{2t}	O_t	$E_{1t} = N_{1t}*(O_t/N_t)$	$E_{2t} = N_{2t}*(O_t/N_t)$
8	10	10	20	1	0	1	0.5	0.5
12	8	10	18	1	0	1	0.444	0.556
14	7	10	17	1	0	1	0.412	0.588
21	5	10	15	1	0	1	0.333	0.667
26	4	8	12	1	0	1	0.333	0.667
27	3	8	11	1	0	1	0.273	0.727
28	2	8	10	0	1	1	0.2	0.8
33	1	7	8	0	1	1	0.125	0.875
41	0	5	5	0	1	1	0	1
				6	3		2.62	6.38

summary

- What is unique about survival?
- Censoring?
- Non-parametric Kaplan Meier Curve for calculating survival probability
- Use of KM in Python