INWS0038 Longitudinal and Multilevel Modelling II - Event History Analysis

22.04 – 10.05 2024 (3 ECTS)

Linus Andersson[12]

[1]University of Turku, [2]Stockholm University

linus.andersson@utu.fi

INWS0038 Longitudinal and Multilevel Modelling II - Event History Analysis

Lecture II, part 1 – Key estimates in EHA

Wednesday 24.04

Linus Andersson[12]

[1]University of Turku, [2]Stockholm University

linus.andersson@utu.fi

Key estimates in EHA

> I'm 23 now but will I live to see 24?
> The way things is going I don't know
>
> - Coolio, Gangsta's Paradise

Key estimates in EHA

- We covered:
  - The duration to event as the dependent variable

- Now, lets consider time or duration as any other variable
  - A random variable
  - Differs across the unit of analysis, people etc
  - Predicted by X

- How to describe the time to event?

Scheduling

Activity	Lesson topic	Keywords	Homework
Lesson 1.	Introduction Key Concepts	Process time, Cencoring, Time-to-Event, Continous and discrete time	
Lesson 2.	Key estimates Descriptive models	Kaplan-meyer, Density, Cum. Distribution function, Survival and Hazard funciton, Kaplan-Meyer, Life tables	✓
Lesson 3.	Key Concepts and estimates for Parametric models	Exponential and Piece-wise exponential models, Shape parameter, the proportional hazard assumption, hazard ratios	✓
Lab 1. Non-Parametric models			
Lesson 4	Discrete and Continous models Data structure	Time-varying variables, Cox, Logit	✓
Lab 2. Parametric models I.			
Lesson 5.	Piecing in together +, other models	Case studies with focus on model choice, Interpretation ation of causality, heterogeneity	
Lab 3. Parametric models II.			
Lesson 6.	Discussion. Presentations.		Presentations

Key estimates in EHA

1. Density, Survival & Hazard Function
2. Life Tables
3. Kaplan-Kaplan-Meier
4. Descriptive analysis

Density, Cumulative probability, Survival & Hazard

Survival function S(t)

Probability of not having had an event up to time t

Cumulative distribution function F(t)

Probability of ever having had the event up to t

Probability density function f(t)

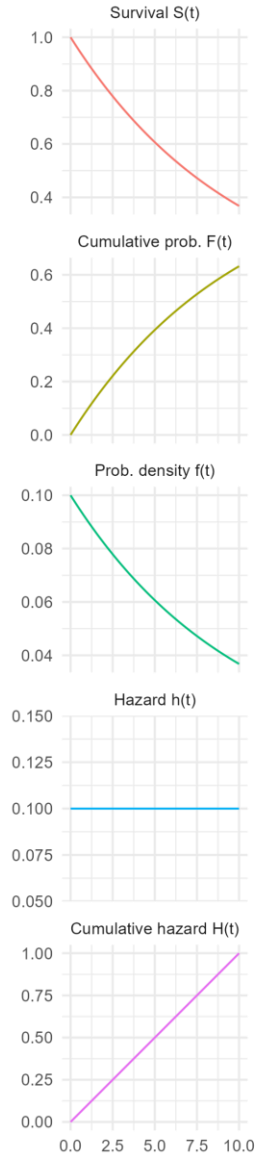
Nr of event at any given time t

Hazard function h(t)

Having a transition at time t, conditional on survival

Cumulative hazard function h(t)

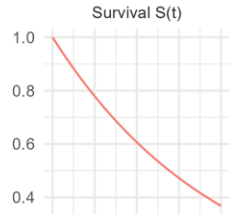
Sum of hazards. Expected nr of events until t



Density, Cumulative probability, Survival & Hazard

Survival function S(t)

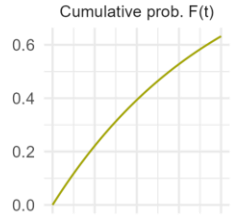
Probability of not having had an event up to time t



F(t-1)

Cumulative distribution function F(t)

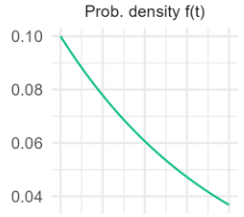
Probability of ever having had the event up to time t



1-S(t)

Probability density function f(t)

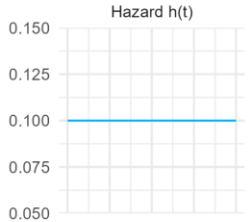
Nr of event at any given time t



Integral of F(t)

Hazard function h(t)

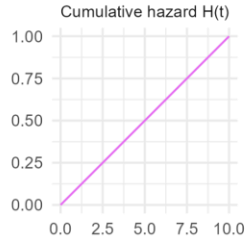
Having a transition at time t, conditional on survival



F(t)/S(t)

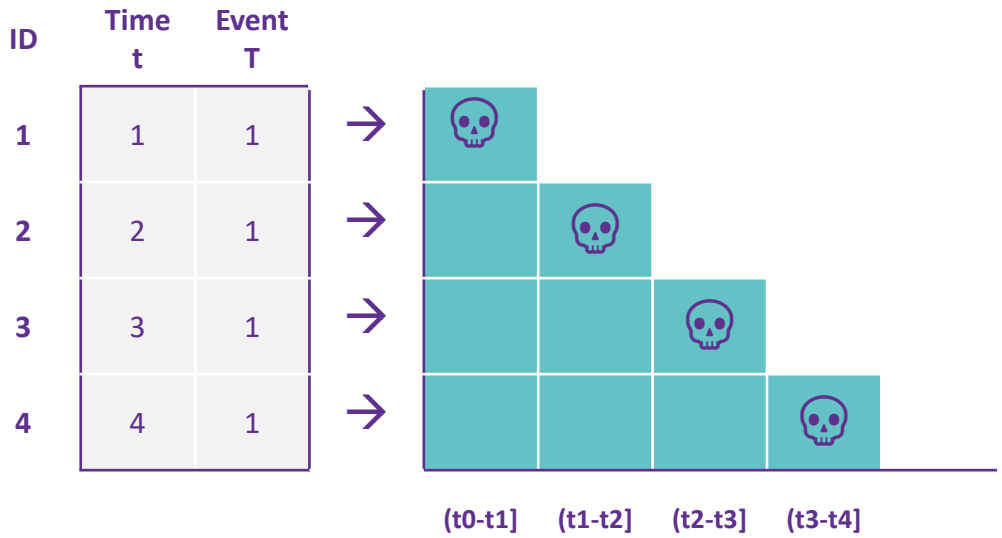
Cumulative hazard function h(t)

Sum of hazards. Expected nr of events until t



Integral of h(t)

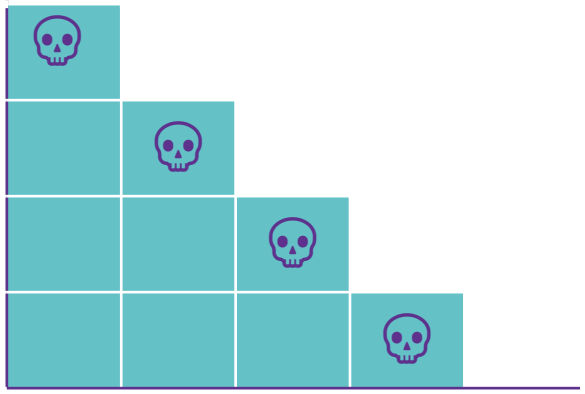
Density, Cumulative probability, Survival & Hazard



A Population of 4
Event transition d = 1 = death
All 4 die
1 death each year

Intuition – Probability density function $f(t)$

ID	Time t	Event T
1	1	1
2	2	1
3	3	1
4	4	1



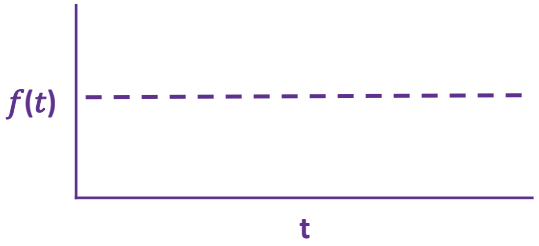
"Probability of death at time t"

$f(t) \sim P(T = t)$

$$\frac{Events}{\sum(events)}$$

$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$

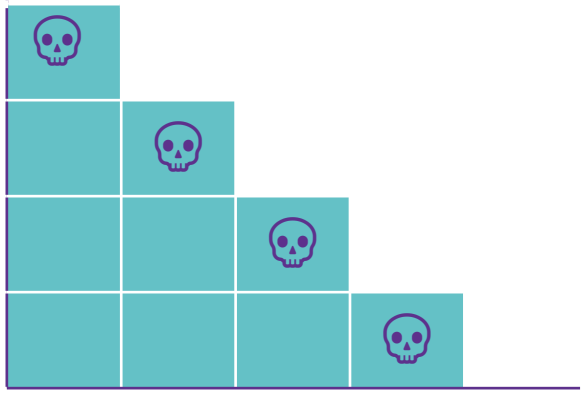
0.25 0.25 0.25 0.25



"25% of death occur at time t1 (and also in t2, t3, and t4)"

Intuition – Survival function $S(t)$

ID	Time t	Event T
1	1	1
2	2	1
3	3	1
4	4	1



t1 t2 t3 t4

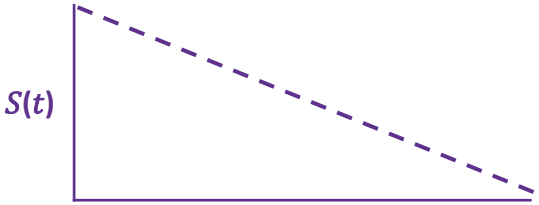
"Probability of survival within time t years"

$S(t) \sim P(T \geq t)$

$$\frac{\sum(events\ up\ t\ time\ t)}{\sum(events)}$$

$\frac{4}{4}$ $\frac{3}{4}$ $\frac{2}{4}$ $\frac{1}{4}$

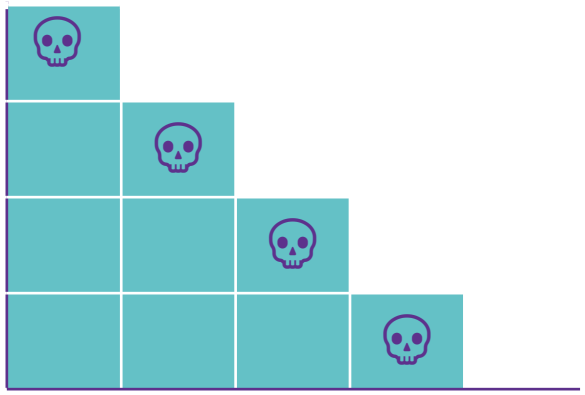
1 0.75 0.50 0.25



"75% of persons survive to time t2"

Intuition – Hazard function $h(t)$

ID	Time t	Event T
1	1	1
2	2	1
3	3	1
4	4	1



"Risk of transition conditional on survival to time t"

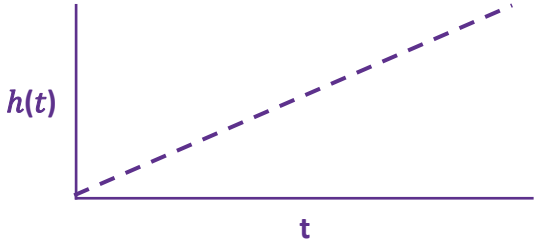
$P(T = t \mid T \geq t)$

$\sim h(t) = f(t) / S(t)$

$$\frac{Events\ at\ time\ t}{\sum(Survive\ to\ time\ t)}$$

$\frac{1}{4}$ $\frac{1}{3}$ $\frac{1}{2}$ $\frac{1}{1}$

0.25 0.33 0.50 1.00



"100% of persons who survive to time t4 dies"

More on the hazard function

- h(t) The Hazard fuction – a key dependent variable
  - The risk of event transition
  - Also know as the Intensity function, transition rate, hazard rate, λ(t), or μ(t)
  - Range from 0 to infinity
  - A rate, not an probability – but conditions on the risk set, alike a conditional probability

$$h(t) = \frac{occurence}{exposure\ time}$$

=

$$h(t) = \frac{occurence\ (event\ transitions\ during\ time\ t)}{time\ spells\ of\ the\ risk\ set\ at\ time\ t}$$

More on the hazard function

- h(t) in the abstract, seemingly straighforward

$$h(t) = \frac{occurence}{exposure\ time}$$

=

$$h(t) = \frac{occurence\ (event\ transitions\ during\ time\ t)}{time\ spells\ of\ the\ risk\ set\ at\ time\ t}$$

More on the hazard function

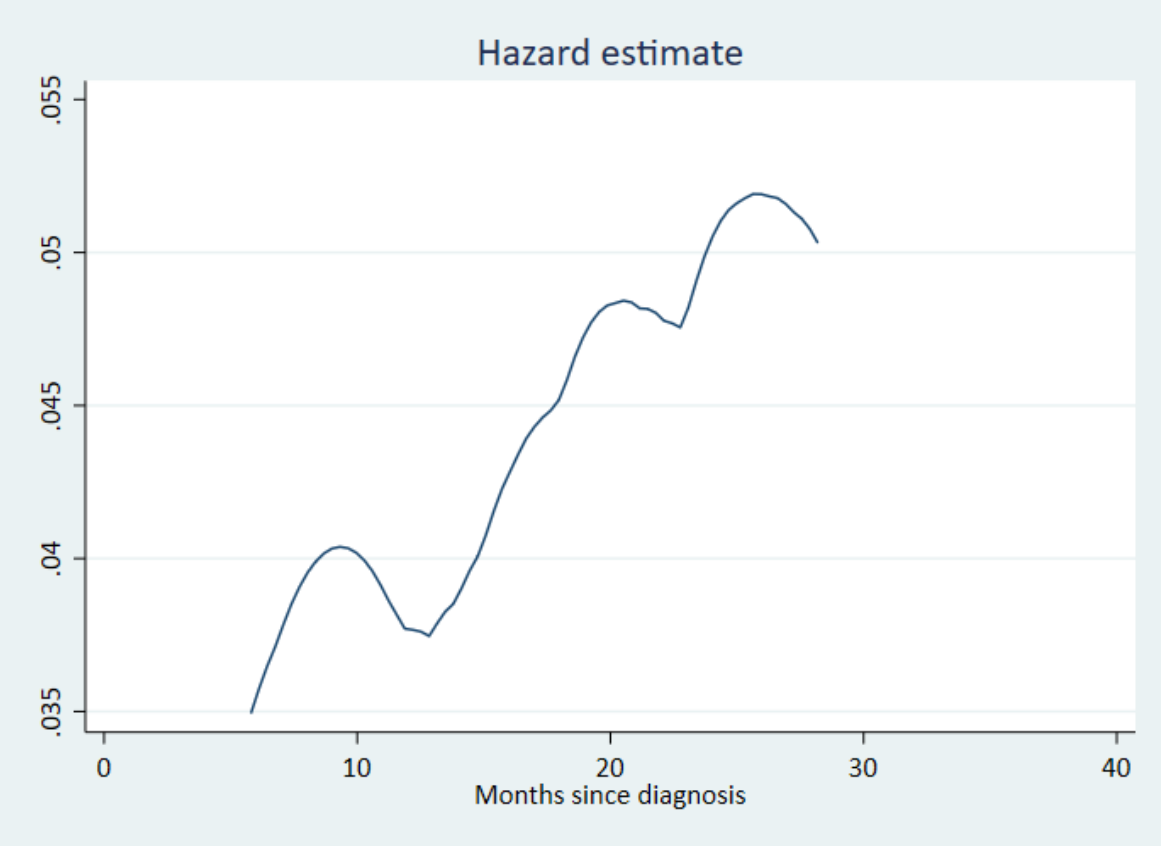
Hazard rate: the point estimate

Category	Total
Number of subjects	48
Number of records	48
Entry time (first)	
Exit time (final)	
Subjects with gap	0
Time on gap	0
Time at risk	744
Failures	31

```
. disp 31/744
.04166667
```

_t	Hazard	Std. err.	z	P>|z|
_cons	.0416667	.0074836	-17.69	0.000

Hazard: the hazard's shape over time



Key estimates in EHA

1. Density, Survival & Hazard Function ✓
2. Life Tables
3. Kaplan-Kaplan-Meier
4. Descriptive analysis

Life tables

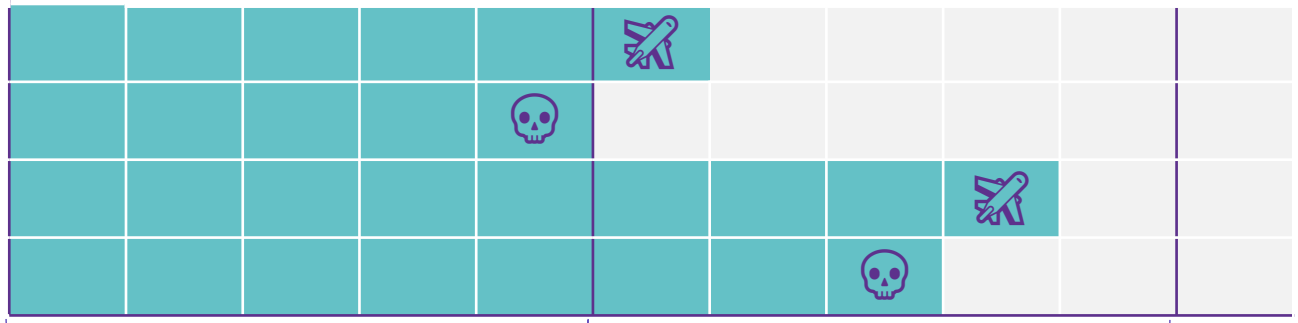
- Often used to analyse mortality, fertility and population change over time
- Divides time into intervals (of age, years, etc)
- Censors individuals at end of time intervals
- Calculates survival and hazard function, among other things

Life tables

ID	Time	Event	
1	5	0	 Right censored by attrition at t5
2	4	1	 Censored at event at t4
3	8	0	 Right censored by attrition at t8
4	7	1	 Censored at event at t7

Life tables

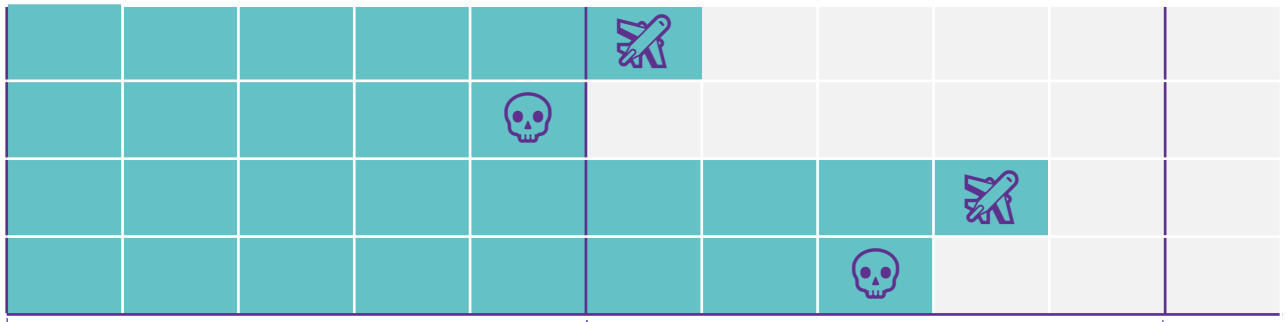
ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



[0,5) [5,10)

Life tables

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



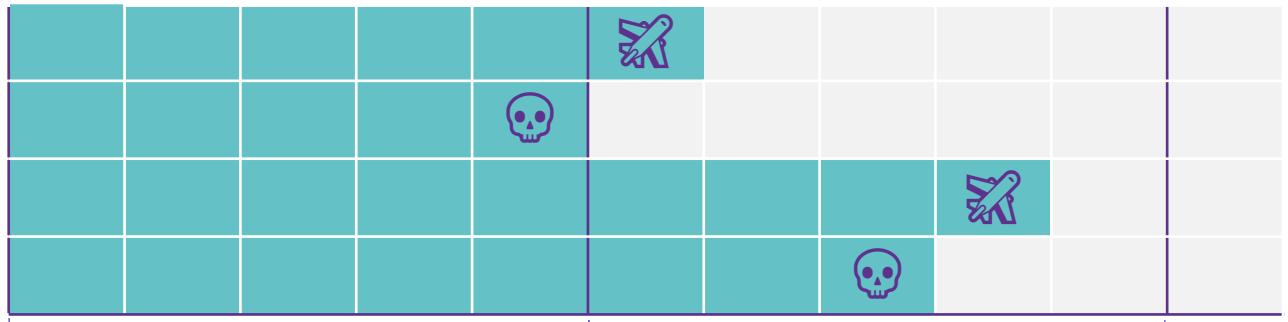
(0,5]

(5,10)

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]						
(5,10)						

Life tables – risk set (n)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



(0,5] (5,10)

4 ID exist in the first 5-year interval (0,5]
n = 4

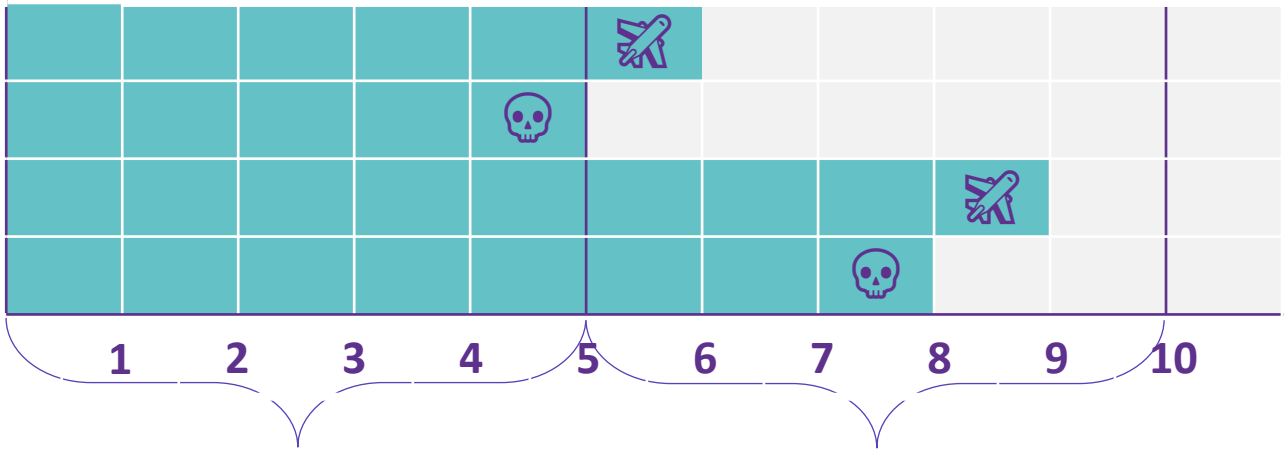
3 ID exist in the second 5-year interval (5,10)
n= 3

life table

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	**4**					
(5,10)	**3**					

Life tables – censored (m)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



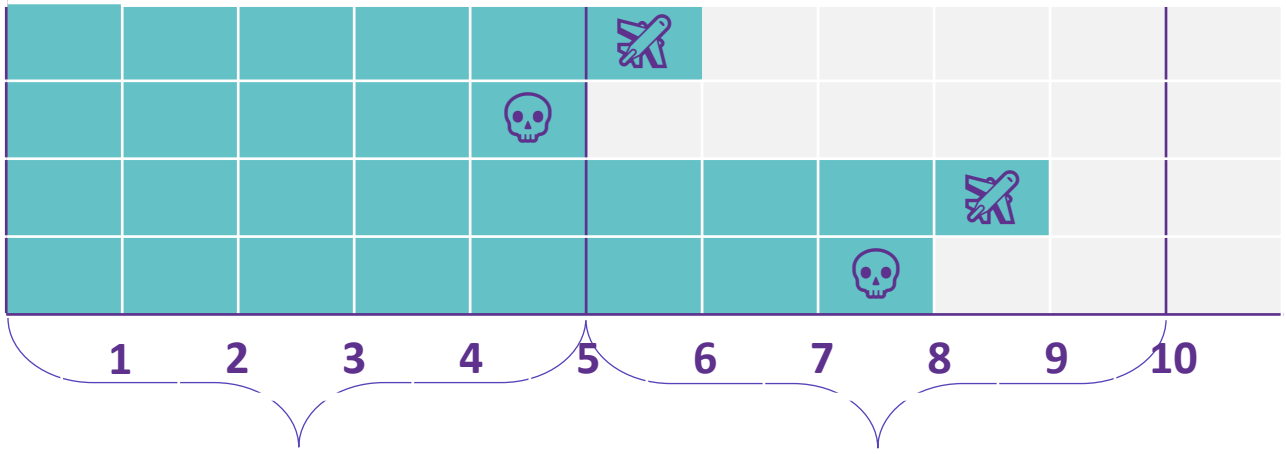
0 ID have event = 0 (censored) in the first 5-year interval (0,5]
m = 4

2 ID have event = 0 (censored) in the second 5-year interval (5,10)
m = 3

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	**4**	**0**				
(5,10)	**3**	**2**				

Life tables – events (d)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



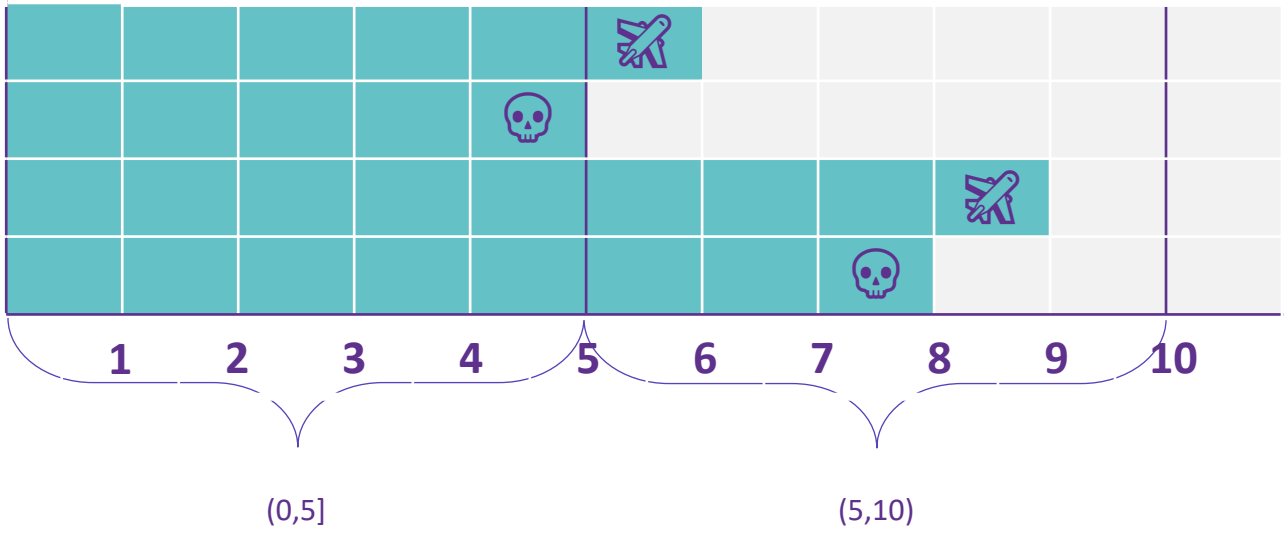
1 ID have event = 1 (death) in the first 5-year interval (0,5]
d = 1

1 ID have event = 1 (death) in the second 5-year interval (5,10)
d = 1

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	**4**	**0**	**1**			
(5,10)	**3**	**2**	**1**			

Life tables – interval span (Δ_t)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1

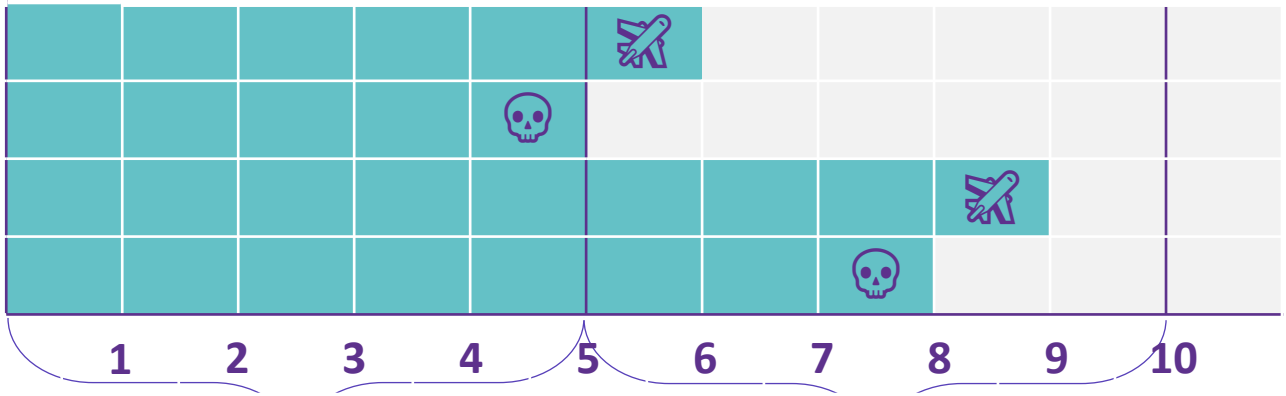


Both time intervals are in 5-year increments
$\Delta_t = 5$

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	**4**	**0**	**1**	**5**		
(5,10)	**3**	**2**	**1**	**5**		

Life tables – Survival function S(t)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



(0,5] (5,10)

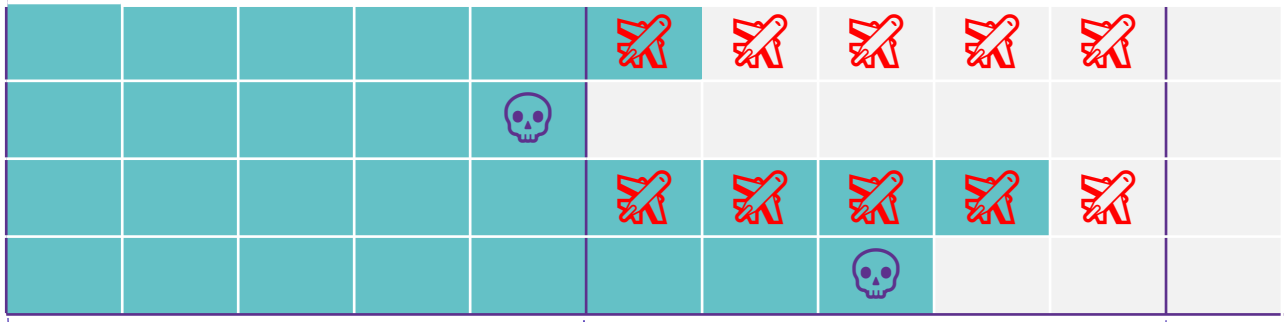
First interval =
The numerator is the risk set at interval start n_k, minus haft of the censored $\frac{m_k}{2}$ and all deaths
d_k = 4 - 0 - 1 = 3.
Divides by denominator n_k minus $\frac{m_k}{2}$
=
3/ 4 = 0.750

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	4	0	1	5	**0.750**	
(5,10)	3	2	1	5		

$$S(t_j) = \prod_{k=1}^{j} \frac{\left(n_k - \frac{m_k}{2}\right) - d_k}{\left(n_k - \frac{m_k}{2}\right)}$$

Life tables – Survival function S(t)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



(0,5] (5,10)

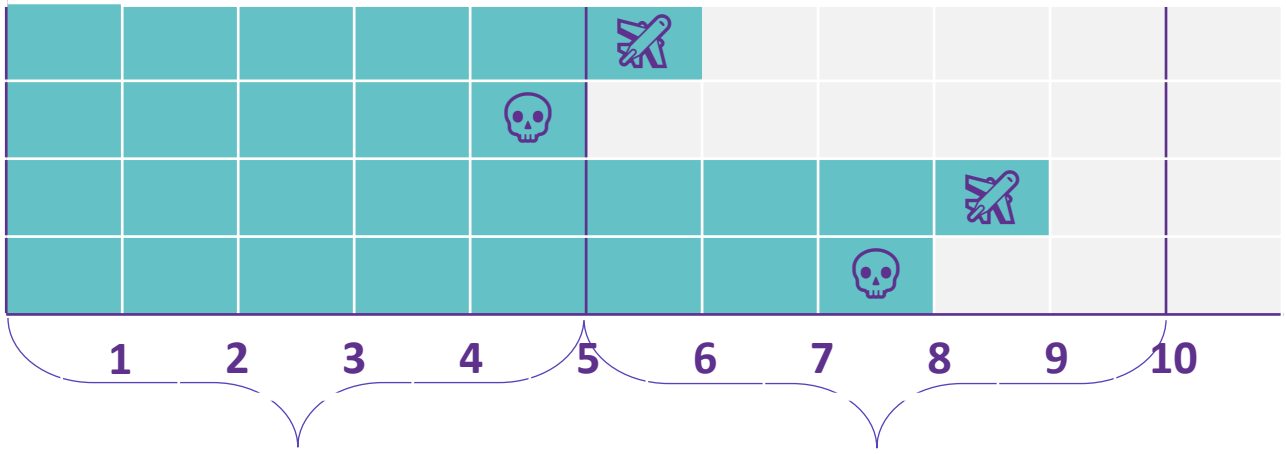
$$S(t_j) = \prod_{k=1}^{j} \frac{\left(n_k - \frac{m_k}{2}\right) - d_k}{\left(n_k - \frac{m_k}{2}\right)}$$

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	4	0	1	5	**0.750**	
(5,10)	3	2	1	5		

Time of censoring within interval "unknown". Therefore divide the censored events by 2 ($\frac{m_k}{2}$) averaging effect on denominator of censoring at start and end of episode

Life tables – Hazard function h(t)

ID	Time	Event
1	5	0
2	4	1
3	8	0
4	7	1



(0,5] (5,10)

First interval
Numerator = Deaths within interval
Denomonator = the risk set at beginning of interval - riskset at end/2 * interval width
=
1/17.5 = 0.057

$$h(t_j) = \frac{d_j}{\left(n_j - \left(\frac{d_j}{2} + \frac{m_j}{2}\right)\right)\Delta t_j}$$

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]	4	0	1	5	0.750	0.057
(5,10)	3	2	1	5		

Key estimates in EHA

1. Density, Survival & Hazard Function ✓
2. Life Tables ✓
3. Kaplan-Kaplan-Meier
4. Descriptive analysis

Kaplan-Meier

- Uses the time of event – not events within intervals as life tables
- Gives a measure for each duration with events only
- adjust risk set to censoring
- Very similar estimates to life tables when n is large

Kaplan-Meier

ID	Time	Event	
1	5	0	 Right censored by attrition at t5
2	4	1	 Censored at event at t4
3	8	0	 Right censored by attrition at t8
4	7	1	 Censored at event at t7

Kaplan-Meier

ID	Time	Event	
1	5	0	Right censored by attrition at t5
2	4	1	Censored at event at t4
3	8	0	Right censored by attrition at t8
4	7	1	Censored at event at t7
5	5	0	Right censored by attrition at t5
6	4	0	Right censored by attrition at t5
7	8	0	Right censored by attrition at t5
8	7	0	Right censored by attrition at t5

Kaplan-Meier

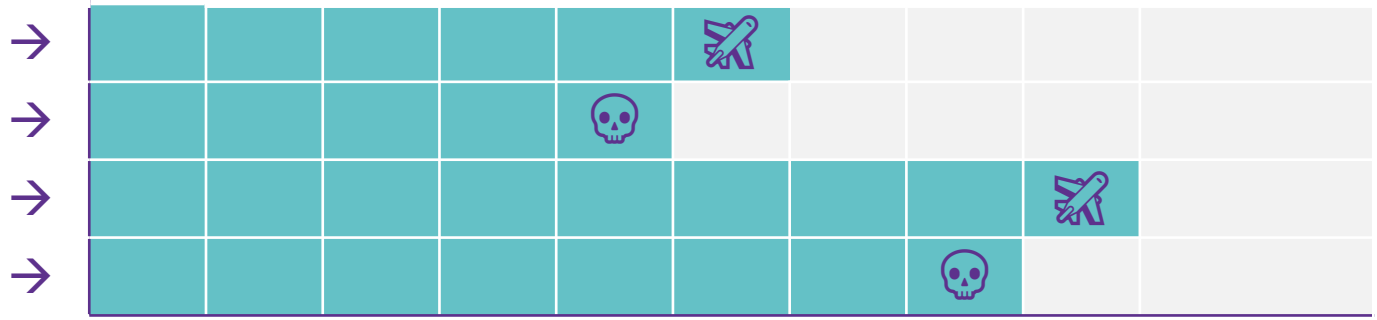
ID	Time	Event	
1	5	0	Right censored by attrition at t5
2	4	1	Censored at event at t4
3	8	0	Right censored by attrition at t8
4	7	1	Censored at event at t7
5	5	0	Right censored by attrition at t5
6	4	0	Right censored by attrition at t5
7	8	0	Right censored by attrition at t5
8	7	0	Right censored by attrition at t5

→

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1

Kaplan-Meier

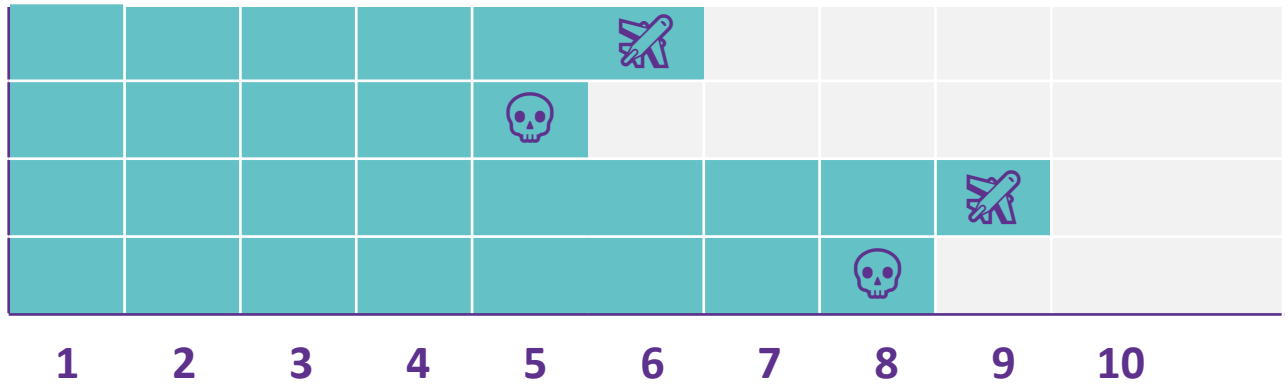
n	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



1 2 3 4 5 6 7 8 9 10

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



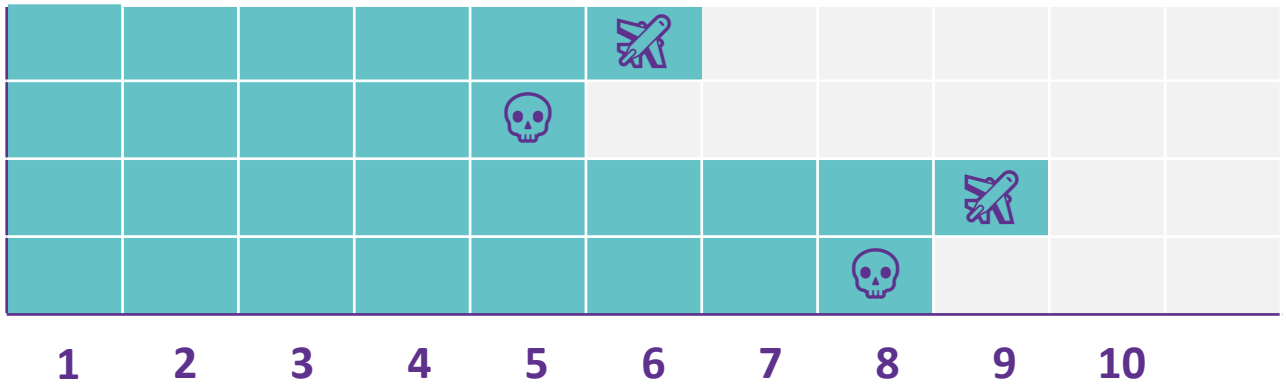
But the table rows should no longer be the interval as in life tables…

life table

t	n	m	d	Δ_t	S(t)	h(t)
(0,5]						
(5,10)						

Kaplan-Meier

n	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



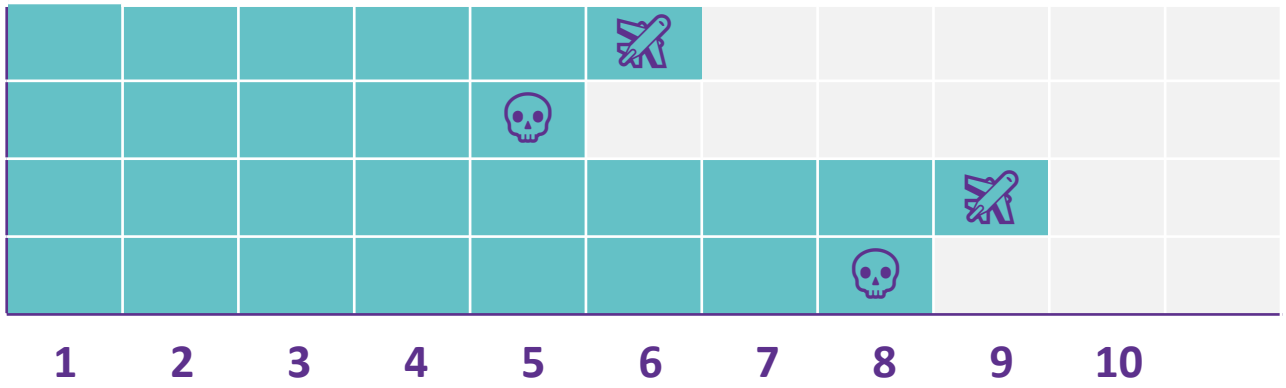
t = our time points where something happens

The rows are simply each unique duration time
Sorted in ascending order

t	n	m	d	S(t)
4				
5				
7				
8				

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1

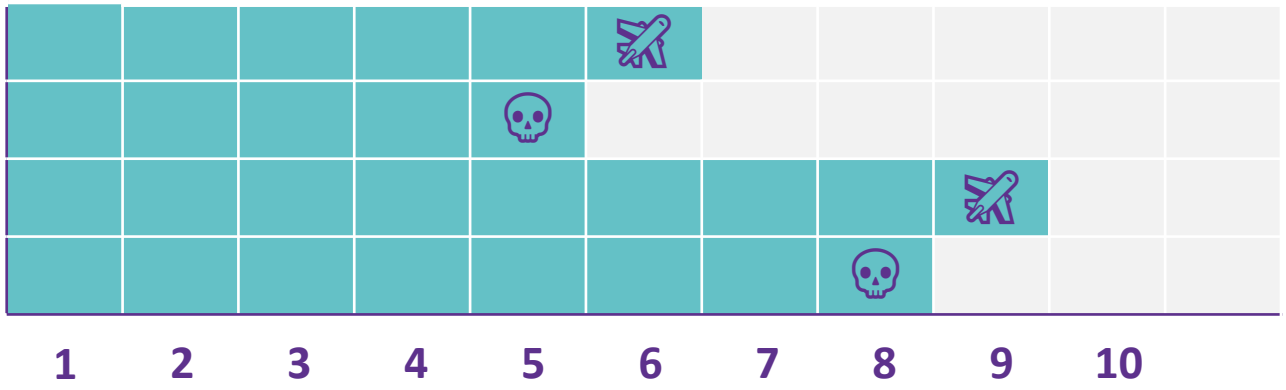


Add up the d deaths of each duration spell

t	n	m	d	S(t)
4			**1**	
5			**0**	
7			**1**	
8			**0**	

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1

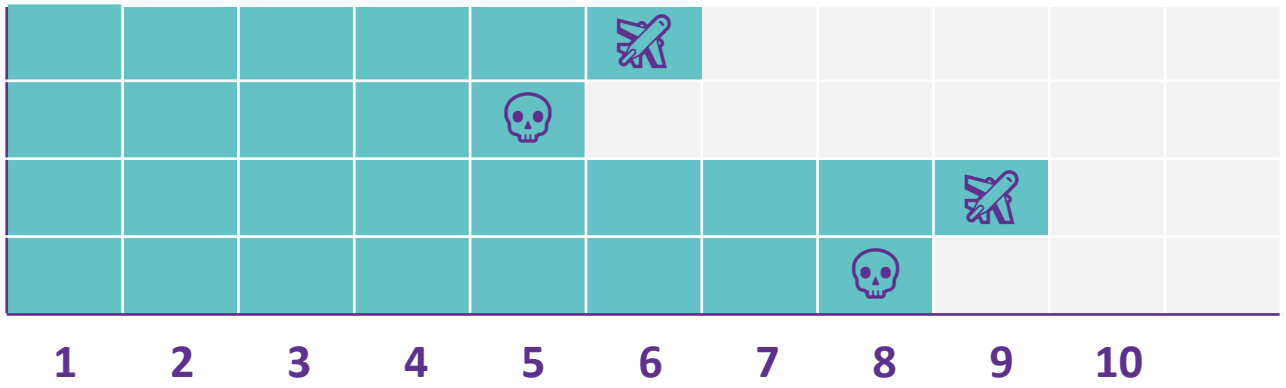


Add up the M censored of each duration spell

t	n	m	d	S(t)
4		0	1	
5		1	0	
7		0	1	
8		1	0	

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



Calculate n = be the number of individuals at the very start of spell t.

For n in the first spell/row t=4,that's the initial total population = 8.

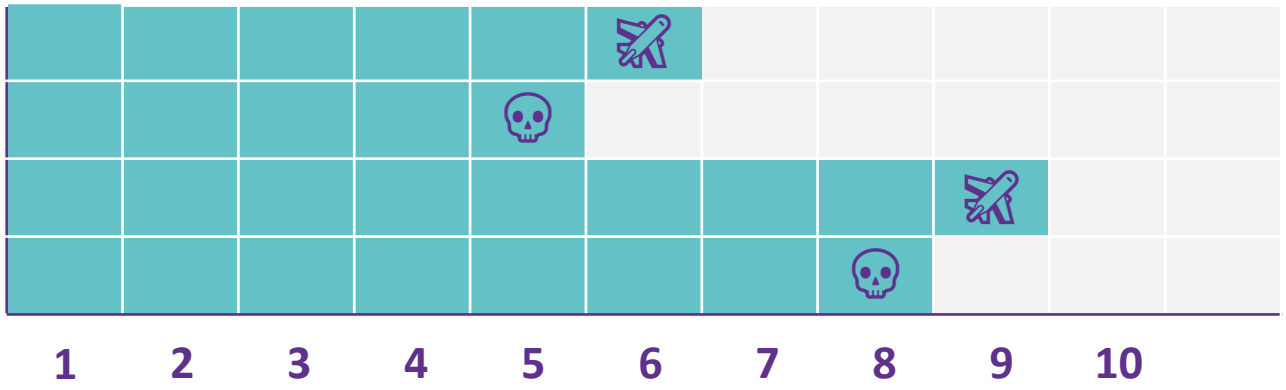
For n in the second spell/row t=5, we substract the dead and the censored of t=4.

8-1 dead = 7

t	n	m	d	S(t)
4	**8**	**0**	**1**	
5	**7**	**1**	**0**	
7	**6**	**0**	**1**	
8	**5**	**1**	**0**	

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



Survival function S(t)
(for the first row t=4)

Numerator: risk set at end of episode (n=8 – d=1 = 7).

Denominator is initial riskset = 8.

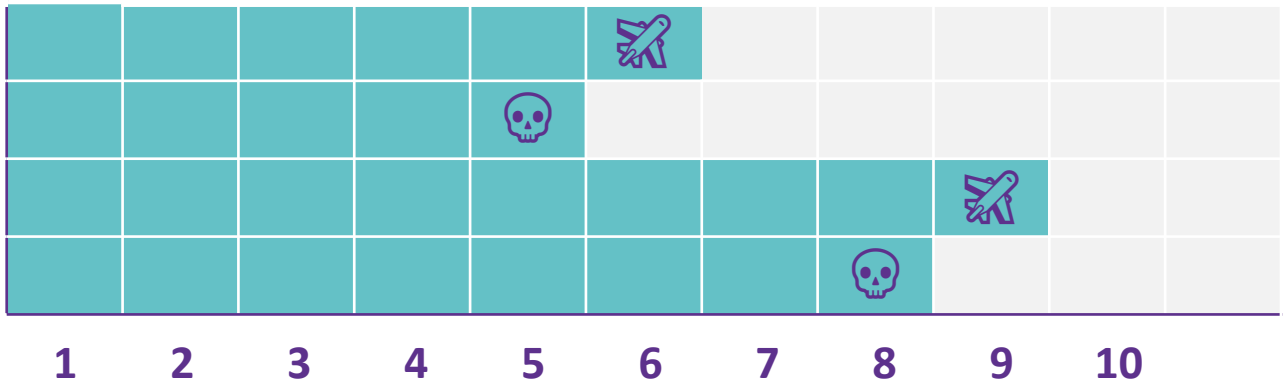
7/8 = 0.87

$$S(t_j) = \prod_{l=1}^{j} \frac{(n_l - d_l)}{n_l}$$

t	n	m	d	S(t)
4	**8**	**0**	**1**	**0.87**
5	**7**	**1**	**0**	
7	**6**	**0**	**1**	
8	**5**	**1**	**0**	

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



Survival function S(t)
(Second row t=5)

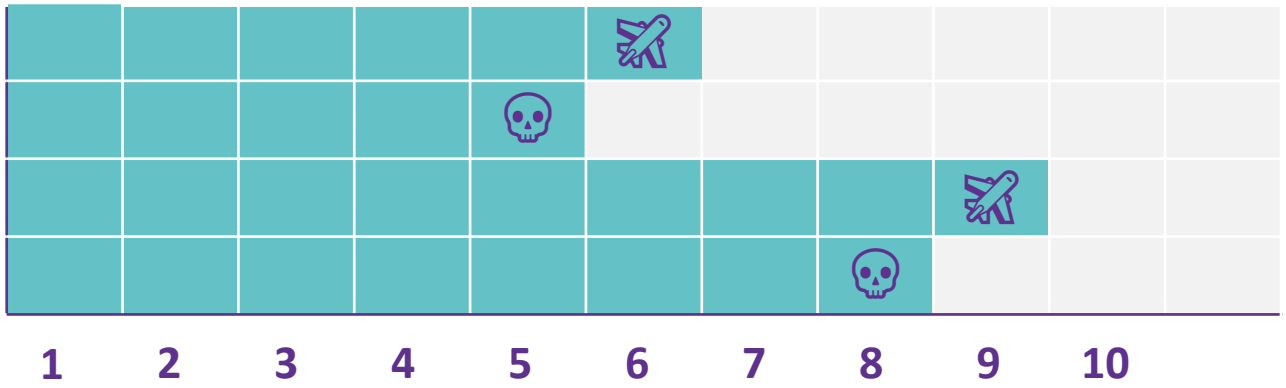
No death = do nothing to S(t)

$$S(t_j) = \prod_{l=1}^{j} \frac{(n_l - d_l)}{n_l}$$

t	n	m	d	S(t)
4	8	0	1	0.87
5	7	1	0	---
7	6	0	1	
8	5	1	0	

Kaplan-Meier

N ID	Time	Event
2	5	0
2	4	1
2	8	0
2	7	1



In a nutshell:
compress rows by event

Adjust risk set to death and censoring

While using only deaths as the numerator

t	n	m	d	S(t)
4	8	0	1	0.87
5	7	1	0	---
7	6	0	1	x.xx
8	5	1	0	---

Key estimates in EHA

1. Density, Survival & Hazard Function ✓
2. Life Tables ✓
3. Kaplan-Kaplan-Meier ✓
4. Descriptive analysis

15 minute break

INWS0038 Longitudinal and Multilevel Modelling II - Event History Analysis

22.04 – 10.05 2024 (3 ECTS)

Linus Andersson[1,2]

[1]University of Turku, [2]Stockholm University

linus.andersson@utu.fi

fluxconsortium.fi

INWS0038 Longitudinal and Multilevel Modelling II - Event History Analysis

Day 2. – Key estimates in EHA

Wednesday 24.04

Linus Andersson[12]

[1]University of Turku, [2]Stockholm University

linus.andersson@utu.fi

fluxconsortium.fi

Key estimates in EHA

- We covered
  - S(t) and h(t)
  - Life tables & Kaplan-Meier

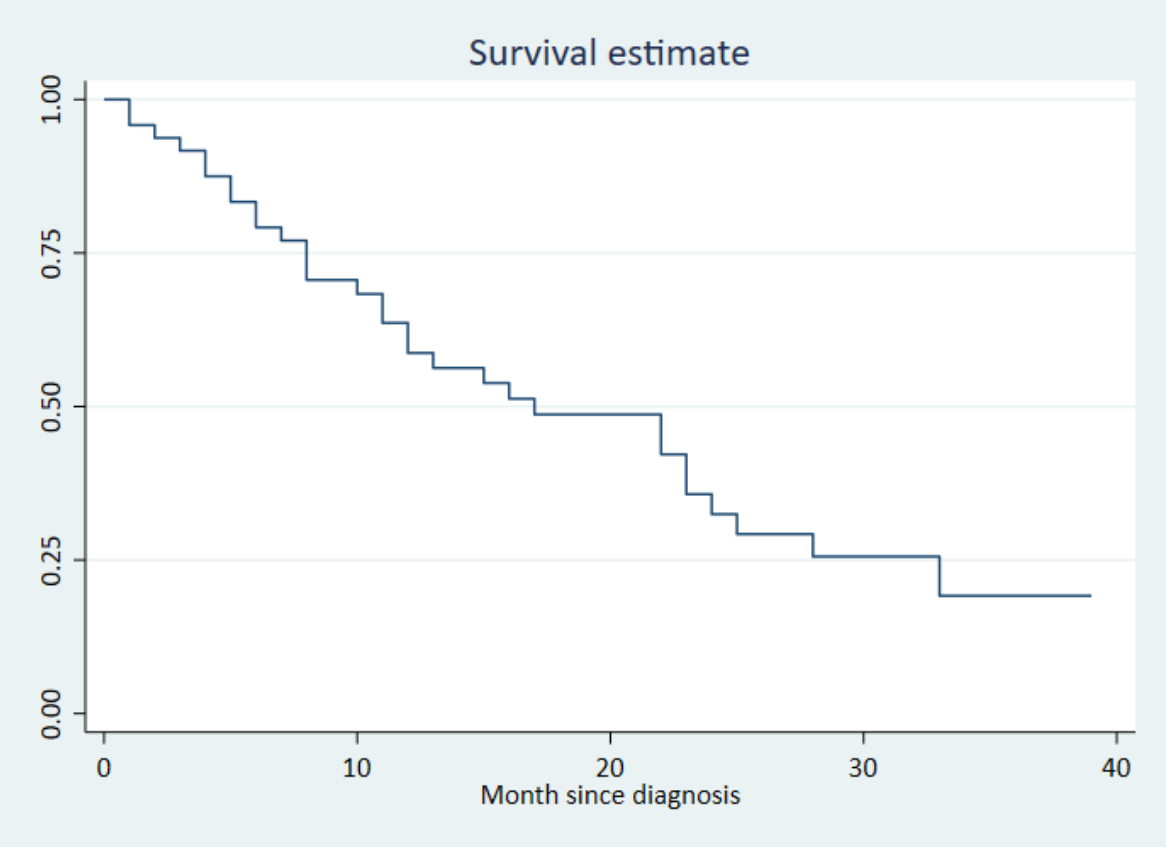
- Now, lets move on to describe data using these estimates
  - Focus on Kaplan-Meier
  - Vizual interpretation – no more tables
  - Stata - No more by-hand calculations
  - Foreshadow in-depth lab excercises

Key estimates in EHA

1. Kaplan-Meier curves
2. Test statistics
3. Pros and cons of non-parametric analysis

Survival function

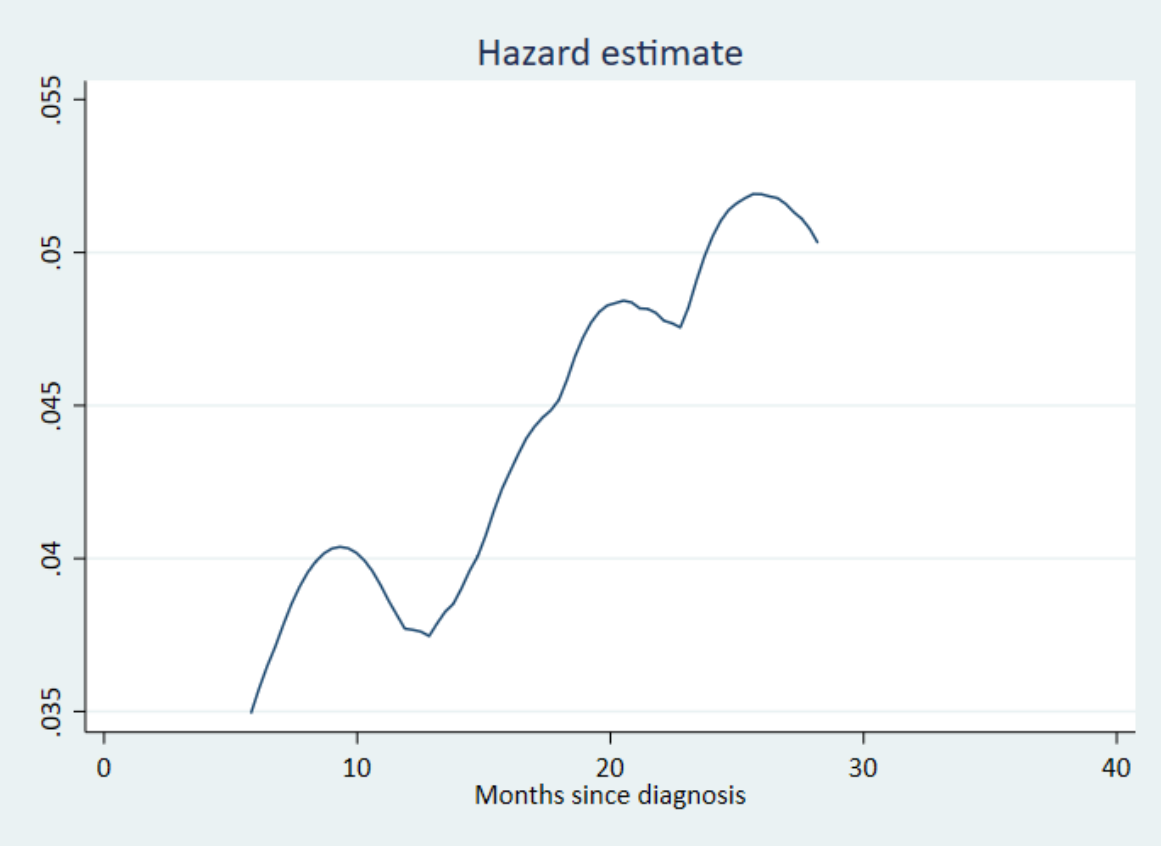
- Note the step-wise function. Flat line means no people die during this time.
- Decreasing only
- Range 1 to 0



"Among the diagnosed, probability of survival of 10 months is somewhat less than .7"

Hazard function

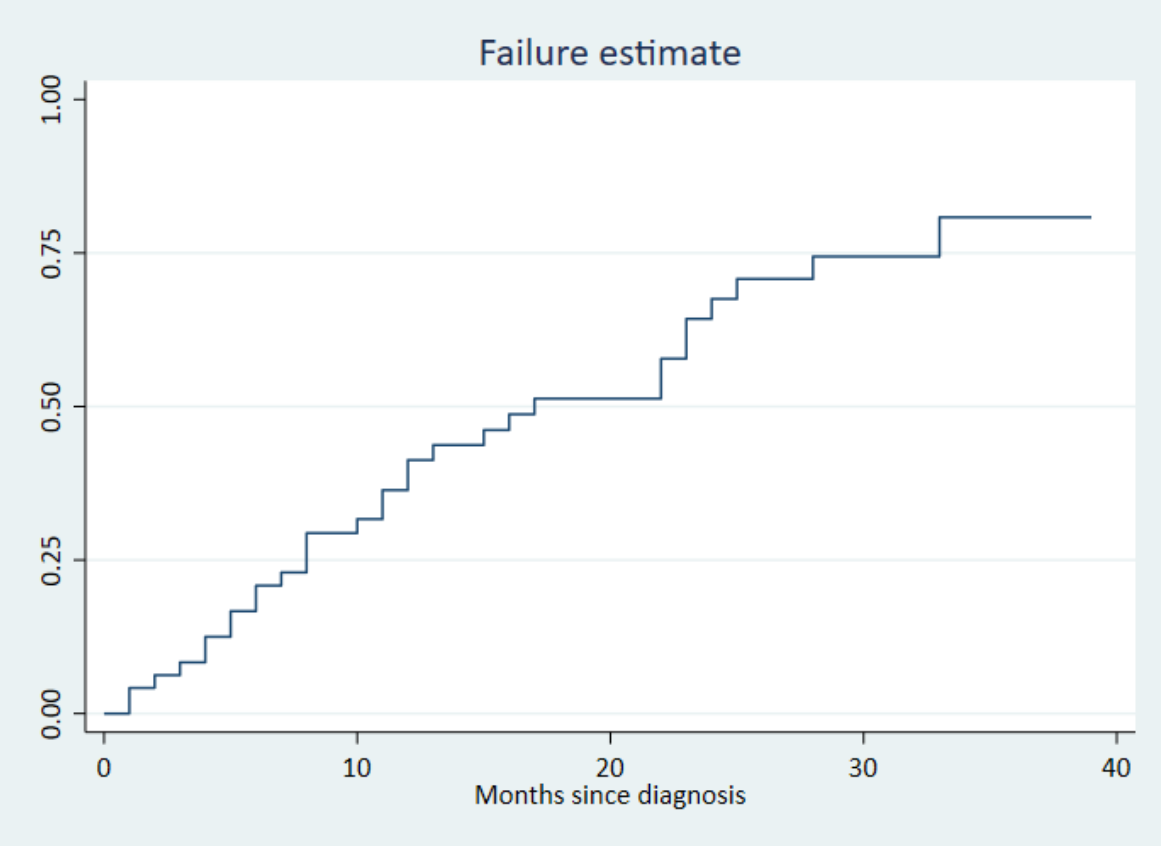
- Kaplan-Meier does not produce an actual hazard rate, but a useful approximation
- Smoothed – will depend on kernels and bandwidth
- Note that the hazard can increase or decrease over time



"The risk of death at a given timepoint, conditional on surviving to this timepoint, reaches a peak about 27 months after diagnosis"

Failure function

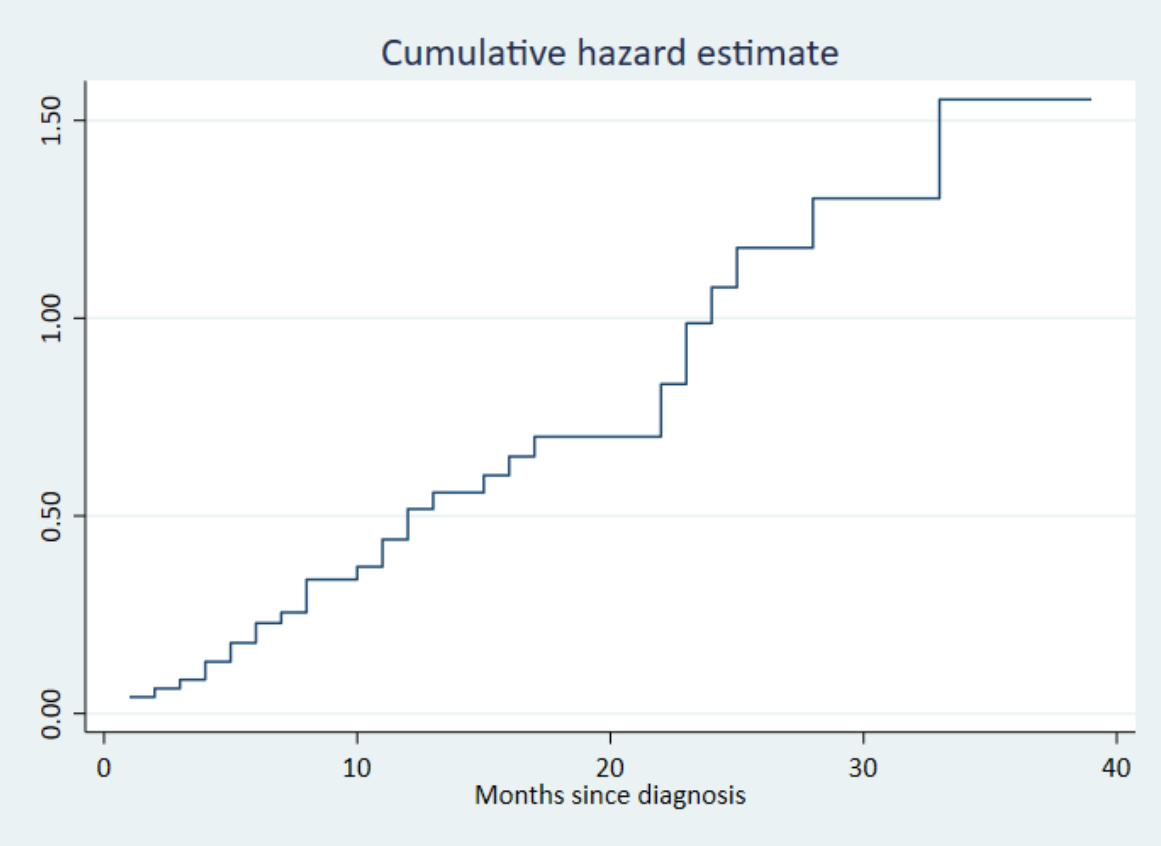
- “The probability of death by time t”
- Inverse of Survival function.
- cumulative probability function
- Always increasing
- Range 0 - 1



“since time of diagnosed, probability of death within 10 months somewhat higher than .25”

Cumulated hazard estimate

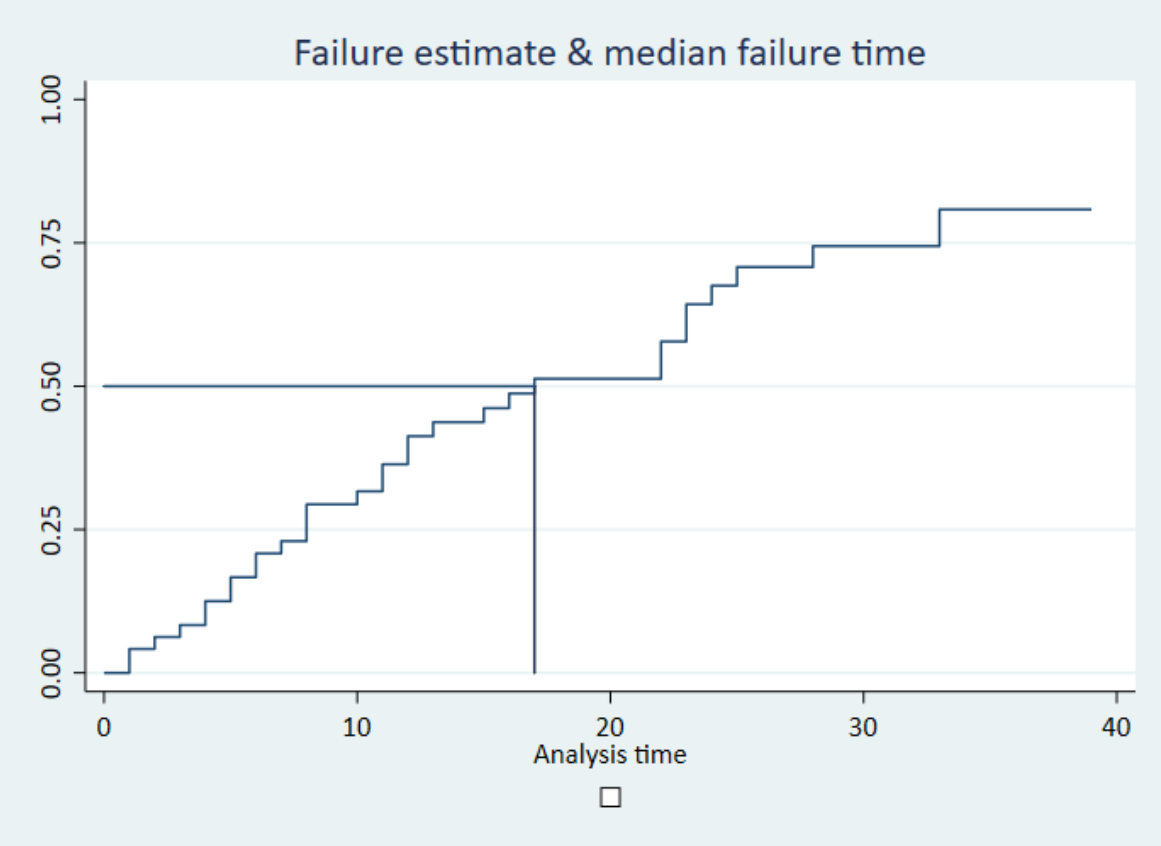
- "The number of times a subject is expected to die given the risk at every instance."
- A risk that accumulates.
- With non-recurring events: hypothetical
- Always increasing
- 1 to infinity



"at 40 moths, a person would be expected to be dead (more than once) given the hazard of death at all preceding time-points"

Median survival time

- Identical to median failure time (see picture)
- A “restriced median”, as right censoring is unknown – under-estimated
- The solution to restricted means - “restriced median” - is worse. Do not extrapolate median survival.
- Only use if 50% or more of sample experience event
- “Half of the initial risk-set died”



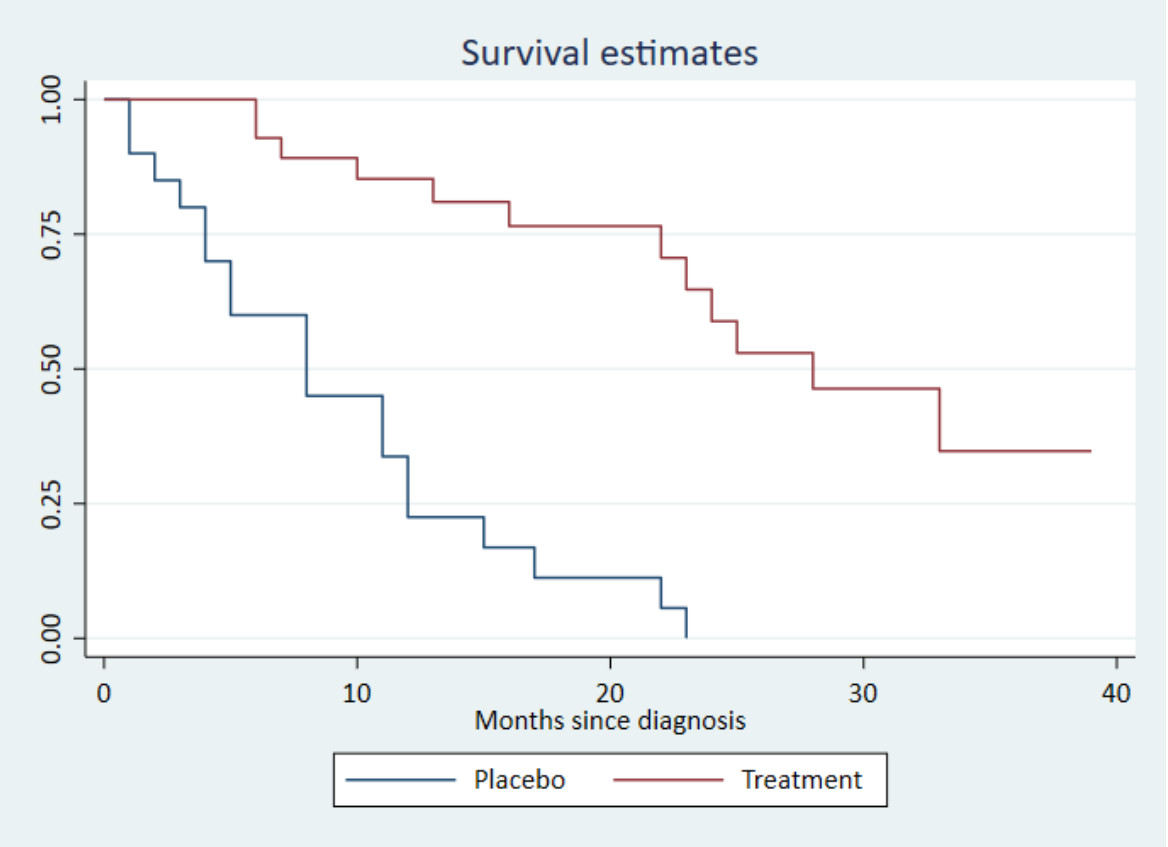
Key estimates in EHA

1. Kaplan-Meier curves ✓
2. Test statistics
3. Pros and cons of non-parametric analysis

Test statistics in EHA

- H_0
  - Survival functions of two groups equal at time t (local test)
  - Survival functions of two groups equal overall (global test)

- Two most common approaches
  - Log-Rank test
  - Wilcoxon-Breslow test

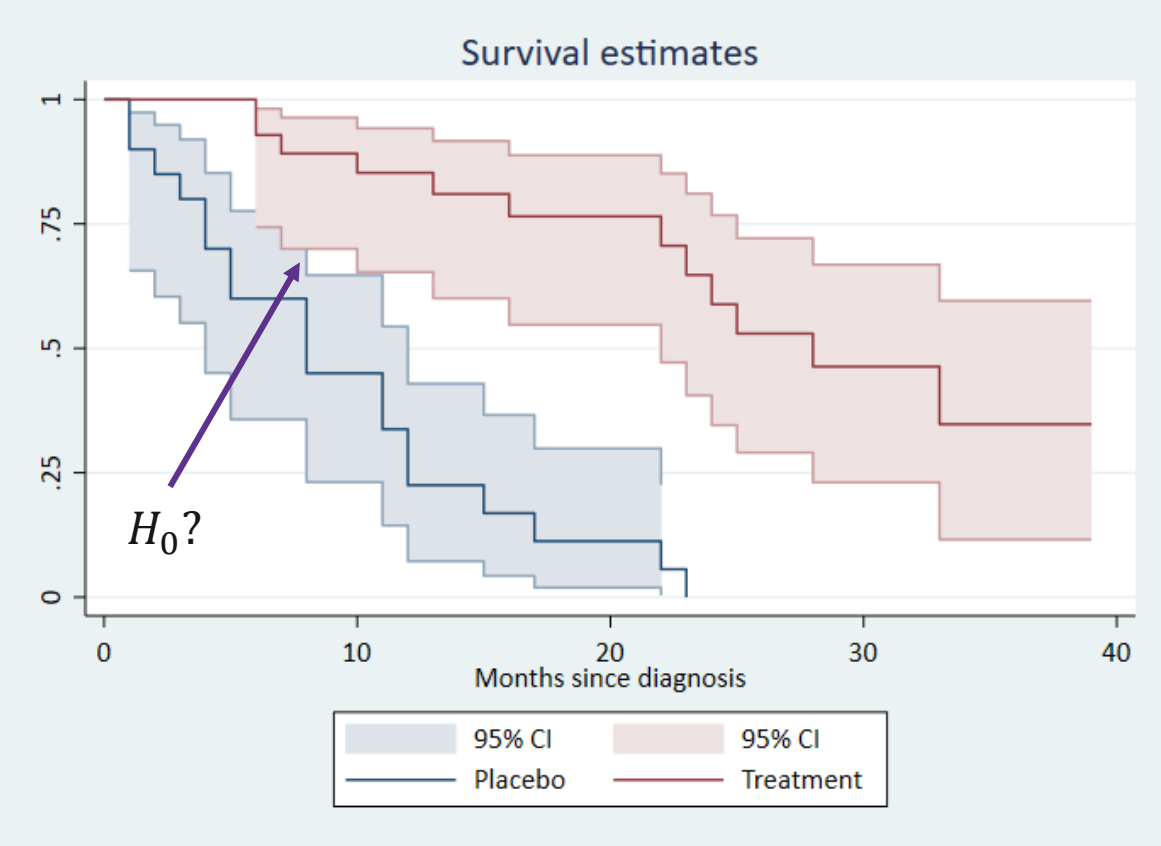


Test statistics in EHA

- Local test of differences at time t
  - H_0 : Survival functions of two groups equal at 8 months after diagnosis

- Quick approach
  - Estimate S(t) and SE at months 6 & 8 by treatment status
  - Use S(t) & SE of placebo and treatment to calculate z-score by hand

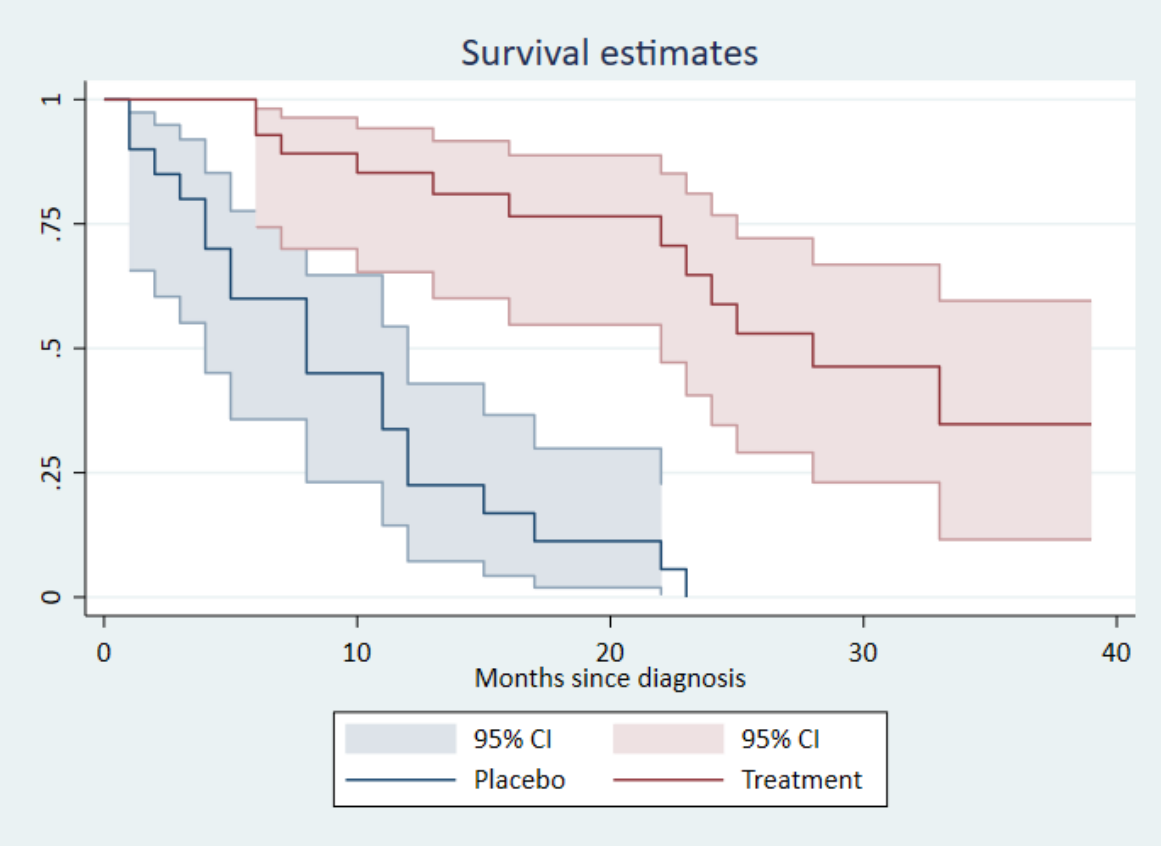
$$z = \frac{S(t8)_{Treatment} - S(t8)_{Placebo}}{\sqrt{SEt8_{Treatment} + SEt8_{Placebo}}}$$



Test statistics in EHA

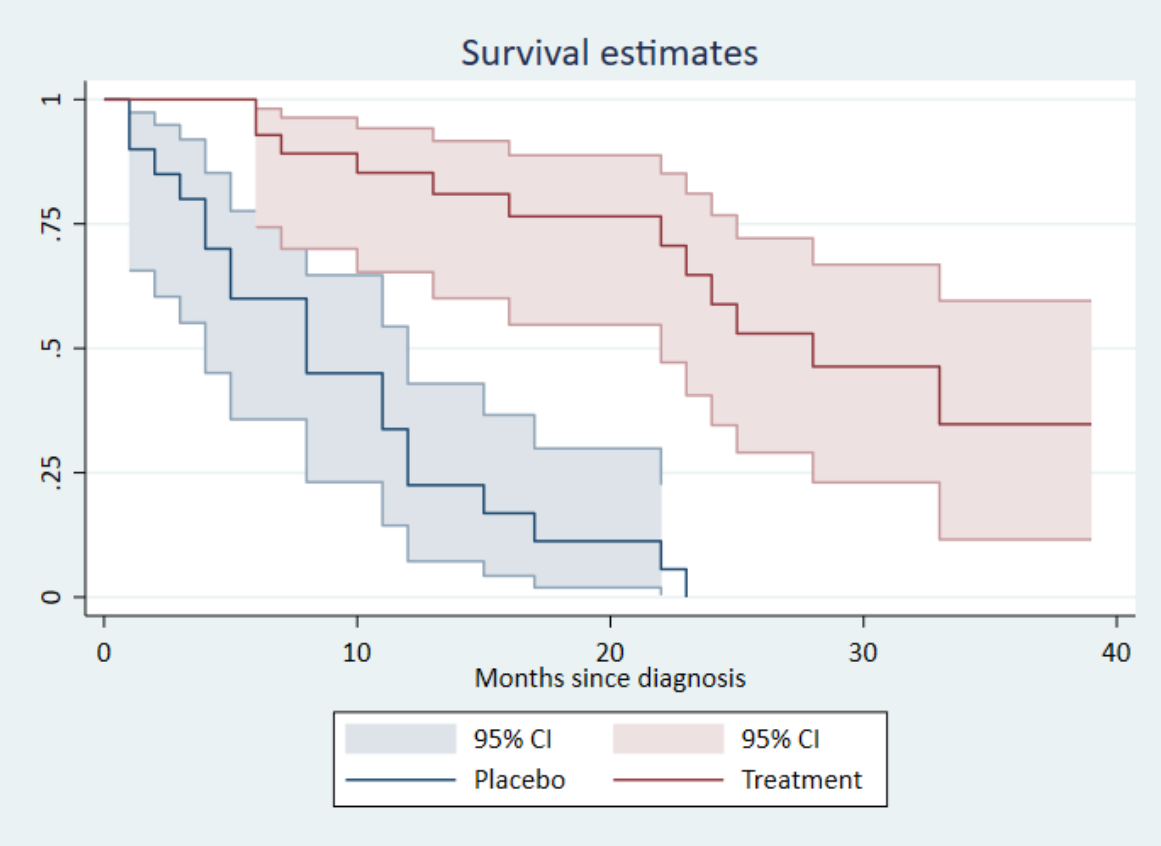
- global test of group differences
  - H_0 : Survival functions of placebo and treatment are overall equal

- Log-rank & Wilcoxon-Breslow test
  - Assumes Non-informative censoring
  - Log-rank sensitive to diverging survival function
  - WB sensitive to converging survival function



Test statistics in EHA

- Global test of trends
  - Survival functions of each age-at-diagnosis group is the same
  - Age-groups are ordinal so we could test (absence of) trends by age-at-diagnosis
  - Both Log-rank & Wilcoxon-Breslow test useful here



Key estimates in EHA

1. Kaplan-Meier curves ✓
2. Test statistics ✓
3. Pros and cons of non-parametric analysis

Pros and cons of non-parametric analysis

- Non-parametric models are robust
  - Data driven, not theory/model driven – no statistical artifacts
- Non-parametric models are limited
  - conditioning on covariates very unpractical
  - Theoretically driven explanation and prediction difficult
- Think about your RQ and data
  - What represents the key time process that you are interested in?
  - ”Figure nr 1” is likely a non-paramatetric description of this.

Key estimates in EHA

1. Kaplan-Meier curves ✓
2. Test statistics ✓
3. Pros and cons of non-parametric analysis ✓