Survival Analysis, Lecture 2 Non-parametric analysis of survival data

Ruth Keogh

Department of Medical Statistics
London School of Hygiene and Tropical Medicine

Aims

- In the last lecture we focused on general concepts
- We now move on to being able to perform some analyses of real survival data
- Unlike at the end of the last lecture, in this lecture we do not assume a particular parametric form for the distribution of the survival times
- As such, the methods we use are described as non-parametric

Aims

Part 1

- Kaplan-Meier method: Estimating survivor functions non-parametrically
- Including estimating uncertainty in the non-parametric estimates

Part 2

- Comparing survival in different groups of individuals
- The log rank test

Part 3

- Estimating the cumulative hazard
- The life table approach

Why use non-parametric methods

Non-parametric methods are a relatively simple starting point for most analyses of survival data.

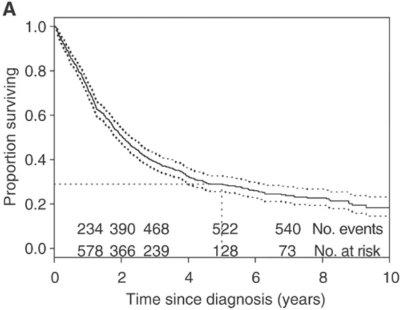
- We can estimate survivor functions without having to make parametric assumptions.
- Non-parametric methods provide a nice way of graphically displaying survival data, including when there is censoring.
- Non-parametric methods provide a simple way of comparing patterns of survival in two (or more) groups of individuals.
- Non-parametric methods can be used to inform more complex modelling of survival data.

Estimating the survivor function: The Kaplan-Meier approach

Example

Clark et al. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer 2003; 89: 232–238.

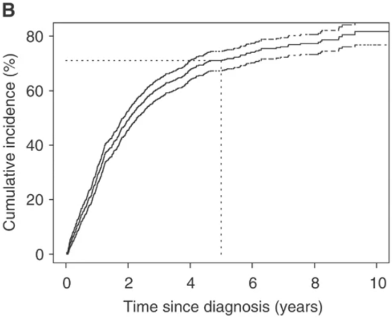
- Survival after diagnosis with ovarian cancer
- 825 patients diagnosed between Jan 1990 and Dec 1999
- Follow-up was until December 2020



Example

Clark et al. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer 2003; 89: 232–238.

- Survival after diagnosis with ovarian cancer
- 825 patients diagnosed between Jan 1990 and Dec 1999
- Follow-up was until December 2020



Reminder of the survivor function

Survivor function: definition

$$S(t) = \Pr(T > t)$$

Suppose that we had no censored survival times in our data and we observed the outcome at times

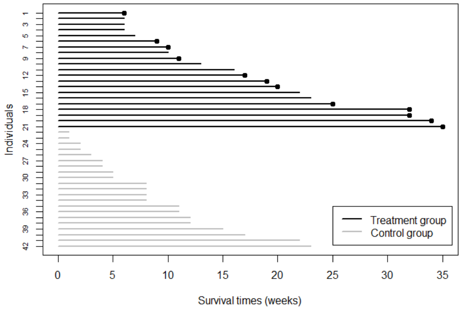
$$t_1 < t_2 < \ldots < t_K$$

An intuitive estimate of the survivor function at time t_j

$$\widehat{S}(t_j) = \frac{\text{Number of individuals with } t > t_j}{\text{Total number of individuals}}$$

- This estimate only exists at the observed times $\hat{S}(t_1), \ldots, \hat{S}(t_K)$
- ... because do not have information in our sample data about what happens in between the observed survival times

Example: Time to death in leukaemia patients



Group	Survival and censoring (*) times
Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23
Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

We don't yet know how to estimate the survivor function when there is censoring so we focus for now on the control group.

Example continued...

Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23

t_j	Number of events d_j	Survivor function estimate $\widehat{S}(t_j)$
1	2	
2	2	
3	1	
4	2	
5	2	
8	4	
11	2	
12	2	
15	1	
17	1	
22	1	
23	1	

Example continued...

Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23

t_j	Number of events d_j	Survivor function estimate $\widehat{S}(t_j)$
1	2	19/21=0.90
2	2	
3	1	
4	2	
5	2	
8	4	
11	2	
12	2	
15	1	
17	1	
22	1	
23	1	

Example continued...

Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23

t_j	Number of events d_j	Survivor function estimate $\widehat{S}(t_j)$
1	2	19/21=0.90
2	2	17/21=0.81
3	1	
4	2	
5	2	
8	4	
11	2	
12	2	
15	1	
17	1	
22	1	
23	1	

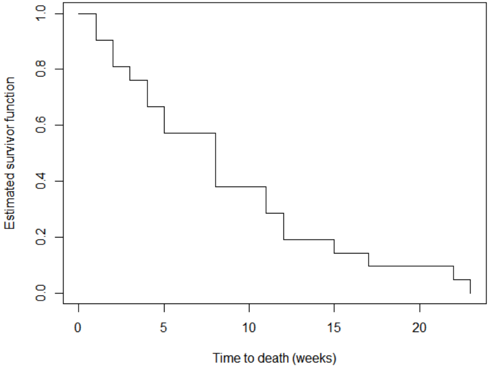
Example continued...

Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23

t_j	Number of events d_j	Survivor function estimate $\widehat{S}(t_j)$
1	2	19/21=0.90
2	2	17/21=0.81
3	1	16/21=0.76
4	2	14/21=0.67
5	2	12/21=0.57
8	4	8/21=0.38
11	2	6/21=0.29
12	2	4/21=0.19
15	1	3/21=0.14
17	1	2/21=0.10
22	1	1/21=0.05
23	1	0

Example continued...

t_j	$\widehat{S}(t_j)$
1	0.90
2	0.81
3	0.76
4	0.67
5	0.57
8	0.38
11	0.29
12	0.19
15	0.14
17	0.10
22	0.05
23	0



Incorporating censoring: estimating the hazard

- We observe the outcome at times $t_1 < t_2 < \ldots < t_K$
- ...but there are also some censoring times
- Define the hazard at each survival time: $h_1, h_2, \ldots, h_K$

Estimating the time-specific hazard from the data

$$\hat{h}_j = d_j / n_j$$

d_j: number of events at time t_j

n_j: number of people at risk at time t_j

Concept of being 'at risk'

- A person is at risk at time t_j if they have not yet had the outcome of interest and have not been censored

Incorporating censoring: estimating $S(t_j)$

1. We have hazards at each survival time: $h_1, h_2, \ldots, h_K$
2. The probability that a person who has survived until time t_1 does not have the event at time t_1 is

$$1 - h_1$$

3. The probability that an individual who does not have the outcome at time t_1 then also survives until time t_2 and does not have the outcome at time t_2 is

$$(1 - h_1)(1 - h_2)$$

4. The survival probability is the probability that an individual does not have the outcome at any time at which they are eligible to have the outcome

$$S(t_j) = \prod_{k=1}^{j} (1 - h_j)$$

Incorporating censoring: estimating $S(t_j)$

Estimating the time-specific hazard from the data

$$\hat{h}_j = d_j/n_j$$

d_j: number of events at time t_j

n_j: number of people at risk at time t_j

Kaplan-Meier estimate of the survivor function

This gives, from the previous slide:

$$\widehat{S}(t_j) = \prod_{k=1}^{j} (1 - d_k/n_k)$$

$$\widehat{S}(t) = \prod_{j|t_j \leq t} (1 - d_j/n_j)$$

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	
7	1	0		
9	0	1		
10	1	1		
11	0	1		
13	1	0		
16	1	0		
17	0	1		
19	0	1		
20	0	1		
22	1	0		
23	1	0		

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	(1-3/21)=0.857
7	1	0		
9	0	1		
10	1	1		
11	0	1		
13	1	0		
16	1	0		
17	0	1		
19	0	1		
20	0	1		
22	1	0		
23	1	0		

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	(1-3/21)=0.857
7	1	0	17	
9	0	1		
10	1	1		
11	0	1		
13	1	0		
16	1	0		
17	0	1		
19	0	1		
20	0	1		
22	1	0		
23	1	0		

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	(1-3/21)=0.857
7	1	0	17	(1-3/21)(1-1/17)=0.807
9	0	1		
10	1	1		
11	0	1		
13	1	0		
16	1	0		
17	0	1		
19	0	1		
20	0	1		
22	1	0		
23	1	0		

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

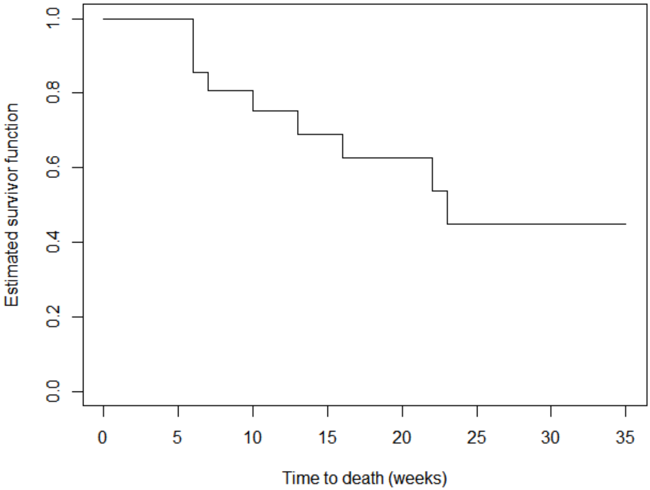
Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	(1-3/21)=0.857
7	1	0	17	(1-3/21)(1-1/17)=0.807
9	0	1		
10	1	1		
11	0	1		
13	1	0		
16	1	0		
17	0	1		
19	0	1		
20	0	1		
22	1	0		
23	1	0		

Example: Time to death among leukaemia patients in the treatment group

Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Survival and censoring times	No. events	No. censorings	No. at risk	Kaplan-Meier estimate $\widehat{S}(t)$
6	3	1	21	(1-3/21)=0.857
7	1	0	17	(1-3/21)(1-1/17)=0.807
9	0	1	16	-
10	1	1	15	0.753
11	0	1	13	-
13	1	0	12	0.690
16	1	0	11	0.627
17	0	1	10	-
19	0	1	9	-
20	0	1	8	-
22	1	0	7	0.538
23	1	0	6	0.448

Example: Time to death among leukaemia patients in the treatment group



Estimating uncertainty in the Kaplan-Meier estimate

- As usual, we want to be able to say something about the precision of our Kaplan-Meier estimates of the survivor function
- This enables us to add confidence intervals to the Kaplan-Meier plots

To do all this we need to find the variance of the Kaplan-Meier estimates

$$\text{var}\{\widehat{S}(t)\} = \text{var}\left\{\prod_{j|t_j \le t}(1-\hat{h}_j)\right\}$$

This variance is estimated by using a series of approximations

Estimating uncertainty in the Kaplan-Meier estimate

Step 1: We start by considering the variance of $\log \hat{S}(t)$

$$\begin{aligned}
\text{var}\{\log \widehat{S}(t)\} &= \text{var}\left\{\log \prod_{j|t_j \le t} (1-\hat{h}_j)\right\} \\
&= \text{var}\left\{\sum_{j|t_j \le t} \log(1-\hat{h}_j)\right\} \\
&= \sum_{j|t_j \le t} \text{var}\left\{\log(1-\hat{h}_j)\right\}
\end{aligned}$$

Step 2: Use a linear approximation

$$\log(1-\hat{h}_j) \approx \log(1-h_j) + (\hat{h}_j - h_j)/(1-h_j)$$

$$\text{var}\left\{\log(1-\hat{h}_j)\right\} \approx \frac{\text{var}(\hat{h}_j)}{(1-h_j)^2}$$

Estimating uncertainty in the Kaplan-Meier estimate

Step 2: Use a linear approximation

$$\text{var}\left\{\log(1-\hat{h}_j)\right\} \approx \frac{\text{var}(\hat{h}_j)}{(1-h_j)^2}$$

Step 3: Use what we know about $\hat{h}_j = d_j/n_j$

$$d_j \sim \text{Binomial}(n_j, h_j)$$

$$\text{var}(\hat{h}_j) = \text{var}\left(\frac{d_j}{n_j}\right) = \frac{\text{var}\left(d_j\right)}{n_j^2} = \frac{h_j(1-h_j)}{n_j}$$

Step 4: Put it all together to give...

$$\text{var}\left\{\log \widehat{S}(t)\right\} = \sum_{j|t_j \leq t} \text{var}\left\{\log(1-\hat{h}_j)\right\} = \sum_{j|t_j \leq t} \frac{h_j}{n_j(1-h_j)}$$

Estimating uncertainty in the Kaplan-Meier estimate

Step 4: Put it all together to give...

$$\mathrm{var}\left\{\log \widehat{S}(t)\right\} = \sum_{j|t_j \le t} \mathrm{var}\left\{\log(1-\hat{h}_j)\right\} = \sum_{j|t_j \le t} \frac{h_j}{n_j(1-h_j)}$$

Final step: We want $\mathrm{var}\left\{\widehat{S}(t)\right\}$

$$\log \widehat{S}(t) \approx \log S(t) + \left\{\widehat{S}(t) - S(t)\right\}/S(t)$$

$$\mathrm{var}\left\{\log \widehat{S}(t)\right\} = \mathrm{var}\left\{\widehat{S}(t)\right\}/S(t)^2$$

Greenwood's formula

$$\mathrm{var}\left\{\widehat{S}(t)\right\} = \widehat{S}(t)^2 \mathrm{var}\left\{\log \widehat{S}(t)\right\} = \widehat{S}(t)^2 \sum_{j|t_j \le t} \frac{h_j}{n_j(1-h_j)}$$

Confidence intervals

95% CI for the Kaplan-Meier estimate using Greenwood's Formula

$$S(t) \pm 1.96\sqrt{\text{var}\{\widehat{S}(t)\}}$$

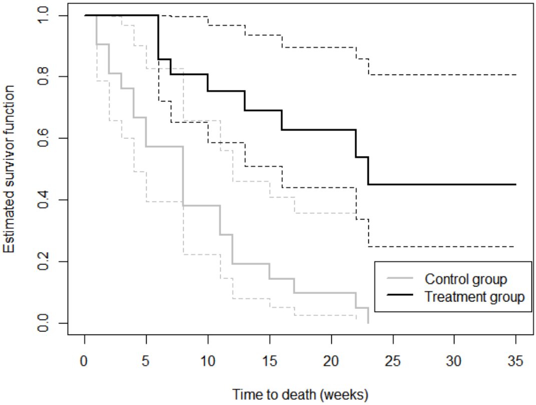
- This can give values outside the range 0 to 1

Alternative confidence intervals

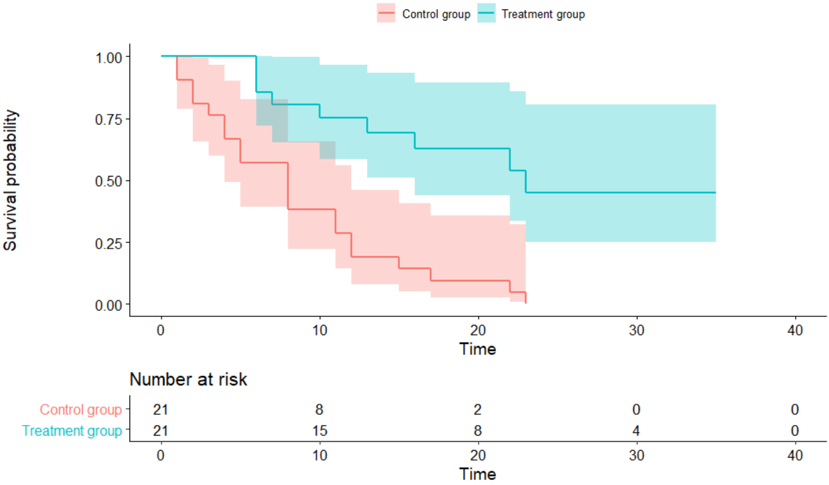
$$\text{var}\left\{\log\left(-\log\widehat{S}(t)\right)\right\} \approx \frac{\text{var}\left\{\log\widehat{S}(t)\right\}}{\{\log S(t)\}^2} = v(t)^2$$

$$S(t)^{\exp\{\pm 1.96 v(t)\}}$$

Example continued: Adding 95% CIs



Example continued: Adding 95% CIs

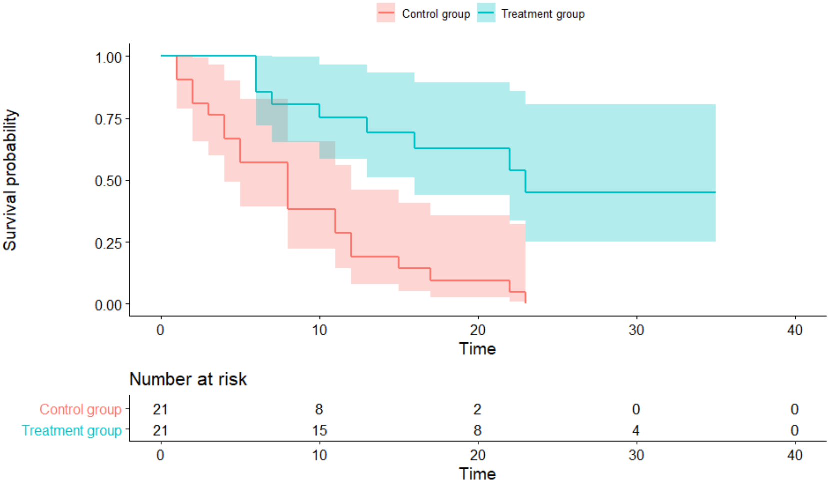


Comparing survival in two or more groups

Extending Kaplan-Meier

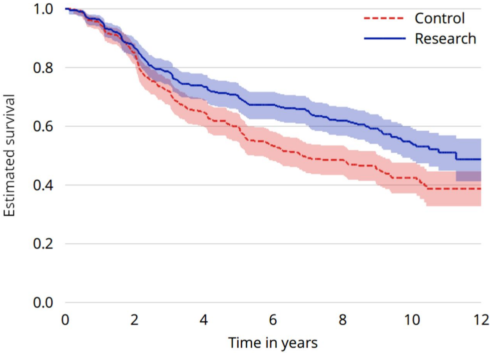
- Very often we want to compare patterns in survival in two groups of individuals
  - placebo and active treatment groups in a randomized trial
  - smokers and non-smokers in an observational study
  - takers and non-takers of statins in an observational study
- The Kaplan-Meier approach can be extended to two or more groups of individuals
  - Simply follow the procedure separately with each group
  - Plot the estimated survivor curves on the same graph to compare

Example continued: Adding 95% CIs



Example

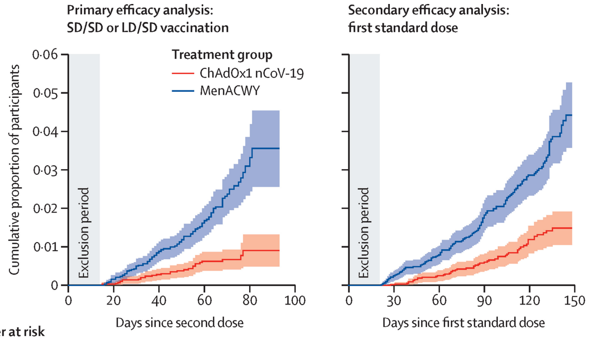
Tim Morris, Chris Jarvis, et al. Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. BMJ Open 2019;9:e030215. doi: 10.1136/bmjopen-2019-030215



Control													
At risk	418	394	351	289	255	224	190	165	154	120	71	27	6
Censored	0	2	5	13	18	30	40	51	60	85	127	167	188
Events	0	22	62	116	145	164	188	202	204	213	220	224	224
Research													
At risk	418	397	351	313	283	263	239	218	198	165	99	39	8
Censored	0	5	12	17	27	35	48	62	71	96	151	208	237
Events	0	16	55	88	108	120	131	138	149	157	168	172	173

Example

Voysey et al. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. The Lancet 2020; 397: 99-111.



Number at risk (number censored)												
ChAdOx1 nCoV-19	5807 (0)	5639 (189)	4779 (1162)	3181 (2620)	499 (5300)	0 (5777)	6307 (0)	5732 (645)	4857 (1443)	3681 (2636)	2490 (3811)	0 (6256)
MenACWY	5829 (0)	5657 (182)	4765 (1164)	3146 (2636)	435 (5322)	0 (5728)	6297 (0)	5718 (639)	4836 (1424)	3652 (2599)	2452 (3760)	0 (6156)

Making formal comparisons

- We can make various observations about the difference between survival curves by looking at plots
- But it is desirable to make a more formal comparison of the survivor curves in two groups
- This can be done using a test called the log rank test
- Also sometimes called the Mantel-Haenszel test

The log rank test: focus on two groups

Group	Survival and censoring (*) times
Control group	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23
Treatment group	6*,6,6,6,7,9*,10*,10,11*,13,16, 17*,19*,20*,22,23,25*,32*,32*,34*,35*

Table of events and non-events at each survival time t_j

Group	No. events at t_j	No. surviving beyond t_j	Total (No. at risk)
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

The log rank test: focus on two groups

Table of events and non-events at each survival time t_j

Group	No. events at t_j	No. surviving beyond t_j	Total (No. at risk)
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

- Under the null hypothesis that the risk of having the event does not differ in the two groups, d_{1j} (or equivalently d_{2j}) has a hypergeometric distribution
- Under the null hypothesis, using information about this distribution, the expected number of events in group 1 at time t_j is

$$e_{1j} = \frac{n_{1j} d_j}{n_j}$$

The log rank test: focus on two groups

- Under the null hypothesis we expect to see no difference in the observed and expected total number of events in the two groups:

$$\sum_j (d_{1j} - e_{1j}) = 0, \qquad \sum_j (d_{2j} - e_{2j}) = 0$$

- Using the hypergeometric distribution

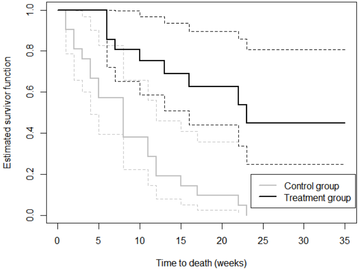
$$v_{1j}^2 = \text{var}(d_{1j}) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

The log rank test

Null hypothesis: the survival curves in the two groups are the same

$$\frac{\left\{\sum_j (d_{1j} - e_{1j})\right\}^2}{\sum_j v_{1j}^2} \sim \chi_1^2$$

Example: Comparing survival curves in treatment and control groups of leukaemia patients



Log rank test

$$\sum_j (d_{1j} - e_{1j}) = 10.3, \sum_j v_{1j}^2 = 6.56$$

Test statistic 16.8 $\Rightarrow$ p-value < 0.0001

Estimating the cumulative hazard

Estimating the cumulative hazard: $H(t)$

- Sometimes it is of interest to estimate the cumulative hazard

Recall the relationships

$$H(t) = -\log S(t)$$

$$H(t) = \int_0^t h(u)du$$

- These relationships give rise to two estimators for the cumulative hazard

Estimating the cumulative hazard: $H(t)$

Recall the relationships

$$H(t) = -\log S(t)$$

$$H(t) = \int_0^t h(u)du$$

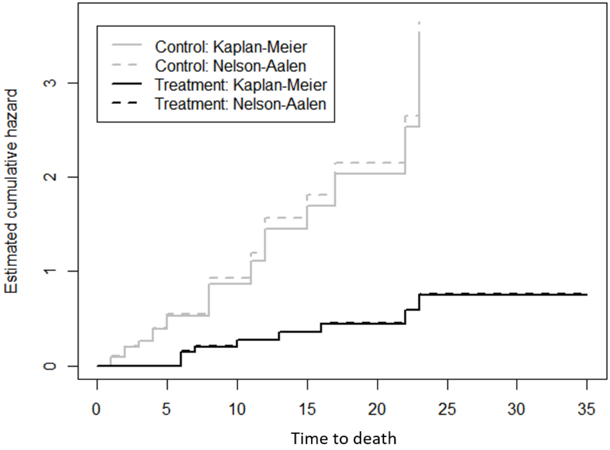
Kaplan-Meier estimate

$$\widehat{H}(t) = -\log \widehat{S}(t)$$

Nelson-Aalen estimate

$$\widetilde{H}(t) = \sum_{j|t_j \le t} \hat{h}_j = \sum_{j|t_j \le t} d_j/n_j$$

Example: Leukaemia patient data: Comparing Kaplan-Meier and Nelson-Aalen estimates of the cumulative hazard



The life-table approach

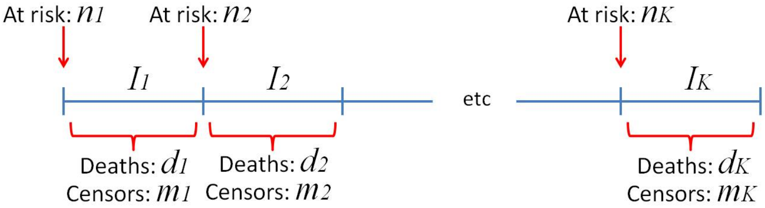
The life table approach

- The Kaplan-Meier approach requires ‘exact’ survival or censoring times
- Sometimes instead of observing individual times, we observe the number of events or censorings within a series of time ranges

Example: Death among men with angina

Year	Number at risk	Number deaths	Number censored
0-1	2418	456	0
1-2	1962	226	39
2-3	1697	152	22
3-4	1523	171	23
4-5	1329	135	24

The life table approach



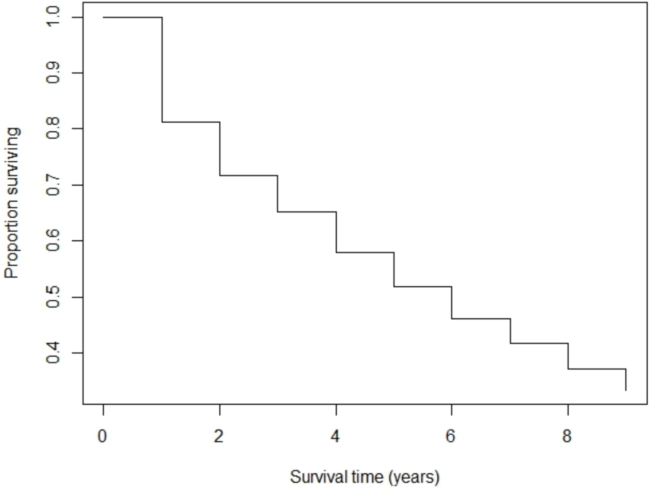
Survivor function estimate

Estimated probability of having the event in interval j for a person at risk at the start of the interval:

$$p_j = \frac{d_j}{n_j - m_j/2}$$

$$\widehat{S}(t) = \prod_{k=1}^{j} (1 - p_k) \qquad t_j \leq t < t_{j+1}$$

Example: Death among men with angina: the life table approach



Extensions

- We can compare survival in more than 2 groups by plotting several survivor curves on the same graph.
- It is often of interest to control for potential confounders in our analyses.
  - We can look at survival curves for the main exposure within strata defined by confounding variables.
  - There is a stratified version of the log rank test.
- This approach becomes increasingly cumbersome as the number of confounders increases.

Extensions

- A drawback of non-parametric methods is that they do not provide an easy way of quantifying differences in survival between groups.
- Non-parametric methods do not allow us to investigate the impact of continuous variables on survival (e.g. blood pressure).
- This is where we need to start thinking about regression-based methods.