Medical Statistics: Survival Analysis

Dr Nick Fieller
Department of Probability & Statistics
University of Sheffield

MAS6012/MAS361/MAS461
2011/2012

Contents

Contents

- **Organization of Course**
  - Two Components
    - Clinical trials
      - Experiments on human (and animal) subjects
      - Ethical issues, efficient use of subjects, etc
    - Survival Analysis
      - (analyzing data on length of lifetimes, e.g. times of remission in leukaemia)
  - Approximately 10 lectures each

- **Organization of course material**
  - Two sets of lecture notes (Survival & Clinical)
  - Clinical Chapters 1 – 10 (~ 1 per lecture)
  - Survival Chapters 1 – 4 main part of course
  - Appendix 0: background maths
    - Maximum Likelihood Estimation
      - (but used only in a couple of places)
  - Appendix 1 use of computer packages
    - SAS, SPSS, Minitab, S-PLUS
  - Exercises & Task Sheets are in Course Booklet
    - Solutions follow later as appropriate

- **Task Sheets & Exercises**
  - Task sheets:–
    - ~ each week
    - simple quick short exercises/reading
    - reinforce / consolidate lecture material
  - Exercises:–
    - 3 sets during semester in weeks 5,8,10
    - Work submitted within 2 weeks will be marked and returned
  - **See Study Guide**
    - **recommendations on time to spend**

Task Sheets & Exercises

- Task sheets:–
  - are designed for you to test **your own understanding** of the course material
  - *you are responsible* for your own learning on the course — task sheets help you in self-assessment
- Exercises:–
  - **Prime route for individual feedback**
  - Task sheets often provide guide for exercises
- Unacceptable reasons for not submitting anything
  - I did not have enough time
  - I knew I could do them so I did not need to submit
  - I could not do anything so did not think it was worth it

Solutions to Task Sheets & Exercises

- Exercises:–
  - Solutions available on web soon after submission
    - Printed solutions will be provided to those who submit
- Task sheets:–
  - are designed for you to test your own understanding of the course material
  - if necessary go back to lecture notes (etc) & re-read relevant sections
    - (and if necessary re-read again & … …)
  - Solutions will be provided on web pages in due course (for revision etc)
    - but ***deliberately*** these will not appear very quickly

- Course web page

  http://nickfieller.staff.shef.ac.uk/
  - Click on Teaching & then on MAS6012/MAS461/MAS361 Medical Statistics
- Lecture notes, task sheets, solutions & data sets available here after distribution in lectures
  - (I don't keep back copies)

Books

- **Campbell, M. J. (2001)**
  Statistics at Square Two. BMJ
- **Collett, D. (2003)**
  Modelling Survival Data in Medical Research. (2nd Edition) Chapman & Hall
- **Everitt, Brian & Rabe-Heskith, Sophia (2001)** *Analyzing Medical Data Using S-PLUS*. Springer. Support material at

http://web1.iop.kcl.ac.uk/IoP/Departments/BioComp/splusBook.shtml

Contents

- **Objectives**
  - statistical modelling & analysis of *lifetime data*.
  - Lifetime data arise especially in medical statistics and in reliability studies.
    - *survival time:-*
      - time from diagnosis to death of a patient
      - *time to recovery* or *remission* of a patient or
      - *time to failure* of an electronic component.

- **Outline**
  - Types of survival data
  - Censoring
  - Parametric & non-parametric approaches
- **Single Sample Methods**
  - survivor & hazard functions
  - Lifetables
  - Kaplan-Meier estimators

- **Two Sample Comparisons:-**
  - log rank test
  - maximum likelihood test
  - likelihood ratio test
  - proportional hazards

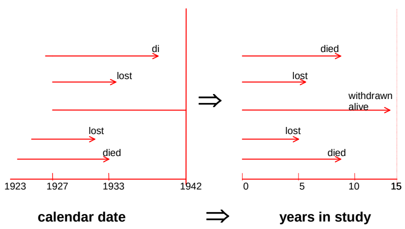
- **Regression Models**
  - parametric models
    - exponential & Weibull
  - non-parametric methods
    - proportional hazards or Cox regression
    - partial likelihood

- **Aims of survival analysis**
  - describe/model a single sample
    - making inferences on a single population
  - compare two or more groups
    - effect of treatments on survival time
  - investigate relationship with covariates
    - effects on survival time of covariates
    - adjust for covariates in comparisons

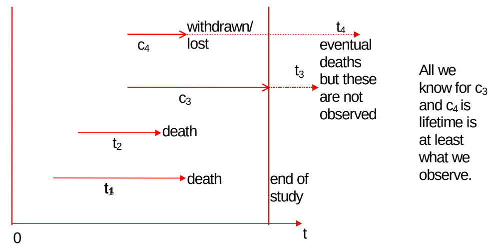
Contents

- **Censoring**
  - event of interest not yet occurred
  - observe only survival time is ***at least*** t
  - some observations t_i (i=1,…,n) are ***censored***
  - censored observations provide information on survival so cannot be ignored

Censoring



Censoring



Censoring

- ***Right censoring***
  - lifetime exceeds some value
- ***Left censoring***
  - lifetime less than some value
- ***Interval censoring***
  - failure occurred during an interval

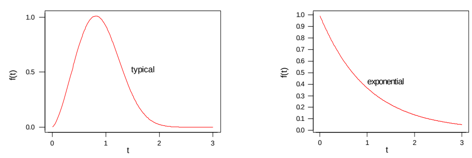
- **Aims**:
  - **estimate lifetime distributions**
    - ⇒ estimate properties of distribution
    - (median lifetime, prob of surviving > 5 years,…)
  - **Censoring**
    - non-parametric — lifetables, Kaplan-Meier
    - parametric — exponential, Weibull.

- survival time is a random variable T
- T>0, continuous variable
  - **p.d.f.** $f(t)$ (t>0)
  - **d.f.** $F(t)=P[T\le t]$
  - (so $f(t)=F'(t)$ and $F(t)=\int_0^t f(u)\,du$)

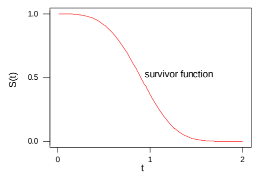
- **Survivor function**
  - $S(t)=P[T>t]=1-F(t)$
  - (so $S'(t) = -f(t)$, $S(t) = \int_t^\infty f(u)du$)
- **Hazard function**
  - $h(t)=\lim_{\delta t\to 0}\left[\dfrac{P[t\le T<t+\delta t\mid T\ge t]}{\delta t}\right]$
  - "P[die at time T given survived until T]"

- **Cumulative hazard function**
  - $H(t)=\int_0^t h(u)du=-\log_e S(t)$
  - f(t), S(t), H(t) and h(t) are equivalent characterizations of a survival distribution
  - all inter-related

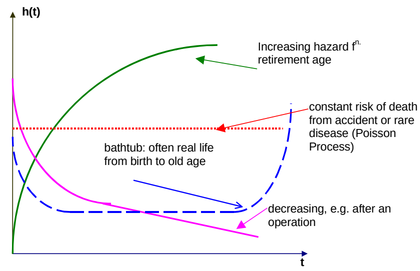
- **Typical patterns**



- **Survivor function**



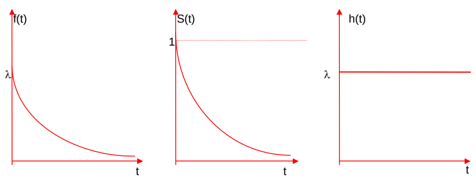
Hazard functions:



- Choose appropriate family of models by recognizing form of hazard function
  - practical situation
  - initial investigation of the data
- then can estimate parameters in model

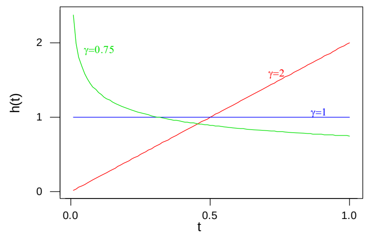
- **Example:-** exponential
  - $f(t) = \lambda e^{-\lambda t}$
  - $S(t) = e^{-\lambda t}$
  - $h(t) = \lambda$ (**NB** constant)
    - [= f(t)/S(t)]
- ***only*** constant for exponential survival distribution

- exponential



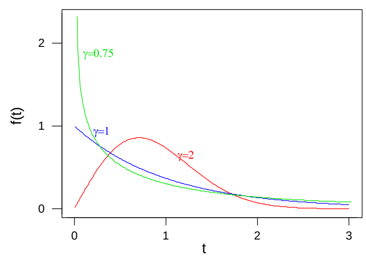
- **Example:-**Weibull
  - $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$
  - $S(t) = \exp(-\lambda t^{\gamma})$
  - $h(t) = \lambda\gamma t^{\gamma-1}$
    - $\gamma > 1$: increasing
    - $\gamma = 1$: constant (exponential)
    - $\gamma < 1$: decreasing

- e.g. $\lambda = 1$; Weibull hazard functions



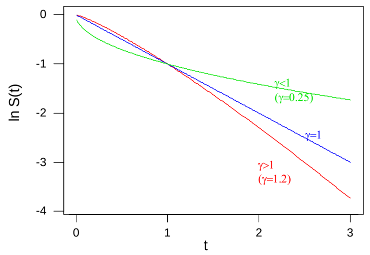
Weibull hazard functions

- e.g. $\lambda = 1$; Weibull density functions



Weibull Density Functions

- e.g. λ = 1; Weibull log-survivor functions



Weibull log survivor functions

- Weibull family
  - very flexible models
  - allows both increasing (γ>1) & decreasing (γ<1) hazard functions
- Difficult to estimate if γ close to 1
  - (numerical instability)

Contents

- **Lifetables**
  - 3 types :-
    - **Population** (from census or survey)
    - **Cohort** (follow a group throughout lifetimes)
    - **Clinical** (or follow-up)
      survival pattern of specific group of individuals.

- **Example**
  - every patient followed up after treatment either until death or up to the end of 1992
  - aim is to estimate probability of surviving for k years in separate steps:
    - first estimate probability of surviving another year given survived up to start of year, for each year
    - then multiply conditional probabilities together

Year of treatment	number treated	Number alive on each anniversary				
		1st	2nd	3rd	4th	5th
1987	62	58	51	46	45	42
1988	39	36	33	31	28	
1989	47	45	41	38	73	
1990	58	53	48	115		
1991	42	40	173			
	248	232				

- prob survive first year =232/248 =0.936

Year of treatment	number treated	Number alive on each anniversary				
		1st	2nd	3rd	4th	5th
1987	62	58	51	46	45	42
1988	39	36	33	31	28	
1989	47	45	41	38	73	
1990	58	53	48	115		
1991	42	40	173			
	248	**232**				

- prob survive second year, given alive at start 173/(232 – **40**) = 0.901

Year of treatment	number treated	Number alive on each anniversary				
		1st	2nd	3rd	4th	5th
1987	62	**58**	51	46	45	42
1988	39	**36**	33	31	28	
1989	47	**45**	41	38	73	
1990	58	**53**	48	115		
1991	42	**40**	**173**			
	248	**232**				

- prob survive third year, given alive at start 115/(173 – 48) = 0.920

Year of treatment	number treated	Number alive on each anniversary				
		1st	2nd	3rd	4th	5th
1987	62	58	**51**	46	45	42
1988	39	36	**33**	31	28	
1989	47	45	**41**	38	73	
1990	58	53	**48**	**115**		
1991	42	40	**173**			
	248	232				

Year after treatment	Prob. of surviving each year	Prob. of dying each year	Lifetable (per 1000) Number alive on each anniversary	Number dying during each year
x	p_x	q_x	l_x	d_x
0	0.936	0.064	1000	64
1	0.901	0.099	936	93
2	0.920	0.080	843	67
3	0.948	0.052	776	40
4	0.933	0.067	736	49
5			687	

0.901 × 936 = 843

- n_x — number alive at start of (x,x+1)
  d_x — number dying in (x,x+1)
  estimate of conditional probability of dying in (x,x+1), given alive at x, is
  $p_x=d_x/n_x$
- What if subjects are lost to follow-up?
  - (and not known if still alive or not)

- w_x— number lost to follow-up
  - includes those who have disappeared
    - (i.e. last report last year)
  - +'withdrawn alive'
- assume withdrawals are uniformly spread over (x,x+1)
- adjusted number at risk $n_x'=n_x–½w_x$

- adjusted number at risk $n_x'=n_x–½w_x$
- adjusted estimate of conditional probability of dying in (x,x+1), given alive at x, is $p_x=d_x/n_x'$
- Then $n_{x+1}=n_x– d_x– w_x$

Interval since operation years x to x+1	Last reported during this interval: Died d_x	withdrawn w_x	Living at start of interval n_x	Adjusted number at risk n_x'	Estimated probability of death q_x	Estimated probability of survival p_x	% of survivors after x years l_x	Estimate of p.d.f. $\hat{p}_{x+\frac{1}{2}}$	Estimate of hazard function $\hat{h}_{x+\frac{1}{2}}$
0 – 1	90	0	374	374.0	0.2406	0.7594	100	0.241	0.274
1 – 2	76	0	284	284.0	0.2676	0.7324	75.9	0.203	0.309
2 – 3	51	0	208	208.0	0.2452	0.7548	55.6	0.136	0.279
3 – 4	25	12	157	151.0	0.1656	0.8344	42.0	0.070	0.181
4 – 5	20	5	120	117.5	0.1702	0.8298	35.0	0.059	0.186
5 – 6	7	9	95	90.5	0.0773	0.9227	29.1	0.023	0.080
6 – 7	4	9	79	74.5	0.0537	0.9463	26.8	0.014	0.055
7 – 8	1	3	66	64.5	0.0155	0.9845	25.4	0004	0.016
8 – 9	3	5	62	59.5	0.0504	0.9496	25.0	0.013	0.052
9 – 10	2	5	54	51.5	0.0388	0.9612	23.7	0.009	0.040
10 –	21	26	47	—	—	—	22.8		

120 = 157 – 25 – 12

Interval since operation years x to x+1	Last reported during this interval Died d_x	withdrawn w_x	Living at start of interval n_x	Adjusted number at risk n_x'	Estimated probability of death q_x	Estimated probability of survival p_x	% of survivors after x years l_x	Estimate of p.d.f. $\hat{p}_{x+\frac{1}{2}}$	Estimate of hazard function $\hat{h}_{x+\frac{1}{2}}$
0 – 1	90	0	374	374.0	0.2406	0.7594	100	0.241	0.274
1 – 2	76	0	284	284.0	0.2676	0.7324	75.9	0.203	0.309
2 – 3	51	0	208	208.0	0.2452	0.7548	55.6	0.136	0.279
3 – 4	25	12	157	151.0	0.1656	0.8344	42.0	0.070	0.181
4 – 5	20	5	120	117.5	0.1702	0.8298	35.0	0.059	0.186
5 – 6	7	9	95	90.5	0.0773	0.9227	29.1	0.023	0.080
6 – 7	4	9	79	74.5	0.0537	0.9463	26.8	0.014	0.055
7 – 8	1	3	66	64.5	0.0155	0.9845	25.4	0004	0.016
8 – 9	3	5	62	59.5	0.0504	0.9496	25.0	0.013	0.052
9 – 10	2	5	54	51.5	0.0388	0.9612	23.7	0.009	0.040
10 –	21	26	47	—	—	—	22.8		

120 = 157 – 25 – 12

117.5 = 120 – ½ x 5

- Estimated survivor function is

  $\hat{S}_x = p_0 p_1 \ldots p_{x-1}$

- estimate of pdf is

  $\hat{f}_{x+\frac{1}{2}} := \hat{S}_x - \hat{S}_{x+1} = \hat{S}_x q_x$

- Estimate of hazard is $\hat{h}_{x+\frac{1}{2}} := \dfrac{2q_x}{1+p_x}$

- Estimate of hazard is

$$\hat{h}_{x+\frac{1}{2}} = \frac{\hat{f}_{x+\frac{1}{2}}}{\hat{S}_{x+\frac{1}{2}}} = \frac{\hat{S}_x q_x}{\frac{1}{2}(\hat{S}_x + \hat{S}_{x+1})} = \frac{2q_x}{1+p_x}$$

noting $\hat{S}_{x+1} = p_x \hat{S}_x$

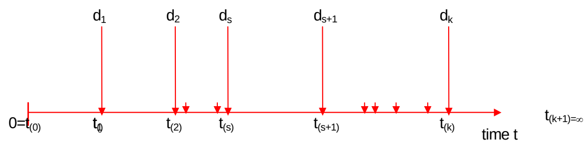
- assumed
  - withdrawals have same probability of death as non-withdrawals.
    - Is loss to follow-up connected with condition?
  - p_x and q_x constant over study
    - (estimated by combining data from several years)
- estimates are subject to sampling error:

$$\text{Var}(\hat{S}_x) = \hat{S}_x^2 \sum_{j=1}^{x-1} \frac{d_j}{n_j(n_j - d_j)}$$

- **Clinical life tables suggest the form of the hazard function**

- Lifetable methods take data in groups. Information is lost if actual lifetimes (perhaps censored) are available & have been grouped

- **Kaplan–Meier estimates**
  - k ordered distinct lifetimes $t_{(1)} < t_{(2)} < .. < t_{(k)}$
  - d_i — number of deaths at $t_{(i)}$ (so $\Sigma d_i = n$)

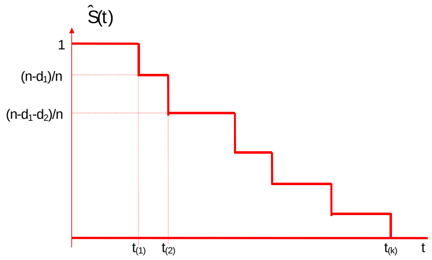


- $\hat{F}(t)$ = proportion of lifetimes < t

  $= \frac{1}{n}\sum_{j=1}^{s} d_j$ for $t_{(s)} \leq t < t_{(s+1)}$

- so $\hat{S}(t) = 1 - \hat{F}(t) = \dfrac{n - \sum_1^s d_j}{n}$

  for $t_{(s)} \leq t < t_{(s+1)}$



- Let r_j be the number at risk (≡number alive) just before $t_{(j)}$,

  Then $r_{j+1}=r_j-d_j$, so

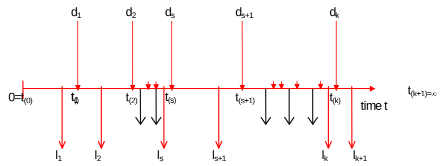
$$\hat{S}(t) = \frac{n-d_1}{n}\cdot\frac{n-d_1-d_2}{n-d_1}\cdot\frac{n-d_1-d_2-d_3}{n-d_1-d_2}\cdots\frac{n-d_1-d_2-\ldots-d_s}{n-d_1-\ldots-d_{s-1}}$$

$$= \left(1-\frac{d_1}{r_1}\right)\left(1-\frac{d_2}{r_2}\right)\cdots\left(1-\frac{d_s}{r_s}\right)$$

$$= \prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \text{ for } t_{(s)} \leq t < t_{(s+1)}$$

- What if some observations censored?
  - adjust numbers at risk r_j by allowing for censoring

- l_1 l_2 l_3 .... l_k numbers censored before time $t_{(j)}$



l_j = # censored in the previous interval

now $r_1=n-l_1$; $r_{j+1}=r_j-d_j-l_{j+1}$ for j=1,2,...,k-1

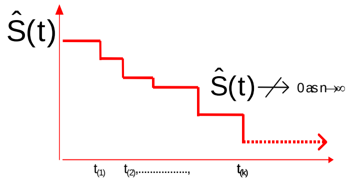
[or $r_j=n-(d_1+d_2+\ldots+d_{j-1})-(l_1+l_2+\ldots+l_j)$ for j≥2]

⇨ Kaplan-Meier product limit

$$\hat{S}(t) = \prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \text{ for } t_{(s)} \leq t < t_{(s+1)}$$

- **Notes**
  - assumes that the l_j censorings survive up to just after $t_{(j-1)}$ and then are removed
  - uncensored case is just a special case with l_j=0 all j
  - If l_{k+1}>0 then, since r_k>d_k

$$\hat{S}(t) = \prod_{j=1}^{k}\left(1 - \frac{d_j}{r_j}\right) > 0 \quad \text{when } t>t_{(k)}$$



But we know that $S(\infty) = 0$

⇨ Kaplan-Meir estimates are ***biased*** if maximum observation is censored

- $\hat{S}(t)$ is subject to sampling error

$$\mathrm{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j=1}^{s} \frac{d_j}{r_j(r_j - d_j)} \quad \text{for } t_{(s)} \le t < t_{(s+1)}$$

- so can get confidence bands for S(t) with +/- 2 x st.error

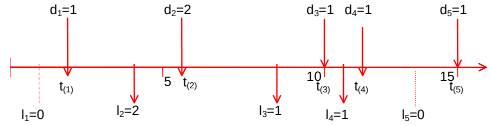
- estimate cumulative hazard H(t) by

$$\hat{H}(t) = -\log_e(\hat{S}(t))$$

or simpler is to use

$$\tilde{H}(t) = \sum_{j=1}^{s} \frac{d_j}{r_j} \quad \text{for } t_{(s)} \le t < t_{(s+1)}$$

- **Example**
  - Remission times for 10 patients
  - 6 relapse 3.0, 6.5,6.5,10,12,15 months
  - **1** lost to follow-up at 8.4 months
  - **3** still in remission at end after 4.0, 5.7, 10.1 months
- **i.e. 4 censored observations**



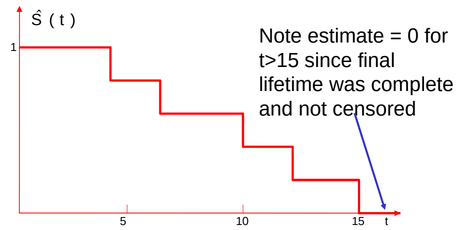
j	$t_{(j)}$	l_j	r_j	d_j	$\hat{S}(t)$		notes
					1	$0\leq t<3.0$	
1	3.0	0	10	1	0.9	$3.0\leq t<6.5$	9/10
2	6.5	2	7	2	0.643	$6.5\leq t<10.0$	9/10x5/7
3	10.0	1	4	1	0.482	$10.0\leq t<12.0$	9/10x5/7x3/4
4	12.0	1	2	1	0.241	$12.0\leq t<15.0$	9/10x5/7x3/4x1/2
5	15.0	0	1	1	0	$15\leq t$	

j	$t_{(j)}$	l_j	r_j	d_j	$\hat{S}(t)$		notes
					1	$0\leq t<3.0$	
1	3.0	0	10	1	0.9	$3.0\leq t<6.5$	9/10
2	6.5	2	7	2	0.643	$6.5\leq t<10.0$	9/10x5/7
3	10.0	1	4	1	0.482	$10.0\leq t<12.0$	9/10x5/7x3/4
4	12.0	1	2	1	0.241	$12.0\leq t<15.0$	9/10x5/7x3/4x1/2
5	15.0	0	1	1	0	$15\leq t$	

7 = 10 – 1 –2

j	$t_{(j)}$	l_j	r_j	d_j	$\hat{S}(t)$		notes
					1	$0\leq t<3.0$	
1	3.0	0	10	1	0.9	$3.0\leq t<6.5$	9/10
2	6.5	2	7	2	0.643	$6.5\leq t<10.0$	9/10x5/7
3	10.0	1	4	1	0.482	$10.0\leq t<12.0$	9/10x5/7x3/4
4	12.0	1	2	1	0.241	$12.0\leq t<15.0$	9/10x5/7x3/4x1/2
5	15.0	0	1	1	0	$15\leq t$	

4 = 7 – 2 –1



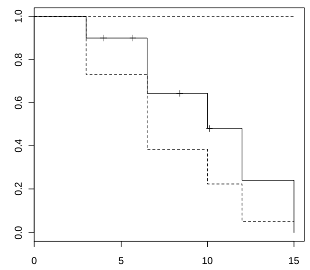
- Implementation in **R**:
  - Need to load library `survival`
    - `> library(survival)`
  - Need to create 'survival object' with `Surv()` `Surv(time,censor,type='right')`
  - t hen use `survfit()` to estimate a survival curve
  - and `plot()` and `summary()` to see details


```
> library(survival)
Loading required package: splines
> load("tumour.Rdata")
> attach(tumour)
> tumour
   time censor
1   3.0      1
2   4.0      0
3   5.7      0
4   6.5      1
5   6.5      1
6   8.4      0
7  10.0      1
8  10.1      0
9  12.0      1
10 15.0      1
```

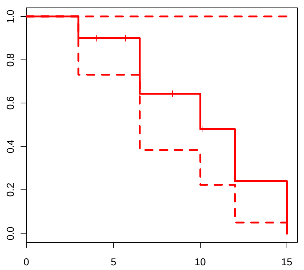
```
# create survival object in tumour.sv:
> tumour.sv <—Surv(time, censor, type = "right")
# estimate survival curve in tumourSurv:
> tumourSurv <-survfit(tumour.sv ~1, data=tumour)
# note ‘regressing’/’relating’ survival object on a
  constant with ~ 1
# look at summary calculations
> summary(tumourSurv)
Call: survfit(formula = tumour.sv, data = tumour)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3.0     10       1    0.900  0.0949       0.7320            1
  6.5      7       2    0.643  0.1679       0.3852            1
 10.0      4       1    0.482  0.1877       0.2248            1
 12.0      2       1    0.241  0.1946       0.0496            1
 15.0      1       1    0.000     NaN           NA           NA
# Note confidence interval for the K-M estimates
  and that in this small data set the interval is
  truncated at 1
```



```
# plot curve:
> plot(tumourSurv)
```



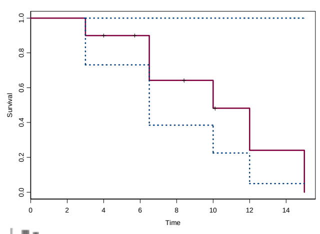

```
# plot curve:
> plot(tumourSurv,lwd=3,col=2)
```



- S-plus implementation:
  - Statistics>Survival>Nonparametric Survival…
  - Need to create formula in dialogue box such as `Surv(time,censor,type='right')~1`

S-plus output:

```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
 3.0     10       1    0.900  0.0949       0.7320            1
 6.5      7       2    0.643  0.1679       0.3852            1
10.0      4       1    0.482  0.1877       0.2248            1
12.0      2       1    0.241  0.1946       0.0496            1
15.0      1       1    0.000      NA           NA           NA
```



Note confidence intervals by default

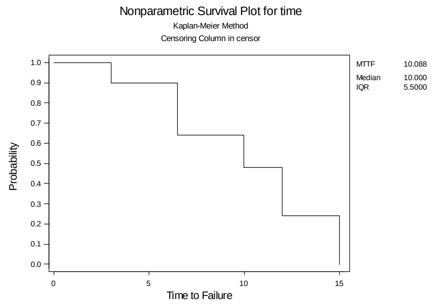
- Minitab implementation:
  - Stat>Reliability/Survival>Nonparametric Dist Analysis-Right Censoring
  - Need to specify value indicating **censored** observations

Minitab output:

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
3.0000	10	1	0.9000	0.0949	0.7141	1.0000
6.5000	7	2	0.6429	0.1679	0.3137	0.9720
10.0000	4	1	0.4821	0.1877	0.1142	0.8501
12.0000	2	1	0.2411	0.1946	0.0000	0.6225
15.0000	1	1	0.0000	0.0000	0.0000	0.0000

Minitab graph:



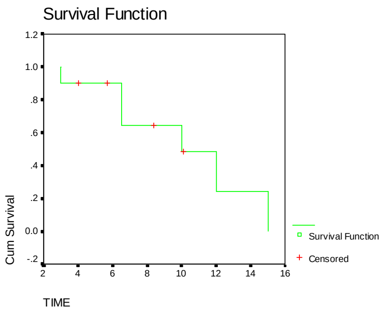
- SPSS implementation:
  - Analyze>Survival>Kaplan-Meier
  - Need to specify value indicating **uncensored** values
  - Note indication of censored values
    - Also indicated by S-PLUS

SPSS Output:

Survival Analysis for TIME

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
3	0	.9000	.0949	1	9
4	1			1	8
6	1			1	7
7	0			2	6
7	0	.6429	.1679	3	5
8	1			3	4
10	0	.4821	.1877	4	3
10	1			4	2
12	0	.2411	.1946	5	1
15	0	.0000	.0000	6	0

SPSS Graph



- **Summary**
  - Life tables
    - Grouped data
    - Allow for censoring by adjusting # at risk
  - Kaplan-Meier
    - Individual data
    - Uncensored case = 1 – empirical CDF
    - Express as a product of terms [$1 - d_i/r_i$]
    - Censored case by adjusting # at risk

- Kaplan-Meier estimates are the key exploratory tool for censored data
  - Cannot draw histograms of censored data
    - Don't know which bin to put a censored in
- Kaplan-Meier Plot is always step 1 in analysing censored data