Survival Data Analysis

Course Booklet

Dr Nick Fieller

Probability & Statistics, SoMaS

University of Sheffield

visiting

UNIVERSITY
OF TAMPERE

2012

(this page left blank for notes)

Contents

Analysis of Survival Data

0. Introduction

0.1 Books

★Altman, D.G. (1991) *Practical Statistics for Medical Research.* Chapman and Hall.

★Campbell, M. J. (2001) *Statistics at Square Two.* BMJ

★Collett, D. (2003) *Modelling Survival Data in Medical Research (2nd Ed.).* Chapman and Hall.

Cox, D.R. & Oakes, D. *Analysis of Survival Data*. Chapman and Hall.

Crowder, M.J., Kimber, A.C., Sweeting, T.J., & Smith, R.L. (1991) *Statistical Analysis of Reliability Data*. Chapman and Hall.

Everitt, Brian & Rabe-Heskith, Sophia (2001) *Analyzing Medical Data Using S-PLUS*. Springer.

Gross, A.J. & Clark, Y.A. (1975) *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley.

Kalbfleisch, J.D. & Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* Wiley.

Lee, E.T. (1980) *Statistical Methods for Survival Data Analysis.* Wadsworth.

Marubini, E. and Valsecchi, M. G. (1995) *Analysing Survival Data from Clinical Trials and Observational Studies.* Wiley.

Miller, R. Jr. (1984) *Survival Analysis*. Wiley.

Swinscow, T. D. V. (1996) *Statistics at Square One (9th Ed.).* BMJ

★ Indicates texts at the appropriate level for this course

0.2 Objectives

The objective of this course is to provide an introduction to the statistical modelling and analysis of *lifetime data*. Lifetime data arise especially in medical statistics as well as in studies of reliability. A lifetime might refer to a *survival time*, (i.e. time to death of a patient from diagnosis) or *time to recovery* or *remission* of a patient or *time to failure* of an electronic component.

0.3 Organization of course material

The notes in the main Chapters 1– 4 are largely covered in the two highlighted books in the list of recommended texts above and are supplemented by various examples and illustrations. These range from simple 'quick problems' to more substantial exercises. These task sheets are designed for you to test your own understanding of the course material. If you are not able to complete the simpler problems then you should go back to the lecture notes (and other course material) and re-read the relevant section (and if necessary re-read again & …). Solutions will be provided to these on the course web pages in due course.

Lectures will consist of introducing the material covered in these notes, filling in details of items such as **R** implementation (including specific commands and menu choices), demonstrating computer analyses and going through key parts of the various example sheets. The lectures will be based on PowerPoint presentations and copies of the slides will be made available on the course webpage at:

http://www.nickfieller.staff.shef.ac.uk/tampere12/index.html

These will be placed there some time after the lecture.This page is also where to look for copies of exercises and solutions. Any typing (or other) mistakes in the notes or the exercises and solutions that are brought to my attention will be noted and corrected in a ***Corrections and Clarifications*** section on this page.

0.4 A Note on R, S-PLUS and MINITAB

The main statistical package for this course is **R.** It is very similar to the copyright package S-PLUS and the command line commands of S-PLUS are [almost] interchangeable with those of **R**. Unlike S-PLUS, **R** has only a very limited menu system which covers some operational aspect but no statistical analyses. A brief guide to getting started in **R** is available from the course homepage.

R is a freely available programme which can be downloaded over the web from http://cran.r-project.org/ or any of the mirror sites linked from there for installation on your own machine. It is available on University networks. **R** and S-PLUS are almost identical except that **R** can only be operated from the command line apart from operational aspects such as loading libraries and opening files. Almost all commands and functions used in one package will work in the other. However, there are some differences between them. In particular, there are some options and parameters available in **R** functions which are not available in S-PLUS. Both S-PLUS and **R** have excellent help systems and a quick check with help(*function*) will pinpoint any differences that are causing difficulties. A key advantage of **R** over S-PLUS is the large number of libraries contributed by users to perform many sophisticated analyses. These are updated very frequently and extend the capabilities substantially. If you are considering using multivariate techniques outside this course (e.g. for some other substantial project) then you would be well advised to use **R** in preference to S-PLUS. Command-line code for he more substantial analyses given in the notes for this course have been tested in **R**. In general, they will work in S-PLUS as well but there could be some minor difficulties which are easily resolved using the help system.

MINITAB is package with a very flexible menu system and a full command-line facility. Some examples of MINITAB code and output are given in the notes since some of those taking the course are already familiar with the package.

Some participants may be familiar with SAS. An alumnus of the course (Kriss Harris) has provided translations into SAS of some of the numerical **R** examples discussed in this course. These can be found at http://krissharris.co.uk/survival/ .

0.5 Data sets

Data sets used in this course are available in a variety of formats on the course web pages.

0.5.1 R data sets

Those in **R** are given first and they have extensions **.Rdata**; to use them it is necessary to copy them to your own hard disk. This is done by using a web browser to navigate to the course web, clicking with the right-hand button and selecting ‘save target as…’ or similar which opens a dialog box for you to specify which folder to save them to. Keeping the default **.Rdata** extension is recommended and then if you use Windows explorer to locate the file a double click on it will open **R** with the data set loaded and it will change the working directory to the folder where the file is located. For convenience all the **R** data sets for Medical Statistics are also given in a WinZip file.

NOTE: It is not possible to use a web browser to locate the data set on a web server and then open R by double clicking. The reason is that you only have read access rights to the web page and since **R** changes the working directory to the folder containing the data set write access is required.

0.5.2 Data sets in other formats

Most of the data sets are available in other formats (Minitab, SPSS etc). It is recommended that the files be downloaded to your own hard disk before loading them into any package but in most cases it is possible to open them in the package *in situ* by double clicking on them in a web browser. However, this is not possible with **R.**

0.6 R libraries required

Most of the statistical analyses described in this book use functions within the `base` and `stats` packages and the `MASS` package. It is recommended that each **R** session should start with

```
library(MASS)
```

The `MASS` library is installed with the base system of **R** and the `stats` package is automatically loaded.

0.5 Outline of Course

1. Introduction:– types of survival data, censoring, outline of parametric and non-parametric approaches.
2. Single sample methods:– Basic concepts of survivor & hazard functions. Lifetables, (population, cohort and clinical). Kaplan-Meier product limit estimator of the survivor function, including censored data. Parametric models (exponential, Weibull and log-Normal).
3. Two Sample Comparisons:– Log rank test, parametric tests (maximum likelihood and likelihood ratio tests), proportional hazards.
4. Regression Models:– exponential regression, covariates and prognostic factors, exponential and Weibull models. Proportional Hazards Model, outline of estimation procedures by partial likelihood.

1 Background and Basic Concepts

1.1 Preliminary Discussion

The objective of a survival data analysis may be just to describe (and model) a single sample of data to describe the lifetimes of a single population or it may be to compare the lifetimes of two or more groups of subjects; for example the two groups may have received different medical treatments and the lengths of survival time measure how effective the treatments are.

A distinctive feature of survival data is that some observations may be ***censored***: often the event of interest (e.g. death, of patient, failure of component, recovery of patient) has not occurred by the time of recording so that all is known is that the lifetime for that subject is *at least* some value (and may well be greater than this value). Such censoring cannot be ignored (i.e. the censored observations cannot just be omitted) since they carry important information about the effectiveness of the treatment (and indeed one hopes that many patients are alive at the end of a medical study!). This introduces a complication in the statistical description and analysis of the data.

1.1.1 Example: Survival of angina pectoris

Data on the survival times of patients with angina pectoris are given by Gehan (1969: J.Chronic Disease). These patients form part of a large group of patients examined at the Mayo Clinic during the 15 year period January 1, 1927 — December 31, 1941.

Survival time (years)	Number of patients known to survive at beginning of interval	Number of patients lost to follow up
0 — 1	2418	0
1 — 2	1962	39
2 — 3	1697	22
3 — 4	1523	23
4 — 5	1329	24
5 — 6	1170	107
6 — 7	938	133
7 — 8	722	102
8 — 9	546	68
9 — 10	427	64
10 — 11	321	45
11 — 12	233	53
12 — 13	146	33
13 — 14	95	27
14 — 15	59	23
15 — 16	30	

This example illustrates: Follow-up study

Grouping

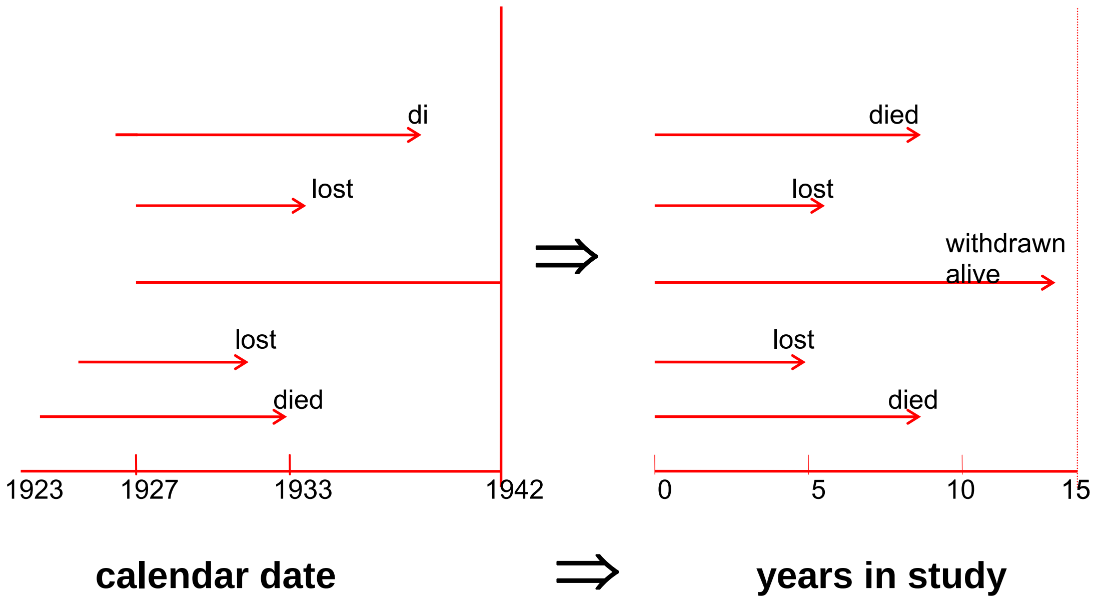
‘Lost’ patients (censoring)

Change in time axis (which is hidden)

— measure from entry into study.

We might also consider (i) several possible causes of death

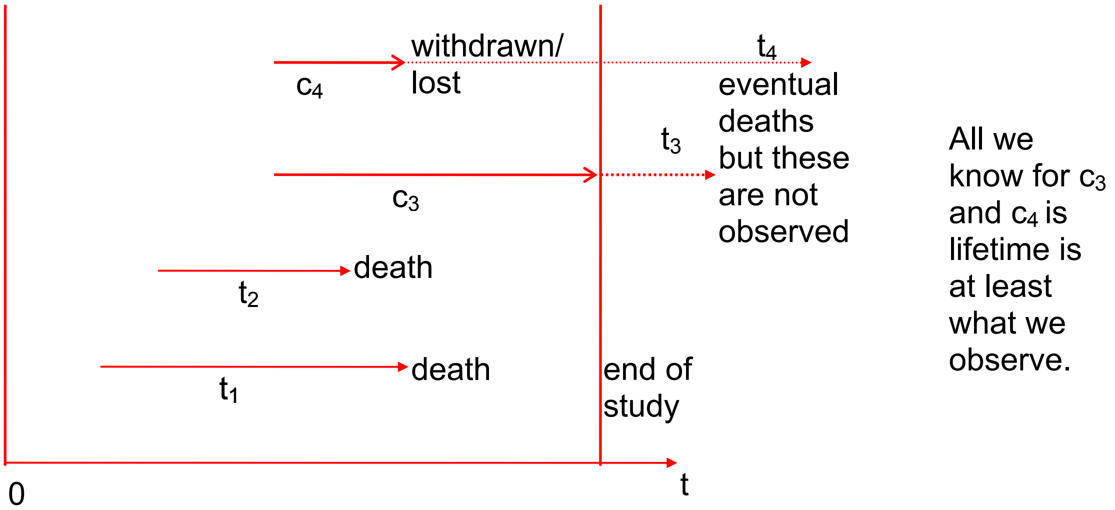
(ii) use of covariates which influence lifetime distributions, e.g. age, sex



1.2 Censoring

This is the complicating feature which identifies the need for a different type of analyses. Sometimes we do not observe the exact lifetime but only know that it exceeds some value (***right censoring***), or even, rarely, only that it is less than some value (***left censoring)***. There are several common censoring schemes.

For example, 'Time' or 'Progressive' Censoring (as in Example 1)



i.e. individuals are subjected to limited periods $c_1,c_2,....,c_n$ of observation and the ith individual lifetime t_i is observed only if $T_i \leq c_i$.

1.2.1 Notes

(i) **Type I censoring** if identical starting points and subjects are observed for a fixed time c_i (then typically c_1=c_2=, ...,=c_n). The number of censorings is then random.

(ii) **Type II censoring** assumes n patients in study at the start and the trial finishes after r deaths (do not specify the end of the trial initially — carry on until r out of n patients are dead, i.e. record $t_{(1)}$,$t_{(2)}$,...,$t_{(r)}$ up to a time $t_{(r)}$ when r have died). This type of censoring is less common in medical studies but is widely used in electronic component testing and reliability studies. The numbering of censorings is not random in Type II censoring but is fixed in advance.

(iii) ***Left censoring*** might occur if subjects are only observed at fixed appointments, and only then is it discovered that death occurred sometime before then, so survival time is *less* than the period of observation. Another example is when the 'lifetime observed' is the time to recurrence of a tumour observable only during surgery.

(iv) ***Interval censoring*** occurs when failure is only known to have occurred during an interval.

(This course will consider only *right censoring* in detail)

1.3 Approaches

Aims:

- estimate lifetime distributions
- predict survival times:–
    - non-parametric — lifetables, Kaplan-Meier
    - parametric — exponential, Weibull.

1.4 Summary

- **Censoring:** lifetime not observed exactly
    - *Right Censoring:* actual lifetime exceeds observation
    - *Left Censoring:* actual lifetime less than observation
    - *Interval Censoring:* observation gives upper and lower bound on lifetime
    - *Type I Censoring:* observation time fixed, # censorings random
    - *Type II Censoring:* # censorings fixed observation time random

This course considers primarily Type 1 right censoring and assumes generally that censoring is ***independent*** of lifetime.

2 Single Sample Models

2.1 Basic concepts

The random variable T measures survival time;

$T>0$, a continuous variable.

The actual survival time, t, of an individual is the value of the variable T.

T has **p.d.f.** (probability density function) $f(t)$ $(t>0)$

and **d.f.** (distribution function) $F(t) = P[T \leq t]$

(so $f(t) = F'(t)$ and $F(t) = \int_0^t f(u)du$)

2.1.1 Survivor function

$$S(t) = P[T \geq t] = 1 - F(t).$$

(so $S'(t) = -f(t)$, $S(t) = \int_t^\infty f(u)du$)

2.1.2 Hazard function

We often model the lifetime through the hazard function, h(t), which measures the 'risk' or 'proneness' to death at time t, given survival up to time t. It is the probability that an individual dies at time t, conditional on (s)he having survived to that time. The hazard function represents the instantaneous death rate for an individual surviving to time t.

$$h(t) = \lim_{\delta t \to 0} \left[\frac{P[t \le T < t + \delta t \mid T \ge t]}{\delta t} \right]$$

(This is also known as the **hazard rate** or **failure rate**.)

2.1.2.1 Interpretation:

Suppose time units are days, Take $\delta t=1$ (small in relation to times considered), $h(t)=P[t \le T<t+1]$, the probability of dying on day t.

2.1.3 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 ways of defining a specific survival pattern uniquely and they are all inter-related.

Clearly, $h(t) = \lim_{\delta t \to 0}\left[\frac{P[t \leq T < t + \delta t, T \geq t]}{P[T \geq t]\delta t}\right]$

$$= \lim_{\delta t \to 0}\left[\frac{P[t \leq T < t + \delta t}{P[T \geq t]\delta t}\right] = \lim_{\delta t \to 0}\left[\frac{F(t + \delta t) - F(t)}{S(t)\delta t}\right]$$

$$= \lim_{\delta t \to 0}\left[\frac{f(t)\delta t}{S(t)\delta t}\right]$$

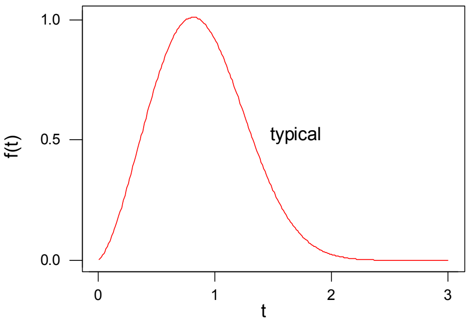
$$= \frac{f(t)}{S(t)}$$

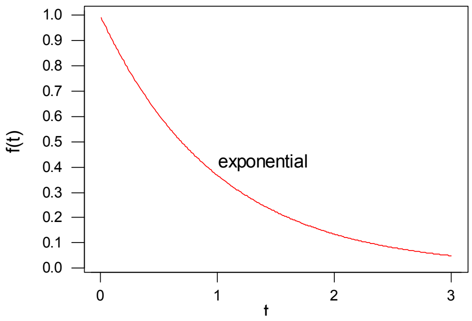
Also f(t) = F′(t) = –S′(t), so h(t) = –S′(t)/S(t) = $-{}^{d\log S(t)}\!/_{dt}$

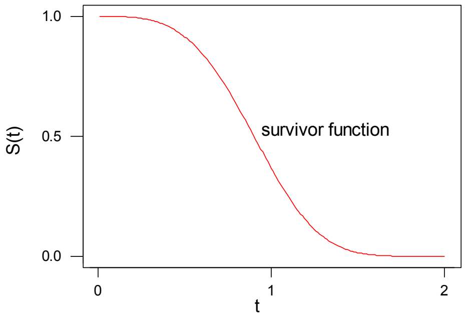
Thus $S(t) = \exp\{-\int_0^t h(u)du\} = \exp\{-H(t)\}$

The survivor function and hazard function are estimated from the observed survival times.

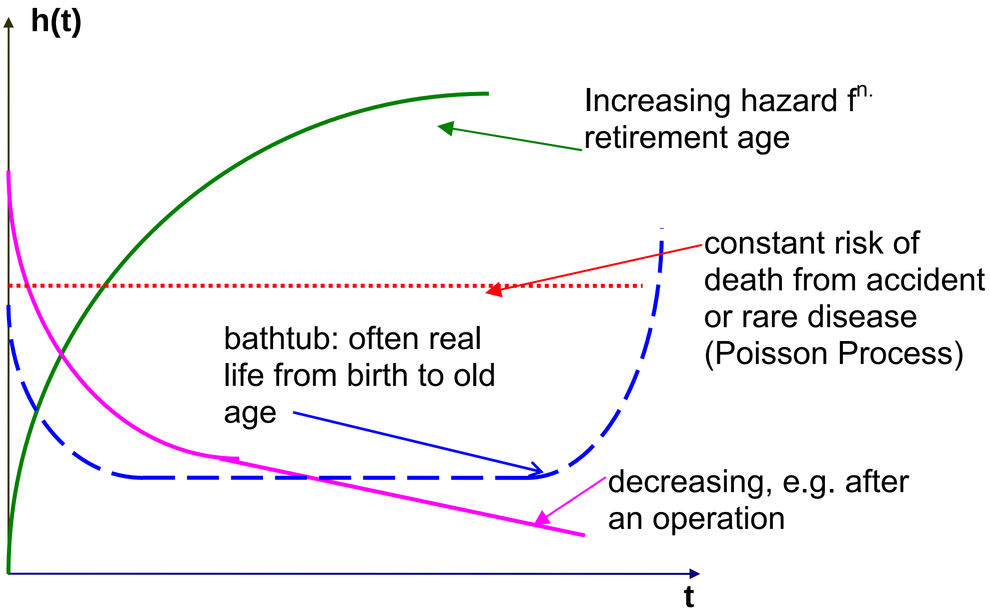
2.2 Typical patterns







Hazard functions:



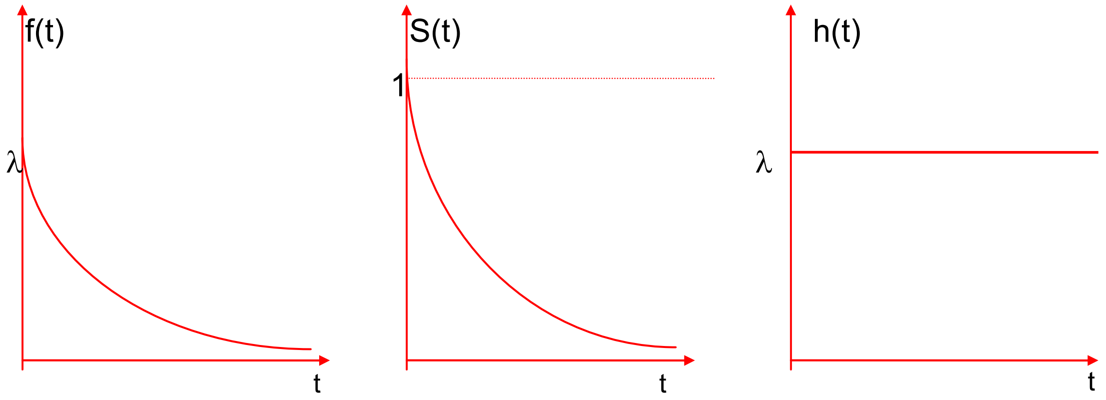
Often in a practical situation we can guess at the form of the hazard function and so recognise an appropriate family of models to try to estimate. Sometimes we can determine the general form of the hazard function from an initial investigation of the data.

2.2.2 Example : exponential

Exponential $f(t) = \lambda e^{-\lambda t}$

$S(t) = e^{-\lambda t}$

$h(t) = \lambda$



The exponential survival distribution is the *only* one with a constant hazard function.

2.2.3 Example: Weibull

$$\text{Weibull: } h(t) = \lambda\gamma t^{\gamma-1}$$

$\gamma > 1$: increasing

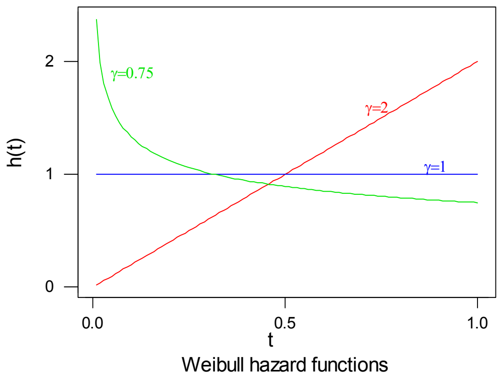
$\gamma = 1$: constant (exponential)

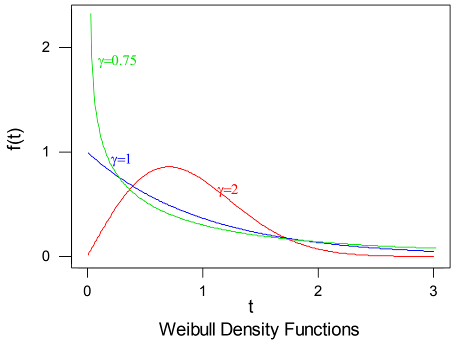
$\gamma < 1$: decreasing

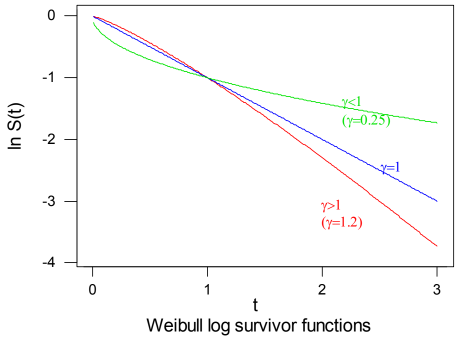
$\therefore \quad S(t) = \exp(-\lambda t^{\gamma})$

$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$

e.g. $\lambda = 1$







Alternative parameterization:

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \quad \text{i.e. } \lambda^{\gamma}\gamma t^{\gamma-1}$$

The Weibull distribution provides a very flexible family of survival distributions with both increasing (γ>1) and decreasing (γ<1) hazard functions. It can be difficult to estimate, particularly if γ is close to 1.

2.3 Lifetables

Before trying to fit a formal statistical, model an initial non-parametric investigation is sensible — often it provides sufficient information for the study and it will always give useful information to help in selecting a suitable family of distributions.

A *lifetable* is a way of expressing or tabulating the death rates experienced by some particular population during a particular period of time.

There are 3 types of lifetable:—

(i) Population (or current)

Obtained from a census or survey. It gives the survival pattern of a group of individuals subject to the age–specific death rates currently observed in the population. It is an artificial population — it gives the pattern of mortality or what would happen if individuals were subjected throughout their lifetime to the present death rates.

(ii) Cohort

Follow a group of individuals throughout their lifetimes.

(iii) Clinical (or follow-up)

Of more relevance in clinical studies — survival pattern of a specific group of individuals. Source of data is usually from a follow-up study.

2.3.1 Example: no censoring

In this study every patient has been followed up after treatment, either until death or up to the end of 1992.

Survival after treatment:

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				

			Lifetable (per 1000)	
Year after treatment	Prob. of surviving each year	Prob. of dying each year	Number alive on each anniversary	Number dying during each year
x	$\mathbf{p_x}$	$\mathbf{q_x}$	$\mathbf{l_x}$	$\mathbf{d_x}$
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	

Censoring: ‘withdrawn alive’ at the end of the study

Notes: 0.936 = 232/248

0.901 = 173/(232–40)

0.920 = 115/(173–48) etc.

The p_x are calculated from the numbers surviving from one year to the next and strictly are [estimates of] conditional probabilities of surviving for that year, conditional on surviving up until the start of the year.

2.3.2 Example: lost to follow up

Complications begin to arise when patients are lost to follow-up — and we do not know if they have died or not ⇒ considered as ‘withdrawn’. (From Armitage, 1971)

Interval since operation years	Last reported during this interval		Living at start of interval	Adjusted number at risk	Estimated probability of death	Estimated probability of survival	% of survivors after x years	Estimate of p.d.f.	Estimate of hazard function
	Died	withdrawn							
x to x+1	d_x	w_x	n_x	n_x’	q_x	p_x	l_x	$\hat{f}_{x+\frac{1}{2}}$	$\hat{h}_{x+\frac{1}{2}}$
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		

d_x: number died during (x, x+1)

w_x: includes those who have disappeared (i.e. last report last year) +'withdrawn alive'

n_x: number living at start of interval (x,x+1) → accumulate d_x+w_x from bottom.

n_x': assume withdrawals are uniformly spread over interval

$\Rightarrow n_x'= n_x-\tfrac{1}{2}w_x$

q_x: conditional probability of dying in (x,x+1), $q_x=d_x/n_x'$

p_x: $=1-q_x$

l_x: life survival rates, $l_0=100$; $l_x=l_0p_0p_1.....p_{x-1}$

$$l_x=100\hat{S}_x \quad ; x=0,1,2,...$$

$$\hat{f}_{x+\frac{1}{2}} := \hat{S}_x - \hat{S}_{x+1} = \hat{S}_x q_x \quad x=0,1,2,..$$

$$\hat{h}_{x+\frac{1}{2}} := \frac{2q_x}{1+p_x}$$

2.3.2.1 Notes:

(i) We have assumed that withdrawals are subjected to the same probability of death as non-withdrawals. This is reasonable if they are 'withdrawn alive' but possibly it is not for 'lost to follow up' (since the reason they may be 'lost' could also affect their chance of dying).

(ii) We have assumed that p_x and q_x remain constant over the study. This is a relatively short period.

(iii) Difficult to calculate expectations of life with censored data and we begin to think in terms of a parametric model.

$$\hat{S}_x = {}^{l_x}\!/_{100}; \quad \hat{f}_{x+\frac{1}{2}} = \hat{S}_x - \hat{S}_{x+1}; \quad \hat{h}_{x+\frac{1}{2}} = \frac{\hat{f}_{x+\frac{1}{2}}}{P[T \geq x+\frac{1}{2}]} = \frac{\hat{f}_{x+\frac{1}{2}}}{\frac{1}{2}(\hat{S}_x + \hat{S}_{x+1})}$$

Clinical life tables can help us decide
what the hazard function might look like

Note: These estimates are subject to sampling error:
Greenwood (1926) showed that approximately

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

2.4 Kaplan–Meier product limit estimate of S(t)

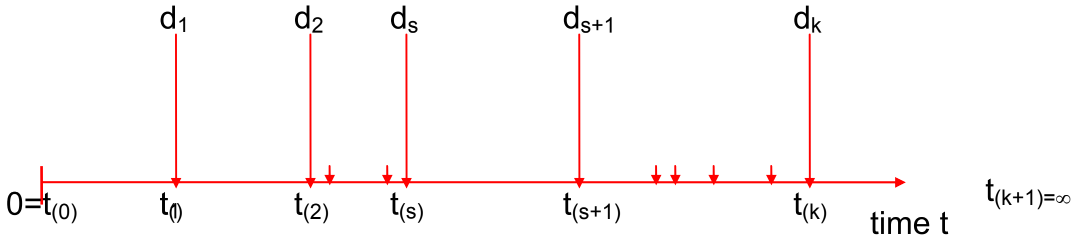
The lifetable methods all consider the data in groups. If the actual lifetimes (perhaps censored) are available, grouping will lose information.

2.4.1 Simple Case, no censoring:

n observations of lifetimes at $t_1, t_2, \ldots, t_n$,

⇨ order: $t_{(1)} < t_{(2)} < .. < t_{(k)}$ (assuming k distinct lifetimes)

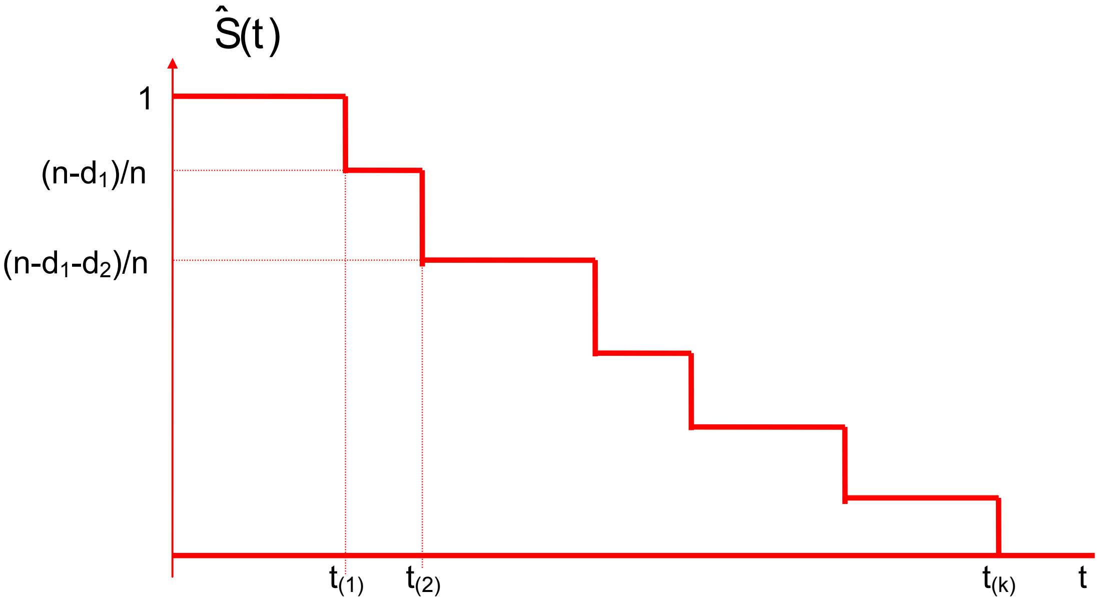
Let d_i be the number of deaths at $t_{(i)}$ (so $\Sigma d_i = n$)



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

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

$$\hat{S}(t) = 1 - \hat{F}(t) = \frac{n - \sum_1^s d_j}{n} \quad \text{for } t_{(s)} \le t < t_{(s+1)}$$



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

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

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

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

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

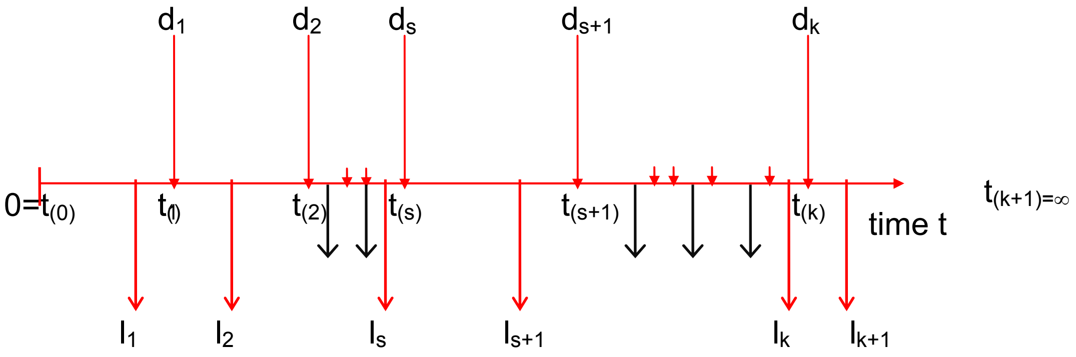
2.4.2 Standard Case, censoring:

Kaplan & Meier suggested using the same type of estimate based on a product when we have censoring.

$t_{(1)}<t_{(2)}<\ldots<t_{(k)}$ are the k distinct lifetimes

$d_1 \quad d_2 \quad \ldots \quad d_k$ number of lifetimes = $t_{(j)}$

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



l_j = number censored in the previous interval

= number with observed times $c_1, c_2, \ldots, c_n$;

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

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

So we have the Kaplan-Meier product limit

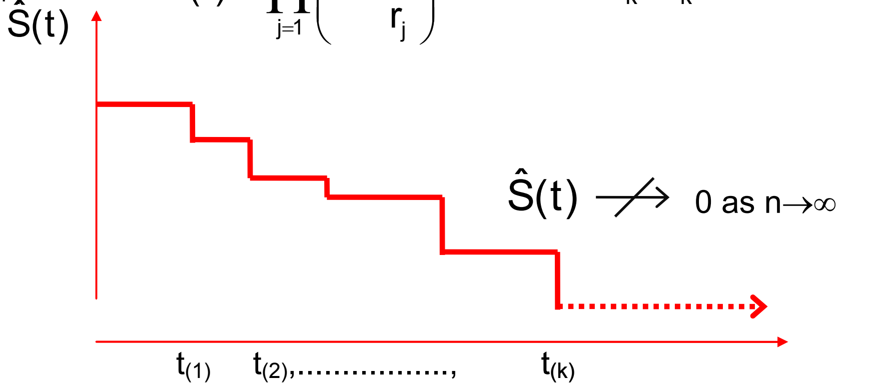
$$\hat{S}(t) = \prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \quad \text{for } t_{(s)} \leq t < t_{(s+1)}$$

2.4.2.1 Notes

(i) Assumes that the l_j censorings survive up to $t_{(j)}$ and then are removed

(ii) Uncensored case is just a special case with l_j=0 all j

(iii) If $l_{k+1}>0$ then $\hat{S}(t)=\prod_{j=1}^{k}\left(1-\frac{d_j}{r_j}\right)>0$ since $r_k>d_k$



(iv) Again $\hat{S}(t)$ is subject to sampling error.

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

(v) Similarly estimate H(t) by $\hat{H}(t)=-\log_e\hat{S}(t)$

slightly simpler estimate is to use $\tilde{H}(t)=\sum_{j=1}^{s}\frac{d_j}{r_j}$ for $t_{(s)}\leq t<t_{(s+1)}$

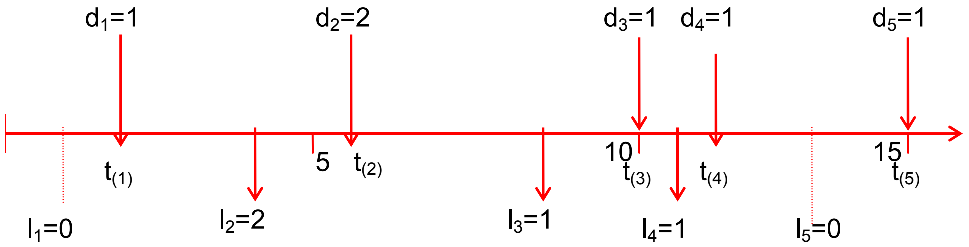
2.4.3 Example: tumour remission timea

Remission times for 10 patients with tumours

6 relapse after 3.0, 6.5,6.5,10,12,15 months

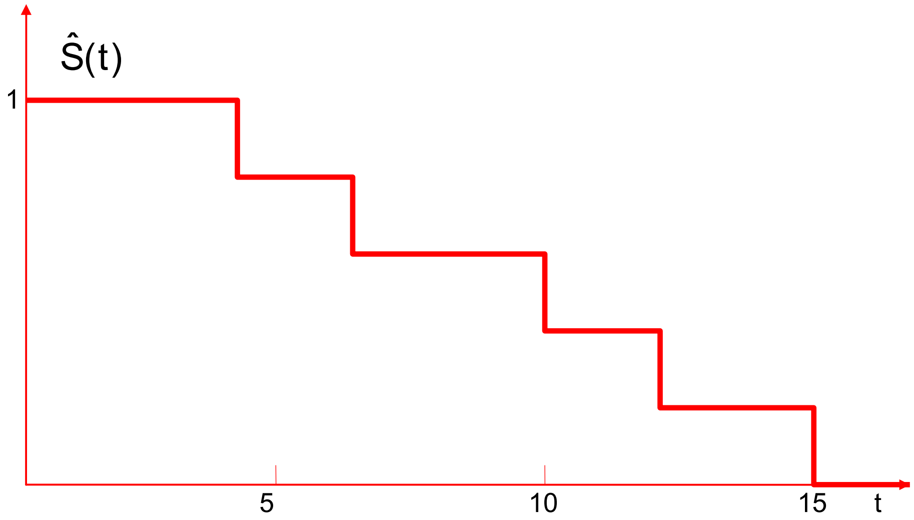
1 lost to follow-up at 8.4 months

3 still in remission at end of study after 4.0, 5.7, 10.1 months



(see data tumour.Rdata)

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



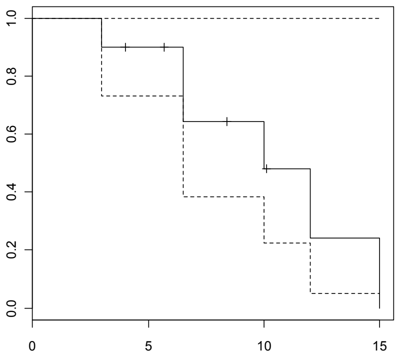
2.4.4 Computer Implementation

2.4.4.1 R

In **R** the functions for analysing survival data are provided in a package called `survival`. It is necessary to load this package with `library(survival)` before any of the commands below can be used. This package is bundled with the base system together with `MASS` etc so there is no need to download it separately from the CRAN site and install it. The first step is create a 'survival object' with the function `Surv()` (note the capitalization). The 'survival object' produced by `Surv()` will be used as the response variable in fitting a model. It contains the information on which observations are censored and which are fully observed events. The next step is to estimate the survivor curve with the function `survfit()`. Technically this step models the survival object produced by `Surv()` on a constant response. Essentially this is regressing the actual survival times on a constant response but making appropriate allowance for the censoring of some observations (information contained in the object produced by `Surv()`). This contains all the information needed for producing Kaplan-Meier plots (including the actual Kaplan-Meier estimate of the survivor function) and in assessing any model that has been fitted. Using the generic function `plot()` will produce the Kaplan-Meier plot, the function `summary()` will give the actual estimate of the survivor function and other details of the model fitting.

This is illustrated below on the data set of tumour remission times. It is necessary to make sure that you have downloaded the data set to a folder on your own hard disk and made this the working directory for your **R** session, perhaps navigating to it using `File>Change dir`… in the **R** menu. Alternatively you need to give the full pathname when loading the data set with something such as `load("C:\\...\\My Documents\\...\\tumour.Rdata")` or navigate to the file with `File>Load Workspace...` from the menus.

```
> 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
> tumour.sv <-Surv(time, censor, type = "right")
> tumourSurv <-survfit(tumour.sv ~ 1, data=tumour)
> 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
> plot(tumourSurv)
```



It is not necessary to separate all the steps, they can be nested into one command:

```
tumourSurv<-survfit(Surv(time,censor,type = "right")~1, data=tumour)
```

The `help()` system will give more details. Line styles and colours can be changed in the `plot()` command. Type `help(par)` to find out more.

2.4.4.2 S-PLUS

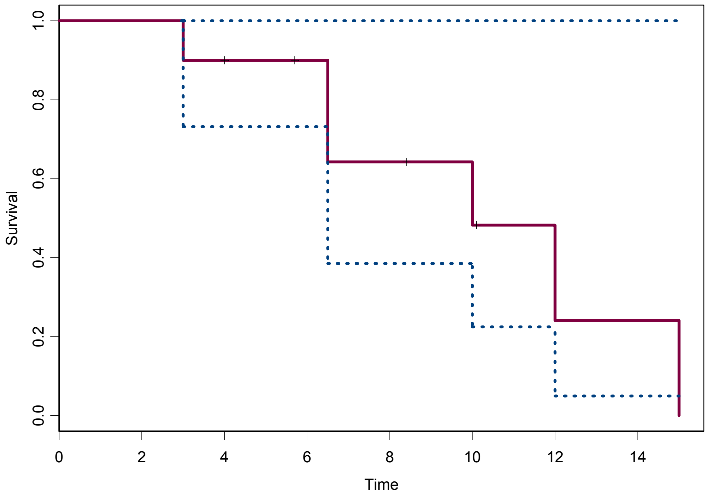
Kaplan-Meier plots are available via the menus in

Statistics>Survival>Nonparametric Survival…

First you need to create a formula by clicking the appropriate button and then select the appropriate variables for Time 1 and Censor codes by highlighting the variables in the Choose Variables box. Then click the Add Response button and you should have a formula of the form `Surv(time,censor,type='right')~1` in the formula box. Click OK and then choose appropriate Options, Results and Plots by clicking on these tabs. With Long Output in results you obtain

```
                *** Nonparametric Survival ***
Call: survfit(formula = Surv(time, censor, type = "right") ~ 1, data
= tumour, na.action = na.exclude, conf.int = 0.95,
     se.fit = T, type = "kaplan-meier", error = "greenwood",
conf.type = "log", conf.lower = "usual")

  n events mean se(mean) median 0.95LCL 0.95UCL
 10      6 10.1     1.39     10     6.5      NA
Call: survfit(formula = Surv(time, censor, type = "right") ~ 1, data
= tumour, na.action = na.exclude, conf.int = 0.95,
     se.fit = T, type = "kaplan-meier", error = "greenwood",
conf.type = "log", conf.lower = "usual")

 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 that the default option is to provide confidence intervals which necessarily have upper and lower limits of 1 and 0 which are plotted in the graphical output (if selected) and that the censoring times are indicated.



Kaplan-Meier plot from S-PLUS

Note that you can produce the analysis with from the command line with the function `survfit(.)` and an example of the call statement is given above. This function does not itself produce the graph but it produces an *object* (in the S-PLUS sense) which can then be plotted using the generic `plot(.)` function.

```
> tumourSurv<-survfit(Surv(time, censor, type = "right") ~1,
                                                     data=tumour)
> plot(tumourSurv)
```

The `help()` system will give more details. Note that this is identical to the **R** command line version.

2.4.4.3 MINITAB

Kaplan-Meier plots are available through the menus:

```
Stat>Reliability/Survival>Nonparametric Dist
                                Analysis-Right Censoring
```

It is necessary to specify the value indicating censored observations, i.e. which value indicates that the observation is censored.

2.4.4.4 SPSS

Kaplan-Meier plots are available through the menus:

```
Analyze>Survival>Kaplan-Meier
```

It is necessary to specify the value indicating uncensored values and the graphical output indicates the censored values as with S-PLUS.

2.4.5 Summary of Non-Parametric Methods

- ♦ 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 r_i

2.6 Parametric Models

2.6.1 Introduction

Lifetime T, p.d.f. f(t) with t>0, d.f. F(t),
survival function S(t)=1 – F(t), hazard function h(t).

Typically the pdf depends on an unknown parameter θ that needs to be estimated from the data. There are many methods of estimation but we concentrate on maximum likelihood estimation (m.l.e.) whose justification relies on asymptotic properties (i.e. large samples). Some details of likelihoods, maximum likelihood estimation and likelihood ratio tests are given in the Appendix 0.

In summary, the likelihood of a parameter θ for data $x_1,\ldots,x_n$ is 'the probability of observing the data $x_1,\ldots,x_n$' . this probability is calculated in terms of the unknown quantity θ and so will be a function of it, L(θ) say. We can now maximize L(θ) wrt θ (by differentiating wrt θ and setting = 0) and the value that produces the maximum, $\hat{\theta}$ say, is the maximum likelihood estimate of θ. It can be thought of as the '*most probable*' value of θ in the light of the data just obtained.

[For small samples we need to look at alternative methods, e.g.
Bayesian methods such as in Smith & Naylor (1987) *Applied Statistics*]

2.6.2 Exponential

$f(t)=\lambda e^{-\lambda t}$ $(t>0)$

$S(t)=e^{-\lambda t}$ $(=1-F(t))=1-(1-e^{-\lambda t}))$

$h(t)=\lambda$

2.6.2.1 Uncensored data

Observe $t_1,t_2,...,t_n$

$\text{Lik}(\lambda;t_1,t_2,...,t_n)=L(\lambda)=\Pi f(t_i)=\lambda^n e^{-\lambda\sum_1^n t_i}$

$\text{Log}_e(L)=\ell(\lambda)=n\log(\lambda)-\lambda\Sigma t_i$

$$\Rightarrow\hat{\lambda}=\frac{n}{\sum_1^n t_i}$$

<u>Confidence Interval for λ</u>

$Y=\sum_1^n T_i\sim\Gamma(n,\lambda)$ with p.d.f. $f(y)=\lambda^n y^{n-1}e^{-\lambda y}/\Gamma(n)$.

If we let $Z=2\lambda Y$ then Z has p.d.f. $f(z)=(½)^n z^{n-1}e^{-½z}/\Gamma(n)$

i.e. $Z\sim\chi^2_{2n}$

So $P[\chi^2_{2n;\alpha/2}<2\lambda\sum T_i<\chi^2_{2n;1-\alpha/2}]=1-\alpha$

and so a $100(1-\alpha)\%$ confidence interval for λ given by

$$\left(\frac{\chi^2_{2n;\alpha/2}}{2\sum t_i},\frac{\chi^2_{2n;1-\alpha/2}}{2\sum t_i}\right)$$

Similarly, MLE of S(t) is $\hat{S}(t)=e^{-\hat{\lambda}t}$

2.6.2.2 Censored Data

'Time' censored, n patients, (potential lifetimes i.i.d. Ex(λ)).

We observe either the lifetime t_i or the fact that $t_i>c_i$, for each individual, (i=1,2,...,n). The simplest case is to assume that the c_i are fixed and given, i.e. non-random

Thus the contribution of each individual to the likelihood is either

$$\lambda e^{-\lambda t_i} \quad \text{(if } t_i \leq c_i\text{)} \qquad \text{("}=P[T_i=t_i]\text{")}$$

or

$$e^{-\lambda c_i} \quad \text{(if } t_i>c_i\text{)} \qquad \text{("}=P[T_i>c_i]\text{")}$$

Define $\delta_i=1$ if $t_i \leq c_i$ (i.e. uncensored) and $\delta_i=0$ if $t_i>c_i$ (i.e. censored)

Then Likelihood = $L(\lambda)=\prod_{i=1}^{n}[\lambda e^{-\lambda t_i}]^{\delta_i}[e^{-\lambda c_i}]^{1-\delta_i}$

so $\log_e[\text{lik }(\lambda)]=\ell(\lambda)=\log_e\lambda\Sigma\delta_i - \lambda\Sigma t_i\delta_i - \lambda\Sigma(1-\delta_i)c_i$

$$\frac{\partial\ell(\lambda)}{\partial\lambda}=\frac{\sum\delta_i}{\lambda}-(\sum\delta_i t_i+(1-\delta_i)c_i)$$

$$\Rightarrow \hat{\lambda}=\frac{\sum_1^n\delta_i}{\sum_1^n\{\delta_i t_i+(1-\delta_i)c_i\}} \quad \text{(noting } \frac{\partial^2\ell(\lambda)}{\partial\lambda^2}>0\text{)}$$

However, the exact distribution of $\hat{\lambda}$ is not now straightforward.

Instead we have to use the asymptotic properties of maximum likelihood estimates, i.e.

$$\hat{\lambda} \approx N\left(\lambda, \left\{-E\left[\frac{\partial^2 \ell}{\partial\lambda^2}\right]_{\lambda=\hat{\lambda}}\right\}^{-1}\right)$$

Now $\dfrac{\partial^2 \ell}{\partial\lambda^2} = \dfrac{-\sum\delta_i}{\lambda^2}$ and $E[\delta_i]=1.P[T_i \leq c_i] + 0.P[T_i > c_i]$

$$=1.(1- e^{-\lambda c_i}) + 0.\ e^{-\lambda c_i}$$

$$=(1- e^{-\lambda c_i})$$

(it is implicit here that the c_i are considered non-random).

So $\mathrm{var}(\hat{\lambda}) \approx \dfrac{\hat{\lambda}^2}{\sum_1^n (1-e^{-\hat{\lambda}c_i})}$.

[Alternatively, $\hat{\lambda} \approx N\left(\lambda, \left\{-\left[\dfrac{\partial^2 \ell}{\partial\lambda^2}\right]_{\lambda=\hat{\lambda}}\right\}^{-1}\right)$, giving $\mathrm{var}(\hat{\lambda}) \approx \dfrac{\hat{\lambda}^2}{\sum_1^n \delta_i}$.]

A 100(1-α)% Confidence Interval for λ is $\hat{\lambda} \pm z_{1-½\alpha} \times$ s.e.($\hat{\lambda}$) where s.e.($\hat{\lambda}$) is the standard error of $\hat{\lambda}$, i.e. $\sqrt{}$var($\hat{\lambda}$).

Interest may be in other aspects

e.g. $\mu=\lambda^{-1}=E[T]$, the mean lifetime or

the age S_α beyond which $100\alpha\%$ survive.

For these we use the result that

$$\mathrm{var}\{g(\hat{\lambda})\} \approx \left[[g'(\lambda)]^2\,\mathrm{var}(\hat{\lambda})\right]_{\lambda=\hat{\lambda}}$$

where g(.) is any [differentiable, monotonic] function.

So $\hat{\mu}=\dfrac{1}{\hat{\lambda}}$ and $\mathrm{var}(\hat{\mu})\approx\dfrac{\hat{\mu}^2}{\sum_1^n(1-e^{-c_i/\hat{\mu}})}$ or $\mathrm{var}(\hat{\mu})\approx\dfrac{\hat{\mu}^2}{\sum_1^n\delta_i}$

and for S_α we have $\alpha = P[T\geq S_\alpha] = S(S_\alpha) = e^{-\lambda S_\alpha}$,

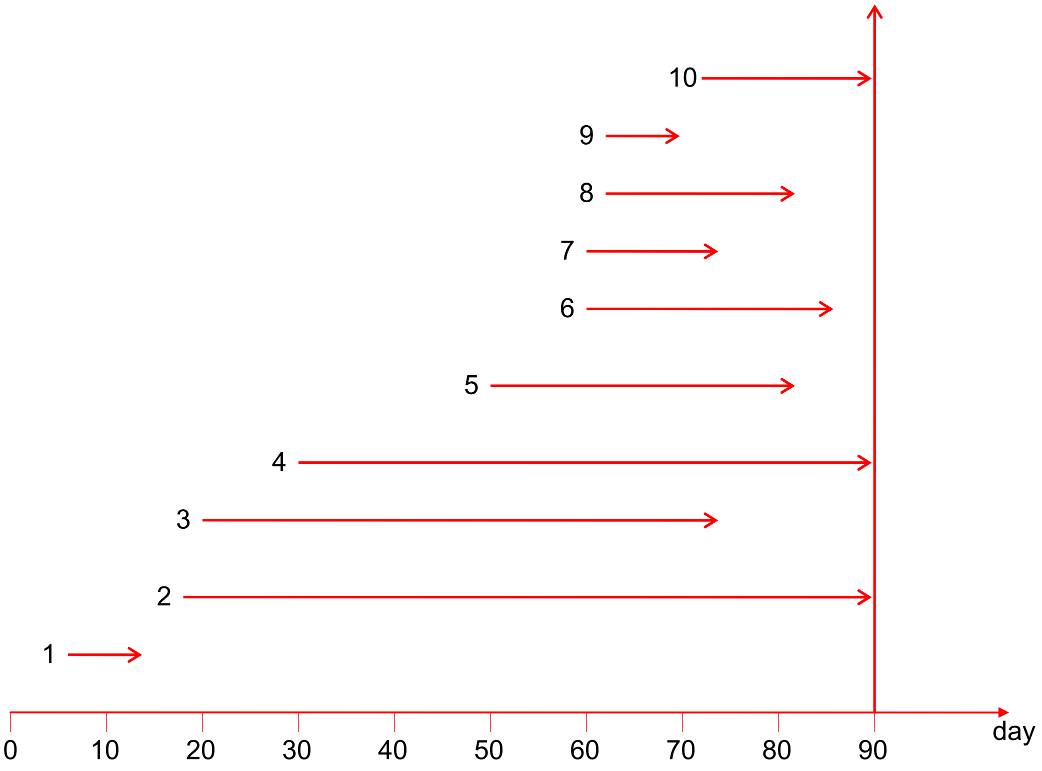
so $S_\alpha = -\lambda^{-1}\log_e(\alpha)$ and $\hat{S}_\alpha = -\hat{\lambda}^{-1}\log_e(\alpha)$

with $\mathrm{var}(\hat{S}_\alpha)=\mathrm{var}(-\hat{\lambda}^{-1}\log_e(\alpha))=[-\log_e(\alpha)]^2\,\mathrm{var}(\hat{\lambda}^{-1})$

2.6.3 Example: survival times of lung cancer

Survival times (in days) of 10 patients with advanced lung cancer. Study terminated after 90 days, (see data set lcancer.Rdata).

Patient no.	1	2	3	4	5	6	7	8	9	10
Entry time	9	18	20	30	49	59	59	60	61	69
Survival time t_i	2	.	51	.	33	27	14	24	4	.
max possible c_i	81	72	70	60	41	31	31	30	29	21
δ_i	1	0	1	0	1	1	1	1	1	0



Thus $\Sigma\delta_i$ = 7 deaths on study.

$\Sigma\delta_i t_i$ = 155, $\Sigma(1-\delta_i)c_i$ = 153

thus $\hat{\lambda} = \frac{7}{308} = 0.0227$ per day

$\hat{\mu} = \frac{308}{7} = 44.0$ days

$\text{s.d.}(\hat{\lambda})=0.00859$

and so a 95% C.I. for λ is $\hat{\lambda} \pm 1.96\,\text{s.d.}(\hat{\lambda}) \Rightarrow$ (0.00586, 0.0395)

2.6.4 Notes

(i) If we have data $t_1,t_2,\ldots,t_n$ (—observation times)

$\delta_1,\delta_2,\ldots,\delta_n$ (—censoring indicators)

where the lifetime T has a distribution depending on a parameter θ

$$L(\theta) = \prod_{\text{deaths}} f(t_i) \prod_{\text{censored}} S(t_i)$$

$$= \prod_{\text{deaths}} h(t_i)S(t_i) \prod_{\text{censored}} S(t_i) \qquad [\text{since } h(t) = f(t)/S(t)\ ]$$

$$= \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i)$$ remembering that some of the t_i's correspond to censoring.

In this notation, for exponential lifetimes, mean λ^{-1}, we have

$$\hat{\lambda} = \frac{\sum_1^n \delta_i}{\sum_1^n t_i} = \frac{\text{total number of deaths observed}}{\text{total time alive of all patients in the study}}$$

(ii) for Type II censoring (i.e. wait until r deaths) it can be shown that

$L_{II}(\theta) = \frac{n!}{(n-r)!} L(\theta) \quad \Rightarrow$ exactly the same estimates as before.

2.6.5★ Refinements: modelling the censoring distribution

A full treatment here is beyond the scope of these notes but a summary of the simplest approaches and results are stated with derivation. The simplest case to consider is the common situation where observations are collected over a fixed time T_{max}. To introduce the randomness of the censoring in a simple way suppose subjects arrive uniformly over the interval $(0, T_{max})$ and that there is no other loss to follow up, so the censoring is caused only because the study ends before the subject experiences the event.

As before, let δ_i be the censoring indicator taking values 0 or 1 but now it can be shewn that the maximum likelihood estimate of λ is the solution of the equation

$$\sum_i \left(\log(\lambda)\delta_i + \lambda t_i + \lambda(1-\delta_i)T_i\right) = 0\,,$$

where T_i is the total time on the study if lost to follow up.

$$E[\delta_i] = 1 - \frac{1-\exp(-\lambda T)}{\lambda T}$$

and that then the standard error of $\hat{\lambda}$ becomes (c.f.§2.6.2.2)

$$\frac{\hat{\lambda}}{n^{\frac{1}{2}}\left(1-\frac{1-\exp(-\hat{\lambda}T)}{\lambda T}\right)^{\frac{1}{2}}}$$

If recruitment is over a shorter period R but the study lasts for a time T so the arrival times are uniformly spread over (0, R) then we have

$$E[\delta_i] = 1 - \frac{1-\exp(-\lambda(T-R))-\exp(-\lambda T)}{\lambda R}$$

with similar modifications to the likelihood equation and standard error.

If further, there is loss to follow up in addition to the censoring caused by a fixed length of study at an exponential rate, with rate η then we have

$$E[\delta_i] = 1 - \frac{\exp(-(\lambda+\eta)(T-R))-\exp(-(\lambda+\eta T)}{(\lambda+\eta)R}$$

2.6.6 Other Distributions

2.6.6.1 Weibull

See §2.2.1: gives a flexible range of hazard functions.

2.6.6.2 Lognormal

(Beware that estimation can be unstable if there are short lifetimes)

$\log(T)\sim N(\mu,\sigma^2)$, $f(t)=(2\pi)^{-½}(\sigma t)^{-1}\exp[-½(\log(t)-\mu)^2/\sigma^2]$

$F(t)=\Phi\{(\log(t)-\mu)/\sigma\}$, $S(t)=1-\Phi\{(\log(t)-\mu)/\sigma\}$

and $h(t)=f(t)/S(t)$ requires numerical evaluation.

2.6.6.3 Others

e.g. gamma, Gumbel, mixtures — most of these require numerical evaluation of the hazard function.

2.6.7 Computer Implementation in R

Estimation of these models an be performed in **R,** S-PLUS and MINITAB but not directly in SPSS. Here we give only guidance on implementation in **R.** The basic function is `survreg()` and one of the arguments specifies which distribution amongst the options “`weibull`”, `"exponential"`, `"gaussian"`, `"logistic"`,`"lognormal"` and `"loglogistic"`. The default is `weibull`. It is also possible to use other distributions if the distribution function and density function are specified. The help system describes how to do this.

Care needs to be taken with the parameterization in `survreg()`. Firstly, the time variable is incorporated as log(time), so for example when extracting the mean survival time the result from `survreg()` is actually the logarithm of the mean survival time. Next, in the Weibull model (in the usual parameterization), the shape parameter is given as the reciprocal of the `survreg()` intercept and the logarithm of the scale parameter is given as the intercept in `survreg()`. Fo the exponential model with rate parameter λ (or mean λ^{-1}) the maximum likelihood estimate of λ is given as 1/exp(intercept). It is perhaps easiest to calculate a confidence interval for the intercept (which actually is for the log(mean survival time) and then transform this to a confidence interval for the true value of λ.

2.6.7.1 Illustration on lung cancer survival times:

As with calculation of the non-parametric Kaplan-Meier estimate the first step is to calculate a survival object which contains the information on censoring. This object can then be used in fitting any of the available survival regression models. To illustrate this on fitting an exponential distribution to the lung cancer survival times given in lcancer.Rdata, first the data needs to be loaded and attached:

```
> load("lcancer.Rdata")
> attach(lcancer)
> time
 [1]  2  4 14 21 24 27 33 51 60 72
> sum(time)
[1] 308
> sum(censor)
[1] 7

> lcancer.regexp<-survreg(lcancer.sv~1,dist="exponential")
> summary(lcancer.regexp)

Call:
survreg(formula = lcancer.sv ~ 1, dist = "exponential")
            Value Std. Error    z        p
(Intercept)  3.78      0.378 10.0 1.35e-23

Scale fixed at 1

Exponential distribution
Loglik(model)= -33.5   Loglik(intercept only)= -33.5
Number of Newton-Raphson Iterations: 4
n= 10
```

Thus the estimate of λ is 1/exp(intercept) = 1/exp(3.78) = 0.0228 and a 95% confidence interval for λ is

$1/\{\exp(3.78 \pm 1.96\times0.378)\} = (0.0108, 0.0479)$ (compare §2.6.3)

These values are rather different from calculating the approximate standard error using the formula given towards the end of §2.6.2.2 .

This would give an approximate standard error of the estimate of λ as 0.378/exp(3.78) = 0.008627 and a confidence interval of (0.00589, 0.0389). These illustrate that calculations of standard errors and confidence intervals can only be approximate, most especially for such small illustrative data sets and none is particularly ‘more accurate’ than any other and the differences apparent here are of little practical importance. If using **R,** as in most practical cases would be the case, then the results from **R** are perfectly adequate.

2.6.7.2 Illustration on tumour remission times:

As with calculation the non-parametric Kaplan-Meier estimate to first step is to calculate a survival object which contains the information on censoring. This object can then be used in fitting any of the available survival regression models. Since the default is available it is not necessary to specify `dist=”weibull”`.

```
> library(survival)
> load("tumour.Rdata")
>
> tumour.sv <-Surv(time, censor, type = "right")
> tumourSurvWeib <-survreg(tumour.sv ~ 1, data=tumour)
>
> summary(tumourSurvWeib)

Call:
survreg(formula = tumour.svweib, data = tumour)
            Value Std. Error     z        p
(Intercept)  2.42      0.147 16.41 1.63e-60
Log(scale)  -1.02      0.312 -3.26 1.12e-03

Scale= 0.361

Weibull distribution
Loglik(model)= -18.3   Loglik(intercept only)= -18.3
Number of Newton-Raphson Iterations: 6
n= 10
```

The survival object tumour.sv can be used for fitting another model, for example the exponential:

```
> tumourSurvExp<-survreg(tumour.sv, data=tumour, dist="exponential")
>
> summary(tumourSurvExp)

Call:
survreg(formula   =   tumour.sv,   data   =   tumour,   dist   =
"exponential")
            Value Std. Error    z        p
(Intercept)  2.61      0.408 6.38 1.76e-10

Scale fixed at 1

Exponential distribution
Loglik(model)= -21.6   Loglik(intercept only)= -21.6
Number of Newton-Raphson Iterations: 4
n= 10
```

Note that it is possible to fix (or constrain to a specific value) the scale parameter of the Weibull distribution. This parameter is often the most difficult to estimate, especially if it is actually close to 1. Fixing it to 1 reduces the Weibull to an exponential model:

```
> tumourSurvWeib1<-survreg(tumour.svweib,data=tumour, scale=1)
>
> summary(tumourSurvWeib1)

Call:
survreg(formula = tumour.svweib, data = tumour, scale = 1)
            Value Std. Error    z        p
(Intercept)  2.61      0.408 6.38 1.76e-10

Scale fixed at 1

Weibull distribution
Loglik(model)= -21.6   Loglik(intercept only)= -21.6
Number of Newton-Raphson Iterations: 4
n= 10
```

which gives identical estimates etc to fitting an exponential model.

2.8 Summary

- Parametric Models
  - Estimate parameters by MLE
  - Uncensored observations contribute $f(t_i)$
  - & censored contribute $S(t_i)$ to likelihood
  - Use MLE theory for standard errors
  - Plug in MLEs for other functions of θ
  - Use formula for s.e.[g(θ)]

- Noting $f(t_i) = h(t_i)S(t_i)$ allows the likelihood to be written concisely with a censoring indicator $\delta_l = 1$ for uncensored, 0 for censored, as $L(\theta) = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i)$

- For exponential data with mean lifetime λ^{-1}

$$\hat{\lambda} = \frac{\sum_1^n \delta_i}{\sum_1^n t_i} = \frac{\text{total number of deaths observed}}{\text{total time alive of all patients in the study}}$$

- Other models mentioned:
  - Weibull
  - log-Normal
  - gamma
  - Gumbel

  - All generally require numerical estimation

  - Easy to do in **R**

3 Two-Sample Comparisons

3.1 Introduction

A common problem is the comparison of two (or more) survival distributions, e.g. which treatment is better? Is the pattern of survivals/deaths for Males different from that for Females?

A simple comparison is to plot the Kaplan-Meier estimates ***for each group***. Is there any difference?

3.2 Logrank Test (non-parametric)

3.2.1 Example: brain tumour survival times

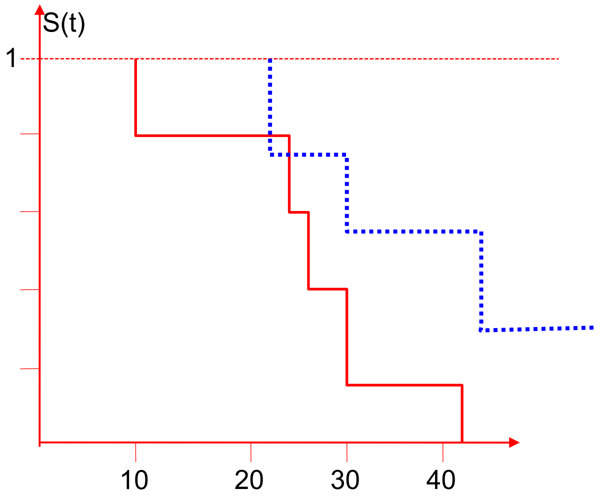
12 brain tumour patients randomized to radiation or radiation+chemotherapy. One year after the start of the study the survival times in weeks are:

Group 1 RT:	10	26	28	30	41	12*
Group 2 RT+CT:	24	30	42	15*	40*	42*

(* denotes censored)

Kaplan Meier (check)

$\hat{S}_1(10)=0.833$	$\hat{S}_2(24)=0.800$
$\hat{S}_1(26)=0.625$	$\hat{S}_2(30)=0.600$
$\hat{S}_1(28)=0.417$	$\hat{S}_2(42)=0.300$
$\hat{S}_1(30)=0.208$	
$\hat{S}_1(41)=0$	



Test: H_0: $S_1(t)=S_2(t)$

H_1: $S_1(t) \neq S_2(t)$ (for some t)

Order times of death for two groups combined $t_{(1)}<t_{(2)}<.......$

Expected number of deaths: (i) Assume H_0 (no difference in groups) is true

(ii) first death at t=10;
of 12 at risk, 1 died

If H_0 true we would expect $1\times{}^6/_{12}$ of these to be in group 1 and $1\times{}^6/_{12}$ of these to be in group 2.

Next death at t=24 , so $1\times{}^4/_9$ in group 1, $1\times{}^5/_9$ in group 2.

		Number at risk			Number of deaths			Expected no. of deaths	
i	$t_{(i)}$	r_{1i}	r_{2i}	r_i	d_{1i}	d_{2i}	d_i	e_{1i}	e_{2i}
1	10	6	6	12	1	0	1	1/2	1/2
2	24	4	5	9	0	1	1	4/9	5/9
3	26	4	4	8	1	0	1	1/2	1/2
4	28	3	4	7	1	0	1	3/7	4/7
5	30	2	4	6	1	1	2	2/3	4/3
6	41	1	2	3	1	0	1	1/3	2/3
7	42	0	2	2	0	1	1	0	1
					O_1=5	O_2=3		E_1=2.87	E_2=5.13

Log rank statistic: $(O_1–E_1)^2/E_1 + (O_2–E_2)^2/E_2 \sim \chi^2_1$ under H_0

Here =2.46 < $\chi^2_{1,0.95}$ = 3.84

i.e. no significant difference in survivor functions at 5% level

3.2.2 Notes

- Obvious generalization to 3 or more groups and then χ^2 has k–1 degrees of freedom
- Several other non-parametric tests are used — generalize the Wilcoxon-Mann-Whitney test ideas, e.g. Gehan, Cox-Mantel, Peto and Peto, Mantel-Haenszel etc. See references and computer packages.

3.2.3 Computer Implementation

3.2.3.1 R

In **R** the function for performing logrank tests is `survdiff()` . The procedure is similar to calculating a non-parametric (i.e. Kaplan-Meier) survival model. The first step, as always with censored survival times, is to create a survival object using `Surv().` Next we need to indicate which group each observed or censored survival time comes from. This is done by regressing the `Surv` object on a factor which indicates the groups. The example below gives uses the brain tumour survival times used above.

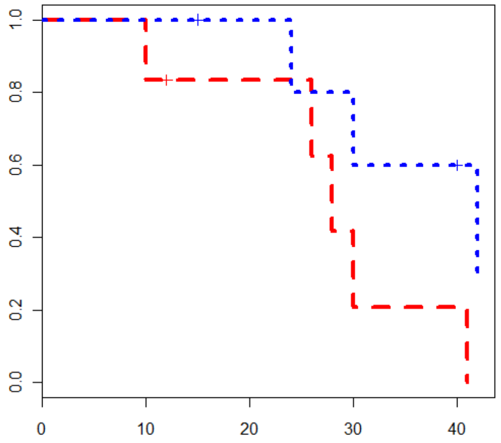
```
> library(survival)
Loading required package: splines
load("braintu.Rdata")

brain.sv<-Surv(time, censor, type = "right")
survdiff(brain.sv ~ group, data=braintu)
Call:
survdiff(formula = brain.sv, data = braintu)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 6        5     2.87     1.575      2.88
group=2 6        3     5.13     0.882      2.88

 Chisq= 2.9  on 1 degrees of freedom, p= 0.0896
>
```

Note that the numerical value is slightly different from above since **R** handles ties in a more sophisticated way. Indeed other packages also differ slightly from each other since they may have different methods.

The survival object brain.sv can be used to fit non-parameteric (Kaplan-Meier) survival models to the data separately for each group and then produce Kaplan-Meier plots)

```
plot(survfit(brainsv,data=braintu),
                        lty=c(2,3),lwd=4,col=c("red","blue"))
```

Note that this uses line styles 2 and 3 (`lty` parameter) and colours red and blue for group 1 and group 2 (`col` parameter) and thickness 4 (`lwd` parameter). Type `help(par)` to find out more details.



3.2.3.2 S-PLUS

In S-PLUS there is ***no*** menu facility for log rank tests but they can be obtained from the command line in just the same way as in R:

```
survdiff(Surv(time,censor,type='right')~group, data=braintu)
```

(note the capitalization of `Surv`) which produces:

```
Call:
survdiff(formula = Surv(time, censor, type = "right")
~ group,
    data = braintu)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 6        5     2.87     1.575      2.88
group=2 6        3     5.13     0.882      2.88

 Chisq= 2.9  on 1 degrees of freedom, p= 0.0896
```

It is possible to draw separate Kaplan-Meier plots from the menus by including the grouping variable as a main effect when creating the formula (see §2.4.4.2).

3.2.3.3 MINITAB

Log-rank tests are obtained as an option in the Kaplan-Meier plot above (§2.4.4.3): click **By variable** and alter default choice in **Result**s

3.2.3.4 SPSS

Log-rank tests are obtained as an option in the Kaplan-Meier plot above (§2.4.4.3): under Options click **Factor** and **Compare Factor**

3.3 Parametric Tests

Generally need to use asymptotic properties of maximum likelihood estimates of parameters or likelihood ratios. The exponential lifetime model is given here as an illustration, but all other models could be handled similarly with numerical estimation of parameters and the asymptotic variances.

3.3.1. M.L.E. Test

Setup: n_1 observations on T_1~ Ex(λ_1), and n_2 on T_2~Ex(λ_2).

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n_j} \delta_{ji}}{\sum_{i=1}^{n_j} t_{ji}} = \frac{\Delta_j}{\Im_j} \text{ for j=1,2}$$

Δ_j=number of deaths in group j, $\Im_j$=total 'time' on test in group j

From 2.7.2 $\hat{\lambda}_j \approx N\left(\lambda_j, \lambda_j^2 / \Delta_j\right)$ for j=1,2

so $\hat{\lambda}_1 - \hat{\lambda}_2 \approx N\left(\lambda_1 - \lambda_2, \lambda_1^2 / \Delta_1 + \lambda_2^2 / \Delta_2\right)$ and so to test the hypothesis

$H_0 : S_1(t) = S_2(t)$ *vs.* $H_1 : S_1(t) \neq S_2(t)$ for some value of t

$\Leftrightarrow H_0 : \lambda_1 = \lambda_2$ *vs.* $H_1 : \lambda_1 \neq \lambda_2$

We have that under H_0 :

$$W = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{\lambda}_1^2 / \Delta_1 + \hat{\lambda}_2^2 / \Delta_2}} \approx N(0,1)$$

3.3.1.1 Example: brain tumour survival times

$n_1=6$ $\Delta_1=5$ $\Im_1=147$

$n_2=6$ $\Delta_2=3$ $\Im_2=193$

So $\hat{\lambda}_1 = 5/147=0.034$ and $\hat{\lambda}_2 = 3/193=0.0155$,

giving W=1.02 which is not significant at 5%,

so no evidence at the 5% level that the survivor functions differ.

3.3.2 Likelihood Ratio Test

Likelihood $L(\lambda_1,\lambda_2) = L(\lambda_1)L(\lambda_2)$,

maximizing $L(\lambda_1,\lambda_2)$ with respect to λ_1 and λ_2 gives

$$L_{max}(\lambda_1,\lambda_2) = L(\hat{\lambda}_1, \hat{\lambda}_2) = \hat{\lambda}_1^{\Delta_1}e^{-\hat{\lambda}_1\Im_1}\hat{\lambda}_2^{\Delta_2}e^{-\hat{\lambda}_2\Im_2}$$

If H_0 is true then the likelihood is $L(\lambda,\lambda)=\lambda^{\Delta_1+\Delta_2}e^{-\lambda(\Im_1+\Im_2)}$

so $\text{Log}_e\{L(\lambda,\lambda)\} = \ell(\lambda,\lambda) = (\Delta_1+\Delta_2)\log_e\lambda - \lambda((\Im_1+\Im_2)$

$$\Rightarrow \hat{\hat{\lambda}} = \frac{\Delta_1+\Delta_2}{\Im_1+\Im_2}$$

$$\text{so } L(\hat{\hat{\lambda}}, \hat{\hat{\lambda}}) = \hat{\hat{\lambda}}^{\Delta_1+\Delta_2}e^{-\hat{\hat{\lambda}}(\Im_1+\Im_2)}$$

and, using the generalized likelihood ratio test we have, under H_0,

$$2\{\ell(\hat{\lambda}_1, \hat{\lambda}_2) - \ell(\hat{\hat{\lambda}}, \hat{\hat{\lambda}})\} \approx \chi_1^2$$

i.e. $2\{\Delta_1\log_e\left(\frac{\Delta_1}{\Im_1}\right) + \Delta_2\log_e\left(\frac{\Delta_2}{\Im_2}\right) - (\Delta_1+\Delta_2)\log_e\left(\frac{\Delta_1+\Delta_2}{\Im_1+\Im_2}\right)\} \approx \chi_1^2$

3.3.2.1 Example: brain tumour survival times

Test statistic = 1.20 < 3.84=$\chi_1^2(5\%)$, i.e. not significant at the 5% level, so again no evidence at the 5% level that the survivor functions differ.

3.4 Computer Implementation

Equivalents of the MLE test can be achieved by estimating parametric regression models with the group indicator as a dummy variable. This is available in **R** and S-PLUS and MINITAB and is described below in §4.1.

3.4.1 Illustration on brain Tumour times in R

Many of the calculation in §3.3.1.1 can be performed easily in R:

```
> library(survival)
> load("braintu.Rdata")
> attach(braintu)
> sum(time[group==1]); sum(time[group==2])
[1] 147
[1] 193
> sum(censor[group==1]); sum(censor[group==2])
[1] 5
[1] 3
```

From these the MLE test statistic can be calculated. A better way is to fit a model of the [censored] survival times on the group indicator using `survreg()` and an exponential distribution:

```
> brain.sv<-Surv(time,censor)
> brain.regexp<-survreg(brain.sv~group, dist="exponential")
> summary(brain.regexp)

Call:
survreg(formula = brain.sv ~ group, dist = "exponential")
            Value Std. Error    z       p
(Intercept) 2.598       1.06 2.44   0.0147
group       0.783       0.73 1.07   0.2836
Scale fixed at 1

Exponential distribution
Loglik(model)= -37.4   Loglik(intercept only)= -38
Chisq= 1.2 on 1 degrees of freedom, p= 0.27
Number of Newton-Raphson Iterations: 4
n= 12
```

Now note that 1/exp(3.381) = 0.034 (i.e. the ML estimate of λ_1) and 1/exp(3.381+0.783) = 0.0155 (i.e. the ML estimate of λ_2)

It can be checked that the standard error of the ML estimate of λ_1 is approximately 0.447/exp(3.381) (using the formula in §2.6.2.2) and that of the ML estimate of λ_2 is approximately $(0.447^2+0.730^2)^{½}$/exp(3.381+0.783) but better is to note that the p-value for testing whether the parameter indicating the group (i.e. whether the groups differ in their survival curves) is 0.284 which is close to the p-value of the MLE test statistic given in §3.3.1.1 of 1.02 which is 0.308 (given by `2*(1-pnorm(1.02)`in R). Further, the chi-squared value of 1.2 given above is precisely the value of the likelihood ratio test statistic.

3.5 Notes

(i) These two tests are asymptotically equivalent.

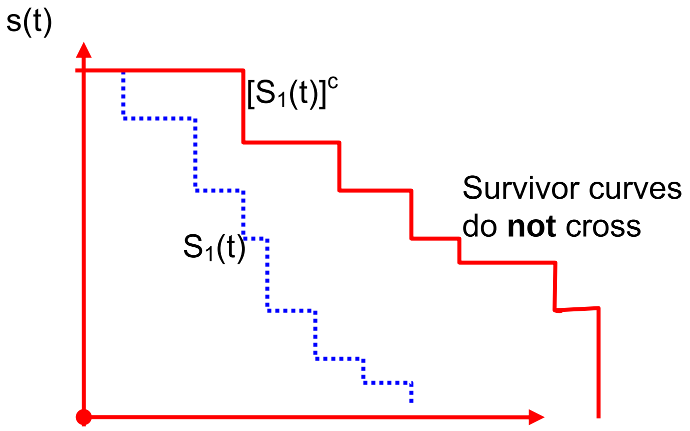
The Likelihood Ratio test is probably better for small samples.

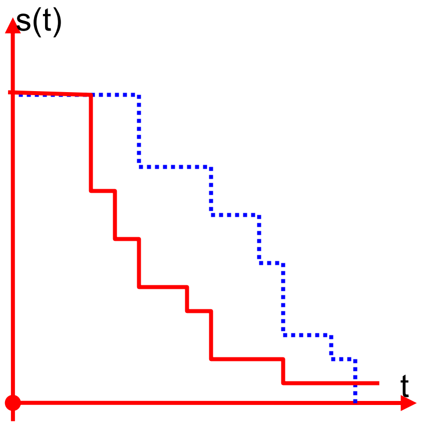
(ii) In large samples little is gained usually in terms of power by using paramteric methods instead of the log-rank test.

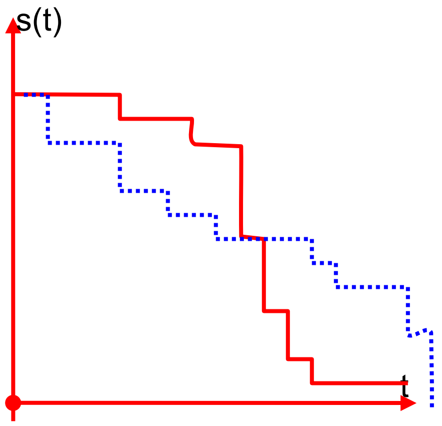
(iii) The log-rank test (devised by Peto & Peto, 1972, JRSS) assumes that $S_2(t)=[S_1(t)]^c$ or equivalently, $h_2(t)=ch_1(t)$

i.e. **proportional hazards** (see diagrams below)

If the two survivor functions differ but not in this way then the log-rank test could give misleading results.







low power

This is worse — the log-rank statistic is almost zero

In the case of the third diagram it could be that a modification of the log-rank test is required and a *generalised log-rank test* would be appropriate. This is achieved in **R** by including an extra parameter `rho` (i.e. ρ) in the call to `survidiff()`. This gives weight $[S(t)]^{\rho}$ to each event (i.e. 'death') in the calculation of the log-rank statistic. Thus more weight can be given to differences in the survival curves in the short term by taking a value of $\rho > 0$ (and less weight if $\rho < 0$). This version of the test is a form of the generalized Gehan-Wilcoxon log-rank test. The simple choice of $\rho = 1$ is the Peto & Peto modification of it.

The test would be particularly appropriate when considering a class of model known as *accelerated failure models* and this topic will be returned to in §4.5.

3.6 Summary

- Log-rank test (2 or more groups)
  - Check assumptions with Kaplan-Meier plots first
  - Best when proportional hazards
  - Based on $\Sigma(O–E)^2/E \sim \chi^2_{k–1}$
  - Calculations similar to K-M (already done)
  - Packages may give slightly different answers because of using $\Sigma(O–E)^2/\text{var}(E)$
  - Implemented in **R** with function `surdiff()`
  - Generalised (Peto & Peto) log-rank test obtained with `surdiff(.,., rho=1)`

- **MLE Test**
  - 2 groups, single parameter
  - based on two-sample Normal test with s.e.s from general ML theory

- **Likelihood Ratio Test**
  - k groups, any # of parameters
  - difference in sum of maximized likelihoods and pooled likelihood
  - test statistic ~ χ^2_r where r = (k – 1) × # of parameters

- **All of these can be performed in R**

4 Regression Models

4.1 Introduction

Each individual may be subject to their

own individual hazard rate $h_i(t)$

e.g. $T_i \sim Ex(\lambda_i)$ with $f_i(t) = \lambda_i e^{-\lambda_i t}$; i=1,2,...,n

Clearly, for the uncensored case $\hat{\lambda}_i = \dfrac{1}{t_i}$; i=1,2,...n

More useful models are obtained if we can inter-relate the λ_i's, e.g. incorporate covariates in a regression model.

4.2 Parameteric Regression Models

4.2.1 Exponential Regression Model

Suppose $t_i \sim Ex(\lambda_i)$

We take $E[T_i] = \lambda_i^{-1} = \alpha + \beta x_i$ (i=1,2,...,n)

where the x_i are known values of a covariate.

Then $L(\alpha,\beta) = \prod_{i=1}^{n}\left\{\dfrac{1}{\alpha+\beta x_i} e^{-\frac{t_i}{\alpha+\beta x_i}}\right\}$

so $\log_e L = \ell(\alpha,\beta) = -\sum_{i=1}^{n}\log_e(\alpha+\beta x_i) - \sum_{i=1}^{n}\dfrac{t_i}{\alpha+\beta x_i}$

$$\text{so } \frac{\partial\ell}{\partial\alpha} = -\sum_{i=1}^{n}\frac{1}{\alpha+\beta x_i} + \sum_{i=1}^{n}\frac{t_i}{(\alpha+\beta x_i)^2} = 0$$

$$\text{and } \frac{\partial\ell}{\partial\beta} = -\sum_{i=1}^{n}\frac{x_i}{\alpha+\beta x_i} + \sum_{i=1}^{n}\frac{t_i x_i}{(\alpha+\beta x_i)^2} = 0$$

and we solve these by iterative maximum likelihood. We would have to obtain initial values for α and β and we might do this by plotting the observed survival times t_i against x_i and fitting a line by eye, measuring slope and intercept or obtaining the least squares estimate. These would be very rough estimates but would serve as a starting point for the iterations. This is the method used in the numerical example below and is perhaps the easiest if one is actually doing this 'by hand' as in the example but it is different from the method used in **R** and S-PLUS and other packages which typically model the logarithm of survival time. An illustration of the same example analysed with **R** and S-PLUS is given below.

The general theory of maximum likelihood estimation gives that

$$\text{asymptotically} \quad \begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} \approx N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{bmatrix} -E\left[\dfrac{\partial^2 \ell}{\partial\alpha^2}\right] & -E\left[\dfrac{\partial^2 \ell}{\partial\alpha\partial\beta}\right] \\ -E\left[\dfrac{\partial^2 \ell}{\partial\alpha\partial\beta}\right] & -E\left[\dfrac{\partial^2 \beta}{\partial\beta^2}\right]\end{bmatrix}^{-1}\right)$$

$$\approx N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{bmatrix} \sum\dfrac{1}{(\alpha+\beta x_i)^2} & \sum\dfrac{x_i}{(\alpha+\beta x_i)^2} \\ \sum\dfrac{x_i}{(\alpha+\beta x_i)^2} & \sum\dfrac{x_i^2}{(\alpha+\beta x_i)^2}\end{bmatrix}^{-1}\right)$$

4.2.1.1 Notes

(i) Estimate λ^{-1} by $\hat{\alpha}+\hat{\beta}x$ for given x

(ii) var($\hat{\alpha}+\hat{\beta}x$) = var($\hat{\alpha}$) + x^2 var($\hat{\beta}$) + 2x cov($\hat{\alpha}$, $\hat{\beta}$)

$\Rightarrow$ e.g. 95% Confidence Interval for λ^{-1}

(using a Normal approximation)

(iii) Strictly we need to impose the condition that $\alpha+\beta x>0$ but this is not a problem in practice if the model is reasonably appropriate.

(iv) Could extend to censored data, bringing in the censored observations to the likelihood in a similar way to that in the simple exponential model (i.e. with the survivor function) but note that this would alter the closed form expressions given for the variances of the estimates given above.

4.2.1.2 Example: myelogenous leukemia

Survival times (from date of diagnosis) of patients with acute myelogenous leukemia.

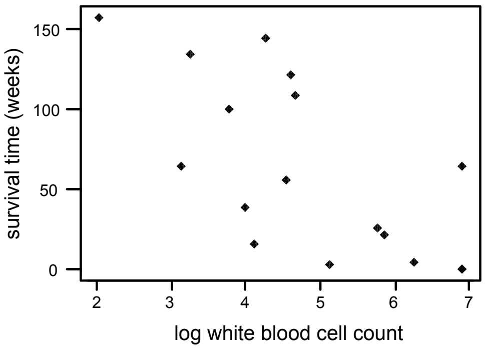
Covariate: white blood cell count.

because of variability take $\lambda_i^{-1}=\alpha+\beta\log(\text{WBC})$

(AG positive) n=17

Patient	WBC×10^2	Survival Time (weeks)
1	23	65
2	7.5	156
3	43	100
4	26	134
5	60	16
6	105	108
7	100	121
8	170	4
9	54	39
10	70	143
11	94	56
12	320	26
13	350	22
14	1000	1
15	1000	1
16	520	1
17	1000	5
Median values	100	56

Survival Time vs. Log White Blood Cell Count



From the graph we can obtain initial estimates of α and β and then iterate.

This gives

$$\hat{\alpha}=240,\hat{\beta}=-44;\hat{\text{var}}(\hat{\alpha})=95.5,\hat{\text{var}}(\hat{\beta})=20.1,\hat{\text{cov}}(\hat{\alpha},\hat{\beta})=-1914$$

and so a 95%CI for $\alpha+\beta x$ ($=\lambda_x^{-1}$) from

$$240-44x-1.96[95.5-3828x+20.1x^2]^{0.5<}=\lambda_x^{-1}<240-44x+1.96[95.5-3828x+20.1x^2]^{0.5}$$

In this example it was fortunate that $\hat{\alpha}+\hat{\beta}x>0$ for the range of values in the data set. To avoid this problem we could instead model the log of the survival term as a linear function of the covariate and this is the way that such parametric regression models are implemented in computer packages. [NB: here the estimates are MLE estimates obtained by substituting for λ_i in terms of α and β and then, from general asymptotic MLE theory the variances and covariance are obtained from the inverse of the matrix of second derivatives. This is beyond the scope of this course]

4.2.2 Computer Implementation

4.2.2.1 R

Parametric models can be fitted using `survreg()` as described in §2.6.6. As always the first step is to create a survival object with `Surv()` to combine the censoring information with the survival times. This 'object' is then regressed on any appropriate covariates. Estimates of the coefficients of the covariates in the regression, together with their standard errors are provided in the results. These can be used to perform [partial] z-tests of hypotheses that the separate covariates have no effect on the survival distribution. The tests are strictly conditional on all the remaining covariates being included in the model and hence are properly termed partial z-tests. Note also that the analysis models the log of the survival times and this needs to be remembered when estimating any quantity from the model such as median survival time for a particular set of covariates.

The procedure is illustrated on the survival times of patients with myelogenous leukemia. In this particular case none of the observations is censored so in the initial `Surv()`step the censoring variable can be omitted.

```
> library(survival)
Loading required package: splines
> load("wbcleuk.Rdata")
> wbcleuk
   patient    wbc survival log.wbc.
1        1   23.0       65 3.135494
2        2    7.5      156 2.014903
3        3   43.0      100 3.761200
4        4   26.0      134 3.258097
5        5   60.0       16 4.094345
6        6  105.0      108 4.653960
7        7  100.0      121 4.605170
8        8  170.0        4 5.135798
9        9   54.0       39 3.988984
10      10   70.0      143 4.248495
11      11   94.0       56 4.543295
12      12  320.0       26 5.768321
13      13  350.0       22 5.857933
14      14 1000.0        1 6.907755
15      15 1000.0        1 6.907755
16      16  520.0        1 6.253829
17      17 1000.0        5 6.907755
18       2    7.5      156 2.014903
> attach(wbcleuk)
> survreg(survival~log.wbc.)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regexp<-survreg(wbcleuk.sv~log.wbc.,dist="exponential")
> summary(wbcleuk.regexp)

Call:
survreg(formula = wbcleuk.sv ~ log.wbc., dist = "exponential")
            Value Std. Error     z        p
(Intercept)  7.84      1.052  7.45 9.05e-14
log.wbc.    -0.89      0.220 -4.05 5.15e-05

Scale fixed at 1

Exponential distribution
Loglik(model)= -84.5   Loglik(intercept only)= -92.9
        Chisq= 16.86 on 1 degrees of freedom, p= 4e-05
Number of Newton-Raphson Iterations: 4
n= 18
```

Because the logarithms of survival times are used this models the logarithm of the mean survival time for a given value of log(wbc). Thus the estimated mean survival time for a patient with a wbc of 54 is exp(7.84-0.89log(54)) = 72.95 days. An approximate standard error for this estimate can be calculated from the standard errors for the intercept and coefficient and using the formula for standard errors of a function of an estimate given in §2.6.2.2.

Since this model is an exponential model, the estimated median survival time for a patient with a wbc of 54 is –72.95log(0.5) = 50.57 days. Note that the dataset contains one observed survival time of a patient (number 9) with a wbc of 54 who survived 39 days.

For further illustration given below is the analysis in fitting a Weibull model to the same data. Since the Weibull is the default distribution for `survreg()` it is not necessary to specify that the distribution is Weibull with `dist=”weibull”` and it is omitted.

```
> summary(wbcleuk.regweib)
Call:
survreg(formula = wbcleuk.sv ~ log.wbc.)
             Value Std. Error      z        p
(Intercept)  7.849      0.986  7.957 1.76e-15
log.wbc.    -0.882      0.207 -4.265 2.00e-05
Log(scale)  -0.105      0.187 -0.562 5.74e-01
Scale= 0.9
Weibull distribution
Loglik(model)= -84.3   Loglik(intercept only)= -92
        Chisq= 15.4 on 1 degrees of freedom, p= 8.7e-05
Number of Newton-Raphson Iterations: 5
n= 18
```

It might be noted that the estimated scale parameter is 0.9, very close to 1.0 which is equivalent to the exponential model. In fact noting that the log(scale) is estimated as –0.105 with standard error 0.187 it is clear that this estimate is not significantly different from zero so there is little evidence provided by these data that the Weibull model fits better than the simpler exponential model.

The estimated mean survival time for a patient with a wbc of 54 is exp(7.849–0.882log(54) = 76 days, little different from the estimate based on the exponential model. Further investigation (not given here) indicates that the standard error of the Weibull estimate is appreciably larger than the exponential estimate, again illustrating the superiority of the exponential fit.

4.2.2.2 S-PLUS

Parametric models can be fitted from the menus under `Statistics>Survival>Parametric Survival …` and the operation of this follows the familiar pattern of first needing to create a formula to declare which is the survival time, what is the censoring variable and what are the explanatory variables. Note that censored values are handled easily.

Note also that the analysis models the log of the survival time so the estimates from S-PLUS given below differ substantially from those obtained by iteration 'by hand' in fitting the exponential model to the actual survival times. However, it is not difficult to see that the two models are essentially equivalent in that estimates of the actual survival times for a given value of the white blood cell count are very close. The 'long results' output is:–

```
Call:
survReg(formula = Surv(survival) ~ log.wbc., data = wbcleuk,
na.action = na.exclude, dist = "weibull", scale = 0,
     control = list(maxiter = 30, rel.tolerance = 1e-005,
failure = 1))
             Value Std. Error      z        p
(Intercept)  7.849      0.986  7.957 1.76e-015
   log.wbc. -0.882      0.207 -4.265 2.00e-005
 Log(scale) -0.105      0.187 -0.562 5.74e-001

Scale= 0.901

Weibull distribution
Loglik(model)= -84.3   Loglik(intercept only)= -92
     Chisq= 15.4 on 1 degrees of freedom, p= 0.000087
Number of Newton-Raphson Iterations: 4
n= 18
```

Other distributions available are Weibull, extreme, Gaussian (i.e. normal), loggausian, logistic and loglogistic. In practice, the exponential model is rarely used and the Weibull the most common model to try first.

4.2.2.3 MINITAB

Parametric models can be fitted from the menus under `Stat>Reliability/Survival>Regression with Life Data …` with the same set of distributions as in S-PLUS.

4.2.2.4 SPSS

There are currently no facilities for fitting parametric models in SPSS (i.e. up to version 15).

4.2.3 Two-Sample Example

Consider the two group exponential model of §3.3

$$T_1 : f_1(t) = \lambda_1 e^{-\lambda_1 t} \qquad h_1(t)=\lambda_1$$

$$T_2 : f_2(t) = \lambda_2 e^{-\lambda_2 t} \qquad h_2(t)=\lambda_2$$

Let x=0 for group 1
=1 for group 2 — a binary indicator variable

Model: $h(t;x) = \lambda e^{\beta x}$ $= \lambda$ for x=0 (group 1)
or λe^{β} for x=1 (group 2)

i.e. a reparameterization: $\lambda=\lambda_1$; $\beta=\log_e(\lambda_2/\lambda_1)$

The sign of β determines whether $\lambda_2>\lambda_1$ or $\lambda_1>\lambda_2$
$(\beta>0)$ $(\beta<0)$

4.2.4 Notes

(i) $\log_e h(t;x)=\log_e(\lambda)+\beta x$

So the model is sometimes called

the *log-linear model* for the hazard function

(ii) If we parameterize in this way it ensures that $h(t;x)>0$

(iii) Could extend this to several groups with $h(t;\underline{x})=\lambda e^{\underline{\beta}'\underline{x}}$

(with the use of dummy variables)

$x_1=1$ if group A, $x_1=0$ otherwise; $x_2=1$ if group B, $=0$ otherwise

the $\underline{x}=(x_1,x_2)$ and (1,0)→A ; (0,1) →B ; (0,0) →C

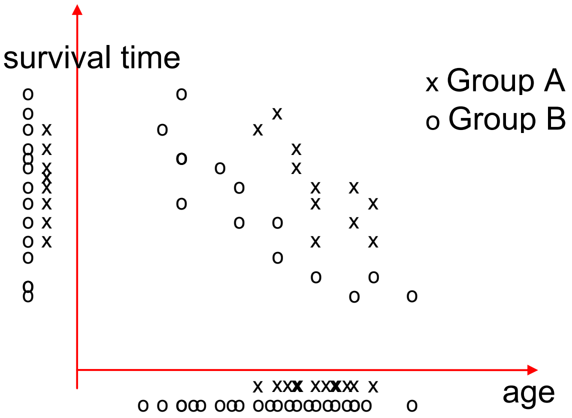
4.3 Covariates and Prognostic Factors

In comparisons of treatments randomization in large samples ensures that factors affecting survival time, e.g. stage of disease, age, sex etc., are balanced between the treatment groups. Regression type models can allow for these factors as well within a randomization trial.

4.3.1 Notes

(i) Resulting procedures can be <u>more sensitive</u> to treatment differences.

If we ignore age factor and collapse values onto one line then no difference between the groups



(ii) Factors themselves may be of interest → prognostic factors

i.e. for predicting survival times

(iii) Randomization may occasionally 'go wrong'

— regression methods help to correct for this.

(iv) Interactions: treatment effect may be different for 'different' patients (according to the prognostic factors)

e.g. x_1=0 for control group, x_1=1 for treatment group

x_2=0 for stage I, x_2=1 for stage II and x_2=2 for stage III.

Then define x_3=$x_1$$x_2$ for the interaction

between the factors age and stage of disease

4.3.2 Modelling

(i) look at each factor separately

(ii) try $E[T] = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \underline{\beta' x}$

→ problems since we need to ensure that $\underline{\beta' x > 0}$ all $\underline{\beta}$ and $\underline{x}$

(iii) try $E[T] = \alpha \exp\{\underline{\beta}'\underline{x}\}$ — log-linear type model

(iv) Try to model through the hazard function

$h(t) \to h(t;\underline{x}) = h_0(t) \exp\{\underline{\beta}'\underline{x}\}$

where $h_0(t)$ is the ‘underlying’ hazard rate

(v) β_j reflects the effect of x_j on survival

if $\beta_j > 0$: increasing $x_j \Rightarrow$ hazard $\nearrow \Rightarrow$ poorer survival prospect

if $\beta_j < 0$: increasing $x_j \Rightarrow$ hazard $\searrow \Rightarrow$ better survival prospect

if $\beta_j = 0$: increasing $x_j \Rightarrow$ no effect on survival.

4.3.3 Exponential Model

$$h(t;\underline{x}) = \lambda\exp\{\underline{\beta}'\underline{x}\}$$

$$f(t;\underline{x}) = \lambda\exp\{\underline{\beta}'\underline{x}\}\exp\{-\lambda t\exp\{\underline{\beta}'\underline{x}\}\}$$

$$S(t;\underline{x}) = \exp\{-\lambda t\exp\{\underline{\beta}'\underline{x}\}\}$$

<u>Data</u>: $(t_1,\delta_1,\underline{x}_1), (t_2,\delta_2,\underline{x}_2),\ldots., (t_n,\delta_n,\underline{x}_n)$ where $\underline{x}_i=(x_{i1},x_{i2},\ldots,x_{ip})$

Estimate the p+1 parameters λ, $\beta_1,\beta_2,\ldots,\beta_p$ by maximum likelihood

$$L(\lambda,\underline{\beta})=\prod_{i=1}^{n}[\lambda\exp\{\underline{\beta}'\underline{x}_i\}]^{\delta_i}\exp\{-\lambda t_i\exp\{\underline{\beta}'\underline{x}_i\}\}$$

$$=\prod_{i=1}^{n}[\lambda\exp\{\underline{\beta}'\underline{x}_i\}\exp\{-\lambda t_i\exp\{\underline{\beta}'\underline{x}_i\}\}]^{\delta_i}[\exp\{-\lambda t_i\exp\{\underline{\beta}'\underline{x}_i\}\}]^{1-\delta_i}$$

$$\ell(\lambda,\underline{\beta}) = \Sigma\delta_i\log_e\lambda + \Sigma\delta_i\,\underline{\beta}'\underline{x}_i - \Sigma\lambda\,t_i\exp\{\underline{\beta}'\underline{x}_i\}$$

$$\frac{\partial\ell}{\partial\lambda}=\frac{\Delta}{\lambda}-\Sigma t_i\exp\{\underline{\beta}'\underline{x}_i\}$$

where $\Delta=\Sigma\delta_i$ the total number of deaths

$$\frac{\partial\ell}{\partial\beta_j}=\Sigma\delta_i x_{ij}-\lambda\Sigma x_{ij}\,t_i\exp\{\underline{\beta}'\underline{x}_i\}\quad (j=1,2,\ldots,p)$$

Setting these two derivatives = 0 and solving iteratively

gives the maximum likelihood estimates of λ and β.

For estimates of variance and standard errors we need

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{\Delta}{\lambda^2}$$

$$\frac{\partial^2 \ell}{\partial \lambda \partial \beta_j} = -\ \Sigma x_{ij}\, t_i \exp\{\underline{\beta}'\underline{x}_i\} \qquad (j=1,2,...,p)$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\ \lambda \Sigma x_{ij} x_{ik}\, t_i \exp\{\underline{\beta}'\underline{x}_i\} \quad (j=1,2,...,p;\ k=1,2,...,p)$$

4.3.4 Other Models

Covariates can be incorporated into any model used for survival data by including a multiplicative term $\exp\{\underline{\beta}'\underline{x}\}$ in the expression for the hazard function, e.g. for the Weibull use $h(t;\underline{x})=\lambda\gamma t^{\gamma-1} \exp\{\underline{\beta}'\underline{x}\}$.

Many packages offer a wide choice of distributional models for regression analysis of survival data

4.4 Proportional Hazards Model

This is a semi-parametric model proposed by Cox (1972).

It is convenient and general model for comparing 2 groups of survival times — very widely applied.

$$h(t,\underline{x}) = h_0(t)\exp\{\underline{\beta}'\underline{x}\}$$

4.4.1 Notes

(i) $h_0(t)$ — baseline hazard function, i.e. corresponds to hazard of a patient with $\underline{x}=\underline{0}$

(ii) Dependence of failure time on explanatory variables is precisely modelled; but actual distribution of failure is not parametrically specified, i.e. $h_0(t)$ is not specified.

(iii) Useful in medical situations

— important to know which prognostic variables have an effect and to what extent

— knowing the actual distribution of survival time is not as important.

(iv) **special cases:**

Exponential: $h(t;\underline{x})=\lambda\exp\{\underline{\beta}'\underline{x}\}$ $h_0(t)=\lambda$

Weibull: $h(t;\underline{x}) =\lambda\gamma t^{\gamma-1}\exp\{\underline{\beta}'\underline{x}\}$ $h_0(t)=\lambda\gamma t^{\gamma-1}$

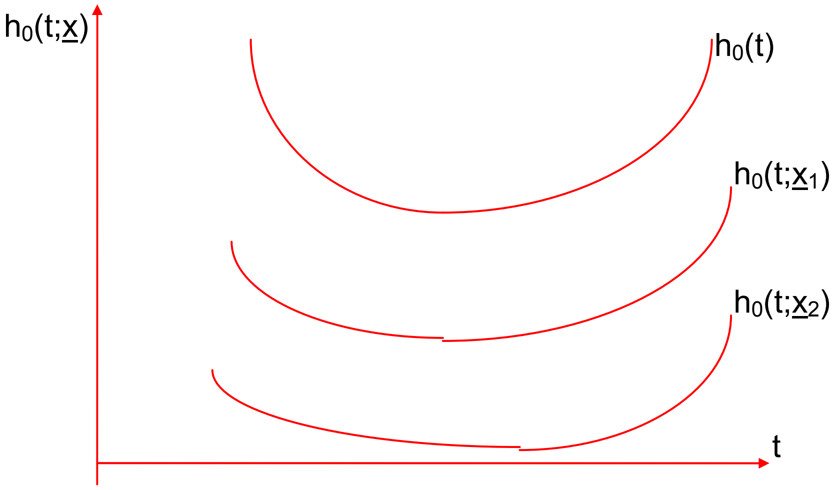
(v) For patients with covariates $\underline{x}_1$ and $\underline{x}_2$ we have

$$h(t;\underline{x}_1)/\, h(t;\underline{x}_2) = h_0(t)\exp\{\underline{\beta}'\underline{x}_1\}/\, h_0(t)\exp\{\underline{\beta}'\underline{x}_2\}$$

$$= \exp\{\underline{\beta}'(\underline{x}_1-\underline{x}_2)\}$$ ***INDEPENDENT OF t***

i.e. hazard functions for any 2 patients are proportional over time,

i.e. the linear component of the model does not vary with time.



The model assumes patients have the same 'shape' of hazard function — but shifted multiplicatively according to $\underline{x}$.

Note that under this model ***they can never cross.***

4.4.2 Parameter Estimation

Observations

Survival times	$t_1,t_2,\ldots,t_n$
Censorings	$\delta_1,\delta_2,\ldots\delta_n$
Covariates	$\underline{x}_1,\underline{x}_2,\ldots\underline{x}_n$

$$h(t;\underline{x}_i) = h_0(t)\exp\{\underline{\beta}'\underline{x}_i\}$$

$$S(t;\underline{x}_i) = \exp\{-\textstyle\int_0^t h_o(u)\ \exp\{\underline{\beta}'\underline{x}_i\}du\}$$

$$= \exp\{-\exp\{\underline{\beta}'\underline{x}_i\}\textstyle\int_0^t h_o(u)\ du\}$$

$$= [S_0(t;\underline{x}_i)]^{\exp\{\underline{\beta}'\underline{x}_i\}} \text{ — the baseline survivor function}$$

$$f(t;\underline{x}_i) = h_0(t)\exp\{\underline{\beta}'\underline{x}_i\}\ S(t;\underline{x}_i)$$

$$\text{likelihood} = \prod_{i=1}^{n}[h(t_i;\underline{x}_i)]^{\delta_i}S(t_i;\underline{x}_i)$$

This involves $h_0(t)$, so to proceed further we need to specify a parametric form for $h_0(t)$, e.g. $h_0(t)=\lambda$ say.

Alternatively we can use the ***partial likelihood approach***

4.4.3 Partial Likelihood Approach

(Called *partial* since it does not make direct use of the actual censored and uncensored survival times)

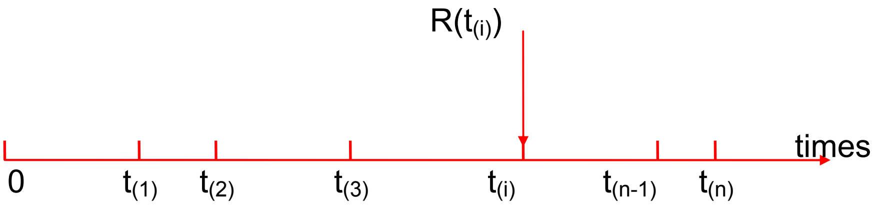
Suppose that there are *no ties* — only one individual dies at each death time:

Ordered (by time) observations:

Survival times	$t_{(1)},t_{(2)},\ldots,t_{(n)}$
Censorings	$\delta_{(1)},\delta_{(2)},\ldots\delta_{(n)}$
Covariates	$\underline{x}_{(1)},\underline{x}_{(2)},\ldots\underline{x}_{(n)}$

<u>Risk set R(t) at t</u>: set of individuals alive and in the trial just before time t.

$R(t_{(i)})$ is the set of individuals who are alive and uncensored at time just before $t_{(i)}$.



Consider time points at which deaths occur:

P[individual (i) dies at $t_{(i)}$ | exactly one patient in
the risk set R($t_{(i)}$) dies at $t_{(i)}$]

$$=\lim_{\delta t\to 0}\left\{\frac{\text{P[death of (i) in }(t_{(i)},t_{(i)}+\delta t)\,|\,R(t_{(i)})}{\text{P[one death in }(t_{(i)},t_{(i)}+\delta t)\,|\,R(t_{(i)})}\right\}$$

$$\underline{\Omega}\ \frac{h(t_{(i)};x_{(i)})\delta t}{\sum\limits_{j\in R(t_{(i)})} h(t_{(i)};x_j)\delta t}$$

{because .....

P[one death in [$t_{(i)}$,$t_{(i)}$+δt] | R($t_{(i)}$)]

$$= \sum_{j\in R(t_{(i)})} \text{P[j dies]P[others don't die]}$$

$$= \sum_{j\in R(t_{(i)})} h(t_{(i)};x_j)\delta t \prod_{\substack{k\in R(t_{(i)})\\ k\neq j}} [1-h(t_{(i)};x_k)\delta t$$

$$\underline{\Omega}\ \sum_{j\in R(t_{(i)})} h(t_{(i)};x_j)\delta t \quad \text{ignoring terms in } (\delta t)^2$$

.... end of explanation}

$$= \frac{h_0(t_{(i)})e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} h_0(t_{(i)})e^{\underline{\beta}'\underline{x}_{(j)}}}$$

$$= \frac{e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} e^{\underline{\beta}'\underline{x}_{(j)}}}$$

NOTE proportional hazards assumption necessary here

Individuals for whom the survival times are censored do not contribute to the numerator but they do enter the summation over the risk set at death times of subject less than the censored time.

Form the ‘likelihood’ by taking products over the observed failure times $t_{(i)}$.

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} e^{\underline{\beta}'\underline{x}_{(j)}}} \right\}^{\delta_{(i)}}$$

4.4.3.1 Notes

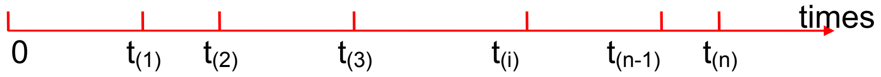
(i) If no censored observations this is a ***conditional*** likelihood, conditional on the observed $t_{(1)},t_{(2)},\ldots,t_{(n)}$.

(ii) With censored observations this is known as a ***partial*** likelihood.

(iii) Use of partial likelihood was justified by Cox (1975) in Biometrika. He showed that the usual likelihood methods apply in this case. So we maximize the (partial) likelihood to estimate β by $\hat{\underline{\beta}}$ and asymptotically

$\hat{\underline{\beta}} \approx \mathrm{N}(\underline{\beta},\left[-\frac{\partial^2\ell}{\partial\beta_i\partial\beta_j}\right]^{-1}_{\hat{\underline{\beta}}})$ where ℓ is the log likelihood.

(iv)



Consider likelihood contributions at observed failure times. No extra information on $\underline{\beta}$ is obtained from the fact that there are no failures between two specific observed lifetimes.

If we had a parametric form for $h_0(t)$ in our model, there would then be contributions to the inferences about $\underline{\beta}$ from the intervals with no failures. Intervals between successive death times convey no information about the effect of explanatory variables on the hazard of death. This is because the baseline hazard has an arbitrary form and so it is conceivable that $h_0(t)$ and hence $h(t)$ is zero in those time intervals in which there are no deaths. This in turn means that those intervals give no information on the $\underline{\beta}$ parameters.

(v) Ties (Peto’s adjustment)

$t_{(1)}<t_{(2)}<\ldots<t_{(k)}$ the k distinct survival times

$d_{(1)}, d_{(2)},\ldots\ldots,d_{(k)}$ numbers of deaths at these times

$D(t_{(1)}), D(t_{(2)}),\ldots D(t_{(k)})$ death set at $t_{(i)}$

Allow each of $d_{(i)}$ deaths at $t_{(i)}$ to contribute a factor (as before) to partial likelihood, each with the same risk set $R(t_{(i)}$

$$\text{‘Likelihood’} = \prod_{i=1}^{k}\left\{\frac{\exp\{\sum\limits_{j\in D(t_{(i)})}\underline{\beta}'\underline{x}_{(j)}}{\left[\sum\limits_{j\in R(t_{(i)})}e^{\underline{\beta}'\underline{x}_{(j)}}\right]^{d_{(i)}}}\right\}$$

Satisfactory provided $d_{(i)}/n_{(i)}$ is small, where $n_{(i)}$ is the number of individuals at risk at $t_{(i)}$.

Other methods for adjustment for ties exist (e.g. by Cox, Breslow, Efron,.....).

4.4.4 Example: atrial fibrillation

The table below gives details of a proportional hazards model fitted to some data obtained from patients being treated for atrial fibrillation. The purpose of the treatment is to maintain normal heart rhythm and ‘survival time’ is in terms of time to relapse.

Variable	Coefficient	Standard Error	χ^2 statistic (using lrt)	coeff/s.e.
treatment (0=A, 1=B)	–1.42	0.64	4.89	**–2.22**
age (years)	–0.004	0.034	0.01	–0.12
sex (1=M,0=F)	0.31	0.72	0.18	0.43
volume of heart (mml)	0.0076	0.0036	4.44	**2.11**
Duration of symptoms (months)	–0.004	0.063	0.00	–0.06
digitalisation	–0.59	0.73	0.66	–0.81

Questions: (a) Describe the effects of treatments and additional covariates on time to relapse.

(b) The above analysis did not consider treatment×covariate interactions. How would this be done?

Answers:

(a) Model is

$h(t;\underline{x}) = h_0(t)\exp\{\beta_1x_1+\beta_2x_2+...+\beta_6x_6\}$

where x_1=0 for treatment A, and x_1=1 for treatment B, i.e.

$h(t;\underline{x}) = h_0(t)\exp\{\beta_2x_2+...+\beta_6x_6\}$ for treatment A, and

$h(t;\underline{x}) = h_0(t)\exp\{\beta_1+\beta_2x_2+...+\beta_6x_6\}$ for treatment B

and use approximation $\hat{\underline{\beta}} \approx N(\underline{\beta}, \left[-\frac{\partial^2\ell}{\partial\beta_i\partial\beta_j}\right]^{-1}_{\hat{\underline{\beta}}})$ for calculating standard errors.

Note that $\exp\{\beta_1\}$ has an interpretation of the *hazard ratio* of treatment B to treatment A seen by dividing the last two expressions above.

To see which factors are shewn to affect the hazard we examine the estimated coefficients and their standard errors. We can use the χ^2 statistic or coeff/s.e. for a factor with only 2 levels or a continuous covariate. If there is a factor with k-levels coded as k–1 dummy variables then the χ^2 statistics for each of the dummy variables should be added together and compared with χ^2 on k–1 d.f.

Note that $\chi^2 \approx (\text{coeff/s.e.})^2$ so if the χ^2 is not given and only the coefficients and their standard errors then it is possible to deduce the χ^2 values and so find the overall χ^2 value for a k-level factor coded as k–1 dummy variables. Note that it is not sensible (i.e. *wrong*) to consider the parameters for each of the levels of a factor in isolation. The k–1 parts of the chi-squared statistic must be combined and an overall assessment made of the factor made.

Treatment: $\hat{\beta}_1$/s.e.($\hat{\beta}_1$)= |–2.22| > 1.96 so we have good evidence of effect of treatment. $\hat{\beta}_1$ < 0 so treatment = 1 decreases hazard, i.e. treatment B is ‘better’

Heart volume:

$\hat{\beta}_4$/s.e.($\hat{\beta}_4$) = +2.11 >1.96 so increased heart volume decreases relapse time.

No evidence that other factors affect relapse time

NB Not shewn that other factors have no effect

It is also useful to calculate confidence intervals for the parameters, not just those where there is evidence that the factor is affecting the survival but also for those where the evidence is not there. This allows assessment of how big the effect could be. For example, the 95% CI for β_3 (M/F) is 0.31 ± 2 × 0.72 =(–1.13, 1.75), i.e. there could be a large difference between M & F and the data do not exclude this possibility. It may also be useful to calculate a confidence interval for exp{β_3} which has the interpretation of the *hazard ratio* of males to females. This would give (0.323, 5.75) so that the hazard for men could be just less than a third that for women or nearly six times as much; the data exclude neither possibility.

(b) Interaction terms would be handled by creating a new variable as the product of the treatment and the covariate values. In this case the treatment is coded as 0 for treatment A and 1 for B, so the value of this interaction term would be 0 for all subjects receiving A and the same as the covariate for those on B. In the example above Treatment is variable x_1 and age is variable x_3 and there are six variables in all. We create a new variable $x_7 = x_1 \times x_3$ and then our model is

$h(t;\underline{x}) = h_0(t)\mathrm{exp}\{\beta_2x_2+\beta_3x_3+\ldots+\beta_6x_6\}$ for treatment A, and

$h(t;\underline{x}) = h_0(t)\mathrm{exp}\{\beta_1+\beta_2x_2+(\beta_3+\beta_7)x_3+\ldots+\beta_6x_6\}$ for treatment B

and β_7 reflects the interaction effect, (note that x_7 is identical to x_3 for those on treatment B but 0 for those on A).

Exactly the same method is appropriate for handling interactions between two continuous covariates and between two 2-level factors. Interactions involving a k-level factor can only be handled by converting the factor into k–1 dummy binary variables. In this case the interaction term has k–1 degrees of freedom if it is a k-level factor×covariate interaction or (k–1)(j–1) degrees of freedom for an interaction between a k-level and a j-level factor. This also means that the separate parts of the chi-squared statistic must be combined before assessing significance.

4.4.5 Computer Implementation

4.4.5.1 R

Cox proportional hazards models can be fitted using the function `coxph()`. The operation of this follows the familiar pattern of first needing to create a survival object combining the survival time with censoring information using Surv() and then regressing this on the explanatory variables. This is illustrated with the survival times of subjects with acute myelogenous leukaemia which has no censoring and only one explanatory variable (log white blood cell count):–

```
> library(survival)
Loading required package: splines
> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regcox<-coxph(wbcleuk.sv~log.wbc.)
> summary(wbcleuk.regcox)
Call:
coxph(formula = wbcleuk.sv ~ log.wbc.)

  n= 18

           coef exp(coef) se(coef)     z Pr(>|z|)
log.wbc. 1.1753    3.2392   0.3244 3.623 0.000292 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
log.wbc.     3.239     0.3087     1.715     6.118

Rsquare= 0.669   (max possible= 0.982 )
Likelihood ratio test= 19.89  on 1 df,   p=8.192e-06
Wald test            = 13.12  on 1 df,   p=0.0002916
Score (logrank) test = 17.39  on 1 df,   p=3.043e-05
```

Primary interest will be in the estimated coefficients and their standard errors to perform partial z-tests and estimate hazard ratios.

4.4.5.2 S-PLUS

Cox proportional hazards models can be fitted from the menus under `Statistics>Survival>Cox Proportional Hazards …` and the operation of this follows the familiar pattern of first needing to create a formula to declare which is the survival time, what is the censoring variable and what are the explanatory variables. Note that censored values are handled easily. The 'long results' output for the very simple example of survival times with acute myelogenous leukaemia which has no censoring and only one explanatory variable (log white blood cell count):–

```
     *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(survival) ~ log.wbc., data = wbcleuk,
na.action = na.exclude, method = "efron", robust = F)

  n= 18

         coef exp(coef) se(coef)    z       p
log.wbc. 1.18      3.24    0.324 3.62 0.00029

         exp(coef) exp(-coef) lower .95 upper .95
log.wbc.      3.24      0.309      1.72      6.12

Rsquare= 0.669   (max possible= 0.982 )
Likelihood ratio test= 19.9  on 1 df,   p=8.19e-006
Wald test            = 13.1  on 1 df,   p=0.000292
Score (logrank) test = 17.4  on 1 df,   p=0.0000304
```

It is possible to produce a Kaplan-Meier plot of the survival time calculated for a subject with mean values of all the covariates.

4.4.5.3 MINITAB

There are currently no facilities in MINITAB for Cox regression (i.e. up to version 15)

4.4.5.4 SPSS

Cox proportional hazards regression is available through the menus:–

```
Analyze>Survival>Cox Regression ...
```

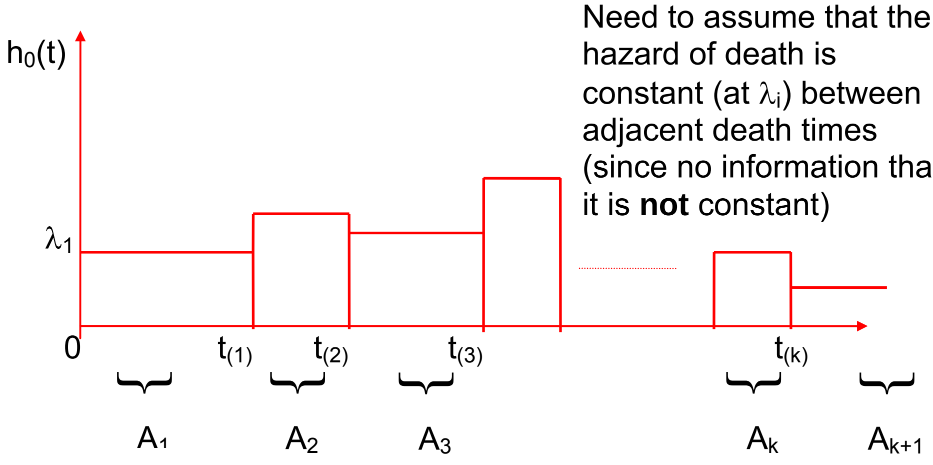
Need to specify value indicating uncensored values and the graphical output indicates the censored values as with S-PLUS.

4.4.6 Estimation of h(t)

We can obtain $\hat{\underline{\beta}}$, can we also estimate $h_0(t)$ to get $\hat{h}_0(t)$ and then $\hat{S}_0(t)$ and thus $\hat{S}(t,\underline{x}) = \hat{S}_0(t)\exp(\hat{\underline{\beta}}'\underline{x})$

There several methods available for the estimation of the baseline hazard function, e.g. Breslow (1974, Biometrics).

Let $t_{(1)}<t_{(2)}<\ldots<t_{(k)}$ be the k distinct death times.



Need to assume that the hazard of death is constant (at λ_i) between adjacent death times (since no information that it is **not** constant)

The full likelihood is

$$L = \prod_{i=1}^{n} \{h_0(t_i) e^{\underline{\beta}'\underline{x}_i}\}^{\delta_i} \exp\{-\textstyle\int_0^{t_i} h_0(u) e^{\underline{\beta}'\underline{x}_i} du\}$$

and since $h_0(t)=\lambda_j$ if $t \in A_j=(t_{(j-1)},t_{)j)}]$ we have

$$\int_0^{t_i} h_0(u) = \int_0^{t_{(1)}} \lambda_1 du + \int_{t_{(1)}}^{t_{(2)}} \lambda_2 du + \ldots + \int_{t_{(j-1)}}^{t_i} \lambda_j du$$

so

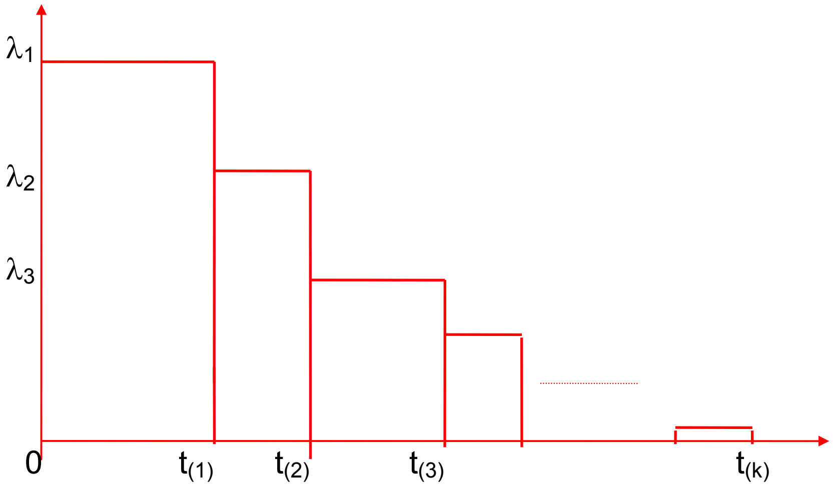
$$L=\prod_{j=1}^{k+1} \prod_{i \in A_j} \{\lambda_j e^{\underline{\beta}'\underline{x}_i}\}^{\delta_i} \exp\left[-e^{\underline{\beta}'\underline{x}_i}[\lambda_1 t_{(1)} + \lambda_2(t_{(2)} - t_{(1)}) + \ldots + \lambda_j(t_i - t_{(j-1)})\right]$$

replace $\underline{\beta}$ by $\hat{\underline{\beta}}$,, then find estimates $\hat{\lambda}_1, \hat{\lambda}_2$,..... by MLE.

$$\Rightarrow \hat{\lambda}_j = \frac{d_j}{\sum_{i \in R(t_{(j)})} e^{\hat{\underline{\beta}}'\underline{x}_i}(t_{(j)} - t_{(j-1)}) + \sum_{i \in A_j} e^{\hat{\underline{\beta}}'\underline{x}_i}(t_i - t_{(j-1)})},$$

where d_j = number of deaths at $t_{(j)}$ and noting that the term in the denominator summing over $i \in A_j$ does not appear if all lifetimes are observed.

If the estimates of λ_j suggest a pattern, e.g. roughly constant or exponentially decreasing or..... then this might suggest a parametric form for $h_0(t)$.



4.4.7 Model checking

4.4.7.1 log–log plots

e.g. two treatments, $x_1=0$ and $x_1=1$;

$h(t;x)=h_0(t)e^{\beta x_1}$, so $h_1(t)=h_0(t)$ and $h_2(t)=h_0(t)e^{\beta}$

$$-\log_e S_1(t)=H_1(t)=\int_0^t h_1(u)du$$

$$-\log_e S_2(t)=H_2(t)=\int_0^t h_2(u)du$$

so $\log_e[-\log_e S_2(t)]=\beta+\log_e[-\log_e S_1(t)]$

so if we plot $\log_e[-\log_e\{\hat{S}_j(t)\}]$ *vs.* t for both groups we should get parallel curves a distance β apart — if the curves cross then a proportional hazards model is not appropriate.

If this diagnostic test fails and it is concluded that a proportional hazards model is not appropriate then there are two options. One is in the special case where proportionality only breaks down for one particular factor in which case a *stratified proportional hazards model* might be considered (see details below in §4.4.7.3 and Collett (2003) §11.1.1) and the other option is to consider a parametric model which does not have the proportional hazards property.

4.4.7.2 Residuals

Various types of residuals can be defined for survival regression models. Full discussions are given in Collett (2003) §4 and §7 for Cox and parametric regression respectively and Everitt & Rabe-Heskith (2001) §17.5. Of particular note are *Schoenfeld Residual*s which can be obtained with the **R** function `cox.zph(.)`. Schoefeld residuals are defined only for non-censored observations and there is a separate set for each of the covariates. They are defined as the difference between the value of the covariate of interest for the subject experiencing the event (say death) and the expectation over all members of the risk set of the covariate:

$$r_{ik} = x_{ik} - \sum_{j\in R(t_i)} x_{jk}\hat{p}_j ,$$

where r_{ik} is the Schoenfeld residual for individual i for the k^{th} covariate, x_{jk} are the values for that covariate for the individuals j in the risk set for $R(t_i)$ of individuals at time t_i , the time of death for the i^{th} individual and $\hat{p}_j$ is the estimated probability that the j^{th} person dies by time t_i.

These should be independent of time so a plot of these versus time should shew no dependence. Deviation from independence will indicate inadequacy of the model.

4.4.7.3 Implementation in R

The non-obvious step for log-log plots needed is that to produce separate plots of the estimated survivor function for different levels of a factor the factor needs to be used in the model as a stratum indicator using the function **strata(.)**, i.e. the proportional hazards model is fitted [almost] separately within each level of the factor. This step cannot be by-passed. For example, for the lymphoma data where there are two levels of the variable `stage` (which must be converted to a factor) the following:–

```
> library(survival)
Loading required package: splines
> load(“lymphoma.Rdata”)
> attach(lymphoma)
> stage<-factor(stage)
> lymph.cox<-coxph(Surv(time,censor)~strata(stage))
> lymph.cox
Call:coxph(formula=Surv(time,censor) ~ strata(stage))

Null model
  log likelihood= -17.77164
  n= 18
 > plot(survfit(lymph.cox))
```

–:will produce two separate K-M plots.

To produce a plot of the log survivor function the vertical scale needs to be changed appropriately:–

```
>plot(survfit(lymph.cox),fun="cloglog",lty=2:3,col=4:5)
```

where the line styles and colours have been chosen with the parameters `lty` and `col` (this also uses a log scale for the horizontal axis which is useful since it tends to ‘straighten’ the plot and so makes it easier to judge parallelism). To make the plot more visible it is usually best to adjust the thickness of the lines with the `lwd` parameter, perhaps `lwd=3`. To find out more about plotting parameters type `help(par)`.

4.4.7.4 S-PLUS implementation

In S-PLUS log-log plots can be drawn using the same commands as in **R** given in §4.4.7.3.

4.4.8★ Time-dependent covariates

It is possible to extend the model to 'time-dependent covariates'

e.g. $h(t;x)=h_0e^{\beta x+\gamma xt}$

which is equivalent to allowing β to change with time. This may be appropriate if x=0 means receiving treatment, x=1 means receiving no treatment (so without treatment patient gets worse and worse).

More generally, we have if $\underline{x}=\underline{x}(t)$ then $h(t;\underline{x})=h(t;\underline{x}(t))=h_0(t)e^{\underline{\beta}'\underline{x}(t)}$ and this gives

$$S(t)=\exp\left\{-\int_0^t e^{\underline{\beta}'\underline{x}(u)}h_0(u)du\right\}$$

so the survivor function depends not only on the baseline hazard function $h_0(t)$ but also on the values of the time-dependent covariates over the interval (0,t).

One application is to transplant studies. Define

$z(t)=0$ before transplant

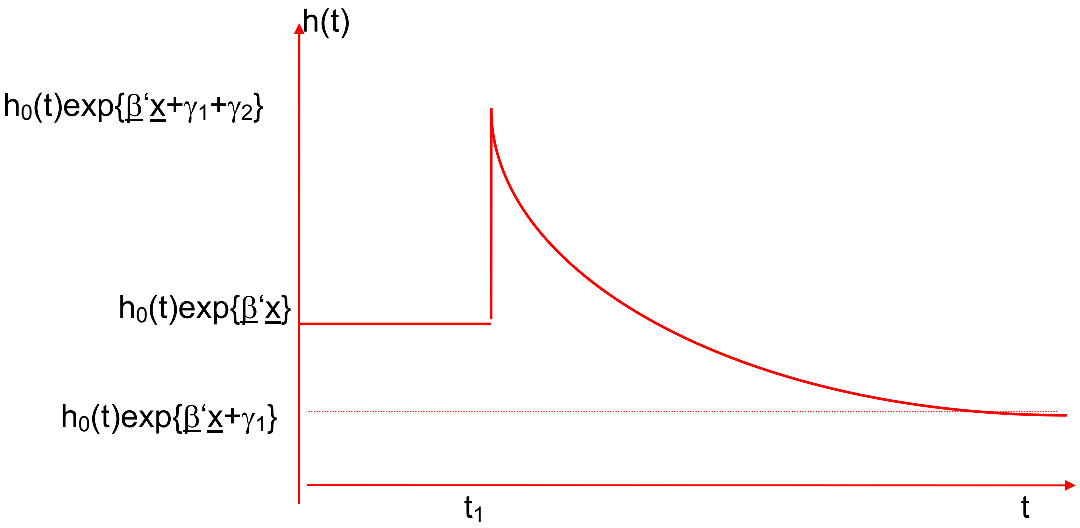
$z(t)=1$ after transplant

and then $h(t)=h_0(t)e^{\underline{\beta}'\underline{x}+\gamma z(t)}$ where $\underline{x}$ are non-time-dependent covariates. Then the effect of the transplant is measured by γ.

A more sophisticated model is given by Cox & Oakes (1984) with

$$h(t;\underline{x}) = h_0(t)\exp\{\underline{\beta}'\underline{x} + z(t)[\gamma_1 + \gamma_2 e^{\gamma_3(t-t_1)}]\}$$

with z(t) = 0 for t < t_1, 1 for t > t_1



In this model the effect of the transplant is to increase the hazard to $h_0(t)\exp\{\underline{\beta}'\underline{x}+\gamma_1+\gamma_2\}$ (assuming $\gamma_1+\gamma_2>0$) from which it decreases exponentially to $h_0(t)\exp\{\underline{\beta}'\underline{x}+\gamma_1\}$ which is less than the initial hazard if $\gamma_1<0$.

4.5★ Accelerated Failure Time Regression Models

4.5.1★ Introduction

In situations where the proportional hazard assumption is violated (as illustrated in the notes in §3.5) the alternatives are to use a parametric model (which does not have the proportional hazards property of course) or to consider a class of models termed *accelerated failure time* ***(AFT)*** *models*. These are models where the survivor function for a subject with covariate **x** takes the form $S(t;\mathbf{x}) = S_0(t.e^{\beta'\mathbf{x}})$ where $S_0(.)$ is some baseline parametric survivor function. Of course for certain choices of distribution this model has the proportional hazards property, for example the exponential where $S_0(t) = \exp\{\lambda_0 t\}$ and the Weibull. For this reason an accelerated failure time model is often taken to refer to a Weibull model.

The name comes from *accelerated-life* testing, typically of electronic components, where the components under test would be subject to enhanced stress, e.g. higher voltages than those under which the components would be used. The objective would be to complete the whole experiment within a shorter time. For example old-fashioned electric light bulbs might be expected to have a mean lifetime of 1000 hours under normal operating conditions but increasing voltage and operating temperature might accelerate the time to failure so that all those under test would fail within a week and thus allow an analysis to be completed without the complication of large numbers of censored observations. In these situations it was plausible to consider that for the stressed components time itself was progressing faster and so in terms of a model that the effect of a covariate (e.g. voltage) could be regarded as multiplying the timescale by some numerical factor.

In recent years these models have been considered for use in medical situations, especially now that software for fitting them to data including censored observations has become widely available. The **R** package eha which has a full set of facilities dates from 2003 and the current version 1.2.18 from February 2010.

Unfortunately some texts within the medical area advocate use of AFT models when ‘it is desired to speed up the time to the event’ — such as time to relief of symptoms and other ‘positive’ events whereas proportional hazards models should be used for situations where it is required to slow down the time to a ‘negative’ event, such as death. This is clearly incorrect**:** – use of a particular statistical model does not influence the time of occurrence of events. The model has to be chosen so that it fits the available data (not vice versa). However, it is true that experience has shown that for experiments on a short time scale (days or a very few weeks) AFT models are often found to work well whilst for long term studies (years) proportional hazards models may be found to be preferable. There is no guarantee that this will hold for any data set that is encountered and there is no substitute for investigating models and checking validity with appropriate diagnostics (log–log plots, residual etc). In part this ‘folk-lore’ belief and sweeping generalisation could be a reflection that in longer term studies the numbers of actual events observed can be relatively small and so it is difficult to discover that proportional hazards models do not fit the data but in short term studies where events are relatively common it is easier to detect deviations from a model.

4.5.2* Implementation in R

The functions for fitting AFT models are in library `eha`. The particular function for fitting a regression is `aftreg()` and it may be noted that he library also contains routines for fitting proportional hazards parametric and non-parametric regression models (`aftreg.fit()` and `coxreg()` respectively. The second of these is an alternative to `coxph()` but care must be used when comparing them since the parameterisation is different. The other functions in the library (especially for diagnostics) can be found by the command `library(help=eha)`and so are not described here.

4.5.3* Example

This is the same example as analysed with `coxph()` in §4.4.5.1.

First, a model with the default Weibull distribution is fitted and then one using a Gompertz distribution. You are invited to compare the results from fitting these two models and that given in §4.4.5.1 using appropriate diagnostics. Not surprisingly, the results from the non-parametric proportional hazards model and from the Weibull accelerated failure time model are close (though parameterized differently) since the Weibll model belongs to both families.

```
> library(eha)
Loading required package: survival
Loading required package: splines

> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regcox<-coxph(wbcleuk.sv~log.wbc.)
> summary(wbcleuk.regcox)
> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regaftW<-aftreg(wbcleuk.sv~log.wbc.)
> summary(wbcleuk.regaftW)
Call:
aftreg(formula = wbcleuk.sv ~ log.wbc.)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
log.wbc.            3.589     0.882     2.415     0.207     0.000

log(scale)                    7.440  1702.521     1.119     0.000
log(shape)                    0.105     1.110     0.187     0.574

Events                    18
Total time at risk          1154
Max. log. likelihood      -84.309
LR test statistic         15.4
Degrees of freedom        1
Overall p-value           8.6824e-05
>
> wbcleuk.regaftG<-aftreg(wbcleuk.sv~log.wbc., dist="gompertz")
> summary(wbcleuk.regaftG)
Call:
aftreg(formula = wbcleuk.sv ~ log.wbc., dist = "gompertz")

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
log.wbc.            3.589     0.872     2.393     0.240     0.000

log(scale)                    8.392  4412.587     1.131     0.000

 Shape is fixed at  1

Events                    18
Total time at risk          1154
Max. log. likelihood      -84.892
LR test statistic         16.5
Degrees of freedom        1
Overall p-value           4.97646e-05
>
```

4.6 Summary & Conclusions

Regression models

- allow individual hazards linked by covariates
- Investigate prognostic factors
    - Factors of interest
    - Allow for covariates
    - More precise analysis
    - Could model mean of survival distribution
        - e.g. exponential regression
    - Better to model hazard function
        - Model $h(t,\underline{x})=h_0(t)\text{exp}\{\underline{bx}\}$ ensures $h(t,\underline{x})>0$
    - Estimation by [numerical] MLE or partial MLE
    - Can estimate survival times for given covariates
    - Parametric regression models are available in **R**, S-PLUS and MINITAB but not SPSS.
        - In these cases the logarithm of the mean is modelled as a linear function of the covariates which ensures that the estimate mean is positive.
    - Semi-parametric proportional hazards Cox regression models are available in **R**, S-PLUS and SPSS but not MINITAB.

Proportional Hazards Models

- Only models dependence on covariates
- No statements about survival times
- Only effect of covariates on hazards
- Estimation by maximum partial likelihood
- Check proportional hazards by log-log plots
- May suggest parametric model
- No allowance for individual variability
- i.e. no term in σ_i^2 for ith individual
- Frailty models do allow for this (some facility in **R** and S+)

Accelerated failure time models

- Accelerated failure time models may provide an alternative in cases where the proportional hazards assumption is untenable. Such models are available in **R**.

Further family of models

- A further family of model is the family of proportional odds models where the odds ratio of surviving beyond a time t for two individuals with different covariates is independent of t,

  i.e. $$\frac{S(t;x)}{1-S(t;x)}=\frac{S(t)}{1-S(t)}\exp\{\beta'x\}$$

  This class of models has been studied by some authors but facilities for fitting them to censored data are currently not available in **R**.

5★ Competing Risks

5.1★ Introduction

The survival models considered up to now consider situations where there is a single event of interest and the analysis is concerned with investigating the properties of the single distribution of the failure time distribution or the dependence of the time to this event upon various covariates. This section provides a brief introduction to more complex situations where individuals are exposed to risks of different types of events and the time measured is that to the first of these to occur. One example might be when there are different ‘causes’ of failure, e.g. death following a liver transplant might be from rejection of the organ or from an infection. The dependence of the time to death from different causes on covariates such as tissue type, blood group and ethnicity might be different for the different causes.

At first sight it is tempting when considering one particular cause (e.g. rejection) to regard observations where the cause of failure is not of interest (e.g. from infection) as censored, together with those subjects who are still alive at the end of the study. However, care must be taken in this. While it is possible to use some tools from survival analysis described in earlier chapters, such as log rank tests and Cox proportional hazards regression extreme care must be used with others (e.g. Kaplan-Meier estimates). The key problem is essentially that in the presence of competing risks the probability of failure from the rth cause by time t does not tend to 1 as t increases in the presence of competing risks other than the rth. Also note that if a failure occurs then we observe which cause is responsible but if an observation is censored then we do not. The essential point is that the censoring distribution is not independent of the time to failure from a competing event. In single event survival analysis an assumption of Kaplan-Meier estimation is that individuals who are censored could in principle fail at some later time but if an individual fails from an event not of interest but is treated as if it were censored from the point of view of the event of interest it certainly cannot fail at some future time from the event of interest. This means that Kaplan-Meier estimation over-estimates the [probability of failure and this bias is greater if the competing events have greater hazards than the one of interest.

5.2* Basic terminology

Let T be the time to failure from any of K different causes, r =1, …, K, and R the actual cause (so completed observations of failure times consist of pairs (T,R) and we may also have uncompleted times, i.e. censored observations after times c_i. Let t_j, j = 1,…,N be the ordered distinct times of failure, and d_{rj} be the number of failures from cause r at time t_j, then $d_j = \Sigma_r d_{rj}$ is the total number of failures at time t_j. Note that $d_{rj} \neq 0$ for at least one r and that typically $d_{rj} \neq 0$ for only one r. $d = \Sigma_j d_j$ is the total number of failures. Let n_j be the number of individuals at risk at time t_j (note that these are at risk of failure from any of the causes).

The *cumulative incidence function* for cause r is

$$I_r = P[T \leq t,\ R = r], \text{ for } r = 1,\ldots,r.$$

The *cause specific hazard function* is $\lambda_r(t)$ given by

$$\lambda_r(t) = \lim_{\delta t \to 0}\left(\frac{P\left[T \leq t + \delta t, R = r \mid T > t\right]}{\delta t}\right)$$

The *cumulative cause specific hazard function* is given by

$$\Lambda_r(t) = \int_0^t \lambda_r(u)du \text{ and define}$$

$$S_r(t) = \exp(-\Lambda_r(t))$$

the r^{th} *competing event survival distribution.* Note that

$$S_r(t) = \int_t^\infty \lambda_r(u) S_r(u) du$$

(c.f. §2.1.3) and that

$$S(t) = \prod_{r=1}^{K} S_r(t) = \exp\left(-\sum_{r=1}^{K} \Lambda_r(t)\right)$$

is the probability of not having failed from any cause at time t. Note that

$$I_r(t) = \int_0^t \lambda_r(u) S(u) du$$

and that $I_r(\infty)$ = P[R = r] so it is not a proper distribution function; it is sometimes referred to as a 'subdistribution function'. Note that the integral above involves S(t) and not $S_r(t)$ because $S_r(t)$ is the probability of failing from cause r after time t but the probability of failing from cause r after t requires not failing from *any cause* before time t. Consequently $I_r(t) \neq 1 - S_r(t) = F_r(t)$. Generally it is $I_r(t)$ that is of interest rather than $S_r(t)$ and this is a key reason that naïve Kaplan-Meier estimates have to be used with care.

Clearly the overall hazard function of the distribution of T is $\lambda(t) = \Sigma\lambda_r(t)$. The conditional probability that failure is from cause r given that failure is before time t is

$$\phi_r^I(t) = P[R = r \mid T \leq t] = \frac{I_r(t)}{\sum_{j=1}^{K} I_j(t)}$$

and the conditional probability of failure from cause r within a short interval given survival up until time t is

$$\phi_r^\lambda(t) = \lim_{\delta t \to 0} P[R = r \mid t < T \leq t + \delta t]$$

$$= \lim_{\delta t \to 0}\left(\frac{P\left[t < T \leq t + \delta t, R = r \mid T > t\right]}{P[T \leq t + \delta t \mid t > t]}\right) = \frac{\lambda_r(t)}{\sum_{j=1}^{K} \lambda_j(t)}$$

5.3★ Estimation of hazard and survivor function

S(t) can be estimated by the usual overall Kaplan-Meier estimate

$$\hat{S}(t)=\prod\nolimits_1^N(1-\tfrac{d_j}{n_j})$$

The naïve Kaplan-Meier estimates

$$\hat{S}_r(t)=\prod\nolimits_1^N(1-\tfrac{d_{rj}}{n_j})$$

are generally not of interest since the comments above shew $1-\hat{S}_r(t)$ is a biased estimate of the probability of failing from cause r by time t. Of more interest is an estimate of $I_r(t)$. We can estimate $\lambda_r(t)$ by

$$\hat{\lambda}_r(t_j)={}^{d_{rj}}\!\big/_{n_j}$$

and if $p_r(t_j)$ is the unconditional probability of failing at t_j then

$$\hat{p}_r(t_j)=\hat{\lambda}_r(t_j)\hat{S}(t_{j-1})\text{, with }\hat{S}(t_0)=1.$$

and the cumulative incidence function for cause r is estimated by

$$\hat{I}_r(t)=\sum_{j;\,t_j\le t}\hat{p}_r(t_j)$$

Standard packages which handle survival data can be used to calculate the naïve Kaplan-Meier estimates but more specialist ones specifically providing competing risks facilities are required or estimation of $I_r(t)$ (see below).

5.4★ Analysis of effects of covariates

Note first that the cause specific hazards are directly estimable standard methods which model the effects of covariates on hazards are readily adapted for use in the presence of competing hazards, provided some care in the interpretation is taken. In the case of a small number (i.e. one or two) factors with two or three levels then it is possible to modify the standard log-rank test to test for equality of groups. For regression on continuous covariates the dependence of the individual hazard rates on covariates can be modelled with Cox proportional hazards techniques. Provided the data are arranged in an appropriate augmented form (see below) it is possible to allow for the effects of covariates to be different or identical for different hazards and to test for equality of effects. The usual arrangement of a data file is that each individual contributes one row, with variables indicating time of failure (or censoring), a cause indicating the cause or censored and the various covariates. Thus the cause variable will have K+1 distinct values. For basic analyses e.g. naïve Kaplan-Meier estimation) the cause variable will be the censoring indicator, defining the value for the event of current interest as the code to indicate the event has occurred (see below). The appropriate augmented data file if there are K competing risks is obtained by repeating each individual's data in K rows but adding a censoring indicator, status, to which has values 0 for all causes other than the one causing failure when it has value1. Censored failure times have 0 for status in all rows. Further details are beyond the scope of this brief introduction.

5.5* Implementation in R

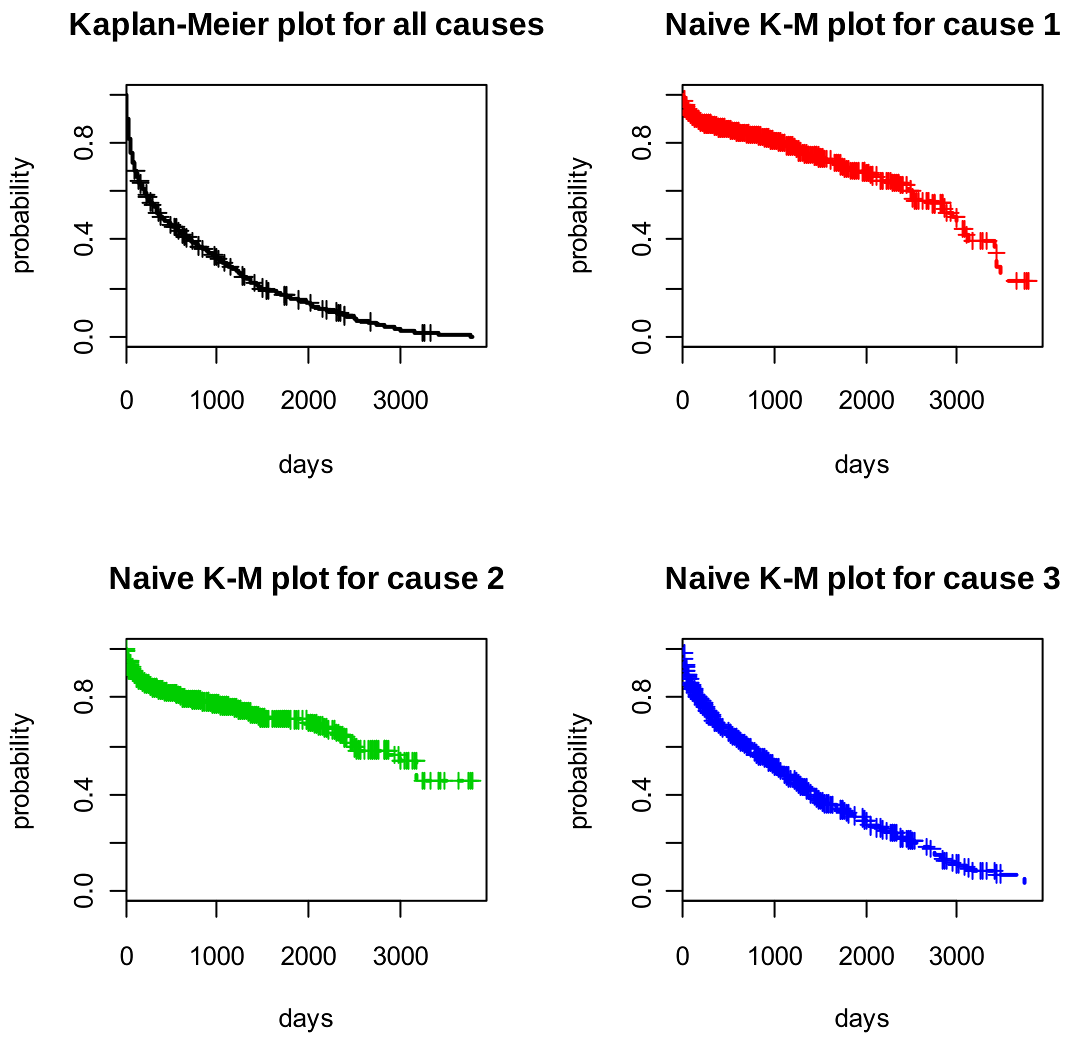
There are a several specialist libraries which provide facilities for handling problems of competing risks. Here we illustrate the library **`cmprsk`**, this has to be downloaded from the CRAN website (or a local mirror site). It automatically loads the **`survival`** library which is bundled with the standard installation of **R** so **`survival`** does not need to be opened first.

5.5.1 Example on organ transplants

These data are semi-artificial. For the purposes of illustration they have been adapted from a much larger data set from a ten year study on survival rates following an organ transplant. The data are provided by NHS Blood and Transplant (www.nhsbt.nhs.uk); codes have been changed and covariates removed. The data set used has 1086 observations of survival times in days (variable **days**) following an organ transplant with three different possible causes of fatality (coded as 1, 2 and 3). The 144 censored observations are indicated by a value 0 for variable **cause**.

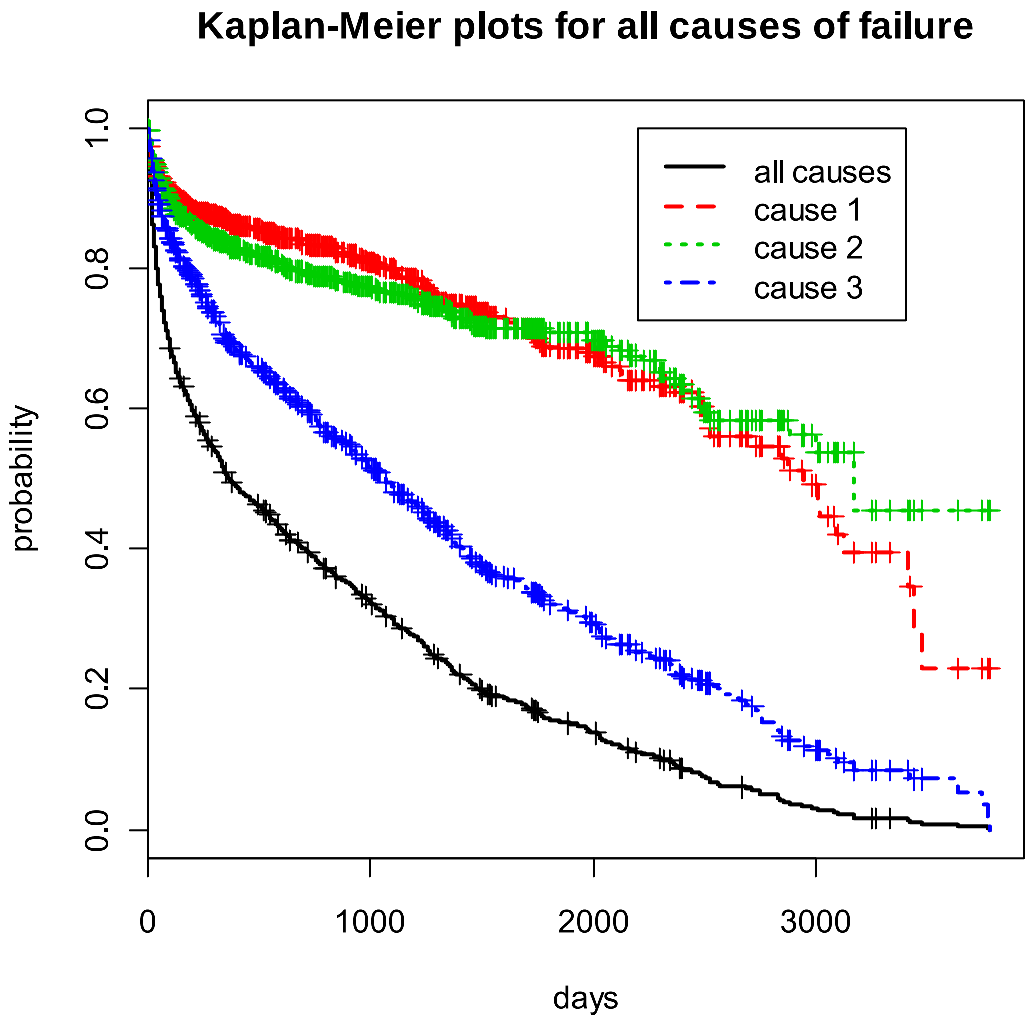
```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

. . . . . . . . . .
. . . . . . . . .
. . . . . . . .

[Previously saved workspace restored]

> library(cmprsk)
Loading required package: survival
Loading required package: splines
> attach(organ)
> organ[1:5,]
  cause days
1     0 1754
2     0   70
3     0   29
4     0 2671
5     0  522
```

```
> par(mfrow=c(2,2))
> plot(survfit(Surv(days,cause!=0)~1),col=1,lty=1,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Kaplan-Meier plot for all causes")
>
> plot(survfit(Surv(days,cause==1)~1),col=2,lty=2,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 1")
>
> plot(survfit(Surv(days,cause==2)~1),col=3,lty=3,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 2")
>
>plot(survfit(Surv(days,cause==3)~1),col=4,lty=4,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 3")
>
```



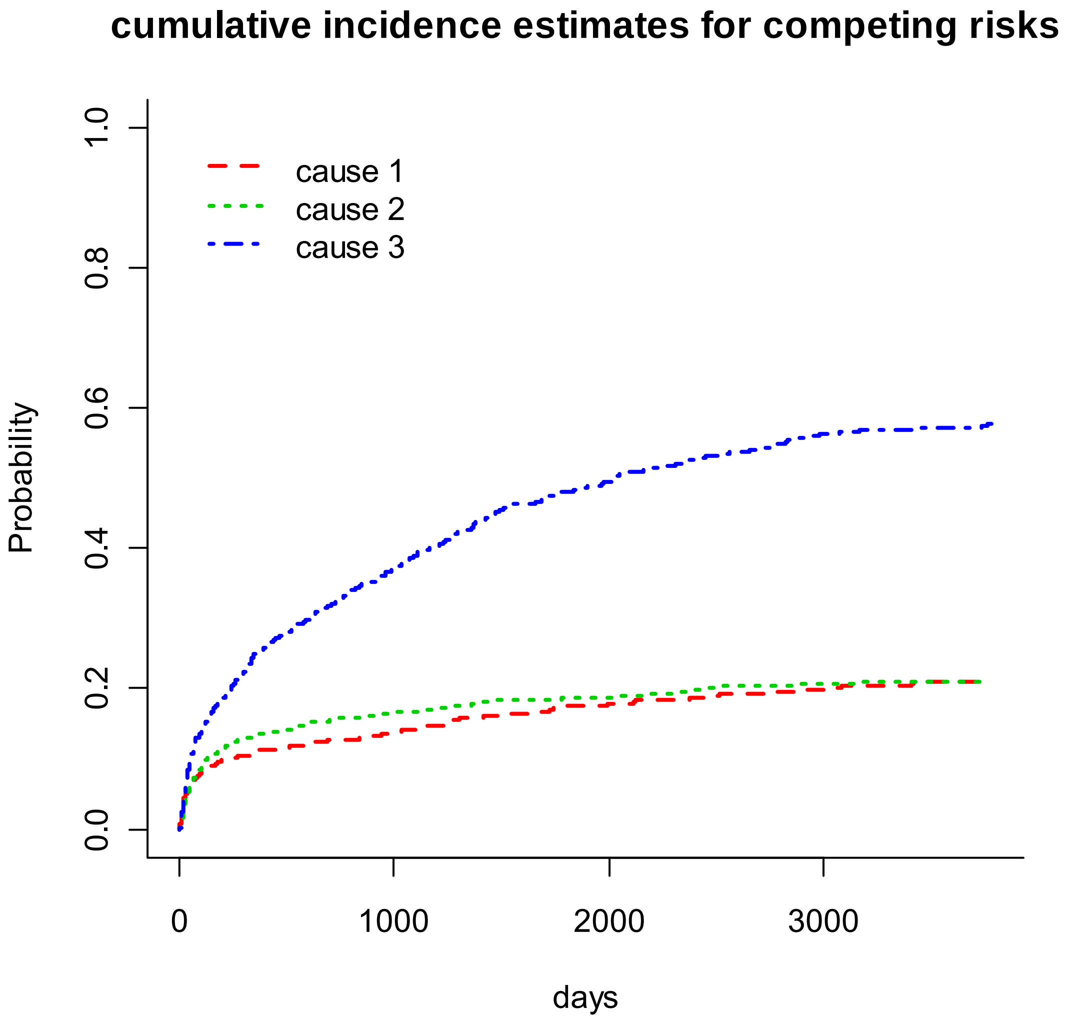
It is clear from the plots above that there is a difference in the survival patterns for the three causes but it is dangerous to draw very specific conclusions on the nature of the patterns since these naïve estimates are biased, possibly to different degrees.

```
>
> par(mfrow=c(1,1))
>
> plot(survfit(Surv(days,cause!=0)~1),col=1,lty=1,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Kaplan-Meier plots for all causes of failure")
>
> lines(survfit(Surv(days,cause==1)~1),col=2,lty=2,lwd=2,conf.int=0)
> lines(survfit(Surv(days,cause==2)~1),col=3,lty=3,lwd=2,conf.int=0)
> lines(survfit(Surv(days,cause==3)~1),col=4,lty=4,lwd=2,conf.int=0)
> leg.txt=c("all causes","cause 1","cause 2", "cause 3")
> legend(2200,1.0,leg.txt,col=1:4,lty=1:4,lwd=2)
>
```



Now we look at the cumulative incidence functions

```
> organ.cmprsk<-cuminc(days, cause)
> plot(organ.cmprsk,
+ main="cumulative incidence estimates for competing risks",
+ xlab="days", lty=2:4,lwd=2,col=2:4,
+ curvlab=c("cause 1", "cause 2", "cause 3"))
>
> curvlab=c("cause 1", "cause 2", "cause 3"))
```



It is clear here that there is a substantial difference between cause 3 and the first two which are similar. Further, it can be seen how different these are from the naïve Kaplan-Meier estimates (which would be reflections about a vertical axis of the plots above).