Chapter 4: Survival Analysis

4.1 What is survival analysis about?

Interest centres on a group or groups of individuals for each of whom there is a defined point event called 'failure' occurring after a length of time called the 'failure time' or 'survival time' T. We assume failure occurs at most once on any individual.

Examples:

- time from diagnosis of cancer to death
- periods of unemployment in economics
- lifetimes of machine components in industrial reliability
- time taken for participants to complete a task in psychology experiments.

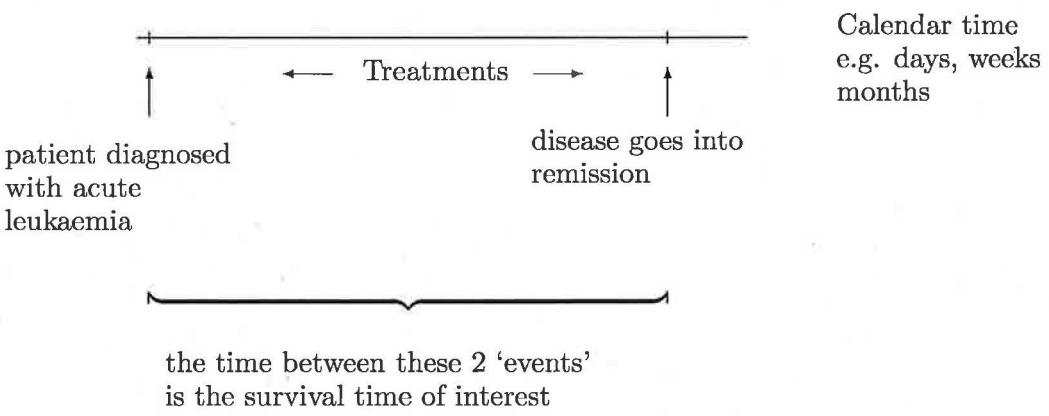
Engineers refer to 'failure time data', the health sciences call it 'survival data', and in sociology, it is called 'event history data'.

Survival analysis is one of the oldest statistical disciplines with roots in demography and actuarial science from the 17^{th} century. See *Encyclopedia of Biostatistics* for numerous articles. Modern developments date from the work by Kaplan and Meier on life-tables in the 1950's, and the Cox model (1972).

Our emphasis will be on biostatistical applications but the theory and methods are more broadly applicable.

4.2 Some Motivating examples

1. Time to remission of leukaemia patients



Questions of interest: What is the probability of patients receiving this treatment going into remission after 6 weeks? Or 12 weeks? Can we compare two or more treatment groups?

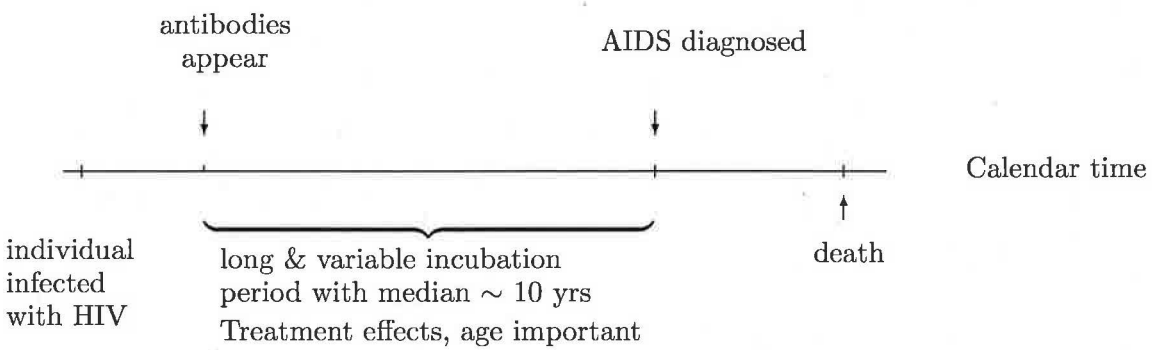
What characteristics of patients are related to the time it takes to go into remission, and whether or not they do go into remission?

Can we predict who will go into remission?

See Example 1.2 on handout from Klein and Moeschberger *Survival analysis: techniques for censored and truncated data*, Second Edition, Springer, 2003. (Hereafter referred to as KM.)

2. Time to AIDS diagnosis

The HIV/AIDS disease process can be described as follows:



We may be interested in the incubation period defined as the time from initial infection to AIDS diagnosis as an alternative to the definition presented in the diagram, and in answering questions about the shape and nature of the incubation distribution. For example, what is the probability of developing AIDS one year following infection? Following 2 years? 5 years? And so on.

How do available treatments affecting this 'AIDS–free' period? We may want to compare survival in two groups. Survival time can also be defined to be the time from AIDS diagnosis to death. Do treatments given in this late stage of the disease process prolong life?

[See Example 1.19 from KM.]

Note that the time scale may not be continuous time.

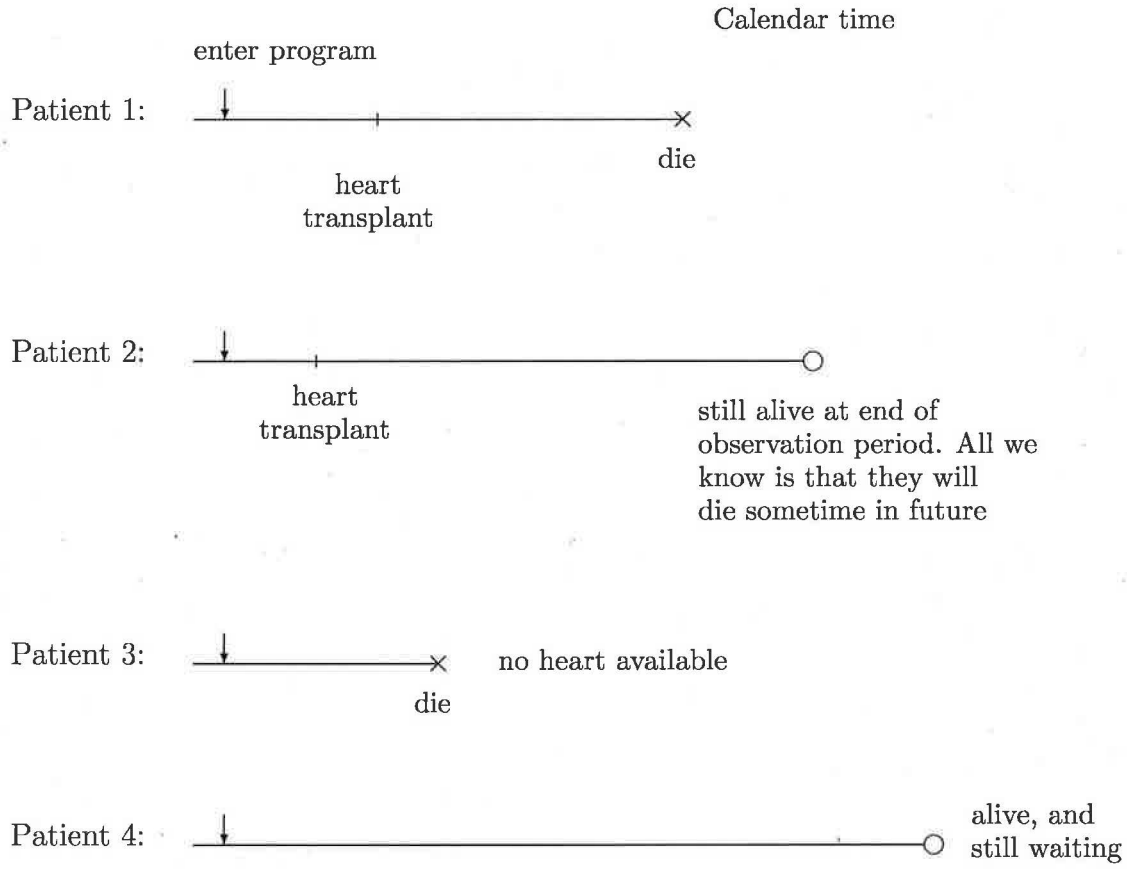
Examples:
- no. of dollars an insurance company pays out in a particular cas
- no. of ovulation cycles before a woman on invitro fertilization treatment achieves pregnancy.

Note too that survival analysis is widely applied to both observational **and** experimental data. Example 1 is of a clinical trial, whereas Example 2 is typically observational data.

3. Time to death following heart transplant

Do heart transplants save lives? This was a big question in the 1980's when they started transplant surgery. It is hard to prove this hypothesis directly.

Eligible patients enter the program and wait... . There are 4 types of patients:



To answer the question, do heart transplants save lives, we need to compare the outcomes of those who do and do not receive transplants.

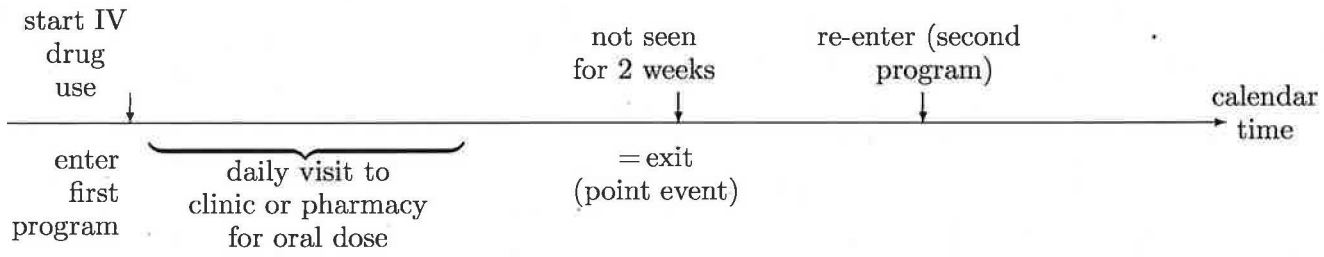
The problem is that patients who do not receive hearts are not a valid 'control group'.
Why not?
Allocation of hearts is not random. A major problem arises with selection effects: people who receive hearts have to survive long enough to get one. Sicker patients may die quickly while waiting, and automatically get pushed into the control group. Moreover, surgeons may choose/select recipients who are more likely to benefit from the transplant, rather than someone who is close to death. Researchers cannot use randomization as it is unethical in this context.

Attempts to get round the problem include the use of time-dependent covariates and matching large databases – these are large registries, and one attempts to compare outcomes with all these who **could** have been selected with similar characteristics. This approach is effective for bone-marrow, kidney and other registries. See the *EOB*.

It is possible that technology may provide 'perfect match' organs via techniques that exploit the nuclear transfer technique that underpins cloning.

4. Multivariate survival times: the South Australian Methadone Program

Methadone is an effective oral substitute for heroin. In past research, I analysed data from 1980-1991 to determine factors affecting retention on the program. Patterns of participation on the program looked like:



How to define survival time? For example, time on P1 alone, or time P2 alone. In fact, we analysed the data as a recurrent event survival problem. How should we adjust for previous time on methadone and other prior events and information, such as age at first injecting drug use?

We will consider *univariate* survival times only.

5. Further examples: see Cox and Oakes (1984).

Section 4.2: Survival time, censoring and truncation

4.2.1 Key notions

Let T be survival, or failure, time. It is a non-negative, univariate response random variable.

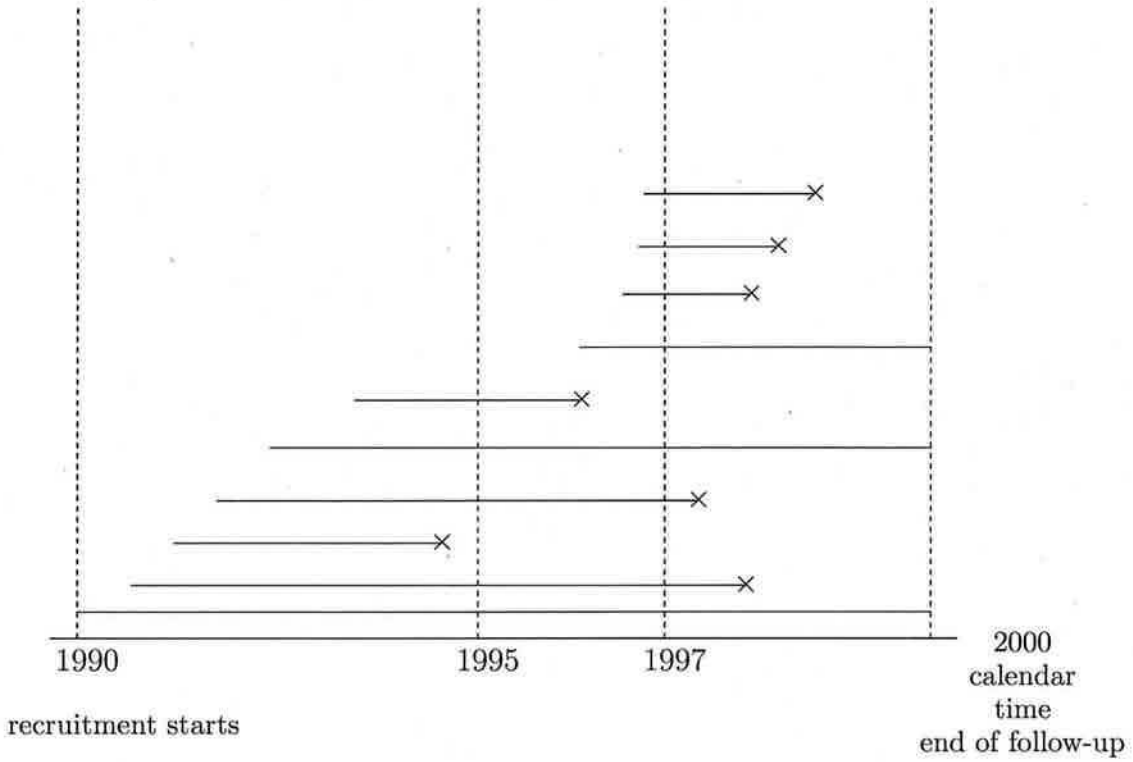
What makes the study of $T \geq 0$ so special? After all, there are many standard statistical techniques for handling non-negative random variables.

1. T is usually highly skewed to the right, and has appreciable dispersion.
2. More importantly, observations are often incomplete, because the study or clinical trial ends before all individuals have experienced the event of interest. The presence of incomplete observations is called censoring, and the most common and important form is called *right censoring.*

The *time origin* should be precisely defined for each individual. It is also desirable that individuals be as comparable as possible at their time origin.

For example, in a randomized clinical trial, the date of randomization is the natural choice. The time origin ($T = 0$) is usually not the same *calendar time* for each individual, and survival time is measured form his/her date of entry to the trial.

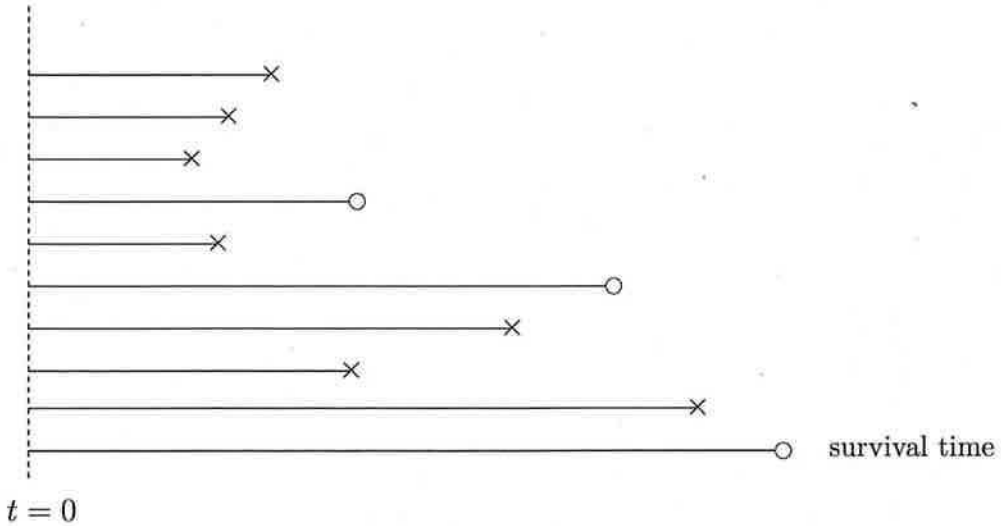
Suppose individuals enter a clinical trial over a substantial period with staggered entry. The experience of 10 individuals followed to the year 2000 might be:



× denotes death

Now translate this to *survival* time. That is, time from randomization (entry into study).

Let o denote censoring.



Note that there are two time scales:

calendar (real) time and survival time.

Note too that censoring presents problems in analysis not discussed in many standard statistical texts and courses.

The time origin need not always be at the point at which the individual enters the study, but if it is not, special methods are needed. For example, in studying the reliability of machine components, some components may already have been in use before observation begins. Another example: in epidemiological studies of the effects on morbidity of occupational exposure to asbestos, the natural measure of time is age since this is such a strong determinant of mortality. However, observation on each individual only commences when he starts work in a job which involves exposure to asbestos. Such data are said to be *left-truncated.*

4.2.2 Other forms of censoring and truncation

(a) Left censoring

Example: See 1.17 on KM handout.
Time to first use of marijuana.

For some boys, all we know is that they started smoking sometime prior to now: survival time is measured as age in years, so this is just saying they started at some (unknown) earlier age. These are known as left-censored observations. All that is known is that the event (start smoking) occurred before some given time/age, τ say.

As another example, in some studies of AIDS survival, all subjects are included, in particular those who died before enrolment in the study.

(b) Interval censoring: here the event of interest occurs in an interval.

e.g. The San Francisco Men's Health Study followed HIV^+ men at 6-monthly intervals. Some men developed AIDS between the 6-monthly clinic visits.

(c) Doubly censored data: these occur if there are combinations of any two types of right, left and interval censoring.

e.g. 1.17 The marijuana data are left and right censored.

Censoring is a form of 'missingness' or 'incompleteness', whereas truncation is a *conditioning* event.

(d) Left truncation: arises where the survival time must *exceed* a certain value to be included in the dataset.

e.g. 1.16 K&M: Ages at death of elderly residents in a retirement community.

Suppose we want to study the survival of residents who are followed to death, or until they leave the home. The life-lengths are left-truncated because an individual must survive

to a sufficient age to enter the retirement community. Individuals who die young are not included. Left truncation does not include everyone, as you don't know who you've missed! It is also called 'delayed entry time', and if ignored can cause a problem known as 'length-biased sampling'.

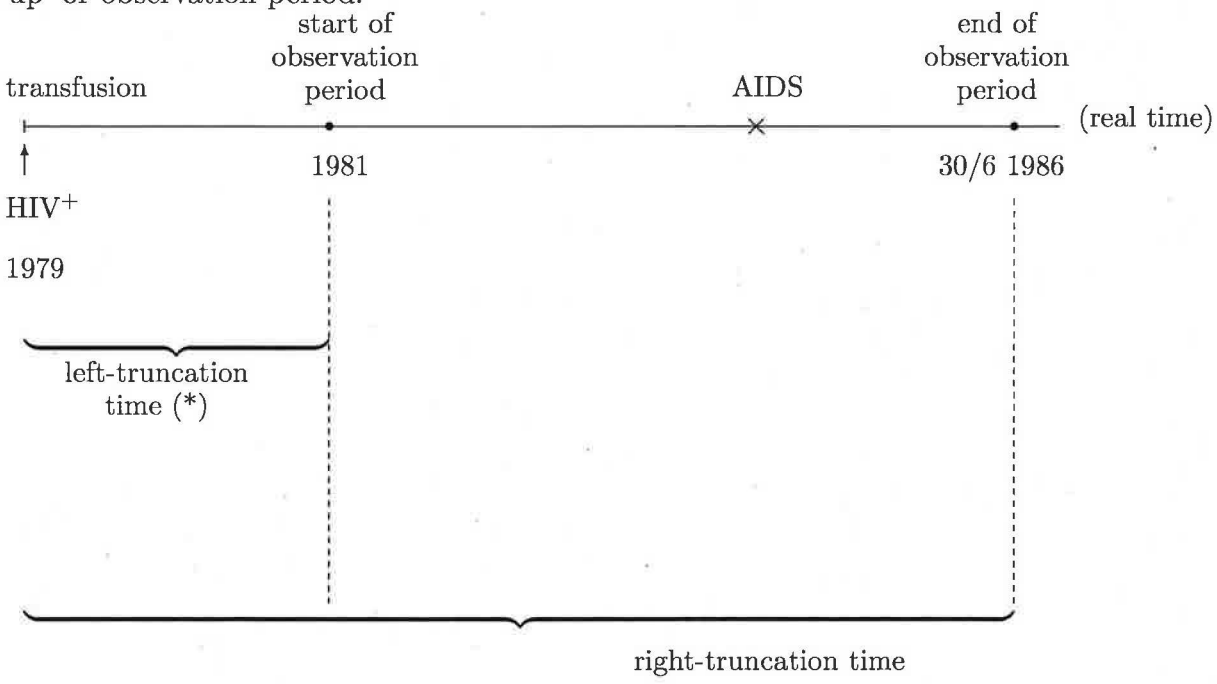
Note that the basic life-table methodology in actuarial studies is what we now consider to be estimation of a survival function from life times with delayed entry (*i.e.* left-truncation) and right censoring.

Note that left censoring involves considering *all* subjects, including those who died (or had the event) before enrolment, whereas left-truncation involves considering only those subjects who survive to enrolment.

(e) Right-truncation: arises when the survival time or event must be smaller than a certain value to be included, *i.e.,* the event has to have occurred by a given time.

Examples: the retrospective determination of transfusion times; for haemophilia patients with AIDS; *Encyclopedia of Biostatistics* p. 4592.

AIDS was not identified until 1981, and the spread of infection was believed to have started in 1979. So as of 1981, only people with incubation periods greater than 2 years would be recorded on the registry. The available data were analysed in mid-1986, the end of the 'follow-up' or observation period:



Why is (*) a left-truncation time? Because people who received a contaminated transfusion but died before 1981, will not necessarily be detected.

The data are also *right-truncated* because the investigator will only know about those contaminated cases that have incubation periods that end prior to mid-1986. The right–truncation time for the above individual is the time between transfusion (*i.e.*, infection) and 30 June 1986. The left-truncation time is the time between transfusion and the start of the observation period, *i.e.* 1981.

§4.3: Distributions of failure time

Consider a homogeneous population. Let T represent continuous failure time for an individual.

There are 3 important requirements:

(i) T is a non-negative random variable;

(ii) the time origin $t=0$ is clearly defined; and

(iii) the time scale is clearly defined.

The probability distribution of T can be specified in a number of mathematically equivalent ways. Three of these are especially useful for survival analysis:

1. Survivor function, $S(t)$.

Probability that T is at least as great as t

ie., $$S_T(t) = P(T \geq t) = 1 - F(t),$$

where $F(t) = \int_0^t f(u)\,du$ is the usual CDF.

Note: $S(0) = 1$, $S(\infty) = 0$; left continuous and monotone decreasing.

2. Probability density function, $f(t)$

$$f_T(t) = -S_T'(t)$$

$$= \lim_{\Delta \to 0^+} P\,\frac{(t \le T < t+\Delta)}{\Delta}$$

ie., $S(t) = \int_t^\infty f(u)\,du$, where

$$f(t) \ge 0, \quad \int_0^\infty f(t)\,dt = 1.$$

3. Hazard function, $h(t)$

Also known as the age-specific failure rate, or the force of mortality. This specifies the instantaneous rate of failure at $T = t$, conditional upon survival to t.

It is defined by

$$h_T(t) = \lim_{\Delta \to 0^+} \frac{P(t \le T < t+\Delta \mid t \le T)}{\Delta}$$

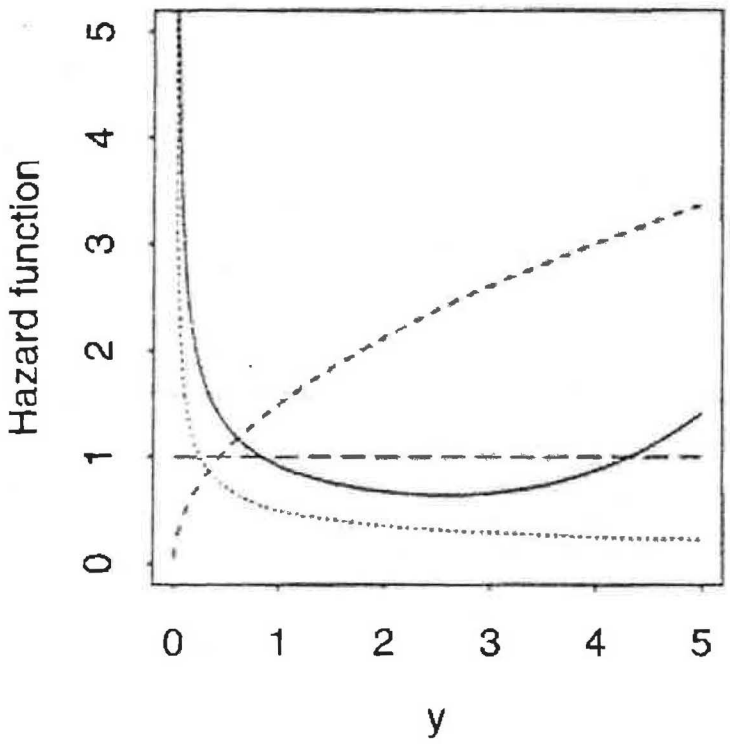
$$= \frac{f_T(t)}{S_T(t)}$$

The hazard function is very useful:

- it describes the way in which the instantaneous probability of failure changes with time;
- it often has direct physical meaning;
- often, prior or external information about the hazard fct. can help select a failure time distribution.

Illustrations: See ACD Figure 8.7 p.188.

'Bathtub-shaped' hazards: often appropriate when individuals are followed from birth to death.



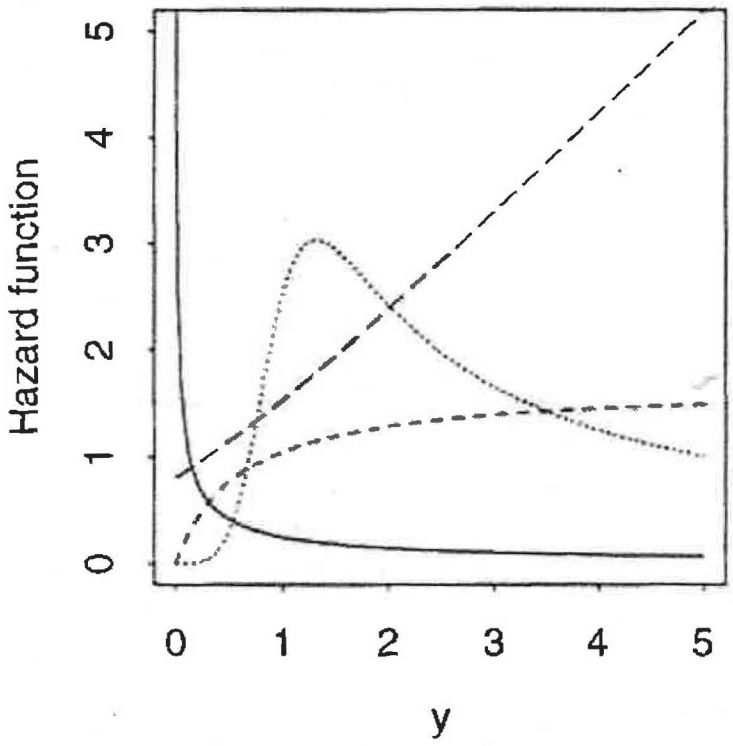


Figure 5.7 Hazard functions. Left panel: Weibull hazards with $\theta = 1$ and $\alpha = 0.5$ (dots), $\alpha = 1$ (large dashes), $\alpha = 1.5$ (dashes), and bi-Weibull hazard with $\theta_1 = 0.3, \alpha_1 = 0.5$, $\theta_2 = \alpha_2 = 5$ (solid). Right panel: Log-logistic hazards with $\lambda = 1$ and $\alpha = 0.5$ (solid), $\alpha = 5$ (dots), gamma hazard with $\lambda = 0.6$ and $\alpha = 2$ (dashes), and standard normal hazard (large dashes).

Monotone increasing hazards : are common. e.g. when interest centres on a period of gradual aging; e.g. period from diagnosis of AIDS to death.

Monotone decreasing hazards : are less common. Certain types of electronic devices have decreasing failure rates in reliability studies. See Weibull dsn for monotone hazards.

There are also other types of hazard:

constant hazard (exponential dsn);

non-monotonic hazards e.g. peak hazard, corresponding to a period of maximum risk

See log logistic, log normal distributions.

Cumulative hazard function, $H(t)$:

Clearly, $h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t)$,

so that $S(t) = \exp\left\{ - \int_0^t h(u)\,du \right\}$

$= \exp\{ - H(t) \}$,

where $H(t) = \int_0^t h(u)\,du$ is called the

cumulative hazard (or integrated hazard) function.

Note: $H(t) = -\log S(t)$ and $f(t) = h(t) \exp\{-H(t)\}$.

Discrete time, T: true discrete survival distributions are rare in practice, but often arise from grouping continuous times.

Now suppose the r.v. T can take values $0 \le t_1 < t_2 < \dots$, and let the probability function be

$$P(T = t_i) = f_i \quad , \quad \sum_i f_i = 1.$$

The survivor function is then

$$S(t) = P(T \ge t) = \sum_{i\,:\,t_i \ge t} f_i$$

$$= f_i + f_{i+1} + \cdots .$$

$S(t)$ is a monotone decreasing, left-continuous fct with $S(0) = 1$, $S(\infty) = 0$.

We may now write the hazard function as

$$h(t) = \sum_i h_i S(t - t_i) , \quad \text{where}$$

$$h_i = P(T = t_i \mid T \geq t_i) = \frac{f_i}{S(t_i)} , \quad i = 1, 2, \ldots$$

(note: $h_i S(y - t_i) = h_i$ when $y = t_i$).

Now, $f_i = S(t_i) - S(t_{i+1})$

$$\Rightarrow h_i = 1 - \frac{S(t_{i+1})}{S(t_i)}$$

$$\Rightarrow S(t_{i+1}) = (1 - h_i) S(t_i) , \quad i = 1, 2, \ldots$$

Substituting recursively gives

$$S(t) = \prod_{i : t_i < t} (1 - h_i) .$$

We define the cumulative hazard as

$$H(t) = -\sum_{i:t_i<t} \log(1-h_i)$$

which as before, gives $S(t) = \exp\{-H(t)\}$.

It can be shown that when the h_i are small,

$$H(t) \simeq \sum_{i:t_i<t} h_i$$

which is the 'natural' definition.

Remarks:

1. $S(t)$ need not $\to 0$ as $t \to \infty$, continuous T, putting positive probability on an infinite survival time. This can happen in practice if, for example, the endpoint for a study is death from a disease, but complete recovery is possible.

2. Survival distributions can be a mix of discrete and continuous random variables e.g., a diagnosis of AIDS after death, but we will not consider these.

Notation: suppose in the absence of censoring that the i^{th} individual has failure time T_i. Suppose observation ceases at time c_i, if the patient hasn't failed by then.

Then observations consist of

$$X_i = \min(T_i, c_i)$$

together with an indicator variable

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq c_i \quad \text{(uncensored)} \\ 0 & \text{if } T_i > c_i \quad \text{(censored)} \end{cases}$$

Types of censoring:

Type I censoring: observation ceases at a pre-determined time, c. This is the simplest form of censoring. Suppose we observe n subjects until c_1 fixed; observe $t_1, \ldots, t_r$ failures ($r < n$) and $(n-r)$ censored values, where all we know is that their failure times lie beyond c. r is the random

variable, and all $c_i = c$.

Type II censoring: observation ceases after a pre-determined number, r, of failures, r fixed. The remaining $(n-r)$ observations are censored, and exceed $T_{(r)}$, the r^{th} observed failure time. The c_i here are random variables. This scheme is typically used in industrial life-testing.

Random censoring: Type II censoring is an example. More generally, suppose the i^{th} of n independent subjects has an associated censoring time C_i drawn from a distribution G, independent of its survival time, T_i.

Write $X_i = \min(T_i, C_i)$, C_i, T_i indep.

Crucial assumption: censoring must be uninformative about failure.

Random censoring often arises in medical studies.

Potential for serious bias if crucial assumption not satisfied: suppose the sickest patients drop out of a trial because the treatment makes them feel worse : this would induce association between survival and censoring variables because patients die soon after they withdraw.

Important remarks:

1. We often use asymptotic procedures in practice which do not require explicit recognition of the censoring process - we analyse the c_i as 'pre-determined constants'.

2. So far, we have considered right censoring. Observations can also be left censored, and left- or right-truncated.

   Left censoring: time of origin is not known exactly e.g. we may observe time to death from AIDS, but the time of infection

is unknown.

Interval-censoring: failure time lies somewhere in an interval between two values.

Censoring is to be distinguished from left or right truncation: these are values which lie outside the range, so they are never seen or never recorded if they are seen. Deaths from blood-transfusion-caused cases of AIDS in the 1970's were truncated values.

See §4.2.

§4.4 Likelihood inference

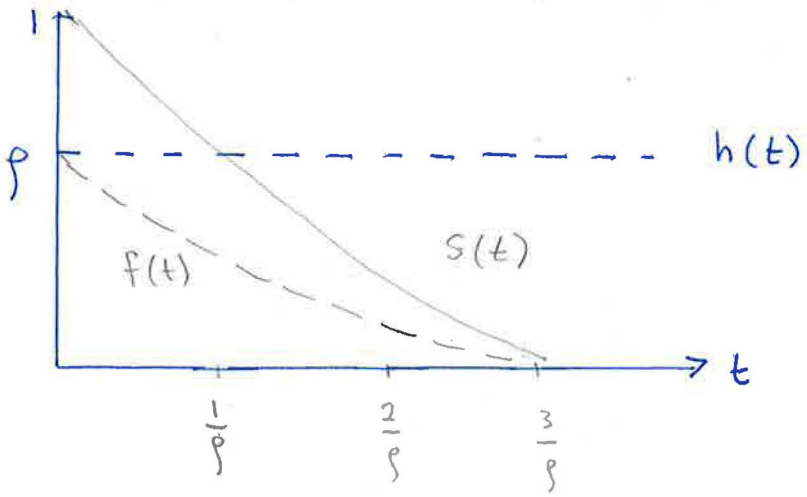
§4.4.1: Exponential distribution, no censoring

When $f(t) = \rho e^{-\rho t}$, $\rho > 0$,

the hazard function is

$$h(t) = \rho, \quad \text{a constant.}$$

Graphically:



Changing ρ changes scale of t.

Consider a single sample: $i = 1, \ldots, n$

In the absence of censoring,

$$L = \prod_{i=1}^{n} f(t_i; \rho) = \prod_{i=1}^{n} \rho e^{-\rho t_i}$$

$$= \rho^n e^{-\rho \sum_{i=1}^n t_i} .$$

The log likelihood is

$$\ell = n \log \rho - \rho \sum_{i=1}^n t_i$$

$$\Rightarrow \frac{\partial \ell}{\partial \rho} = \frac{n}{\rho} - \sum_{i=1}^n t_i$$

so that $\frac{\partial \ell}{\partial \rho} = 0 \Rightarrow \hat{\rho} = \frac{n}{\sum_{i=1}^n t_i}$.

Observe that $\sum_{i=1}^n t_i$ is a minimal sufficient statistic for ρ.

$\sum_{i=1}^n t_i$ is the sum of n i.i.d. exponential random variables with density $\rho e^{-\rho t}$, so that $y = \sum_{i=1}^n t_i$ has the gamma density function

$$\frac{\rho^n}{\Gamma(n)} y^{n-1} e^{-\rho y}$$

and

$$2\rho \sum_{i=1}^n t_i \sim \chi^2_{2n} .$$

Equivalently,

$$2n \frac{\theta}{\hat{\theta}} \sim \chi^2_{2n}$$

and this can be used for hypothesis testing and constructing confidence intervals.

That is, a $(1-\alpha)\%$ confidence interval for θ is

$$\frac{\hat{\theta}}{2n} c^*_{2n,\,1-\frac{\alpha}{2}} < \theta < \frac{\hat{\theta}}{2n} c^*_{2n,\,\frac{\alpha}{2}}$$

where $c^*_{p,\alpha}$ is the upper $\alpha\%$ point of χ^2_p.

§4.4.2 : Type I censoring

Observation stops at a pre-determined time c.

Let $X = \min(T, c)$, $D = I(T \leq c)$.

Then each observation is a pair (x, d)

and contributes a term to the likelihood of the form

$$f(x;\theta)^d \, S(x;\theta)^{1-d} \quad , \; d = 0, 1 .$$

So the likelihood

$$L = \prod_{i=1}^{n} \left[f(x_i;\theta)^{d_i} S(x_i;\theta)^{1-d_i} \right]$$

$$= \prod_{\text{uncens.}} f(t_i;\theta) \prod_{\text{cens.}} S(c;\theta).$$

For r observed failures, we write

$$L = \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(c;\theta).$$

Exponential distribution revisited:

$$L = \prod_{i=1}^{r} \rho e^{-t_i} \prod_{i=r+1}^{n} e^{-\rho c}$$

$$= \rho^r e^{-\rho \sum_{i=1}^{n} t_i} e^{-\rho(n-r)c}.$$

[Note that r is a random variable.]

Then

$$\ell = r \log \rho - \rho \left\{ \sum_{i=1}^{r} t_i + (n-r)c \right\}$$

and

$$\frac{\partial \ell}{\partial \rho} = 0 \Rightarrow \frac{r}{\rho} - \left\{ \sum_{i=1}^{r} t_i + (n-r)c \right\} = 0$$

i.e. $$\hat{\rho} = \frac{r}{\sum_{i=1}^{n} t_i + (n-r)c}$$

The numerator r is the number of observed failures, and the denominator is the observed total time at risk $\sum_{i=1}^{n} x_i$.

Together, $(r, \sum_{i=1}^{n} x_i)$ form a minimal sufficient statistic for ρ. So unless r or Σx_i or some function of them is fixed by design, we have a 2-dimensional statistic for a one-dimensional parameter; this is an example of a curved exponential family.

The exact dsn of $\hat{\rho}$ is difficult to derive, but good approximate procedures for general censoring patterns can be obtained by treating

$$2r\frac{\rho}{\hat{\rho}} \quad \text{as a} \quad \chi^2_{2r} \quad \text{random variable.}$$

(ie. ignoring the fact that r is a r.v.). See shortly.

Aside: for Type II censoring, can show

$$\hat{\rho} = \frac{r}{\sum_{i=1}^{n} t_i + (n-r)\, t_{(r)}} = \frac{r}{W}, \text{ say}$$

where r is now fixed, and $t_{(r)}$ is the r^{th} observed (ordered) failure time.

It is straightforward to show that the exact sampling distribution of $2\rho W$ is χ^2_{2r},

and again exact inference is possible.

§ 4.4.3 : Random censoring

Suppose we have data $(x_1, d_1), \ldots, (x_n, d_n)$ on n individuals. Now assume a random censoring mechanism for which the censoring variables have density and distribution functions g and G, respectively.

Then individual i contributes the following to

to the likelihood

$$f(x_i;\theta)\{1-G(x_i)\} \quad \text{if} \quad d_i = 1 \quad , \text{ and}$$

$$s(x_i;\theta)g(x_i) \quad \text{if} \quad d_i = 0 \; .$$

Thus the likelihood is

$$L = \prod_{i=1}^{n} \left([f(x_i;\theta)\{1-G(x_i)\}]^{d_i} [s(x_i;\theta)g(x_i)]^{1-d_i} \right)$$

with log likelihood

$$\ell(\theta) = \sum_{i=1}^{n} d_i \log \left(f(x_i;\theta)\{1-G(x_i)\} \right) + \sum_{i=1}^{n} (1-d_i) \log \left(s(x_i;\theta)g(x_i) \right).$$

If the censoring mechanism does not depend on θ, then $g(x_i)$ and $G(x_i)$ are constant, and

$$\ell(\theta) \equiv \sum_{\text{uncens.}} \log f(x_i;\theta) + \sum_{\text{cens.}} \log s(x_i;\theta),$$

where the sums are over censored and uncensored individuals. This amounts to treating the censoring pattern as fixed, and includes Type I

censoring, which puts all its probability at c.

The log likelihood function can be written in terms of its hazard and cumulative hazard functions:

$$\ell(\theta) \equiv \sum_{i=1}^{n} \{ d_i \log h(x_i;\theta) - H(x_i;\theta) \}.$$

Note that since the calculation of expected information involves assumptions about the censoring mechanism, standard errors for parameter estimates are based on the observed information.

Example: exponential distribution again

r observed deaths ; $X_i = \min(T_i, C_i)$.

Check: $\ell = \sum_{u} \log \rho - \rho \sum_{all} x_i$

where $\sum_{all} x_i$ is the total of censored and uncensored times, equal to the total time at risk.

Then $\frac{\partial \ell}{\partial \rho} = \frac{r}{\rho} - \sum_{i=1}^{n} x_i$

which gives

$$\hat{\rho} = \frac{r}{\sum_{i=1}^{n} x_i} .$$

The observed information is

$$-\frac{\partial^2 \ell}{\partial \rho^2} = \frac{r}{\rho^2} .$$

Hence the estimate of ρ is zero if there are no failures, and censored data contribute no information about ρ.

Asymptotically, $\text{var}(\hat{\rho}) \simeq \frac{\rho^2}{r}$, with

s.e. $\frac{\rho}{\sqrt{r}}$.

Thus we could construct a normal theory confidence interval for ρ in the usual way. This approx. c.i. is symmetric, which may not be appropriate (see example 4.4.4).

We instead, in general, obtain good approximate procedures for general censoring patterns

by treating $2r\frac{\rho}{\hat{\rho}}$ as a χ^2_{2r} variable,

and constructing confidence intervals as before.

Expected information: consider Type I censoring,

and let $D = I(T \le c)$.

Then $E\left(\sum_{j=1}^{n} D_j\right) = n\,P(X \le c) = n(1-e^{-\rho c})$,

so that $E\left(-\frac{\partial^2 \ell}{\partial \rho^2}\right) = E\left(\frac{r}{\rho^2}\right) = \frac{n(1-e^{-\rho c})}{\rho^2}$.

Letting $c \to \infty$, we can obtain the expected information when there is no censoring,

$$I_\infty(\rho) = \frac{n}{\rho^2}.$$

Therefore, the relative efficiency when there is censoring at c is

$$\frac{I_c(\rho)}{I_\infty(\rho)} = \frac{n(1-e^{-\rho c})/\rho^2}{n/\rho^2} = 1-e^{-\rho c}.$$

This is the proportion of uncensored data (this is not surprising, as we know that censored observations do not contribute to the observed information).

Note that the loss of information becomes more severe as c decreases.

Example 4.4.4 : see handout from Cox & Oakes.

Table 1.1 *Times of remission (weeks) of leukaemia patients (Gehan, 1965, from Freireich et al.)*

Sample 0 (drug 6-MP)	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Sample 1 (control)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

* Censored

Table 1.3 *Cycles to failure (in units of 10^3 cycles) of springs*

Stress (N/mm^2)										
950	225	171	198	189	189	135	162	135	117	162
900	216	162	153	216	225	216	306	225	243	189
850	324	321	432	252	279	414	396	379	351	333
800	627	1051	1434	2020	525	402	463	431	365	715
750	3402	9417	1802	4326	11520*	7152	2969	3012	1550	11211
700	12510*	12505*	3027	12505*	6253	8011	7795	11604*	11604*	12470*

* Censored

the total number of failures divided by the total time at risk. Censored failure times contribute to the denominator but not to the numerator of this ratio.

When there is no censoring, the log likelihood becomes

$$l = n \log \rho - \rho \sum x_i,$$

and the curved exponential family collapses to a full one-dimensional family with the single minimal sufficient statistic $\sum x_i$ for ρ. Here, exact inference for ρ is possible, because $\sum x_i$, the sum of n independent exponentially distributed random variables with the same parameter ρ, has a gamma distribution with index n and scale parameter ρ. Thus $2n\rho/\hat{\rho}$ has a chi-squared distribution with $2n$ degrees of freedom. Interval estimates and hypothesis tests for ρ follow immediately. In particular, a $1-\alpha$ confidence interval for ρ is

$$\frac{\hat{\rho} c^*_{2n,1-\frac{1}{2}\alpha}}{2n} < \rho < \frac{\hat{\rho} c^*_{2n,\frac{1}{2}\alpha}}{2n},$$

where $c^*_{p,\alpha}$ is the upper α point of the chi-squared distribution with p degrees of freedom.

Example 3.1

Consider the leukaemia data of Freireich *et al.* given in Table 1.1. For the control group, with no censoring, $n = 21$ and $\sum x_i = 182$. If an exponential distribution is assumed

$$\hat{\rho} = 21/182 = 0.115,$$

and an exact 95% confidence interval for ρ is (0.071, 0.170), since the 0.025 and 0.975 points of the chi-squared distribution with 42 degrees of freedom are respectively 26.0 and 61.8.

The exact theory holds also with Type II censoring, that is when observation ceases after a predetermined number of failures, d. This is easily seen by noting that if $t_{(i)}$ denotes the ith ordered failure time ($i = 0, 1, \ldots, d$: $t_{(0)} = 0$), then $t_{(i)} - t_{(i-1)}$ has an exponential distribution with parameter $(n-i+1)\rho$ and that

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{d} (n-i+1)(t_{(i)} - t_{(i-1)}),$$

with this censoring mechanism.

With other censoring patterns, the exact sampling distribution of $\hat{\rho}$ is difficult to derive. It has been tabulated for the special case of Type I censoring, when observation on all individuals ceases at a predetermined time c. However, good approximate procedures for general censoring patterns can be obtained by treating $2d\rho/\hat{\rho}$ as a chi-squared variable on $2d$ degrees of freedom, ignoring the fact that d is now a random variable. The resulting confidence intervals are very similar to those obtained from the likelihood ratio.

Example 3.1 (continued)

For the treated group (6-MP), $\sum x_i = 359$, $d = 9$. The maximum likelihood estimator is

$$\hat{\rho} = d/\sum x_i = 9/359 = 0.025.$$

The log likelihood function is plotted in Fig. 3.1 and shows a noticeable lack of symmetry. The 95% confidence interval for ρ from the likelihood ratio is (0.0120, 0.0452). The interval obtained from the upper and lower 0.025 points of the chi-squared distribution with 18

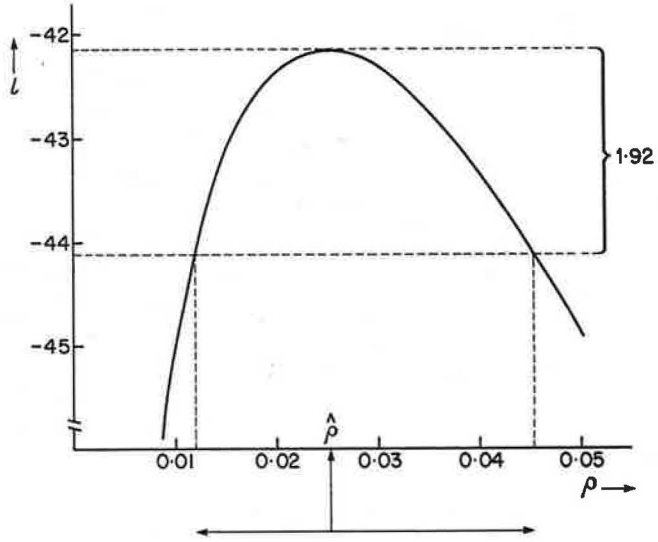


Fig. 3.1. Leukaemia data, 6-MP group. Log likelihood function, l, for exponential parameter, ρ. 95% confidence interval derived from chi-squared distribution. Maximum likelihood estimate, $\hat{\rho}$.

degrees of freedom is (0.0115, 0.0439). The standard error of the maximum likelihood estimate is

$$\left(\left[-\frac{\partial^2 l}{\partial \rho^2}\right]_{\hat{\rho}}\right)^{-1/2} = \left(\frac{\hat{\rho}^2}{d}\right)^{1/2} = 0.00836.$$

This gives a symmetric 95% confidence interval, based on a normal approximation to the distribution of $\hat{\rho}$, of (0.0087, 0.0414). In view of the shape of the likelihood function, this symmetric interval would be an inappropriate choice here.

§4.4.5 Weibull distribution

This 2-parameter distribution is the most widely used failure time distribution.

Survivor function: $S(t) = e^{-(\rho t)^{\kappa}}$, $t > 0$;

scale parameter ρ changes the scale on the horizontal axis; $\rho > 0$

index parameter κ determines shape: usually ranges from 1 to 3, $\kappa > 0$.

Note that $\kappa = 1$ returns the exponential dsn;

t^{κ} has an exponential dsn.

Comparing 2 groups: do they have the same hazard?

Group 1: $\underbrace{t_1, \ldots, t_r}_{\text{obs. deaths}}, \underbrace{t_{r+1}, \ldots, t_m}_{\text{censored}}$ hazard ρ

Group 2: $\underbrace{\nu_1, \ldots, \nu_s}_{\text{obs. deaths}}, \underbrace{\nu_{s+1}, \ldots, \nu_n}_{\text{censored}}$ hazard $\theta\rho$

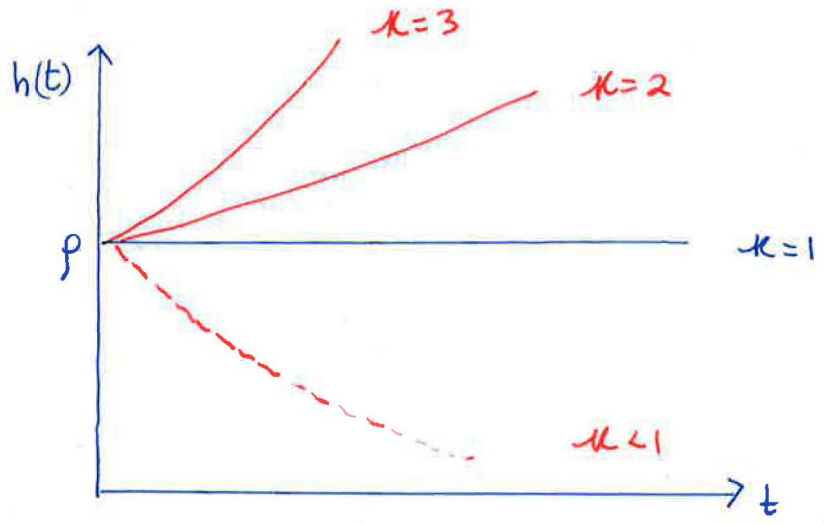
Want to test $H_0: \theta = 1$ i.e. the 2 groups have the same hazard.

The obvious estimate of θ is the hazard ratio

$$\hat{\theta} = \frac{\frac{s}{\sum_{i=1}^{n} \nu_i}}{\frac{r}{\sum_{i=1}^{m} t_i}}$$

with $\operatorname{var}(\log \hat{\theta}) \simeq \frac{1}{r} + \frac{1}{s}$ for 2 independent groups (see Biostatistics III).

Therefore, can construct tests, confidence intervals, etc.



$\kappa > 1$ corresponds to monotone increasing hazards,

$\kappa < 1$ " " " decreasing " .

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

ie. $\frac{d}{dt} \log S(t) = h(t)$.

The integrated hazard

$$H(t) = (\rho t)^{\kappa},$$

so that $h(t) = \frac{d}{dt}(\rho t)^{\kappa} = \kappa \rho^{\kappa} t^{\kappa-1}$.

The density function

$$f(t) = h(t) S(t)$$
$$= \kappa \rho^{\kappa} t^{\kappa-1} e^{-(\rho t)^{\kappa}}.$$

Can transform $z = t^{\kappa}$.

$$f(z) = \varphi^{\kappa} e^{-\varphi^{\kappa} z} \qquad \text{(check!)}$$

$$= \gamma e^{-\gamma z} \quad , \text{ say },$$

where $\gamma = \varphi^{\kappa}$.

So if we know κ, can transform to exponential.

Consider data from a single sample:

$$L = \prod_{u} f(x) \prod_{c} S(x) .$$

$$\Rightarrow \ell = r \log \kappa + \kappa r \log \varphi + (\kappa - 1) \sum_{u} \log(x_i) - \varphi^{\kappa} \sum_{i=1}^{n} x_i^{\kappa} ;$$

check as exercise!

Even in the absence of censoring, there is no fixed dimensional sufficient statistic for (φ, κ); the Weibull dsn is not an exponential family. (Can you show this?).

The first derivatives are

$$\frac{\partial \ell}{\partial \rho} = \frac{\kappa r}{\rho} - \kappa \rho^{\kappa-1} \sum_{i=1}^{n} x_i^{\kappa} ,$$

$$\frac{\partial \ell}{\partial \kappa} = \frac{r}{\kappa} + r \log \rho + \sum_{u} \log x_i - \rho^{\kappa} \sum_{i=1}^{n} x_i^{\kappa} \log(\rho x_i) .$$

If κ is specified, the m.l.e. of ρ can be found from $\left.\frac{\partial \ell}{\partial \rho}\right|_{\kappa} = 0 \Rightarrow \hat{\rho} = \left(\frac{r}{\sum_{i=1}^{n} x_i^{\kappa}} \right)^{1/\kappa}$.

This follows immediately from above from the fact that T^{κ} is exponential with rate ρ^{κ}.

Substitute $\hat{\rho}$ into $\frac{\partial \ell}{\partial \kappa}$ to get

$$0 = \frac{r}{\kappa} + \sum_{u} \log x_i - r \frac{\sum_{i=1}^{n} x_i^{\kappa} \log x_i}{\sum_{i=1}^{n} x_i^{\kappa}}$$

which we solve to get the m.l.e. $\hat{\kappa}$.

Note that this equation does not involve ρ and

can be solved by a one-dimensional iterative scheme in κ.

The second derivatives of l are:

$$I_{\beta\beta} = -\frac{\partial^2 l}{\partial \beta^2} = \frac{\kappa r}{\beta^2} + \kappa(\kappa-1)\beta^{\kappa-2}\sum_{i=1}^{n} x_i^{\kappa},$$

$$I_{\beta\kappa} = -\frac{\partial^2 l}{\partial \beta \partial \kappa} = -\frac{r}{\beta} + \beta^{\kappa-1}(1+\kappa\log\beta)\sum_{i=1}^{n} x_i^{\kappa} + \kappa\beta^{\kappa-1} \times \sum_{i=1}^{n} x_i^{\kappa}\log x_i$$

$$I_{\kappa\kappa} = -\frac{\partial^2 l}{\partial \kappa^2} = \frac{r}{\kappa^2} + \beta^{\kappa}\sum_{i=1}^{n} x_i^{\kappa}\{\log(\beta x_i)\}^2.$$

check these!

The complementary log-log transform:

See shortly.

§4.5 : a test for exponentiality

Suppose we want to test

$H_0 : \kappa = 1$, ρ arbitrary

against

$H_1 : (\rho, \kappa)$ arbitrary.

In other words, assuming a Weibull dsn, is it actually exponential? This is a useful test against alternative hypotheses which specify monotone hazard functions.

Could use (i) Likelihood ratio test.

ie., is -2(difference in likelihoods) $> \chi^2_1$?

For this we would need the joint m.l.e..

(ii) Score test.

ie., find $\frac{\partial l}{\partial \kappa}$ at $\kappa = 1, \hat{\rho}_{\kappa=1}$ and ask if

it's close enough to zero.

The m.l.e. of $\hat{\rho}_\kappa$ of ρ when $\kappa = \kappa_0 = 1$ is

$$\hat{\rho}_{\kappa_0} = \frac{r}{\Sigma x_i} .$$

The score function $U_{\kappa_0} = \left[\frac{\partial}{\partial \kappa} l(\kappa, \rho) \right]_{\kappa_0, \rho = \hat{\rho}_{\kappa_0}}$

$$= r + \sum_u \log x_i - r \frac{\Sigma x_i \log x_i}{\Sigma x_i} .$$

The observed information matrix at $(\kappa_0, \hat{\rho}_{\kappa_0})$ has elements

$$I_{\kappa\kappa} = r + \Sigma (\hat{\rho}_{\kappa_0} x_i) [\log (\hat{\rho}_{\kappa_0} x_i)]^2 ,$$

$$I_{\kappa\rho} = \Sigma x_i \log (\hat{\rho}_{\kappa_0} x_i)$$

$$I_{\rho\rho} = \frac{r}{\hat{\rho}_{\kappa_0}^2} .$$

The inverse matrix v has leading element $(1,1)$

$$v_{\alpha\alpha} = \left(I_{\alpha\alpha} - I_{\alpha\beta}^2 / I_{\beta\beta} \right)^{-1}$$

(check!)

$$= r \left\{ 1 + \frac{\Sigma x_i (\log x_i)^2 - (\Sigma x_i \log x_i)^2}{(\Sigma x_i)^2} \right\}$$

Check as exercise!

Thus, the score test statistic is

$$\frac{U_{\alpha_0}}{\{\text{var}.(U)_{\alpha_0}\}^{\frac{1}{2}}} = U_{\alpha_0} (v_{\alpha\alpha})^{\frac{1}{2}}$$

for $p_w = 1$ (single parameter test), which is approximately $N(0,1)$ under the null hypothesis.

Note that in general, U_w has covariance matrix v_{ww}^{-1}, for parameter vector w.

Leukaemia data revisited: see Assignment 4.