MAS3301 / MAS8311 Biostatistics
Part II: Survival

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2009-10

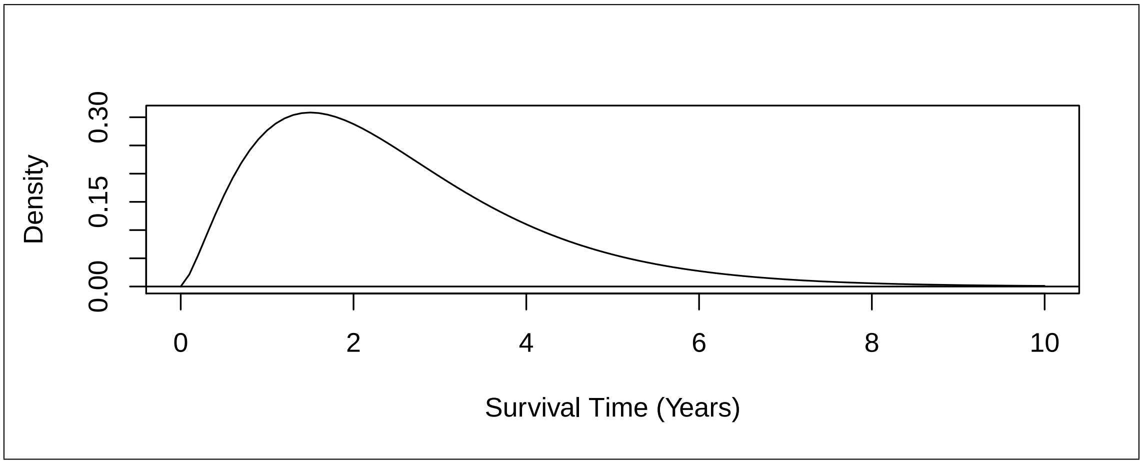


Figure 1: Typical survival pdf.

1 Introduction

1.1 Survival Data

In this part of the course we study methods for analysing data that come from how long people live, or more generally, the length of time taken until a certain event occurs. So *'survival data'* might be data on how long people survive, e.g.

- time till death following heart transplant,
- time till death following diagnosis with AIDS,

although the event which terminates the time interval need not be death:

- length of time in remission for leukemia patients,
- time till rejection of a transplanted organ,
- time till discharge from hospital following an operation.

More formally, the data arise when the time from a defined origin until the occurrence of a predetermined event (often called failure) is measured for each subject. The failure event can occur at most once for each subject. The time taken till failure is referred to as the *failure time* or *survival time*.

What sort of problem does survival analysis address? Suppose there are two different surgical interventions for a certain heart abnormality. Do patients live longer under one or other of the treatments? To assess this we look at the survival times for patients under each of the two treatments and use statistics to decide which one is best.

Why can't we just use methods we've learned on other courses? What's so special about this sort of data?

- Survival time distributions are highly skewed.
- Often the times are censored – the exact time for every individual is not always available to us.

1.2 Basic Features

Let's consider a typical example: a study of survival times of patients following a heart transplant. Patients have transplants at different times, and we recruit patients to our study over a period of a few months (for example). For each patient we know when they had a transplant and when they died. See the graphs drawn in the lecture explaining the difference between *study time* and *patient time*.

Figure 2: Time till death following heart transplant.

Figure 3: Patient time.

Figure 4: Censoring.

1.3 Censoring

In the example above we assumed that we knew exactly when each patient had their transplant and when they died. In practice this will not normally be the case: the data will be incomplete or *censored.* Handling censored data is a key aspect of survival analysis.

How and why are survival data often censored? The simplest reason is because the event we are interested in does not occur within the duration of our study (as above). In medical examples a common reason is 'loss to follow-up' eg. patient X moved away to Australia. However, we often need to think carefully about whether censoring can *bias* the dataset (see informative censoring below). Types of censoring can be classified according to information that they provide:

Right censoring: This is the most common type, and is the type illustrated by the example above. Suppose the patient enters our study at time t_0. and dies at time $t_0 + t$. If the study ends at time $t_0 + c$ where $c < t$ then we know only that the individual's time of death satisfies $t > c$. The death time for this patient is *right censored.* Most of the examples considered in this course will involve right censoring. Possible reasons for right censoring:

- end of study
- 'loss to follow-up' (lost contact with patient)
- death from other cause (e.g. road accident)
- patient has to be withdrawn from study

Left censoring: Here we know only that $t < c$. E.g. time to recurrence of tumour after cancer surgery – patient examined after three months and a tumour is found to be present. Therefore t is less than three months.

Interval censoring: Here we know only that $c_1 < t < c_2$. E.g. time to recurrence of tumour after cancer surgery: Patient examined after 3 months and tumour not present. Patient examined after 6 months and tumour present. Therefore $3 < t < 6$ (months).

Example: In a study of women recovering from breast cancer, patients visited their doctor every 3 months for a period of two years to check for further tumours. What type of censoring do the following individuals represent?

1. Patient A had no further tumours during the trial.
2. Patient B had a tumour identified at the 15 month check-up.
3. Patient C had a tumour present at the first (3 month) check up.

1.4 Informative censoring

In this course, when discussing methods for dealing with censoring, unless stated otherwise we will assume that the censoring is *non-informative.* That is, the actual survival time T is independent of the mechanism which causes an observation to be censored at time c. Put another way, an individual censored at time c must be representative of all similar individuals who survive that long. Censoring might be *informative* if, e.g., the patient is withdrawn from the study because of ill health.

Example 1: Suppose we are looking at time till onset of AIDS in HIV positive patients. Many of the patients in the study are likely to be injecting drug users who are particularly prone to being lost to follow-up. Might the time till onset of HIV be different in injecting drug users than in other patients? Could this bias the data?

Example 2: Often patients are withdrawn from a trial because they are suffering from side-effects or because they become too ill to continue with treatment. However, patients who suffer side-effects from an experimental treatment may well have a different survival experience than those who have no side-effects (e.g. a drug might be less effective in the group with side-effects).

Informative censoring is difficult to deal with. Most statistical models assume no informative censoring occurs, though in practice some form of analysis may be performed to confirm this assumption.

1.5 Typical data

As we've said, survival data consist of a list of survival times, some of which may be censored. In addition, there might be some *covariates* – information on each patient involved in a trial. In the heart transplant example above, we might also have the following information about each patient: age, weight, sex, smoker/non-smoker. All these covariates might reasonably have an effect on a patient's survival time, and later in the course we'll see how such data can be incorporated into an analysis. Typical data are illustrated in this table:

Patient	Survival Time	Died/Censored	Age	Sex	Smoker
1	48	1	62	0	0
2	48	1	58	0	0
3	50	0	51	1	1
4	10	1	64	0	0
5	46	0	61	1	1
6	38	1	60	1	0

2 The survivor and hazard functions

2.1 Key definitions

The survival time t of an individiual can be regarded as the value of a random variable T which is non-negative. The different values T can adopt have some probability distribution (either continuous or discrete) referred to as the *lifetime* distribution. The survival times for a group of individuals can be regarded as a set of independent samples from the lifetime distribution.

Distribution function:

$$F(t) = \Pr(T < t)$$

Probability density function:

$$f(t) = \frac{d}{dt} F(t)$$

Survivor function:

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

It follows from the definition of the survivor function that

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \int_t^\infty f(u) du. \end{aligned}$$

Conditional distribution:

Given an individual lives up to time t_0, the distribution of the future survival time is given by:

$$\begin{aligned} \Pr(\text{dies in } (t_0, t_0 + t) \mid \text{alive at } t_0) &= \Pr(T < t_0 + t \mid T \geq t_0) \\ &= \frac{F(t_0 + t) - F(t_0)}{S(t_0)}. \end{aligned}$$

The probability density of future lifetime is the derivative of this:

$$\frac{d}{dt} \frac{F(t_0 + t) - F(t_0)}{S(t_0)} = \frac{f(t + t_0)}{S(t_0)}.$$

Hazard function:

The hazard function at t_0, denoted $h(t_0)$, is the instantaneous rate of death, i.e. the following limit:

$$h(t_0) = \lim_{\delta t \to 0} \frac{1}{\delta t} \Pr(\text{die in interval } [t_0, t_0 + \delta t) \mid \text{alive at time } t_0)$$

The hazard function is related to the survivor function in the following way:

$$\begin{aligned}
h(t_0) &= \lim_{\delta t \to 0} \frac{1}{\delta t} \Pr(\text{die in interval } [t_0, t_0 + \delta t) \mid \text{alive at time } t_0) \\
&= \lim_{\delta t \to 0} \frac{1}{\delta t} \frac{F(t_0 + \delta t) - F(t_0)}{S(t_0)} \\
&= \frac{1}{S(t_0)} \lim_{\delta t \to 0} \frac{F(t_0 + \delta t) - F(t_0)}{\delta t} \\
&= \frac{1}{S(t_0)} \frac{d}{dt} F(t_0) \\
&= \frac{f(t_0)}{S(t_0)}
\end{aligned}$$

So we have $h(t) = f(t)/S(t)$.
The hazard function gives the following linear approximation

$$\Pr(\text{dies in } (t_0, t_0 + \delta t) \mid \text{alive at } t_0) \approx \delta t \times h(t_0).$$

Cumulative hazard:

$H(t) = \int_0^t h(u)du$
We have written the hazard function in terms of the survivor function, so now we do the converse – and the cumulative hazard comes in handy. Since $S(t) = 1 - F(t)$ it follows that:

$$\begin{aligned} \frac{d}{dt}\log[S(t)] &= \frac{1}{S(t)} \times \frac{d}{dt}S(t) \\ &= -\frac{f(t)}{S(t)} \\ &= -h(t). \end{aligned}$$

Integrating gives

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

so

$$S(t) = \exp[-H(t)]$$

2.2 Examples

Sometimes it's very useful to assume the hazard (or survival) functions have specific forms. Of course, this is equivalent to assuming a specific form for the underlying probability distribution F of the survival time T. We look next at the simplest assumptions we might make.

Exponential Distribution

Suppose that the hazard function is constant: $h(t) = \lambda$. It follows that:

$$H(t) = \lambda t$$
$$S(t) = \exp(-\lambda t)$$
$$f(t) = \lambda \exp(-\lambda t).$$

The probability density function is that for a *exponential* random variable. It has the 'lack of memory property':

$$\Pr(T > t_1 + t_2 \mid T > t_1) = \frac{e^{-\lambda(t_1+t_2)}}{e^{-\lambda t_1}} = e^{-\lambda t_2} = \Pr(T > t_2).$$

Weibull Distribution

More usefully, we would like the hazard function to vary with time: the Weibull distribution is the simplest such example. Suppose that $h(t) = ct^k$ for some constants c and $k > -1$.

$$\begin{gathered} H(t) = c\frac{t^{k+1}}{k+1}, \\ S(t) = \exp\left(-\frac{c}{k+1}t^{k+1}\right), \end{gathered}$$

so

$$\begin{aligned} f(t) &= \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \\ &= ct^k \exp\left(-\frac{c}{k+1}t^{k+1}\right). \end{aligned}$$

This is the *Weibull distribution*. It's more usually written with a different parameterization in the following way. Let $\gamma = k+1$ and $\lambda = c/(k+1)$, $(\gamma > 0, \lambda > 0)$. Then:

$$\begin{gathered} h(t) = \lambda\gamma t^{\gamma-1} \\ H(t) = \lambda t^{\gamma} \\ S(t) = \exp(-\lambda t^{\gamma}) \\ f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}). \end{gathered}$$

Suppose T is a Weibull random variable with parameters λ, γ. It can be shown that the expectation of T is

$$E(T) = \lambda^{-1/\gamma}\Gamma(1+\gamma^{-1}).$$

Here Γ is the gamma function:

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}du.$$

Since the distribution is *skewed*, the median is different from the mean. The median t_m is computed as follows:

$$\begin{aligned} F(t_m) &= 1/2 \\ \exp(-\lambda t_m^{\gamma}) &= 1/2 \\ t_m &= \left(\frac{\log 2}{\lambda}\right)^{1/\gamma}. \end{aligned}$$

The $100p$-th percentile can be computed in a similar way.

Warning! R uses different parameterizations of the Weibull distribution. If you ever fit a Weibull model in R, make sure you know which parameterization is being used. In the R parameterisation the pdf is

$$f(t) = \frac{\gamma}{\beta}\left(\frac{t}{\beta}\right)^{\gamma-1}\exp\left\{-\left(\frac{t}{\beta}\right)^{\gamma}\right\}$$

so

$$\lambda = \beta^{-\gamma} \quad \text{and} \quad \beta = \lambda^{-1/\gamma},$$

the mean is

$$\mathrm{E}(T) = \beta\Gamma(1+\gamma^{-1})$$

and the median is

$$t_m = \beta(\log 2)^{1/\gamma}.$$

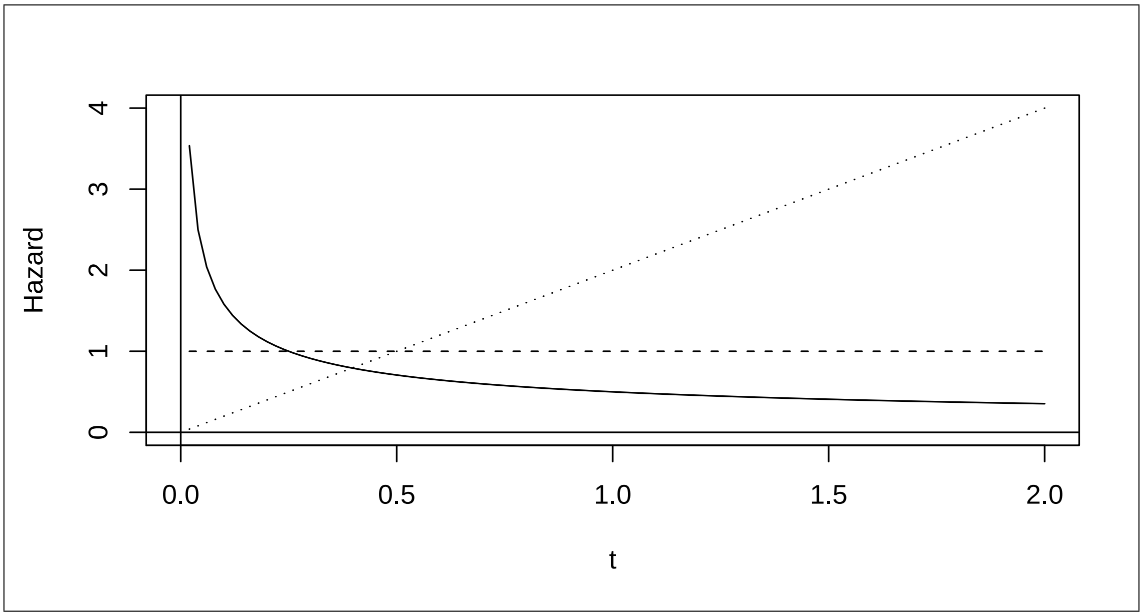


Figure 5: Weibull hazard function with $\lambda = 1$ and $\gamma = 0.5$ (solid), $\gamma = 1$ (dashes), $\gamma = 2$ (dots).

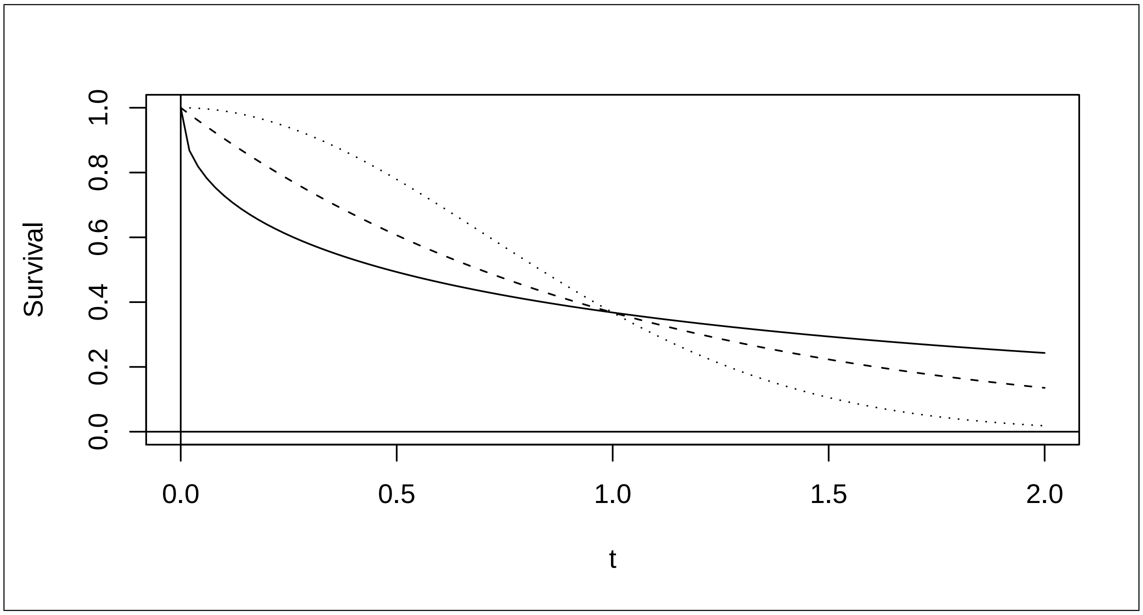


Figure 6: Weibull survivor function with $\lambda = 1$ and $\gamma = 0.5$ (solid), $\gamma = 1$ (dashes), $\gamma = 2$ (dots).

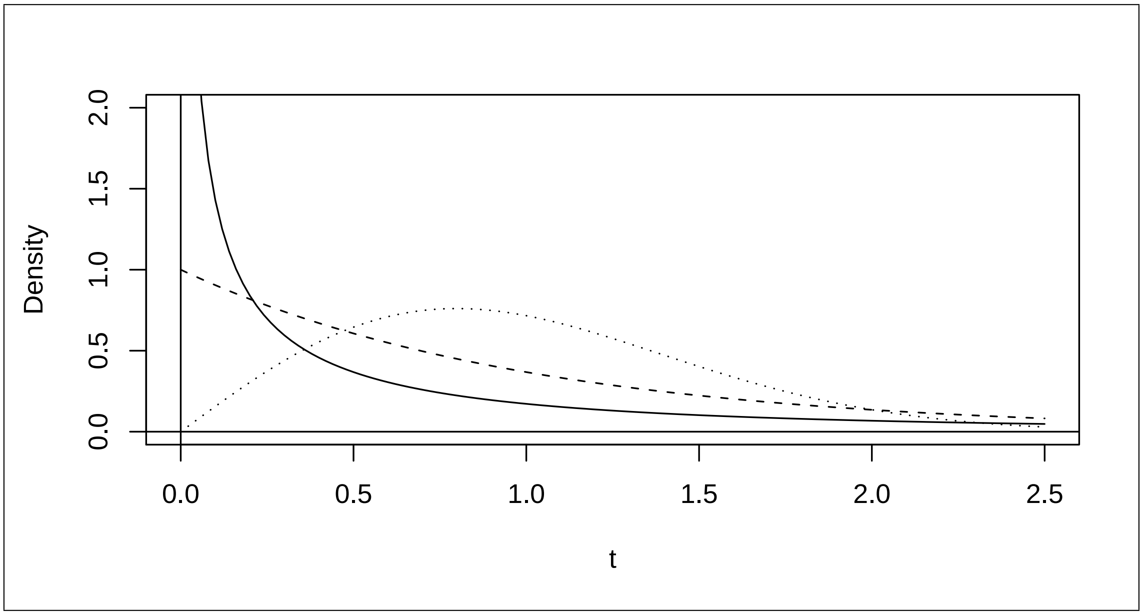


Figure 7: Weibull probability density function with mean 1 and $\gamma = 0.5$ (solid), $\gamma = 1$ (dashes), $\gamma = 2$ (dots).

2.3 Example

Find the survivor function associated with the following hazard function:

$$h(t) = \begin{cases} \lambda + \mu(t_1 - t), & 0 \leq t \leq t_1, \\ \lambda, & t_1 \leq t \leq t_2, \\ \lambda + \mu(t - t_2), & t \geq t_2. \end{cases}$$

where λ, μ, t_1, t_2 are positive constants with $t_1 < t_2$.

Solution

$$\begin{aligned} H(t) &= \begin{cases} \lambda t + \mu(t_1 t - t^2/2), & 0 \leq t \leq t_1, \\ \lambda t, & t_1 \leq t \leq t_2, \\ \lambda t + \mu(t^2/2 - t_2 t), & t \geq t_2. \end{cases} \\ &= \begin{cases} \lambda t + (\mu/2)[t_1^2 - (t_1 - t)^2], & 0 \leq t \leq t_1, \\ \lambda t, & t_1 \leq t \leq t_2, \\ \lambda t + (\mu/2)[(t - t_2)^2 - t_2^2], & t \geq t_2. \end{cases} \end{aligned}$$

Of course we can now find $S(t) = \exp\{-H(t)\}$ and so on.

3 Estimating the survivor function

A very basic first task in any analysis of survival data is to estimate a survivor function. There are two main non-parametric methods: the life table method and the Kaplan-Meier method. The Kaplan-Meier method is the most widely used, and is of fundamental importance in survival analysis.

In the last section we wrote down some distributions on which the survivor and hazard functions might be based – such approaches are called *parameteric*. The methods we present in this section are *non-parametric* as they do not make specific assumptions about distributions.

3.1 The case of no censoring

Let's assume we have a set of survival times with no censoring. Let $N(t)$ denote the number of individuals alive at time t. The *empirical survivor function* $\hat{S}(t)$ is defined by

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the data set}} = \frac{N(t)}{N(0)}$$

and is an estimate of the survivor function. It's easy to see that $\hat{S}(t)$ is a step function. If t_i denotes the time of the i-th failure ($t_1 < t_2 < \ldots$) then

$$\hat{S}(t) = \frac{N(t_i)}{N(0)} \quad \text{for } t_i \leq t < t_{i+1}.$$

(Note: for a given failure time t_i we assume the individual is dead at that exact moment.)

Example: As part of a study to investigate the effect of different climatic conditions on locusts, a population of 22 locusts was studied in the lab and the survival times measured. The (uncensored) times in days are given below, together with the empirical survivor function.

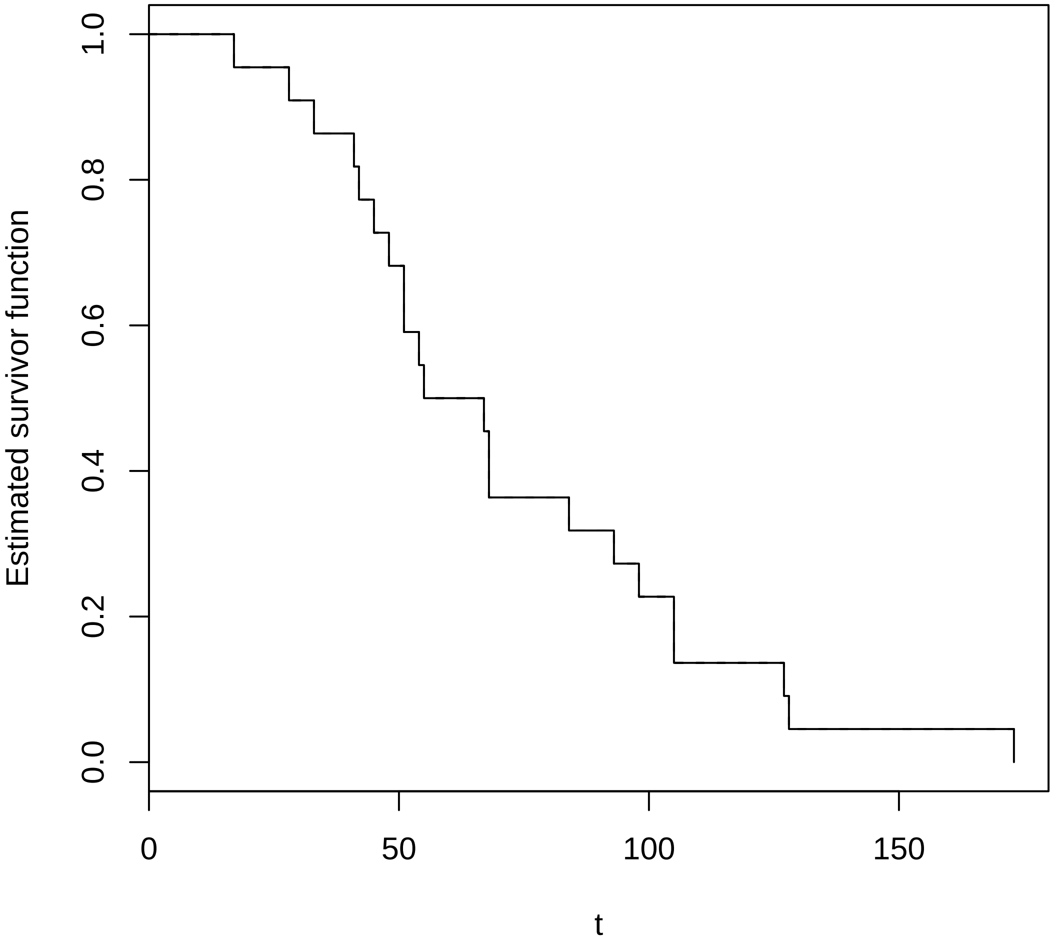
Death times:

$$17, 28, 33, 41, 42, 45, 48, 51, 51, 54, 55$$
$$67, 68, 68, 84, 93, 98, 105, 105, 127, 128, 173$$

Survivor function:

	Failure		$\hat{S}(t)$		Failure		$\hat{S}(t)$
i	Time t_i	$N(t_i)$	$(t_i \leq t < t_{i+1})$	i	Time t_i	$N(t_i)$	$(t_i \leq t < t_{i+1})$.
0	0	22	1.00	10	55	11	0.500
1	17	21	0.955	11	67	10	0.455
2	28	20	0.909	12	68	8	0.364
3	33	19	0.864	13	84	7	0.318
4	41	18	0.818	14	93	6	0.273
5	42	17	0.773	15	98	5	0.227
6	45	16	0.727	16	105	3	0.136
7	48	15	0.682	17	127	2	0.091
8	51	13	0.591	18	128	1	0.045
9	54	12	0.545	19	173	0	0.000

Note that in the table we've given the value of the empirical survivor function for $t_i \leq t < t_{i+1}$ in line i, rather than the value for $t_{i-1} \leq t < t_i$. This is both to conform with the R output format and also to make comparisons with methods considered later more straightforward. Here's the function itself:



Let's consider the variance we might expect in such an estimate. If $S(t)$ denotes the true underlying survivor function then $N(t)$ can be thought of as a binomial random variable $N(t) \sim \text{Bin}(N(0), S(t))$. Such a random variable has variance $N(0)S(t)[1 - S(t)]$ so $\hat{S}(t)$ has variance

$$\frac{S(t)[1 - S(t)]}{N(0)}$$

which is approximately equal to

$$\frac{\hat{S}(t)[1 - \hat{S}(t)]}{N(0)}.$$

The variance can be used to construct confidence intervals for the estimate of the survivor function. This is a simplistic approach, however: $\hat{S}(t_2)$ is not independent of $\hat{S}(t_1)$ when $t_2 > t_1$.

3.2 The life table or actuarial method

The *life table* or *actuarial* method of estimating the survivor function can be used when there is (right) censoring. First we assume that the observation period (the time axis) is split into a number of intervals (not necessarily equal). The j-th interval is $[t_j, t_{j+1})$, for $j = 0, \dots, m$ so that the first interval is $[0, t_1)$ (taking $t_0 = 0$) and the last is $[t_m, \infty)$ (taking $t_{m+1} = \infty$). To compute the life table estimate of the survivor function we do not need the full set of survival times, but instead work with the number of deaths d_j and censored survival times c_j that occur in each interval.
Note: The life table method is primarily used when the full survival data are not available, but we know the number of deaths and censored times in each interval.

Example: Cognitive Behavioural Therapy for depression. 180 patients receiving CBT for depression were followed up for a year to study the duration of their treatment. Every two months the number of patients who completed their treatment was recorded. Censoring occurred when individuals dropped out prior to completing the therapy or transferred to a drug-based treatment. The table below gives the number of patients completing treatment (d_j) and number of censored individuals (c_j) in each 2-month interval.

	$0-2$	$2-4$	$4-6$	$6-8$	$8-10$	$10-12$
d_j	21	23	19	12	7	8
c_j	6	12	17	20	10	6

Let N_j denote the number of individuals *known* to be alive at the start of interval j i.e. those with a death time or censored time with $t \geq t_j$. It follows that

$$N_j = N_{j-1} - c_{j-1} - d_{j-1}.$$

What's the probability of dying during this interval if you were alive at the start? If there was no censoring ($c_j = 0$) it would be d_j/N_j. Now suppose $c_j > 0$. Since the death times for c_j individuals are known to be greater than t_j, not all N_j individuals are at risk of dying across the entire interval. We therefore make the following assumption: We don't know when the c_j individuals with censored times in the interval actually die, so we make the following assumption which is called the *actuarial assumption.*:

Actuarial assumption: The censored times occur uniformly throughout the interval.

The average number of individuals at risk over the interval is therefore

$$N'_j = N_j - \frac{1}{2}c_j.$$

so we take d_j/N'_j as an approximate probability of dying during interval j. The probability of surviving the interval is therefore

$$p_j = 1 - \frac{d_j}{N'_j} = \frac{N'_j - d_j}{N'_j}$$

and the estimate of the survivor function is

$$\hat{S}(t) = \prod_{i=0}^{j} p_i = p_0 \times p_1 \times \cdots \times p_j$$

for $t_j \leq t < t_{j+1}$.

Example continued: The following table gives the life table estimate for the sample data above:

j	Interval	d_j	c_j	N_j	N'_j	$(N'_j - d_j)/N'_j$	$\hat{S}(t)$
0	$[0, 2)$	21	6	180	177.0	0.881	0.881
1	$[2, 4)$	23	12	153	147.0	0.844	0.743
2	$[4, 6)$	19	17	118	109.5	0.826	0.614
3	$[6, 8)$	12	20	82	72.0	0.833	0.512
4	$[8, 10)$	7	10	50	45.0	0.844	0.432
5	$[10, 12)$	8	6	33	30.0	0.733	0.317

We can also estimate the hazard function in a similar way. Fix a time $t \in [t_j, t_{j+1})$. How many individuals are alive at this time? As before, if we assume the censored times are uniformly distributed over the interval, and also that the times of deaths are uniformly distributed, then on average

$$N_j'' = N_j - \frac{1}{2}c_j - \frac{1}{2}d_j$$

are alive at time t. Next recall the definition

$$h(t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \Pr(\text{die in interval } [t, t + \delta t) \mid \text{alive at time } t). \quad (\dagger)$$

Since we assume deaths are uniformly distributed over the interval the number of deaths in the interval $[t, t + \delta t)$ is $\delta t \times d_j/(t_{j+1} - t_j)$. The number of people at risk of dying at time t is N_j'' as described above, so the probability term in (†) is

$$\Pr(\text{die in interval } [t, t + \delta t) \mid \text{alive at time } t) \approx \frac{\delta t \times d_j}{(t_{j+1} - t_j)N_j''}.$$

The estimated hazard function is therefore

$$\hat{h}(t) = \frac{d_j}{(t_{j+1} - t_j)N_j''} \quad \text{when } t_j \leq t < t_{j+1}.$$

The hazard function is another step function, and it's zero on any interval with $d_j = 0$.

4 The Kaplan-Meier Method

4.1 Estimating the survivor function

The Kaplan-Meier (KM) estimator is the most widely used and important nonparametric estimator of the survivor function. It is also known as the *product-limit estimator*.

As with the life-table method, we construct a sequence of time intervals but, in the KM method, each time when we observe a death becomes the start of a new interval. In practice, however, there can be more than one death at the same time. We define a *death time* to be a time at which one or more (uncensored) deaths occur in our survival data

As before we define intervals $[t_j, t_{j+1})$, for $j = 0, \ldots, m$, with $t_0 = 0$ and $t_{m+1} = \infty$. For the KM estimate, however, t_1 is the first death time. Then t_2 is the second death time, and so on. For example, suppose we had the following (uncensored) survival data: $5, 8, 11, 11, 13$. Then the intervals are $[0, 5), [5, 8), [8, 11), [11, 13), [13, \infty)$. The possibility of ties means that we can have $m < N_0$ where N_0 is the number of individuals at t_0.

As before let N_j be the number of individuals that are alive and uncensored just before time t_j. Let d_j be the number of deaths at time t_j and let c_j be the number of censored times in the interval $[t_j, t_{j+1})$. Then $N_{j+1} = N_j - d_j - c_j$. The KM estimate of the probability of surviving the interval $[t_j, t_{j+1})$, given that the individual survives to just before t_j is

$$p_j = \frac{N_j - d_j}{N_j}.$$

For $t_j \leq t < t_{j+1}$, the KM estimate of $S(t)$ is

$$\hat{S}(t) = \prod_{i=1}^{j} p_i.$$

(Note $p_0 = 1$.)

Recall $t_{m+1} = \infty$. If the last observation is a death (at t_m) then

$$p_m = \frac{N_m - d_m}{N_m}$$

but $N_m = d_m$ so $p_m = 0$. If the last observation is a censored lifetime $t^\star$ then $\hat{S}(t)$ is undefined for $t > t^\star$.

Note that, if there is no censoring, then the KM estimate is just the empirical survivor function.

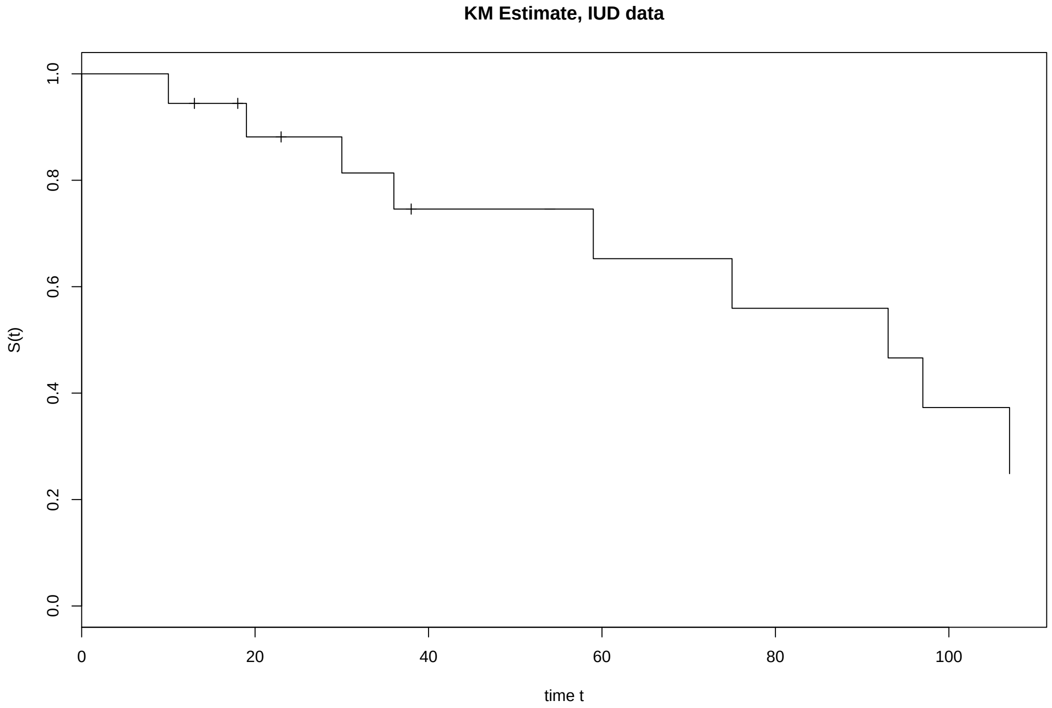
Example: Time till discontinuation of use of an IUD. The survival times (in weeks) are:

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

The asterisks denote censored times. The KM estimate is constructed as follows.

j	Interval	N_j	d_j	c_j	$(N_j - d_j)/N_j$	$\hat{S}(t)$
0	$[0, 10)$	18	0	0	1.00	1.00
1	$[10, 19)$	18	1	2	0.94	0.94
2	$[19, 30)$	15	1	1	0.93	0.88
3	$[30, 36)$	13	1	0	0.92	0.81
4	$[36, 59)$	12	1	3	0.91	0.75
5	$[59, 75)$	8	1	0	0.88	0.65
6	$[75, 93)$	7	1	0	0.86	0.56
7	$[93, 97)$	6	1	0	0.83	0.47
8	$[97, 107)$	5	1	1	0.80	0.37
9	$[107, \infty)$	3	1	2	NA	NA

Here's the function itself:



4.2 Kaplan-Meier estimate of the hazard function

Let t_j be the j^{th} death time. Then the survival probability estimate for the interval beginning at t_j is

$$\hat{p}_j = \frac{N_j - d_j}{N_j}.$$

So the death probability estimate is

$$1 - \hat{p}_j = \frac{d_j}{N_j}.$$

If we suppose that the hazard is constant between t_j and t_{j+1}, then we estimate it as

$$\hat{h}(t) = \frac{d_j}{N_j(t_{j+1} - t_j)}$$

for $t_j \leq t < t_{j+1}$.

Example continued: IUD data as above.

j	Interval	N_j	d_j	d_j/N_j	$\hat{h}(t)$
0	$[0, 10)$	18	0	0.00	0.0000
1	$[10, 19)$	18	1	0.06	0.0067
2	$[19, 30)$	15	1	0.07	0.0064
3	$[30, 36)$	13	1	0.08	0.0133
4	$[36, 59)$	12	1	0.09	0.0039
5	$[59, 75)$	8	1	0.12	0.0075
6	$[75, 93)$	7	1	0.14	0.0078
7	$[93, 97)$	6	1	0.17	0.0425
8	$[97, 107)$	5	1	0.20	0.0200
9	$[107, \infty)$	3	1	NA	NA

4.3 Derivation of the Kaplan Meier Estimate

In this section we'll see why the KM method is 'the right thing to do'. Suppose we divide time into a sequence of fixed intervals, as in a life table. Let the cut points be $t_{(0)}, t_{(1)}, \ldots, t_{(m)}$ where $t_{(0)} = 0$. (We use $t_{(j)}$ to distinguish this from t_j, the j-th death time). The intervals are $[0, t_{(1)})$, $[t_{(1)}, t_{(2)})$, $[t_{(2)}, t_{(3)}), \ldots, [t_{(n-1)}, t_{(m)})$. Suppose that the probability of survival through interval j, i.e. $[t_{(j)}, t_{(j+1)})$, given survival to time $t_{(j)}$, is p_j. Then the probability of survival to time $t_{(k)}$ is

$$\prod_{i=0}^{k-1} p_{(i)}.$$

At time $t_{(k)}$ we observe $N_{(k)}$ alive and uncensored, of whom $d_{(k)}$ die before time $t_{(k+1)}$.

On interval $[t_{(j)}, t_{(j+1)})$, what's the probability of observing d deaths if you start out with $N_{(j)}$ people alive? Well, d is the outcome of a $\text{Bin}(N_{(j)}, p_{(j)})$ random variable. So the probability of observing d deaths is

$$\binom{N_{(j)}}{d} \times (1 - p_{(j)})^d \times p_{(j)}^{(N_{(j)}-d)}.$$

The likelihood of observing $(d_{(0)}, d_{(1)}, \ldots, d_{(m-1)})$ deaths is therefore

$$L = \prod_{i=0}^{m-1} \binom{N_{(i)}}{d_{(i)}} \times (1 - p_{(i)})^{d_{(i)}} \times p_i^{(N_{(i)}-d_{(i)})}.$$

By differentiating the log-likelihood, it can be shown that the maximum likelihood estimate of p_i is

$$\hat{p}_{(i)} = \frac{N_{(i)} - d_{(i)}}{N_{(i)}}.$$

Now, consider what happens as we increase the number of cut points and shorten the intervals. In the limit as the intervals become infinitesimally small, we obtain the KM estimator.

Tutorial Examples 1

1. Find the survivor function associated with the following hazard function:

$$h(t) = \begin{cases} \lambda + \mu(t_1 - t), & 0 \le t \le t_1, \\ \lambda, & t_1 \le t \le t_2, \\ \lambda + \mu(t - t_2), & t \ge t_2. \end{cases}$$

   where λ, μ, t_1, t_2 are positive constants with $t_1 < t_2$. (This is the example from section 2 of the notes.)

2. A hazard function is defined by:

$$h(t) = \frac{2\lambda^2 t}{1 + \lambda^2 t^2}$$

   where $\lambda > 0$ is a constant. Find:

   (a) the survivor function, and

   (b) the expected survival time.

3. We are given the following survival times for 16 individuals, where a $*$ indicates a right censored measurement:

$$20*, 23, 47, 47, 69, 70*, 71, 100*, 101, 110*, 148, 181, 198*, 208*, 212*, 224*$$

   Calculate the Kaplan Meier estimate of the survivor function. Obtain an estimate for the lower (25%) quartile, and find a 95% confidence interval for this estimate.

4. (a) A lifetime distribution has hazard function

$$h(t) = \theta_0 + \theta_1 t + \theta_2 t^2.$$

   Find

   i. the survivor function.

   ii. the probability density function.

   (b) A lifetime distribution has hazard function

$$h(t) = \frac{1}{t+1}.$$

   Find

   i. the survivor function.

   ii. the probability density function.

   iii. the median.

   What happens if you try to find the mean?

   (c) A lifetime distribution has hazard function

$$h(t) = \frac{\rho\theta}{\rho t + 1},$$

   where $\rho > 0$ and $\theta > 1$.

   Find

   i. the survivor function.

   ii. the probability density function.

   iii. the median.

   iv. the mean.

5. A machine has n components, the lifetimes of which are independent. However the whole machine will fail if any component fails. The hazard functions for the components are $h_1(t), \dots, h_n(t)$. Show that the hazard function for the machine is $\sum_{i=1}^{n} h_i(t)$.

6. Suppose that the lifetime distributions of the components in Question 5 are Weibull distributions with scale parameters $\rho_1, \dots, \rho_n$ and a common index (i.e. "shape parameter") γ so that the hazard function for component i is $\gamma\rho_i(\rho_i t)^{\gamma-1}$. Find the lifetime distribution for the machine.

7. Show that, if T is a Weibull random variable with hazard function $\gamma\rho(\rho t)^{\gamma-1}$,

   (a) the median is $\mathrm{M}(T) = \rho^{-1}(\log 2)^{1/\gamma}$,

   (b) the mean is $\mathrm{E}(T) = \rho^{-1}\Gamma(1+\gamma^{-1})$ and

   (c) the variance is $\mathrm{var}(T) = \rho^{-2}\{\Gamma(1+2\gamma^{-1}) - [\Gamma(1+\gamma^{-1})]^2\}$.

   Note that $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}.dx$.

8. The following data are taken from Cameron and Pauling (1978). Patients with advanced cancer of the stomach, bronchus, colon, ovary or breast were treated with ascorbate. The data are survival times in days. Investigate whether the survival times differ with the organ affected. To do this you might use a standard procedure for normally distributed data but, if you do, you should consider whether the data should be transformed and, if so, how. (You may use, for example, R to do the calculations, plots etc.). Present your procedure and conclusions clearly.

Stomach	Bronchus	Colon	Ovary	Breast
124	81	248	1234	1235
42	461	377	89	24
25	20	189	201	1581
45	450	1843	356	1166
412	246	180	2970	40
51	166	537	456	727
1112	63	519		3808
46	64	455		791
103	155	406		1804
876	859	365		3460
146	151	942		719
340	166	776		
396	37	372		
	223	163		
	138	101		
	72	20		
	245	283		

Reference

Cameron, E. and Pauling, L., 1978. Supplemental ascorbate in the supportive treatment of cancer: re-evaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Science* USA, **75**, 4538-4542.