1 Basic concepts in survival analysis

1.1 Examples

Example 1.1

A trial by the World Health Organization (WHO) studied the time from which a women start to use a particular type of contraceptives, called intrauterine device (IUD) until discontinuation because of menstrual bleeding problems. Data are given for 19 women, all of whom were aged between 18 and 35 years and who experienced two previous pregnancies. Discontinuation times that are recorded censored are labeled with an asterisk.

```
-----------------------------------------------------------------------------
     Time in weeks to discontinuation of the use of the IUD
----------------------------------------------------------------------------
    10  13*  18*  19  23*  30  36  38*  54*  56*  59  75  93  97  104
    107* 107* 107*
---------------------------------------------------------------------------
```

In this example, the time origin corresponds to the first day in which a women uses the IUD, and the end-point is discontinuation because of bleeding problems. Some women in the study ceased using the IUD because of the desire for pregnancy, or because they had no further need for a contraceptive, while others were simply lost to follow-up.

Example 1.2

Objective: Effects of Ischemic heart disease on survival.

- Outcome: mortality from any cause within six years of inception date.
- Covariates:
- Age: (min=20 mean=68, max=109)

- CAD : Ischemic heart disease by diagnosis or test
  Yes: 997 and No: 230.

- Q: Is the survival rate for a patient with CAD the same as the survival rate for a patient without CAD?

Tierney, W. M., Takesue, B. Y., Vargo, D. L., and Zhou, X. H. (1997). Using electronic medical records to predict mortality in primary care patients with heart disease: prognostic power and pathophysiologic implications, Journal of General Internal Medicine, in press.

Table 1: Mortality data

PATIENT ID	Survival	Censored	CAD	Age
1	690	1	1	85
2	1207	0	1	91
3	483	1	0	72
6	1593	1	1	64
⋮	⋮	⋮	⋮	⋮
1227	.	.	.	.

1.2 Special features of survival data

- Survival data have generally positively skewed distributed. As a consequence, it will not be reasonable to assume that data of this type have a normal distribution.

- A second feature of survival data which renders standard methods inappropriate is that survival times are frequently censored. The survival time of an individual is said to be censored when the end-point of interest has not been observed for that individual. This may be because the data from a study are to be analyzed at a point in time when some individuals are still alive. Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been lost to follow-up. As an example, suppose that after being recruited to a clinical trial, a patient moves to another part of the country, or to to different country, can no longer be traced. The only information available on the survival experience of that patient is the last date on which he or she was known to be alive. This date may well be the last time that the patient reported to a clinic for a regular check-up.

1.2.1 Three types of censoring

- Right censoring.

  In each of these situations, a patient who entered a study at time t_0 dies at time $t_0 + t$. However, t is unknown, either because the individual is still alive or because he or she has been lost to follow-up. If the individual was last known to be alive at time $t_0 + c$, the time c is called a censored survival time. This censoring occurs after the individual has been entered into a study, is therefore know as *right censoring*. The right-censored survival time is then less than the actual, but unknown, survival time.

- Left censoring

  Another for of censoring is *left censoring*, which is encountered when the actual survival time of an individual is less than that observed. To illustrate this form of censoring, consider a study in which interest centers on the time to recurrence of a particular cancer following surgical removal of the primary tumor. Three months after their operation, the patients are examined to determine if the cancer has recurred. At this time, some of the patients may be found to have a recurrence. For such patients, the actual time to recurrence is less than three months, and the recurrence times of these patients is left-censoring. Left censoring occurs far less commonly than right censoring.

- Interval censoring

Yet another type of censoring is interval censoring. Here, individuals are known to have experienced a failure within an interval of time. Consider again the example concerning the time to recurrence of a tumor. If a patient is observed to be free of the disease at three months, but is found to have had a recurrence when examined six months after surgery, the actual recurrence time of that patient is known to be between three months and six months.

Our focus is on right censoring.

- A censored observation provides partial information on the survival time of the patient. Dropping the censored observations from analysis might bias your results.
- Reasons for censoring:
  In a clinical trial, censoring occurs in the following forms:
  - Loss to follow-up.
  - Drop out.
  - Termination of the study.

1.3 Patient time and study time

In a typical study, patients are not all recruited at exactly the same time, but accrue over a period of months or even years. After recruitment, patients are followed up until they die or lost to follow-up, or until a point in calendar time which marks the end of the study. The calendar time period in which an individual is in the study is known as the study time. As far as each patient is concerned, the trial begins at some time t_0. The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as patient time.

1.4 An important assumption

An important assumption which will be made in the analysis of censored survival data is that the actual survival time of an individual, t, is independent of any mechanism which causes that individual's survival time to be censored at time c, where $c < t$. This means that if we consider a group of individuals, all of whom have the same values of relevant prognostic variables, an individual whose survival time is censored at time c must be representative of all other individuals in that groups who have survived to that time.

However, this assumption cannot be made if, for example, the survival time of an individual is censored through treatment being withdrawn as a result of a deterioration in the physical condition of a patient. This type of censoring is known as informative censoring.

1.5 Basic concepts

In summarizing survival data, there are two functions of central interest, naming the survival function and the hazard function. Let T be the survival time of an individual.

The actual survival time of a patient, t, can be regarded as the value of a variable T, which takes any non-negative value.

- Survival function:

$$\text{Survival function} = S(t) = P(T > t).$$

  The survival function can be used to represent the probability that an individual survives from the time origin to some time beyond t.

- Hazard function:
  the instantaneous rate of death or failure at time t, given that the individual survives up til t.

$$h(t) = \frac{P(t \leq T < t + \delta \mid T \geq t)}{\delta}$$

  for very small δ.

- The relationship:

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

2 Exponential models

A simple and often effective approach to the study of survival data is based on properties of the exponential model. This model is characterized by a constant hazard function

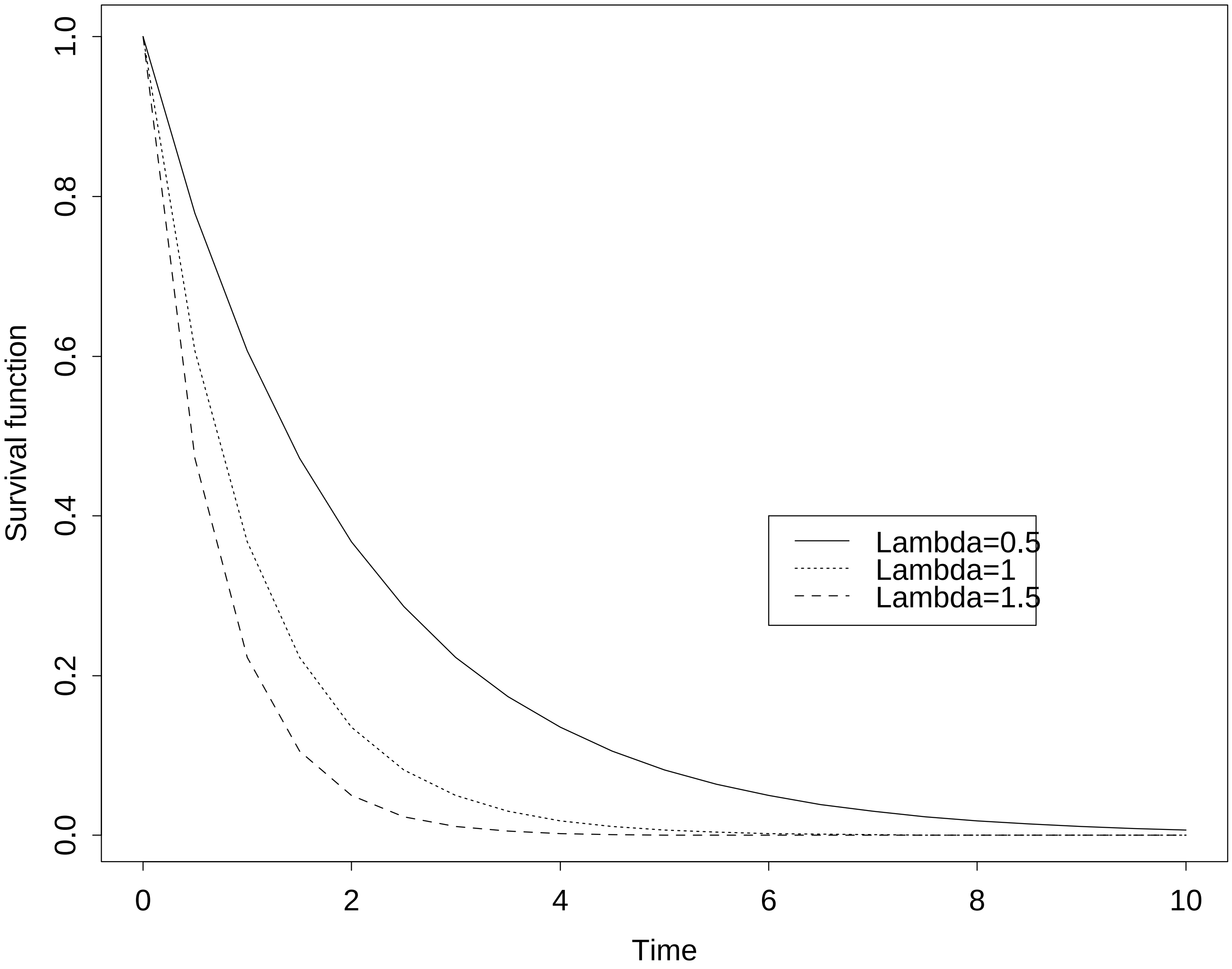
$$h(t) = \lambda$$

where $\lambda > 0$. The survival function is

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

Here, λ is the hazard rate of death. The exponential model has been widely used in areas ranging from studies on the lifetimes of a manufactured items to research involving survival or remission times in chronic disease. The distribution where $\lambda = 1$ is called the standard exponential distribution.

Figure 1: Exponential survival functions for different λ.



$$mean = \mu = \lambda^{-1}.$$

The median of the exponential distribution, $t(50)$, is such that

$$S(t(50)) = 0.5.$$

$$t(50) = \frac{1}{\lambda} log2.$$

References:

1. Epstein, B. (1958). The exponential distribution and its role in life-testing. Ind. Qual. Control, 15, 2-7

2. Feigl, P., and Zelen, M. (1965). Estimation of exponential probabilities with concomitant information. Biometrics, 21, 826-838

A special property of an exponential model:
the population does not "age", or it has a constant hazard rate. A constant rate arises when failure occurs at random. For examples, struck by falling objects, injured by high winds, or attached by wild animals, deaths among very old people or deaths among very ill people. In general, random failure often adequately describes human mortality over a short time interval.

2.1 Inference about λ

Suppose there is a random sample of n individuals with lifetimes $T_1, \ldots, T_n$, but that associated with each individual is also a fixed censoring times $L_i > 0$. We observed T_i only if $T_i \leq L_i$ and the data therefore consist of pairs

$$(t_i, \delta_i), i = 1, \ldots, n$$

where

$$t_i = min(T_i, L_i) \text{ and}$$

$\delta_i = 1$ if $T_i \leq L_i$, and

The estimator for λ is

$$\hat{\lambda} = \frac{r}{T}$$

where $r = \sum_{i=1}^{n} \delta_i$ and $T = \sum_{i=1}^{n} t_i$. The asymptotic variance of $\hat{\lambda}$ is

$$var(\hat{\lambda}) = \lambda^2/r.$$

The standard deviation of $\hat{\lambda}$ is

$$\frac{\hat{\lambda}}{\sqrt{r}}.$$

The limits of $100(1-\alpha)\%$ confidence interval for λ are

$$\hat{\lambda} \pm z_{\alpha/2} s.d.(\hat{\lambda}),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

2.2 Inference about μ and $t(50)$

Estimators for mean and median are

$$\hat{\mu} = \frac{T}{r},$$

and

$$t(\hat{50}) = \frac{T}{r} \log 2.$$

Proposition 1 *The variance of a function $g(\hat{\lambda})$ of $\hat{\lambda}$ is such that*

$$var(g(\hat{\lambda}) = (\frac{dg(\hat{\lambda})}{d\hat{\lambda}}^2)var(\hat{\lambda}).$$

In particular, the standard deviation of the estimated median survival time is

$$s.d.(t(\hat{50})) = t(\hat{50})/\sqrt{(r)}.$$

The limits of $100(1-\alpha)\%$ confidence interval for $t(50)$ are

$$t(\hat{50}) \pm z_{\alpha/2} s.d.(t(\hat{50})).$$

The standard deviation of the mean survival time μ is

$$s.d.(\hat{\mu}) = \mu/\sqrt{(r)}.$$

The limits of $100(1-\alpha)\%$ confidence interval for μ are

$$\hat{\mu} \pm z_{\alpha/2} s.d.(\hat{\mu}).$$

Example:

Consider a situation in which pieces of equipment are installed at different times. At a later date some of the pieces will have failed and the rest will still be in use. The aim is to study the lifetime distribution of this type of equipment. The data are summarized in the following table.

Lifetime in days									
2	72*	51	60*	33	27	14	24	4	21*

where starred quantities denote censored observations.

Assume this lifetime follows an exponential distribution.

- Estimate the mean lifetime and its 95% confidence interval.
- Estimate the median lifetime and its 95% confidence interval.
- Estimate the probability of lifetime of an equipment is greater than 32.

Solution:

$$log(2) = 0.693, r = 7, T = 308$$

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

$$\hat{\lambda} = r/T = 0.023$$

Since mean=λ^{-1} and $median = \lambda^{-1} \log 2$,

$$\hat{\mu} = 44, t(\hat{50}) = 30.5.$$

Their corresponding standard deviations are

$$s.d.(\hat{\mu}) = 44/\sqrt{7} = 16.63, s.d.(t(\hat{5}0)) = 30.5/sqrt7 = 11.53$$

95% confidence intervals are

$$(44 - 1.96 * 16.63, 44 + 1.96 * 16.63) = (11.4, 76.6)$$

for the mean and

$$(30.5 - 1.96 * 11.53, 30.5 + 1.96 * 11.53) = (7.9, 53.1).$$

$$P(T > 32) = S(32) = exp(-0.023 * 32) = 0.479.$$