

CHAPTER THREE

SURVIVAL DATA ANALYSIS (ANALYSIS OF TIME-TO-EVENT DATA)

3.1 INTRODUCTION

Survival analysis is the name for a collection of statistical techniques used to describe and quantify time to event data. Survival analysis concerns analysing the time to the occurrence of an event, e.g. time until a cancer patient dies.

The response is often referred to as a failure time, survival time, or event time. In survival analysis, we use the term ‘failure’ to define the occurrence of the event of interest (even though the event may actually be a ‘success’ such as recovery from therapy).

The term ‘survival time’ specifies the length of time taken for failure to occur. Situations where survival analyses have been used in epidemiology are stated in example 3.1.

Example 3.1

- ▶ Survival of patients after surgery.
- ▶ The time taken for a farm to experience its first case of an exotic disease.
- ▶ Time from diagnosis of cancer to death due to the cancer.
- ▶ Time from diagnosis of cancer to death due to any causes.
- ▶ Time from diagnosis of localized cancer to metastases.
- ▶ Time from remission to relapse of leukemia.
- ▶ Time to re-offending after being released from jail.
- ▶ Time between two attempts to donate a unit of blood for transfusion purposes.
- ▶ Time from HIV infection to AIDS.
- ▶ Time to the first goal (or next goal) in a hockey game.
- ▶ Time from exposure to cancer incidence in an epidemiological cohort study.

The aim in survival analysis is to analysis time to event data. These kinds of dada are encountered in many areas;

- **Medicine:** time to death for patient having a certain deaths (this explains the terminology “survival analysis)
- **Sociology :** time to finding a new job after a period of unemployment
- **Engineering:** time to failure of a machine etc

But survival analysis does not have to be about death: The word ‘survival’ does not necessarily relate to lack of death; it could mean failure to become diseased. In general, ‘survival’ means lack of experience of the event of interest.

Events may include death, injury, onset of illness, recovery from illness, transition above or below the clinical threshold of a meaningful continuous variable (e.g. CD4 counts), exclusive breast feeding, age at first sex, age at first marriage, or any designated experience of interest that may happen to an individual.

Survival time refers to a variable which measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest.

Three basic requirements for time-to-event measurements

- ✓ Agreed scale of measurement for time
- ✓ Unambiguous origin for the measurement of ‘time’
- ✓ Precise definition of ‘response,’ or occurrence of the event of interest

Nature of Survival Data: Censoring

Survival-time data have two important special characteristics:

1. Non Negativity:

Survival times are non-negative i.e we are interested in a non-negative response variable T ($T \geq 0$), consequently are usually positively skewed.

2. Censoring:

Typically, some subjects (i.e., units of observation) have **censored** survival times. In longitudinal studies, exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (those who are disease free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up period.

These individuals are called censored observations. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set. The following terms are used in relation to censoring:

- **Right censoring:** a subject is right censored if it is known that the event of interest occurs sometime after the recorded follow-up period. It is practically common types of censoring.

- **Left censoring:** a subject is left censored if it is known that the event of interest occurs some time before the recorded follow-up period. It is common in social scientists.
- **Interval censoring:** a subject is interval censored if it is known that the event of interest occurs between two times, but the exact time of failure is not known. In effect, we say ‘I know that the event occurred between date A and date B: I know that the event occurred, but I don’t know exactly when.

Censoring may occur in one of the following reasons:

- ☞ Termination of the study before the event occurs (administrative censoring);
- ☞ Death due to a cause not considered to be the event of interest (in cause-specific survival analyses); and
- ☞ Loss to follow-up, for example, if the patient emigrates.

Survival & Hazard Functions

Survival Function $S(t)$

Let ‘T’ denote a nonnegative ($T \geq 0$) random variable representing the lifetimes of individuals in some population. We treat the case where T is continuous. Let $F(t)$ denote the (cumulative) distribution function of T with corresponding probability density function $f(t)$. Note $f(t) = 0$ for $t < 0$. Then,

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

The probability that an individual survives to time t is given by the **survivor** function

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx.$$

This function is also referred to as the reliability function. Note that $S(t)$ is a monotone decreasing (non-increasing) function with $S(0) = 1$ and $S(\infty) = \lim_{t \rightarrow \infty} S(t) = 0$.

Conversely, we can express the probability density function $f(t)$ as;

$$f(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

Hazard Function $h(t)$

The hazard function specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t and is defined as;

$$h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

We see here that $h(t)\Delta t$ is approximately the probability of a death in $(t, t + \Delta t]$, given survival up to time t .

The hazard function is also referred to as the risk or mortality rate, the hazard rate, the instantaneous death rate, the intensity rate or the force of mortality. We can view this as a measure of intensity at time t or a measure of the potential of failure at time t . The hazard is a rate, rather than a probability. It can assume values in $[0, \infty)$.

Integrating $h(u)$ over $(0, t)$ gives the **cumulative hazard** function $H(t)$:

$$H(t) = \int_0^t h(u) du$$

❖ Relationship between $h(t)$, $H(t)$ and $S(t)$

The hazard function is given by:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} [\ln S(t)]$$

Survival function $S(t)$ is equivalent to the negative exponent of cumulative hazard function as:

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

Therefore, the cumulative hazard function is given as:

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

Example: Consider the exponentially distributed survival variable T with density function $f(t)$ as:

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

- ✓ Survival function: $S(t) = e^{-\lambda t}$
- ✓ Hazard function: $h(t) = \lambda$
- ✓ Cumulative hazard function: $H(t) = \lambda t$

3.2 KAPLAN-MEIER ESTIMATES OF SURVIVAL PROBABILITIES

If we have a random sample from the population when there is **no censoring** observation, the survivor function $S(t)$ is defined as the probability that an individual survives for a time greater than or equal to t . This function can be estimated by empirical survivor function, given by:

$$S_n(t) = \frac{\# \text{ of observations } > t}{n} = \frac{\#\{t_i > t\}}{n}$$

A widely used method for estimation of the survival function is the Kaplan Meier method. This method produces the **Kaplan-Meier estimator**, a nonparametric estimator, which does not assume any known algebraic form of the estimated survival function. The Kaplan-Meier estimator is also referred to as the **product-limit estimator**. Suppose k distinct survival times are observed.

Arranged in increasing order, they are $t_1 < t_2 < t_3 \dots < t_k$. At time t_1 , there are n_i subjects who are said to be **at risk**; that is, they survived up to this time (not including it) and were not censored. Denote by d_i the number of subjects who die at time t_i . To simplify notation, let $t_0 = 0$ and $d_0 = 0$. Then, the Kaplan-Meier estimator of the survival function $S(t)$ is:

$$\hat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq 0$$

Example 3.2 A biotech company conducted a 2-year clinical trial testing the efficacy of a new heart valve. The survival times (in months) of 10 patients with the heart valve implants were recorded.

The plus sign "+" next to the observation signifies that the observation is censored. The data are: 24+, 16+, 8, 19, 10, 8+, 5, 17, 20, 10

There are eight distinct survival times, given here in increasing order as:

5, 8, 10, 16, 17, 19, 20, 24

Then, using Kaplan Meier estimation methods, we can find the estimated survival function $S(t)$ in the following table.

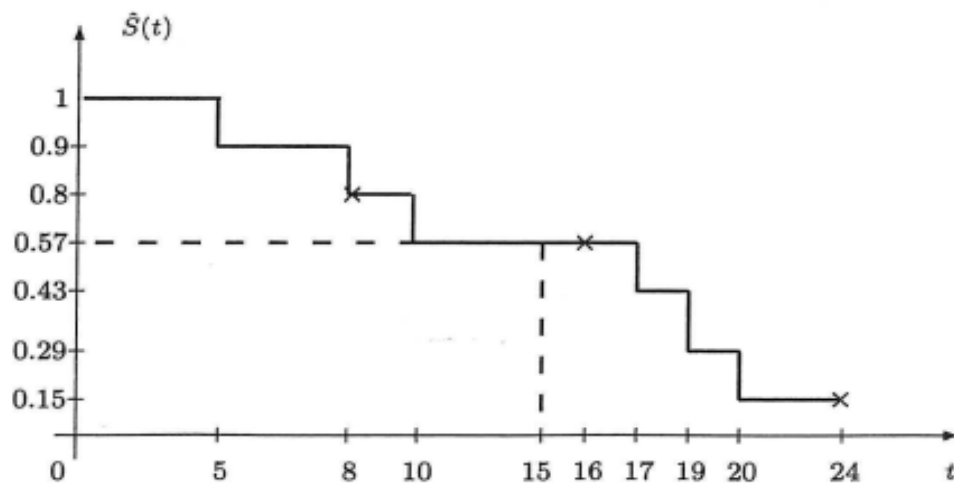
Table 3.1 Estimation of $S(t)$ by the Kaplan–Meier Method in Example 3.2

Time t_i	At Risk n_i	Died d_i	Censored at Time t_i	Survival Rate $\left(1 - \frac{d_i}{n_i}\right)$	Estimator $\hat{S}(t), t_i < t < t_{i+1}$
0	10	0	0	$1 - 0 = 1.00$	1.00
5	10	1	0	$1 - \frac{1}{10} = 0.90$	$(1.00)(0.90) = 0.90$
8	9	1	1	$1 - \frac{1}{9} = 0.89$	$(0.90)(0.89) = 0.80$
10	7	2	0	$1 - \frac{2}{7} = 0.71$	$(0.80)(0.71) = 0.57$
16	5	0	1	$1 - 0 = 1.00$	$(0.57)(1.00) = 0.57$
17	4	1	0	$1 - \frac{1}{4} = 0.75$	$(0.57)(0.75) = 0.43$
19	3	1	0	$1 - \frac{1}{3} = 0.67$	$(0.43)(0.67) = 0.29$
20	2	1	0	$1 - \frac{1}{2} = 0.50$	$(0.29)(0.50) = 0.15$
24	1	0	1	$1 - 0 = 1.00$	$(0.15)(1.00) = 0.15$

The Kaplan-Meier Survival Curve

The *Kaplan-Meier survival curve* is the plot of the Kaplan-Meier estimator of the survival function $\hat{S}(t)$ against time t . This curve is a step-function that decreases at the times of deaths. The censored times are usually marked by a cross (x). If a death and a censoring occur at the same time, a cross for the censored observation is put at the bottom of the step.

Example 3.3 In Example 3.2, the Kaplan-Meier survival curve is a plot of $\hat{S}(t)$ given in the last column of Table 3.1 against time(t). For instance, from this plot; the estimated probability of 15-month survival is 0.57.

**Figure 3.1** The Kaplan–Meier survival curve in Example 3.2

3.3 COMPARING SURVIVAL CURVES IN TWO GROUPS

To compare the efficacy of two treatments, subjects who enter a clinical trial are randomly placed into two treatment groups. The survival data are then recorded for each group. The question of interest is whether the two treatments are equally effective. This translates into testing whether the survival functions for these groups differ significantly. A two-sided test of statistical hypotheses is appropriate. The hypotheses are

$H_0: S_1(t) = S_2(t)$ for all t

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

The most commonly used test for data with censored observations is the **log-rank test**, which derives its name from the fact that it is related to a test that uses logarithms of ranks of observations.

To compute the log-rank statistic, proceed as follows. Denote by $t_1 < t_2 < \dots < t_k$ the ordered uncensored observations (times of deaths) in both samples combined. At each time t_i , the data can be summarized by a 2 x 2 table:

Group	Status of Subject		Total
	Died	Survived	
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Here d_{1i} and d_{2i} are the numbers of subjects who died at time t_i in groups 1 and 2, respectively; $d_i = d_{1i} + d_{2i}$; n_{1i} and n_{2i} are the numbers of subjects at risk at time t_i in groups 1 and 2, respectively; and $n_i = n_{1i} + n_{2i}$

The null hypothesis is equivalent to independence of the "group" and "status of subject" variables in all 2 x 2 tables. Under H_0 , d_{1i} is a hyper geometric random variable with parameters n_i (the population size), n_{1i} (the size of the group of interest), and d_i (the sample size). The expected value of d_{1i} is

$$\mathbb{E}(d_{1i}) = \frac{n_{1i} d_i}{n_i}$$

The variance is

$$\text{Var}(d_{1i}) = \frac{n_{1i} n_{2i} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}$$

Summing over all i , $i = 1, \dots, k$, yields a statistic

$$U = \sum_{i=1}^k (d_{1i} - \mathbb{E}(d_{1i}))$$

where

$$\mathbb{E}(U) = 0 \quad \text{and} \quad \text{Var}(U) = \sum_{i=1}^k \frac{n_{1i} n_{2i} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}$$

Standardizing leads to the log-rank test statistic

$$z = \frac{U}{\sqrt{\text{Var}(U)}}$$

Has an approximately a standard normal distribution with mean zero and variance one ($N(0, 1)$). Alternatively the log-rank statistic is z^2 , which has an approximately chi-squared distribution with one degree of freedom.

Example 3.4 A clinical trial is conducted to evaluate a new nicotine patch. Subjects are randomly assigned to either the treatment group or the control group. The treatment group receives the nicotine patch under study, while the control group receives the best nicotine patch currently available on the market. The measurement is the length of time (in months) that a subject goes without a cigarette. The data for the two groups are as follows:

Treatment	3.4	3.6+	4.1	4.9+	5.8+
Control	2.0	3.7+	4.3	4.9+	

The researchers would like to know whether the two nicotine patches differ significantly, so a log-rank test is performed. The test hypotheses are

$$\begin{aligned} H_0 &: S_{\text{treatment}}(t) = S_{\text{control}}(t) \quad \text{for all } t \\ H_1 &: S_{\text{treatment}}(t) \neq S_{\text{control}}(t) \quad \text{for some } t \end{aligned}$$

The times of events in both groups combined are 2.0, 3.4, 4.1, and 4.3. The 2 x 2 tables corresponding to each of these times follow:

$$t_1 = 2.0$$

Group	Status of Subject		Total
	Died	Survived	
Treatment	0	5	5
Control	1	3	4
Total	1	8	9

$$d_{11} = 0, \mathbb{E}(d_{11}) = \frac{(5)(1)}{9} = \frac{5}{9}, \text{Var}(d_{11}) = \frac{(5)(4)(8)(1)}{(9)^2(8)} = \frac{20}{81}$$

$$t_2 = 3.4$$

Group	Status of Subject		Total
	Died	Survived	
Treatment	1	4	5
Control	0	3	3
Total	1	7	8

$$d_{12} = 1, \mathbb{E}(d_{12}) = \frac{(5)(1)}{8} = \frac{5}{8}, \text{Var}(d_{12}) = \frac{(5)(3)(7)(1)}{(8)^2(7)} = \frac{15}{64}$$

$$t_3 = 4.1$$

Group	Status of Subject		Total
	Died	Survived	
Treatment	1	2	3
Control	0	2	2
Total	1	4	5

$$d_{13} = 1, \mathbb{E}(d_{13}) = \frac{(3)(1)}{5} = \frac{3}{5}, \text{Var}(d_{13}) = \frac{(3)(2)(4)(1)}{(5)^2(4)} = \frac{6}{25}$$

$$t_4 = 4.3$$

Group	Status of Subject		Total
	Died	Survived	
Treatment	0	2	2
Control	1	1	2
Total	1	3	4

$$d_{14} = 0, \mathbb{E}(d_{14}) = \frac{(2)(1)}{4} = \frac{1}{2}, \mathbb{V}ar(d_{14}) = \frac{(2)(2)(3)(1)}{(4)^2(3)} = \frac{1}{4}$$

Consequently,

$$\begin{aligned} U &= \left(0 - \frac{5}{9}\right) + \left(1 - \frac{5}{8}\right) + \left(1 - \frac{3}{5}\right) + \left(0 - \frac{1}{2}\right) \\ &= -0.2806 \\ \mathbb{V}ar(U) &= \frac{20}{81} + \frac{15}{64} + \frac{6}{25} + \frac{1}{4} = 0.9713 \end{aligned}$$

The log-rank test statistic is $Z = -0.2806/\sqrt{0.9713} = -0.2847$. The approximate P-value for the two-sided test is $2\mathbf{P}(Z > 0.2847) = 0.7759$. Alternatively, the test statistic is $Z^2 = 0.081$ and the approximate P-value is $\mathbf{P}(X^2(1) > 0.081) = 0.7759$. Thus the null hypothesis of equal survival functions is not rejected at the 0.05 level of significance, and the conclusion is that the two nicotine patches do not differ significantly. Graphically survival curve for the two groups is shown below.

