UNIT 14 INTRODUCTIVE CONCEPTS OF SURVIVAL ANALYSIS

Structure

14.1 INTRODUCTION

In medical studies, generally, we are interested in how subjects/patients of the study respond to treatment and how treatment influences disease progression. This becomes a far more powerful tool than merely looking at occurrences when we record not only the response of the treatment (event) but also the time range over which the event occurs and how long they survived and what is the risk of their failure. For that, we follow up or track the subjects/patients until an event happens (death) and note down the time of event occurrence. Such types of studies can take weeks, months, or even years to complete. If the event occurred in all subjects, then we may use many methods of analysis that you have studied in previous courses such as mean, median, mode or standard deviation to summarised the data, coefficient of correlation, Spearman's rank coefficient of correlation to examine the association between the response of the treatment and age or gender, etc., regression and logistic regression to explore the relationship between the dependent/outcome variable and explanatory variable, the parametric and non-parametric tests for testing the significance of the impact of the treatment, etc.

But if at the end of follow-up, we could not observe the time of the event of all subjects due to shorter follow-up time, lost to follow up due to certain reasons such as a move to another locality or death from a cause other than the disease of interest or refuse to participate in the study after a certain time, etc. (such observations are called censored observations) then we cannot use such methods for analysing the data. One way is that we can remove censored observations from the study but if we do so we get biased results. Further, survival data are rarely normally distributed, but are skewed and comprise

typically many early events and relatively few late ones. These features of the data make the use of special methods to analyse such type of data. Therefore, the researcher developed a technique that is also used in the analysis of such type of data. This technique is known as **survival analysis**.

Therefore, if we are interested to study:

- the time up to which an event of interest happens in the subjects/patients
- how long a subject/patient will survive
- the risk of failure (hazard rates) of the subjects/patients

then we use survival analysis.

Survival analysis is a collection of statistical techniques used to describe and analyse time to event data.

In this unit, we shall familiarise you with the basic concepts of survival analysis. We start with what survival analysis is? And in which situations, we apply it. In Sec. 14.2, a discussion on the design of the study which is appropriate for the application of survival analysis is given. Further, we also explore the data structure of these designs and the usual analysis in the absence of survival analysis in this section. Censoring is the main concept that is to be needed to apply the survival analysis. We discuss the concept of censoring and its types in Sec. 14.3. The unit will conclude with a discussion on tools for survival analysis under which we explain the survival time, event, survival functions: survival function, cumulative distribution function, probability density function, hazard and cumulative hazard functions of time to event in Sec. 14.4. The survival functions are mathematically equivalent and given any one of the remaining can easily be obtained. Therefore, in Sec. 14.5, we discuss the relationships between them. A brief summary of what we have covered in this unit is described in Sec. 14.6. If you will face some problems in the solution of the exercises then you can go through Sec. 14.7 in which we provide the solutions/answers to the exercises.

Objectives

After studying this unit, you should be able to:

- know why we cannot apply the statistical tools of analysis which we have studied in the previous courses of PGDAST.
- identify situations where survival analysis can be applied;
- explain the importance of survival analysis;
- choose appropriate study designs for survival analysis;
- explain the concept of censoring in survival analysis;
- identify the notations and terminology commonly used in survival analysis;
- define and calculate the survival functions;
- find the proportion of individuals who remain free from the event of interest before/after a certain time;
- compute the risk of the event of interest at a particular point time, among those who have survived until that point; and
- evaluate the risk of the event at a particular time interval.

We now come to the section-wise discussion through which you will get our objectives. Let us start with the first main section design of the study and data

structure of survival analysis and look at them from the data analysis point what are the difficulties in the usual analysis.

14.2 DESIGN OF STUDY AND DATA STRUCTURE OF SURVIVAL ANALYSIS

In the course MSTE-003: Biostatistics-I, you have learned various study designs namely: cross-sectional studies, case-control study, cohort study, randomized controlled trials, etc,. These are the various designs that help us to answer epidemiological questions and the selection of a specific design is based on the question you would like to answer. You also have the knowledge that the data analysis is heavily influenced by the study design, structure and type of data. We will recall two designs namely cohort study and randomized controlled trial in this section. The main idea of this description is not to understand more about these designs but to look at them from the data analysis point and observe what are the difficulties in the usual analysis.

14.2.1 Design and Data Description of a Cohort Study

A cohort study is usually a prospective study where a group (cohort) of subjects/patients/individuals having the similar characteristics has been identified based on their exposure status and divided into two groups out of which one group of exposed subjects/patients/individuals and the other of non-exposed subjects/patients/individuals. After that, both groups follow up for a predefined time to observe the occurrence of an outcome of interest. During the follow-up period, some of them experience the outcome of interest and others do not in both groups. End of the follow-up period, the occurrence of outcome in the exposed and non-exposed groups in terms of rate of occurrence and duration for that occurrence are calculated and compared. The design and structure of a cohort study is shown in Fig. 14.1.

Prospective study is a type of group study in which participants are enrolled in the study before they develop the debases or event of interest.

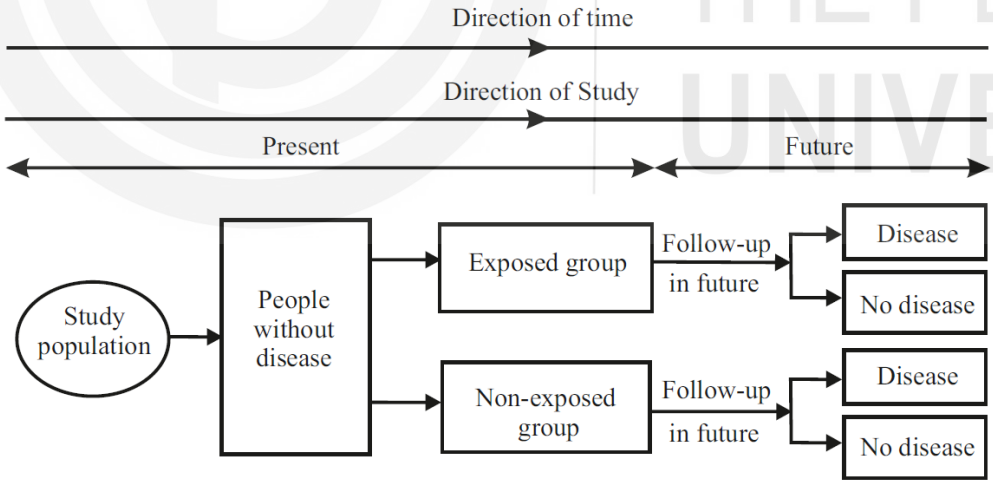

Fig. 14.1: Design of cohort study.

Consider a study where the cohort is defined by the patients/subjects who had a heart attack, were treated in a hospital and were discharged. This cohort of patients was followed up after discharge. The basic objective of the study was to understand the mortality experience of these subjects and further study the gender-wise difference in mortality experiences. There were 500 such subjects and the follow-up period was for 15 years. We will have a look at the partial data presented in Table 14.1. The event status at the end of the follow up is written as yes = 1 and no = 0. Gender male is coded as 0 and females as 1.

Table 1: Partial data of the subjects

Subject No.	Gender (Male = 0, Female = 1)	Death (Yes = 1, No = 0)
1	0	0
2	0	0
3	1	0
4	0	1
5	0	0
6	0	1
7	0	0
8	0	1
9	1	1
10	0	0
11	0	0
12	0	1
13	0	0
14	1	1
15	0	0
16	0	0
17	1	1
18	1	1
19	0	0
20	0	1
.	.	.
.	.	.
.	.	.
491	0	1
492	0	1
493	0	1
494	0	1
495	1	1
496	1	1
497	1	0
498	1	0
499	0	0
500	0	1

One interesting question could be **"is the death rate among males different from females?"** We will try to answer this question statistically.

Here gender is the exposure variable and mortality is the outcome variable. Both are categorical and hence fit for a 2×2 chi-square test.

Exposure	Event (Death)		Total
	Yes	No	
Male	111	189	300
Female	104	96	200
Total	215	285	500

Mortality among males = 111/300 = 37%,

Mortality among females = 104/200 = 52%

We calculate the value of the chi-square statistic as we have calculated in Unit 11 of the course MSTE-003: Biostatistics-I.

Chi square = 11.01 and $p < 0.001$. Since the p-value is less than $\alpha = 0.05$ so deaths also related to gender.

Further, the knowledge of the epidemiology session will prompt us to calculate relative risk (RR).

$$RR = \frac{\text{Mortality among males}}{\text{Mortality among females}} = \frac{0.37}{0.52} = 0.71$$

Are these statistics adequately describing the study results? Will additional information on time to death add any advantage to the description and analysis?

We will have a look at the modified data given in Table 2.

Table 2: Partial data with time to death of the subjects

Subject No.	Gender (Male = 0, Female = 1)	Time to event (in days)	Death (Yes = 1, No = 0)
1	0	2177	0
2	0	2171	0
3	1	2190	0
4	0	296	1
5	0	2132	0
6	0	201	1
7	0	2123	0
8	0	1495	1
9	1	921	1
10	0	2176	0
11	0	2172	0
12	0	1670	1
13	0	2191	0
14	1	866	1
15	0	2164	0
16	0	2168	0
17	1	904	1
18	1	2354	1
19	0	2145	0
20	0	62	1
.	.	.	.
.	.	.	.
.	.	.	.
490	0	475	0
491	0	397	1
492	0	14	1
493	0	10	1
494	0	69	1
495	1	32	1
496	1	10	1
497	1	665	0
498	1	726	0
499	0	531	0
500	0	258	1

Here the entry of an additional variable, time to death, brings us a wealth of information about when exactly the death happened. Infect this is the contribution of cohort follow up. That means, to study the mortality experience we have the following additional information

- Mortality happened or not?
- If mortality happened when it happened?
- If mortality was not happened, till what time do we have this information?

Thus, our outcome variable is no more a binary variable but it is a combination of a binary variable namely **event happened or not** and a continuous variable namely **time to event** or duration of follow up. The combination of columns three and four together form our outcome variable. A closer look discloses that the new outcome variable has many more properties.

Firstly, the event did not happen for all and the event differs from person to person and it mostly has a skewed distribution as shown in Fig. 14.2. Secondly, the follow-up time for people for whom the event did not happen also varies as shown in Fig. 14.3. This is usually known as a loss to follow-up which is a common phenomenon in Cohort study. Time to the event is not complete for people for whom the event did not happen.

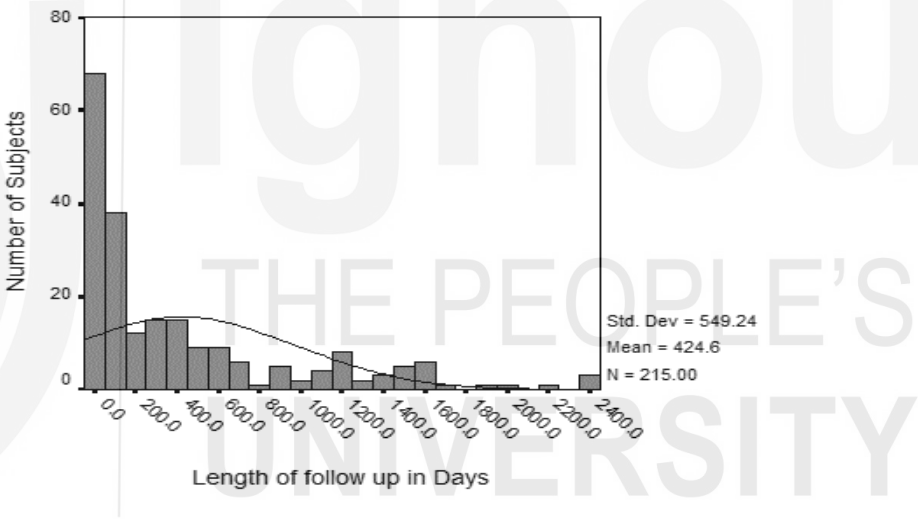

Fig. 14.2: Survival experience for the people who died.

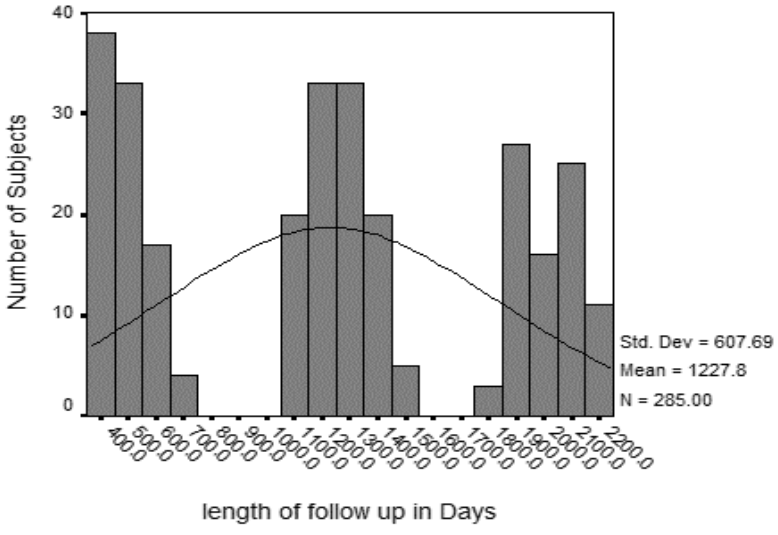

Fig. 14.3: Survival experience for the people who were alive.

The survival time for males and females also varies as shown in Fig. 14.4 and 14.5 respectively.

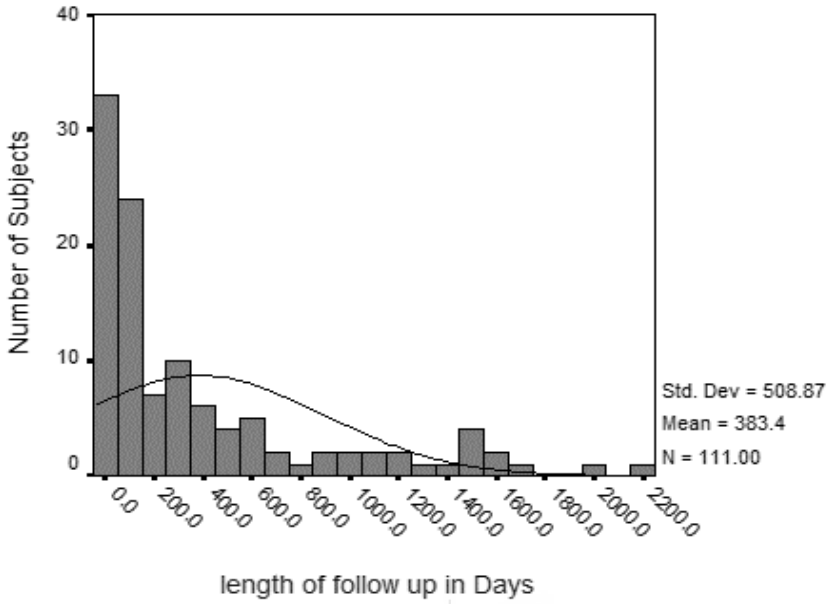

Fig. 14.4: Survival experience for the male who died.

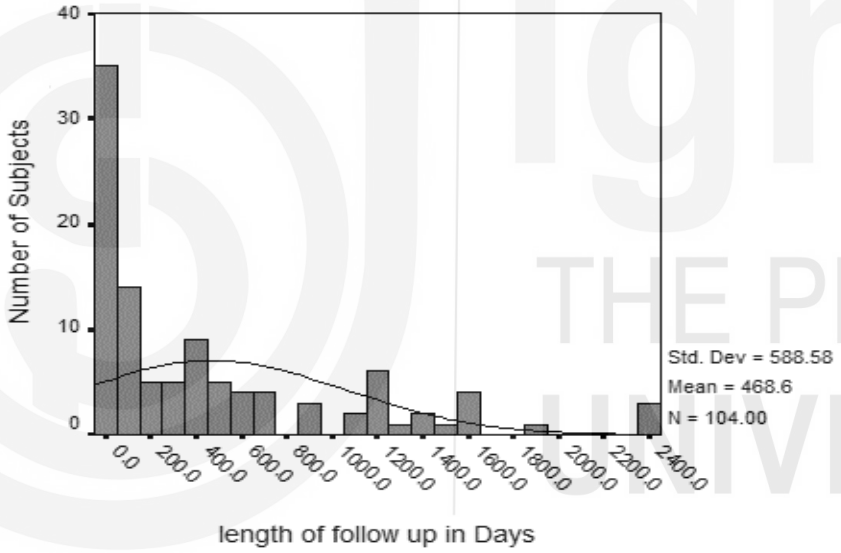

Fig. 14.5: Survival experience for the female who died.

In summary, the routine analysis does not address the following issues:

- The length of follow-up for different subjects is different.
- Does not take care of the duration for an event to happen (time to event).
- Do not consider lost to follow-up.
- End up with the incorrect result.
- Survival rate at the end vs survival experience during the period.

Survival analysis is a special statistical technique which incorporates all the above-mentioned points and provides a summary and inference about the data.

14.2.2 Design and Data Description of a Randomized Controlled Trial

Now let us look at another analytical study design namely an intervention trial. This is again a prospective study where the effect and value of intervention are compared against a control group. This design is usually done to test the treatment efficacy on disease conditions. The most widely used intervention design is the randomized controlled trial where minimum there will be a control group and one intervention group. The control group of patients receives no treatment or the standard available treatment and the intervention group of patients receives new intervention which requires testing. The allocation of treatment to intervention and control groups is done through a process known as randomization. After treatment allocation, the subjects were followed up for a specified time duration to observe the outcome/ event of interest. The usual outcome is the cure of the disease. Let us have a close look at the typical diagrammatic presentation of a randomized controlled trial is shown in Fig. 14.6.

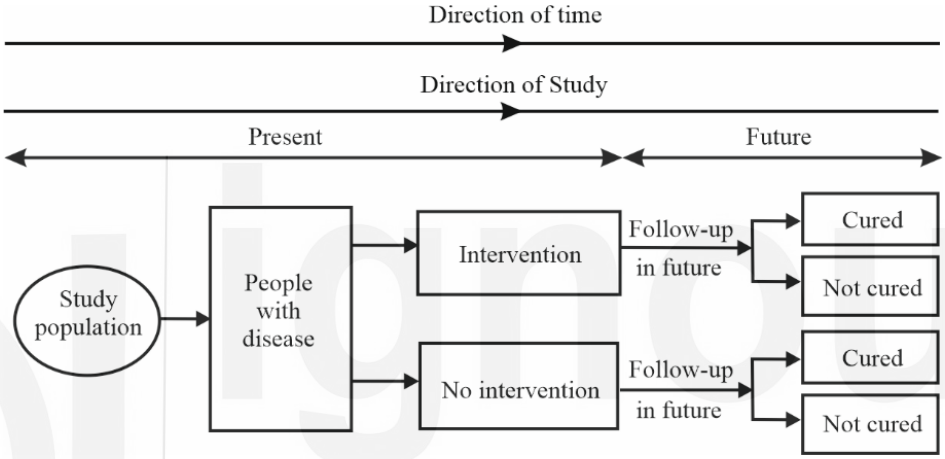

Fig. 14.6: Design of randomized control trial study

Consider a study of breast cancer patients where the control group received the surgery alone and the intervention group received surgery and chemotherapy. The outcome variable is the duration of survival after treatment. We may plan to follow up on these subjects say for 5 years. For all patients, the outcome need not have to happen during this period. However, for the patients for whom the outcome happened we get a true time of the event. In other cases, we have incomplete time for the event. A comparison of the design of the intervention study with that of the cohort study reveals to us that both these designs generate data of the same nature and hence are analysed in the same way. All problems described for the cohort study are applicable to this design also.

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs.

In the above sub-sections, we just give the idea in which situation you will go through the survival analysis. We now define what survival analysis is?

In a general sense, survival analysis is a collection of statistical methods for data analysis for which the outcome (response or dependent) variable is the time until an event occurs. It means that survival analysis is used to analyse data in which the time until the outcome (event) is of interest. We measure the time from we start the study until an event occurs. Originally, survival analysis was concerned with the time from the treatment of a subject until the death of the subject due to which the name survival analysis was to be given. But survival analysis is also applicable to many other kinds of areas such as the time until second heart attack recurrence, time until AIDS for HIV patients, time until a machine part fails, time until a post-graduate person is unemployed, etc.

Now, you may try the following exercises.

E1) Write the situation where we use survival analysis with an example.

E2) Why are the statistical tools you have studied in the previous courses of PGDAST not used in survival analysis?

Now let us discuss some of the properties of these types of data.

14.3 CONCEPT OF CENSORING

The observation of survival data has two components namely the beginning point and endpoint(described in Section 14.4). The beginning point is the time point when the follow-up starts and the end point when the event of interest happened. The beginning point may be the day of diagnosis and the endpoint is the day of death as in our example of breast censor. In the intervention trial, the beginning point is the day of allocation and the endpoint is the day on which the patient got cured. But in real life, we face the major problem to follow up the study for observation of survival time. The follow-up may be incomplete, that is, the event of interest may not be observed on every subject/patient/ individual. If we do not know the survival time exactly for all the subjects/ patients/ individuals under the study then the data so obtained is called censored data. There are three main reasons for censoring:

1. The first is that the study has a shorter follow-up time. In this situation, some of the subjects have not experienced the event of interest at the end of the study (patients Cand D of the example considered in Section 14.4). Due to this, we cannot know the exact survival time for such subjects.
2. Second is the withdrawal of the subject(s) from the study for some reasons such as death from a cause other than the disease of interest as an accident (patient E of the example considered in Section 14.4).
3. The third is lost-to-follow-up. This situation may arise when a subject moves to another locality or refuse to participate in the study after a certain time (patient B of the example considered in Section 14.4). Therefore, we lost to follow up before the experience of the event of interest. We cannot know whether they would have had an event of interest if they had lasted for their complete possible follow-up duration.

Therefore, when the survival time is incomplete due to loss to follow up or subjects in the study have not experienced the event of interest at the end of the study then the survival time is called a censored survival time and the data so obtained is called the censored data.

There are three basic distinct methods of censorship viz, left, right and interval which are as follows:

14.4.1 Left censoring

Sometimes, we face the situation when we know that the event happened but the time at which it happened is not known which makes the observation incomplete.

Left censoring arises when we know only that the event of interest occurred before a certain time but we do not know the exact time of its occurrence. For example, suppose patient suffering from cold and fever comes to the hospital and the doctor suspects that the patient is a COVID-19 patient. For confirmation, the doctor takes the COVID-19 test on the patient. If the test result is positive, then we know that the patient is infected by the COVID-19, but we do not know the exact time when the patient has been infected. This generates left-censored observation.

14.4.2 Right censoring

Right censoring occurs whenever the exact time of occurrence of an event is not known. This can happen either the study terminates before the subject does not experience the event of interest, i.e., he/she will live longer than the duration of the study, or could not be a part of the study completely and left early without experiencing the event of interest, i.e., they left and we could not study them any longer. In this censoring, we follow the subjects from the beginning of the test until they are lost to follow up or end the study. In such circumstances, we know the exact time of the starting point but do not know the exact time of occurrence of an event of interest. The observation is incomplete on the right side. For example, the cancer patient B lost to follow up after a period of study due to moving to another century and the patients C and E were alive at the end of the study. In these cases, we know the starting point but do not know the exact survival times of these patients. Similarly, if a COVID-19 patient example, if the doctor of the hospital starts the treatment of the COVID-19 patient but after some days he/she leaves the hospital due to certain reason then we will know the exact time of starting the treatment but will not observe the exact time of the recovery of the patient from COVID-19. So, we get the right-censored observation.

Right censoring is the most common type of censoring for survival time. So in a general sense, censoring in survival means right censoring.

There are situations where it is known that the event happened but the time at which it happened is not known which makes the observation incomplete. Another situation is the event happened but do not know the beginning of the exposure. These all contribute to an incomplete observation and this incompleteness is more on the left side. These types of observations are called left censored.

Survival analysis mostly dealt with right-censored observation and we also follow the same way.

14.4.3 Interval censoring

If we know only the time interval in which the event of interest has occurred but do not know the exact time of its occurrence then the censoring is called interval censoring. For example, if a group of smokers is being assessed half-yearly for the development of coronary heart disease and after two years, we observe that the disease has been developed in some person then we do not the exact time of its development because we only assess the same after every six months. We know only it developed between 1.5-2nd year. In the case of the COVID-19 patient, if the patient is put under the treatment and after certain days the doctor again takes the COVID-19 test and if the test results in negative then we know only that the patient cured between an interval but do not the exact time of the cure so this will give interval-censored observation.

For better understanding answer the following exercises.

E3) Define censoring.

E4) Describe different types of censoring with the help of examples other than those discussed in this unit.

14.4 TOOLS FOR SURVIVAL ANALYSIS

In MST-002: Descriptive Statistics, you have studied that we generally use the mean and standard deviation to the measures of central tendency and variability respectively. But when we deal with the study for which the outcome variable of interest is time until an event occurs (survival data) and some of the individuals have not had the event of interest, and thus their true time to event is unknown then in such cases we cannot apply mean and standard deviation. Further, survival data are rarely normally distributed, but are skewed and comprise typically many early events and relatively few late ones. These are the features of the data that make the special methods called **survival analysis**. Survival analysis is a collection of statistical procedures for data analysis.

To understand the concepts of survival analysis, you should first understand the notations and terminology of survival analysis in Biostatistics which have been reinvented several times in different disciplines.

Time

In survival analysis, by time, we mean the time from the beginning (Starting) of follow up of a subject/patient until an interesting event occurs. It can measure in days, weeks, months, years or some other measures of time. In a survival study, the time is generally called **survival time** because it gives the time that a subject/patient has survived over some follow up period. It has two components. The first is the **beginning (starting) point** when the follow-up start (this component is called clock starts) and the second is the endpoint when the event of interest happened (called clock ends). Therefore, the interval between a **beginning point** and an **endpoint** of follow up is called the **survival time**. We can easily calculate the survival time using the following formula

Survival Time = End time point – Start time point

For example, suppose an epidemiologist wishes to understand the length of survival of cancer patients. From the date of diagnosis, he treated the patients and followed up. Suppose he had the following information on the six cancer patients. Patient A entered the study in March 2021 and died in June 2021. Patient B entered the study in April 2021 and moved out to another country in September 2021. Patients C and D entered the study in August 2021 and September 2021, respectively. Patient D died in March 2022 and patient C was alive at the end of the study in April 2022. Patient E joined the study in December 2021 and is alive till the end of the study in April 2022. Patient F entered the study in January 2022 and died in March 2022 due to an accident. We can describe the same narration in the following table:

Table 3: Follow up study of cancer patients

Patient ID	Admission Date	Follow Up Date	Survival Time (in months)	Vital Status
A	March 2021	June 2021	4	Died
B	April 2021	September 2021	6	Lost to follow up
C	August 2021	April 2022	9	Alive
D	September 2021	March 2022	7	Died
E	December 2021	April 2022	5	Alive
F	January 2022	March 2022	3	Died due to accident

Patient A entered the study on March 2, 2021. It means that the follow-up start from that date and the **beginning point** is March 2021. Patient A died in June 2021. So the endpoint is June 2021. Hence the survival time of Patient A is 4 months (March 2021- June 2021). Similarly, the **beginning point** of follow up

for patient B is April 2021 and he/she moved out to another country in September 2021. So, we have known that patient B has survived up to September 2021 and after that, we do not have any information about him/her so the endpoint for patient B is September 2021. Therefore, the survival time of patient B is 6 months. Similarly, we can find the survival time of other patients.

In such type of studies, the subjects/ patients entre in the study at the time of diagnosis the disease. As in our example, the subjects enter the study from the date of diagnosis. We can show the survival time of the patients graphically as shown in Fig. 14.7.

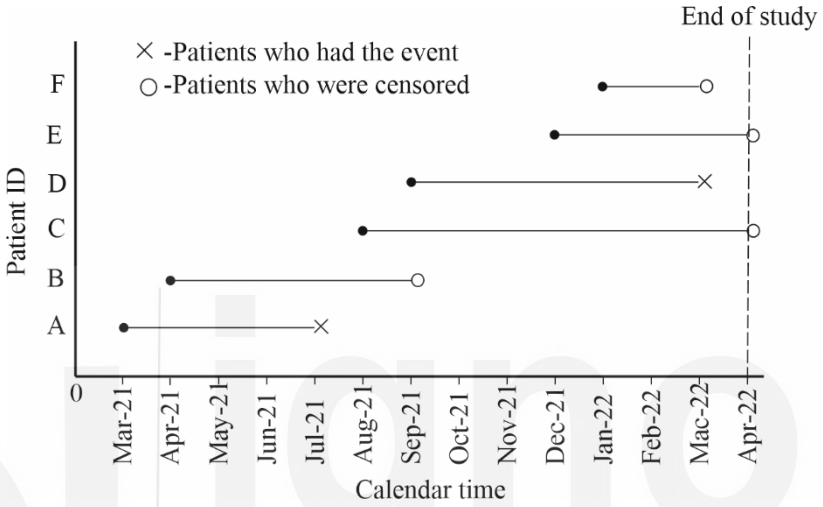

Fig. 14.7: Calendar time of the patients.

This time is known as calendar time. Since observations do not always begin at zero, therefore, it is not easy to compare the survival time of the subject so we convert the calendar time into a comparable format, regardless of when they were enrolled in the study. For that, we bring all subjects to a common starting point where the time t is zero (t = 0). Therefore, we convert the calendar time to survival analysis format as shown in Fig. 14.8.

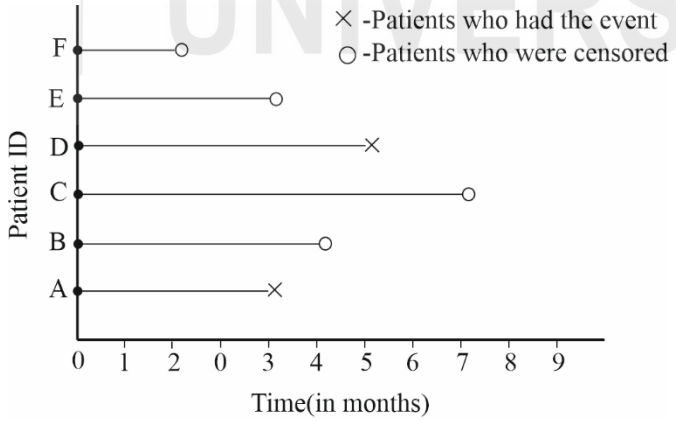

Fig. 14.8: Survival time of the patients in months.

The set of such survival times recorded during a study is called **survival data.**

Event

By event in survival analysis, we mean the outcome of our interest that may happen to a subject/patient/individual. It can be the development of a disease, release from remission, response to treatment, recovery, death, etc. In the above example, the epidemiologist is interested to know the length of survival of cancer patients. So the outcome of interest of the epidemiologist is death.

Therefore, in this case, death is an event. Similarly, if an analyst wishes to know the effect of smoking increases the risk of coronary heart disease. Therefore, in this case, coronary heart disease is the event. In other words, we can be defined an event as

An event is a qualitative change that can be placed in time.

By a **qualitative change**, we mean a transition from one state to another. For example, a "death" is a transition from the state of being **alive** to the state of being **dead**. Coronary heart disease consists of the transition from **non-coronary heart disease** to **coronary heart disease**.

Random Variable

We are now ready to teach the fundamental mathematical concepts and notation used in survival analysis. Since survival time may vary from person to person and is subject to random variations and like any random variable, we denote survival time by a capital letter T. Since it describes the time to event of interest happen therefore it is nonnegative and takes any value equal to or greater than zero. That is T lies between 0 to ∞. To denote any particular (specific) value of survival time T, we use the small letter t. For example, if we are interested whether a brain tumour patient survives more than 12 months then small t equals to 12 and random variable T takes the value as $T > t \Rightarrow T > 12$.

Now, we are ready to define various functions that are used in survival analysis. In survival analysis, we use the following functions to study the survival time.

a) Survival function (survivorship function).

b) Cumulative distribution function

c) Probability density function, and

d) Hazard and cumulative hazard functions.

14.4.1 Survival Function

Imagine that you have a patient who is suffering from a brain tumour and his family members ask you what is the chance of surviving him for more than two years. Then you can answer such questions with the help of the survival function.

We defined survival function as the probability that an individual (a subject) survives more than some stated (specified) time t. It is denoted by S(t). Mathematically, we can say that S(t) gives the probability that the random variable T exceeds a specified time t. Therefore, in symbolical form, we can define the survival function as

$S(t) = P[\text{An individual survives more than time t}]$

$S(t) = P[T > t]$ … (1)

With the help of the survival function, we can obtain the probability of surviving at each point in time. For example, t = 1, 2 , 5..

$S(1) = P[T > 1]$

$S(2) = P[T > 2]$

$S(5) = P[T > 5]$

The survivor function is essential in a survival study because calculating survival probabilities for various values of t offers essential summary information from survival data and in most applied cases, we are generally interested in describing how long the study individual is alive rather than how quickly dies. Thus, in survival analysis, we focus on the estimation of the survival function.

The survival function is also known as the **cumulative survival rate.**

Since survival function S(t) is a probability, therefore it always lies between zero and one ($0 \leq S(t) \leq 1$) for all values of t. Theoretically, the time t ranges from 0 to infinite. The survival function at t = 0 is 1, i.e.

$$S(0) = P[T > 0] = 1$$

It is so because at the starting point of the study no one has experienced the event of interest that is, all subjects/patients survive at time point t = 0.

Similarly at $t = \infty$, the survival function is 0, i.e.,

$$S(\infty) = P[T > \infty] = 0$$

Since nobody would survive forever so the survival function becomes zero.

The curve between the survival function S(t) and time t is called the **survival curve**. It is given by Berkson in 1942. Theoretically, the survival curve is smooth as shown in Fig. 14.9.

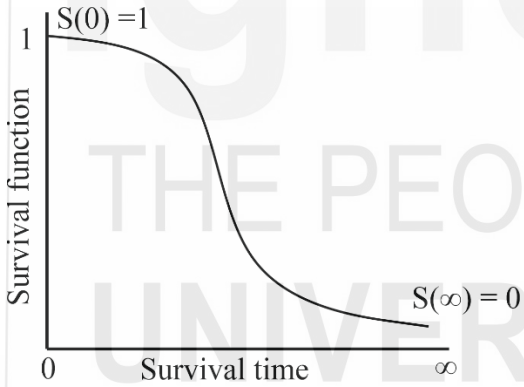

Fig. 14.9: Survival function.

Initially, the curve starts with S(t) = 1 and heads downward as t increases and finally falls to zero at $t = \infty$. If the survival curve is steep towards the centre as shown in Fig. 14.10(a) then it represents the low survival rate or short survival time and if the survival curve is flat as shown in Fig. 14.10(b) then it represents a high survival rate or longer survival.

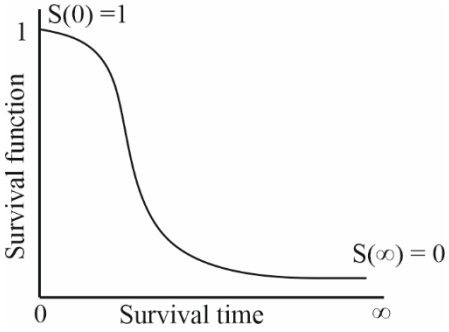

Fig. 14.10(a)

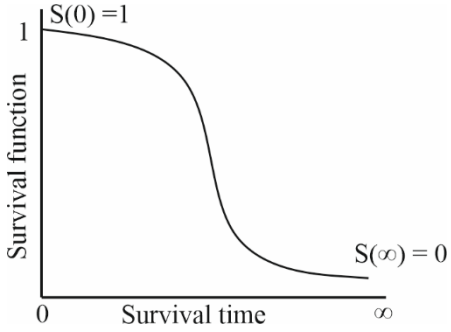

Fig. 14.10(b)

Since the survival data are rarely normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. In such a case, the mean would be strongly affected by outliers. In such a case median is used to describe the central tendency in compression of mean as described in Unit 1 of MST-002: Descriptive Statistics. The survival function or survival curve is also used to find the median or other percentiles and to compare survival distributions of two or more groups. For finding the median we draw a line parallel to the X-axis from the point 0.5 of the Y-axis on the survival curve and then a perpendicular from that point of the curve to the X-axis as shown in Fig. 14.11. The point on the X-axis represents the position of the median.

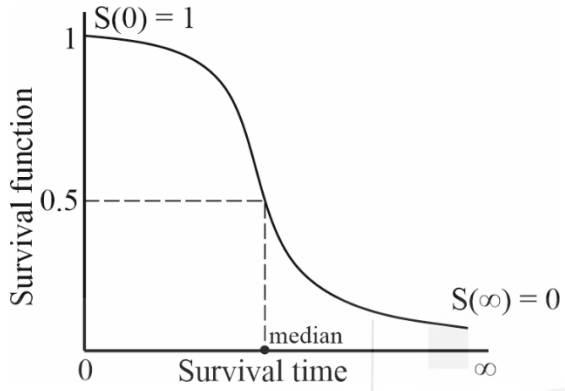

Fig. 14.11: Survival function with Median.

14.4.2 Cumulative Distribution Function

In survival analysis, another useful function is the cumulative distribution function that gives the probability that the survival time is less than or equal to a specific time. For example, if you have a patient who is suffering from COVID-19 and his family members ask you what is the chance of curing him/her before 14 days. Then you can answer such questions with the help of the cumulative distribution function. It is denoted by F(t). Mathematically, we can say that F(t) gives the probability that the random variable T is less than or equal to a specified time t. Therefore, in symbolical form, we can define the cumulative distribution function as

$F(t) = P[\text{An individual survives less than time t}]$

$$F(t) = P[T \leq t] \quad \ldots (2)$$

The cumulative distribution function just gives the oppositive probability of survival function, therefore, we can write F(t) as

$$F(t) = P[T \leq t] = 1 - S(t) \quad \ldots (3)$$

The curve between the cumulative distribution function F(t) and time t is called **CDF curve**. Theoretically, the survival curve is smooth as shown in Fig. 14.12.

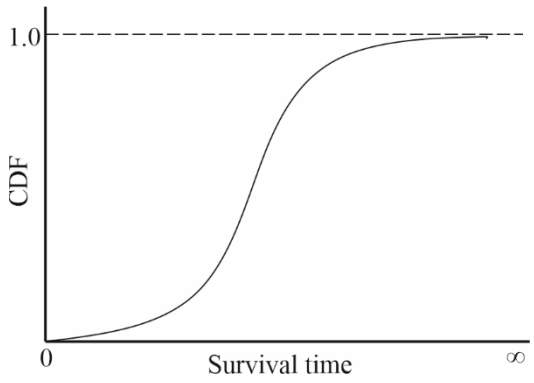

Fig. 14.12: CDF curve.

Initially, the curve starts with F(t) = 0 and heads upward as t increases and finally F(t) = 1 at t = ∞.

14.4.3 Probability Density Function

As we know time is a continuous variable, therefore, the survival time T is a continuous random variable. Like any other continuous random variable, the survival time T has a probability density function(pdf). The pdf is used to find the proportion of the subjects/patients who experience the event of interest (die) in an interval of time and is also used to find the peaks of the high frequency of the event. We can define the pdf of T as a ratio of the probability that a subject/patient dying or an event of interest occurs in the extremely small internal when the time interval approaches zero over the time interval. In a statistical /mathematical sense, it is defined as the ratio that the random variable T lies in the short interval t to $t+\Delta t$ over Δt when Δt tends to zero. It is denoted by $f(t)$. Therefore, in the symbolical form, we can define the survival function as

$$f(t) = \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t]}{\Delta t} \quad \ldots (4)$$

The probability density function is a rate rather than a probability because the expression is a ratio of probability and time interval and we just obtain probability per unit time which is not a probability but a rate. Therefore, unlike probability, it can exceed 1 and lie between 0 and ∞.

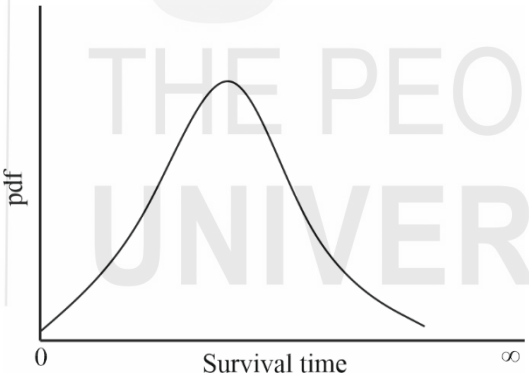

Fig. 14.13: Probability density curve.

The graph of pdf is called a density curve. We prepare a density curve in Fig. 14.13. The figure shows a high failure rate at the beginning of the study and decreasing failure rate as time increases.

14.4.4 Hazard and Cumulative Hazard Functions

The other most important function of survival analysis is the hazard function. When we observed that a subject/patient has lived up to a particular time t, and want to know what is the chance/probability that he/she dies "in the next instant" of time? Then we use the hazard function to know the same. The term "hazard" is used in the survival analysis to describe the risk of failure of a subject/patient in a small interval after time t, given that the subject/patient has survived up to time t. It represents the instantaneous event rate for a subject/patient who has already survived to the time "t."

In other words, the hazard function is the probability that an individual will experience an event (for example, death) within a small-time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time t.

Other names for the hazard function are the **conditional failure rate, the force of mortality, the instantaneous (relative) failure rate, the intensity rate, and the hazard rate**.

We can define hazard function as a ratio of the conditional probability that an individual (subject) dying or an event of interest occurs in the extremely small internal given that the individual has survived to the beginning of the interval when the time interval approaches zero over the time interval. In a statistical /mathematical sense, it is defined as the ratio that the random variable T lies in the short interval t to $t + \Delta t$ given that $T > t$ over Δt when Δt tends to zero. It is denoted by $h(t)$. Therefore, in the symbolical form, we can define the hazard function as

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[t \le T < t + \Delta t \mid T \ge t\right]}{\Delta t} \quad \ldots (5)$$

Like the probability density function, the hazard function is also a rate rather than a probability. Therefore, unlike the probability, it can exceed 1 and lie between 0 and ∞.

As the graphs of survival function and pdf, the graph of hazard function draws against the time t but it does not have to start from 1 and go down to zero as the survival curve. It can start from anywhere and it remains constant, going up and down in any direction over time. We prepare different types of graphs of hazard function in Fig. 14.14.

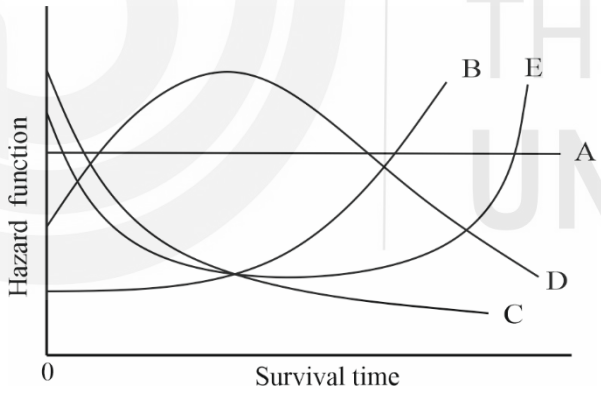

Fig. 14.14: Hazard function with different shapes.

Graph A shows a constant hazard function. Such type of graph exists when h(t) remains the same for all values of t. For example, if a person continues to be healthy throughout the study period, then his instantaneous failure rate/death rate at any time of the study remains constant. Increasing hazard function with time is shown in Graph B. Such type of graph might be expected where the patients do not respond to treatment as in brain tumour patients' cases. Graph C shows a hazard function that decreeing over time. Such a graph may be expected when patients are recovering from typical surgery such as brain tumour surgery. The main cause of death is the operation itself and the chance of death decreases if the operation is successful. Sometimes, the patients have a high risk of death at initial and after treatment, the risk decreases. For such types of situations, the hazard function graph is shown in the Graph D.

Sometimes, we observe that the hazard function first decreases then remains constant and then increases with time as shown in Graph E. Since the shape of such type of pattern is like a U letter of the English alphabet or bathtub so it is called **U-shaped** or **bathtub** curve. This shape of hazard function is of great importance in medical fields where it is used to determine the useful lifetime and the peak-time failure of the patients. Such type of curve describes the organism's life. The shape of the bathtub comprises three parts. At the initial stage, the risk of mortality is high and decreases as time increases as shown in Fig. 14.15.

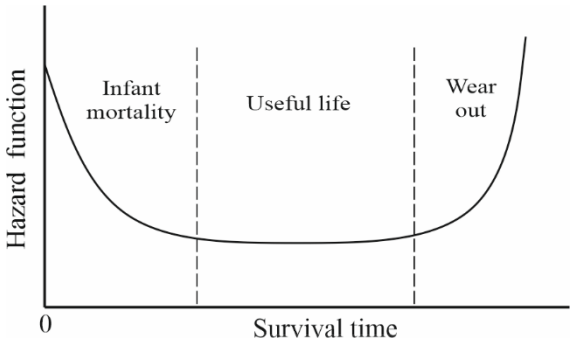

Fig. 14.15: Bathtub shape of the hazard function.

This part is known as **infant mortality.** After that part, the organism may die due to random causes like accidents. So, the second part represents the useful life and it is almost constant and is known as **useful life**. After spanning a long life, the organism starts to wear out and at that time the hazard rate starts to increase as shown in the figure. The third part is called the **wear-out** phase.

Now let us see the term cumulative hazard. The word cumulative hazard is used to represent the total hazard up to time t, which means the sum of hazards at each point up to t. The usual notation for cumulative hazard is H(t).

From the above description, we can write

$$H(t) = \int_0^t h(t)\,dt$$

A cumulative hazard function is an important tool for the analysis of survival time data. H(t) and S(t) are related but in the reverse direction in such a way that as H(t) increases, S(t) decreases. Different investigators follow different approaches to study the survival data. Some are interested in studying through the survival function and others are through the hazard function. But both ways we get the phenomena studied. Here, we will see both approaches to studying the survival data.

Now, you may try the following exercises.

E5) Describe why we use different types of survival functions for the analysis of survival data.

E6) Describe various types of functions used in survival analysis.

14.5 RELATIONSHIPS BETWEEN SURVIVAL FUNCTIONS

So far, we have defined the various functions that are used in survival analysis such as S(t), F(t), f(t) and h(t). These functions are mathematically equivalent

and given any one of the remaining can easily obtain. In this section, we try to understand the relationships between them and how we obtain anyone using others.

As we have studied in Unit 6 of MST-003: Probability Theory that the cumulative distribution function of a continuous variable is defined as

$$F(t) = P[T \le t] = \int_{-\infty}^{t} f(t)dt = \int_{0}^{t} f(t)dt \quad \text{Since time t lies 0 to } \infty. \qquad \ldots (6)$$

So we can express the survival function S(t) in terms of F(t) and f(t) as

$$S(t) = P[T > t] = 1 - P[T \le t] = 1 - F(t) \qquad \ldots (7)$$

$$S(t) = P[T > t] = \int_{t}^{\infty} f(t)dt \qquad \ldots (8)$$

Also, the probability density is the derivative of the cumulative distribution function. Therefore, we can write f(t) in terms of F(t) as

$$f(t) = \frac{d}{dt}F(t) \qquad \ldots (9)$$

$$f(t) = \frac{d}{dt}[1 - S(t)] \quad [\text{as } F(t) = 1 - S(t)]$$

$$f(t) = -\frac{d}{dt}S(t) \qquad \ldots (10)$$

By the definition of hazard, we have

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

$$= \lim_{\Delta t \to 0} \frac{P\left[t \le T < t + \Delta t\right]}{S(t)\Delta t} = \frac{f(t)}{S(t)}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{S(t)}\left[-\frac{d}{dt}S(t)\right] \quad [\text{from equation (10)}]$$

Integrating h(t) with respect to t on both sides within the limits t = 0 to t = t, we get

$$\int_0^t h(t)dt = -\int_0^t \frac{1}{S(t)}\frac{d}{dt}\left[S(t)\right]dt = -\int_0^t \frac{d}{dt}\log_e|S(t)|dt$$

$$= -\left[\log_e|S(t)|\right]_0^t = -\left(\log_e|S(t)| - \log_e|S(0)|\right)$$

$$= -\log_e|S(t)| + \log_e(1) \quad [\because S(0) = 1]$$

$$= -\log_e|S(t)| + 0$$

Since S(t) is nonnegative so we can take S(t) in place of $|S(t)|$.

Hence,

$$-\int_0^t h(t)dt = \log_e S(t)$$

$$\Rightarrow S(t) = e^{-\int_0^t h(t)dt} \qquad \left[\because \log_e a = m \Rightarrow a = e^m\right]$$

$$\text{or } S(t) = \exp\left(-\int_0^t h(t)dt\right)$$

We now put together the relations between S(t), F(t), f(t) and h(t) in Table 4 as follows:

Table 4: Relationships between the functions R(t), F(t), f(t) and h(t)

	S(t)	F(t)	f(t)	h(t)
S(t)	–	1 – F(t)	$\int_t^\infty f(t)dt$	$\exp\left(-\int_0^t h(t)dt\right)$
F(t)	1 – S(t)	–	$\int_0^t f(t)dt$	$1-\exp\left(-\int_0^t h(t)dt\right)$
f(t)	$-\frac{d}{dt}[S(t)]$	$\frac{d}{dt}[F(t)]$	–	$h(t)\exp\left(-\int_0^t h(t)dt\right)$
h(t)	$-\frac{d}{dt}\log_e[S(t)]$	$\frac{\frac{d}{dt}[F(t)]}{1-F(t)}$	$\frac{f(t)}{\int_t^\infty f(t)dt}$	–

Now, we learn how to use these relationships to obtain other functions with the help of examples.

Example 1: If the survival time T (in years) has the probability density function as

$$f(t) = \begin{cases} \theta e^{-\theta t}; & \theta > 0, t \geq 0 \\ 0; & \text{otherwise} \end{cases}$$

Then find

i) Survival function,

ii) Cumulative distribution function,

iii) Hazard function, and

iv) Median when $\theta = 0.2$.

Solution: The probability density function of the survival time T is given as

$$f(t) = \begin{cases} \theta e^{-\theta t}, & t \geq 0, \theta > 0 \\ 0, & \text{otherwise} \end{cases}$$

i) From the definition of survival function, we have

$$S(t) = P[T > t] = \int_t^\infty \theta e^{-\theta t} dt$$

$$= \left[\frac{\theta e^{-\theta t}}{-\theta}\right]_t^\infty \qquad \left[\because \int_a^b e^{kt}dt = \left[\frac{e^{kt}}{k}\right]_a^b\right]$$

$$= -\left[0 - e^{-\theta t}\right] \qquad \left[\because e^{-\theta t} \to 0 \text{ as } t \to \infty\right]$$

$$= e^{-\theta t}$$

ii) By the definition and relation of cumulative distribution function with survival function, we have

$$F(t) = P[T \le t] = 1 - S(t) = 1 - e^{-\theta t}$$

iii) As we know, the hazard function in terms of survival and density functions as

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta$$

iv) Let t_{md} denote the median time to survival time T. Then, we have

$$S(t_{md}) = P\left[T > t_{md}\right] = 0.5$$

$$\int_{t_{md}}^{\infty} f(t)dt = 0.5 \Rightarrow \int_{t_{md}}^{\infty} \theta e^{-\theta t} dt = 0.5$$

$$\left[\frac{\theta e^{-\theta t}}{-\theta}\right]_{t_{md}}^{\infty} = 0.5 \Rightarrow e^{-0.2 t_{md}} = 0.5$$

Taking natural logarithm on both sides

$$-0.2 t_{md} = \log_e(0.5) \Rightarrow -0.2 t_{md} \simeq -0.6931$$

$$\Rightarrow t_{md} \simeq \frac{0.6931}{0.2} \simeq 3.4355 \text{ years}$$

Example 2: The hazard rate of a group of patients is given by

$$h(t) = t$$

Find the survival function and density function of the group.

Solution: Here, we have given the hazard rate and we have to find the survival function, therefore, we use the relationship between h(t) and S(t). From the Table 4, we have

$$S(t) = \exp\left(-\int_o^t h(t)dt\right) = \exp\left(-\int_0^t t\,dt\right) = \exp\left(-\left[\frac{t^2}{2}\right]_0^t\right)$$

$$= \exp\left(-\frac{t^2}{2}\right)$$

Now, we know h(t) and S(t) and have to find f(t) so it is better to use the relationship among them. Therefore,

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

$$= t \exp\left(-\frac{t^2}{2}\right)$$

Now, you may like to solve the following questions.

E7) If the survival time of insect species has the following probability density function

$$f(t) = \frac{1}{\theta} e^{-\frac{t}{\theta}}; \quad \theta > 0,\ t \geq 0$$

Then derive the expression for survival and hazard functions

E8) For the given hazard function

$$h(t) = k$$

Find

i) Survival function,

ii) Cumulative distribution function,

iii) Probability density function,

iv) Median of the survival function T, and

v) Cumulative hazard function.

We end this unit by giving a summary of its contents.

14.6 SUMMARY

1. Cohort study and intervention trials, routine analysis of these designs and problems of these analyses.
2. The concept of time and event in survival analysis.
3. Concept of lost to follow up, survival rate versus survival experience.
4. Concept of censoring, left censoring, right censoring.
5. Survival function, density function of time to event, hazard function and cumulative hazard function.

14.7 SOLUTIONS/ANSWERS

E1) Refer to 14.2.

E2) Refer to 14.2.

E3) Refer to 14.3.

E4) Refer to 14.3.

E5) Refer to 14.4.

E6) Refer to 14.4.

E7) We follow the same procedure as explained in Example 1. Therefore,

$$S(t) = P[T > t] = t = \int_t^{\infty} \frac{1}{\theta} e^{-\frac{t}{\theta}} dt$$

$$= \left[\frac{\frac{1}{\theta} e^{-\frac{t}{\theta}}}{-\frac{1}{\theta}} \right]_t^{\infty} \qquad \left[\because \int_a^b e^{kt} dt = \left[\frac{e^{kt}}{k} \right]_a^b \right]$$

$$= -\left[0 - e^{-\frac{t}{\theta}}\right] \qquad \left[\because e^{-\theta t} \to 0 \text{ as } t \to \infty\right]$$

$$= e^{-\frac{t}{\theta}}$$

As we know, the hazard function in term of survival and density functions as

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\theta} e^{-\frac{t}{\theta}}}{e^{-\frac{t}{\theta}}} = \frac{1}{\theta}$$

E8) Here, the hazard function is given and we have to find the survival and other function using the relationships given in Table 14.2 as

$$S(t) = \exp\left(-\int_o^t h(t)dt\right) = \exp\left(-\int_0^t k\,dt\right) = \exp\left(-k\left[t\right]_0^t\right) = \exp\left(-kt\right)$$

We can obtain the cumulative distribution function as

$$F(t) = 1 - S(t) = 1 - \exp\left(-kt\right)$$

Now, we know h(t) and S(t) and have to find f(t) so it is better to use the relationship among them. Therefore,

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

$$\Rightarrow f(t) = h(t)S(t) = k\exp\left(-kt\right)$$

Let t_{md} denote the median time to survival time T. Then, we have

$$S(t_{md}) = P\left[T > t_{md}\right] = 0.5$$

$$\int_{t_{md}}^{\infty} f(t)dt = 0.5 \Rightarrow \int_{t_{md}}^{\infty} ke^{-kt}dt = 0.5$$

$$\left[\frac{k\,e^{-kt}}{-k}\right]_{t_{md}}^{\infty} = 0.5 \Rightarrow e^{-0.2t_{md}} = 0.5$$

Taking natural logarithm on both sides

$$-0.2t_{md} = \log_e(0.5) \Rightarrow -0.2t_{md} \simeq -0.6931$$

$$\Rightarrow t_{md} \simeq \frac{0.6931}{0.2} \simeq 3.4355$$

Now, we try to find the cumulative hazard function. By the definition of the same, we have

$$H\left(t\right) = \int_0^t h\left(t\right)dt = \int_0^t kdt = k\left[t\right]_0^t = kt$$