Introduction

This introduction to survival analysis gives a descriptive overview of the data analytic approach called **survival analysis.** This approach includes the type of problem addressed by survival analysis, the outcome variable considered, the need to take into account "censored data," what a survival function and a hazard function represent, basic data layouts for a survival analysis, the goals of survival analysis, and some examples of survival analysis.

Because this chapter is primarily descriptive in content, no prerequisite mathematical, statistical, or epidemiologic concepts are absolutely necessary. A first course on the principles of epidemiologic research would be helpful. It would also be helpful if the reader has had some experience reading mathematical notation and formulae.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

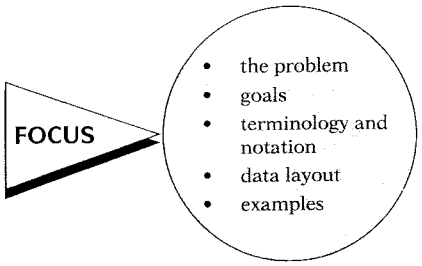
From Kleinbaum
Survival Analysis
A self-learning Text

Objectives

Upon completing the module, the learner should be able to:

1. Recognize or describe the type of problem addressed by a survival analysis.
2. Define what is meant by censored data.
3. Define or recognize right-censored data.
4. Give three reasons why data may be censored.
5. Define, recognize, or interpret a survivor function.
6. Define, recognize, or interpret a hazard function.
7. Describe the relationship between a survivor function and a hazard function.
8. State three goals of a survival analysis.
9. Identify or recognize the basic data layout for the computer; in particular, put a given set of survival data into this layout.
10. Identify or recognize the basic data layout, or components thereof, for understanding modeling theory; in particular, put a given set of survival data into this layout.
11. Interpret or compare examples of survivor curves or hazard functions.
12. Given a problem situation, state the goal of a survival analysis in terms of describing how explanatory variables relate to survival time.
13. Compute or interpret average survival and/or average hazard measures from a set of survival data.
14. Define or interpret the hazard ratio defined from comparing two groups of survival data.

Presentation



This presentation gives a general introduction to survival analysis, a popular data analysis approach for certain kinds of epidemiologic and other data. Here we focus on the problem addressed by survival analysis, the goals of a survival analysis, key notation and terminology, the basic data layout, and some examples.

I. What Is Survival Analysis?

Outcome variable: **Time until an event occurs**

We begin by describing the type of analytic problem addressed by survival analysis. Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs.*

Start follow-up — TIME — Event

By **time,** we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the **age** of an individual when an event occurs.

Event: death
disease
relapse
recovery

By **event,** we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

Assume 1 event

> 1 event — Competing risk

Although more than one event may be considered in the same analysis, we will assume that only one event is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem is generally characterized as a **competing risk** problem, which is beyond the scope of this presentation.

Time ≡ survival time

Event ≡ failure

In a survival analysis, we usually refer to the time variable as **survival time,** because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a **failure,** because the kind of event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event.

EXAMPLE

1. Leukemia patients/time in remission (weeks)
2. Disease-free cohort/time until heart disease (years)
3. Elderly (60+) population/time until death (years)
4. Parolees (recidivism study)/time until rearrest (weeks)
5. Heart transplants/time until death (months)

Five examples of survival analysis problems are briefly mentioned here. The first is a study that follows leukemia patients in remission over several weeks to see how long they stay in remission. The second example follows a disease-free cohort of individuals over several years to see who develops heart disease. A third example considers a 13-year follow-up of an elderly population (60+ years) to see how long subjects remain alive. A fourth example follows newly released parolees for several weeks to see whether they get rearrested. (This type of problem is called a recidivism study.) The fifth example traces how long patients survive after receiving a heart transplant.

All of the above examples are survival analysis problems because the outcome variable is time until an event occurs. In the first example, involving leukemia patients, the event of interest (i.e., failure) is "going out of remission," and the outcome is "time in weeks until a person goes out of remission." In the second example, the event is "developing heart disease," and the outcome is "time in years until a person develops heart disease." In the third example, the event is "death" and the outcome is"time in years to death." Example four, a sociological rather than a medical study, considers the event of recidivism (i.e., getting rearrested), and the outcome is time in weeks until rearrested. Finally, the fifth example considers the event "death," with the outcome being "time until death (in months from receiving a transplant)."

We will return to some of these examples later in this presentation and in later presentations.

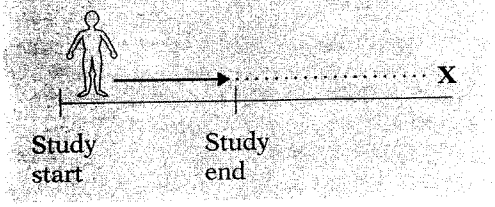
II. Censored Data

Censoring: don't know survival time exactly

Most survival analyses must consider a key analytical problem called **censoring.** In essence, censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly.**

EXAMPLE

Leukemia patients in remission:



Why censor?

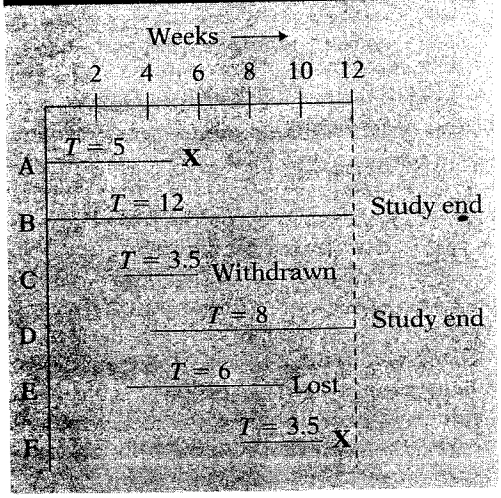
1. study ends—no event
2. lost
3. withdraws

As a simple example of censoring, consider leukemia patients followed until they go out of remission, shown here as **X.** If for a given patient, the study ends while the patient is still in remission (i.e., doesn't get the event), then that patient's survival time is considered censored. We know that, for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission after the study ends, we do not know the complete survival time.

There are generally three reasons why censoring may occur:

(1) a person does not experience the event before **the study ends;**
(2) a person is **lost to follow-up** during the study period;
(3) a person **withdraws from the study** because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction).

EXAMPLE



These situations are graphically illustrated here. The graph describes the experience of several persons followed over time. An **X** denotes a person who got the event.

Person A, for example, is followed from the start of the study until getting the event at week 5; his survival time is 5 weeks and is *not* censored.

Person B also is observed at the start of the study but is followed to the end of the 12-week study period without getting the event; the survival time here is censored because we can say only that it is *at least* 12 weeks.

Person C enters the study between the second and third week and is followed until he/she withdraws from the study at 6 weeks; this person's survival time is censored after 3.5 weeks.

Person D enters at week 4 and is followed for the remainder of the study without getting the event; this person's censored time is 8 weeks.

Person E enters the study at week 3 and is followed until week 9, when he is lost to follow-up; his censored time is 6 weeks.

X Event occurs

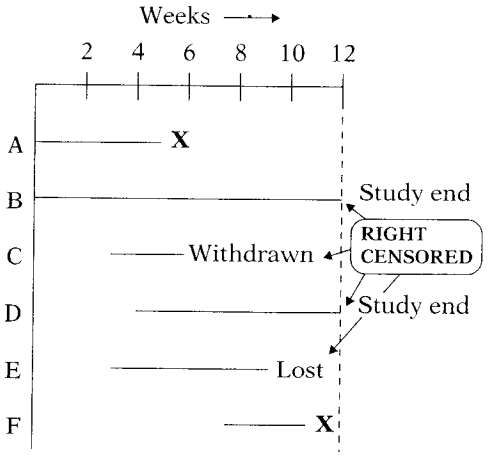
Person F enters at week 8 and is followed until getting the event at week 11.5. As with person A, there is no censoring here; the survival time is 3.5 weeks.

SUMMARY
Event: A, F
Censored: B, C, D, E

In **summary,** of the six persons observed, two get the event (persons A and F) and four are censored (B, C, D, and E).

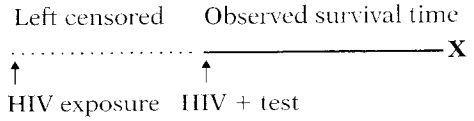
Person	Survival time	Failed (1); censored (0)
A	5	1
B	12	0
C	3.5	0
D	8	0
E	6	0
F	3.5	1

A table of the survival time data for the six persons in the graph is now presented. For each person, we have given the corresponding survival time up to the event's occurring or up to censorship. We have indicated in the last column whether this time was censored or not (with 1 denoting failed and 0 denoting censored). For example, the data for person C is a survival time of 3.5 and a censorship indicator of 0, whereas for person F the survival time is 3.5 and the censorship indicator is 1. This table is a simplified illustration of the type of data to be analyzed in a survival analysis.



Notice in our example that for each of the four persons censored, we know that the person's exact survival time becomes incomplete at the **right** side of the follow-up period, occurring when the study ends or when the person is lost to follow-up or is withdrawn. We generally refer to this kind of data as **right-censored.** For these data, the complete survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side. Although data can also be **left-censored,** most survival data is right-censored. In the remainder of this text, we will consider right-censored data only.

Left-censored data:



Left-censored data can occur when a person's survival time becomes incomplete at the left side of the follow-up period for that person. For example, if we are following persons with HIV infection, we may start follow-up when a subject first tests positive for the HIV virus, but we may not know exactly the time of first exposure to the virus. Thus, the survival time is censored on the left side, because there is unknown follow-up time from the time of first exposure up to the time of first positive HIV test.

III. Terminology and Notation

We are now ready to introduce basic mathematical terminology and notation for survival analysis. First, we denote by a **capital *T*** the random variable for a person's survival time. Since T denotes time, its possible values include all nonnegative numbers; that is, T can be any number equal to or greater than zero.

T = survival time ($T \geq 0$)

random variable

Next, we denote by a **small letter *t*** any specific value of interest for the random variable capital T. For example, if we are interested in evaluating whether a person survives for more than 5 years after undergoing cancer therapy, **small *t*** equals 5; we then ask whether capital T exceeds 5.

t = specific value for T

EXAMPLE

Survives > 5 years?

$T > t = 5$

Finally, we let the Greek letter delta (δ) denote a (0,1) random variable indicating either failure or censorship. That is, $\delta = 1$ for failure if the event occurs during the study period, or $\delta = 0$ if the survival time is censored by the end of the study period. Note that if a person does not fail, that is, does not get the event during the study period, censorship is the **only** remaining possibility for that person's survival time. That is, $\delta = 0$ if and only if one of the following happens: a person survives until the study ends, a person is lost to follow-up, or a person withdraws during the study period.

δ = (0, 1) random variable

$$= \begin{cases} 1 \text{ if failure} \\ 0 \text{ if censored} \end{cases}$$

- study ends
- lost
- withdraws

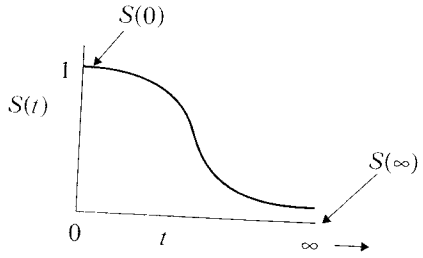
We next introduce and describe two quantitative terms considered in any survival analysis. These are the **survivor function,** denoted by $S(t)$, and the **hazard function,** denoted by $h(t)$.

$S(t)$ = survivor function
$h(t)$ = hazard function

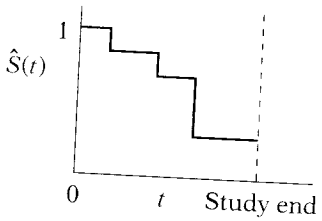
$S(t) = P(T > t)$

t	$S(t)$
1	$S(1) = P(T > 1)$
2	$S(2) = P(T > 2)$
3	$S(3) = P(T > 3)$
•	•
•	•
•	•

Theoretical $S(t)$:



$\hat{S}(t)$ in practice:



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

The survivor function $S(t)$ gives the probability that a person survives longer than some specified time ***t:*** that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t.

The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values of t provides crucial summary information from survival data.

Theoretically, as t ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph here, where t identifies the X-axis, all survivor functions have the following characteristics:

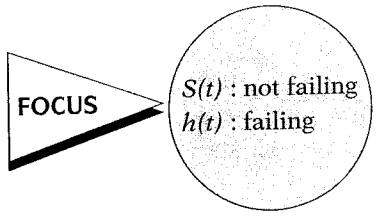
- they are nonincreasing; that is, they head downward as to increases;
- at time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one;
- at time $t = \infty$, $S(t) = S(\infty) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero.

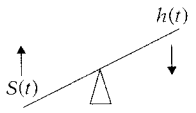
Note that these are **theoretical** properties of survivor curves.

In practice, when using actual data, we usually obtain graphs that are **step functions,** as illustrated here, rather than smooth curves. Moreover, because the study period is never infinite in length, it is possible that not everyone studied gets the event; the estimated survivor function, denoted by a caret over the S in the graph, thus does not go all the way down to zero at the end of the study.

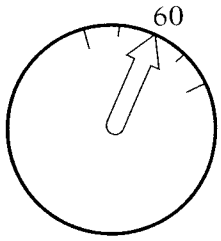
The hazard function, denoted by ***h(t),*** is given by the formula: $h(t)$ equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time. This mathematical formula is difficult to explain in practical terms.

$h(t)$ = instantaneous potential

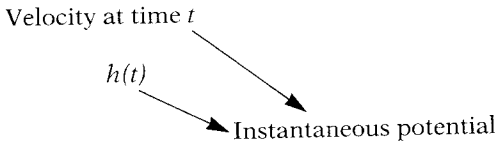




Before getting into the specifics of the formula, we give a conceptual interpretation. **The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.** Note that, in contrast to the survivor function, which focuses on *not* failing, the hazard function focuses on failing, that is, on the event occurring. Thus, in some sense, the hazard function can be considered as giving the negative side of the information given by the survivor function. That is, the higher $S(t)$ is for a given t, the smaller is $h(t)$, and vice versa.



To get an idea of what we mean by instantaneous potential, consider the concept of velocity. If, for example, you are driving in your car and you see that your speedometer is registering 60 mph, what does this reading mean? It means that if in the next hour, you continue to drive this way, with the speedometer exactly on 60, you would cover 60 miles. This reading gives the **potential,** at the moment you have looked at your speedometer, for how many miles you will travel in the next hour. However, because you may slow down or speed up or even stop during the next hour, the 60-mph speedometer reading does not tell you the number of miles you *really* will cover in the next hour. The speedometer tells you only how fast you are going *at a given moment;* that is, the instrument gives your instantaneous potential or velocity.



Similar to the idea of velocity, a hazard function $h(t)$ gives the instantaneous potential at time t for getting an event, like death or some disease of interest, given survival up to time t. The given part, that is, surviving up to time t, is analogous to recognizing in the velocity example that the speedometer reading at a point in time inherently assumes that you have already traveled some distance (i.e., survived) up to the time of the reading.

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

(Given)

In mathematical terms, the given part of the formula for the hazard function is found in the probability statement—the numerator to the right of the limit sign. This statement is a conditional probability because it is of the form, P of A, given B, where the P denotes probability and where the long vertical line separating A from B denotes "given." In the hazard formula, the conditional probability gives the probability that the event will occur in the time interval between t and $t + \Delta t$, given that the survival time, T, is greater than or equal to t. Because of the given sign here, the hazard function is sometimes called a **conditional failure rate.**

Conditional probabilities: $P(A \mid B)$

1.2. $P(t \le T < t + \Delta t \mid T \ge t)$

Hazard function ≡ conditional failure **rate**

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

Probability per unit time

We now explain why the hazard is a **rate** rather than a probability. Note that in the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities. The numerator is the conditional probability we just discussed. The denominator is Δt, which denotes a small time interval. By this division, we obtain a probability per unit time, which is no longer a probability but a rate. In particular, the scale for this ratio is not 0 to 1, as for a probability, but rather ranges between 0 and infinity, and depends on whether time is measured in days, weeks, months, or years, etc.

Rate: 0 to ∞

$P = P(t \le T < t + \Delta t \mid T \ge t)$

P	Δt	$P/\Delta t$ = rate
$\frac{1}{3}$	$\frac{1}{2}$ day	$\frac{1/3}{1/2} = 0.67$ / day
$\frac{1}{3}$	$\frac{1}{14}$ week	$\frac{1/3}{1/14} = 4.67$ / week

For example, if the probability, denoted here by P, is 1/3, and the time interval is one-half a day, then the probability divided by the time interval is 1/3 divided by 1/2, which equals 0.67 per day. As another example, suppose, for the same probability of 1/3, that the time interval is considered in weeks, so that 1/2 day equals 1/14 of a week. Then the probability divided by the time interval becomes 1/3 over 1/14, which equals 14/3, or 4.67 per week. The point is simply that the expression P divided by Δt at the right of the limit sign **does not give a probability. The value obtained will give a different number depending on the units of time used, and may even give a number larger than one.**