1 Introduction to Survival Analysis

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_1,

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

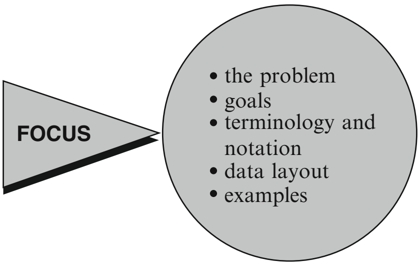
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing the chapter, 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?

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*.

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

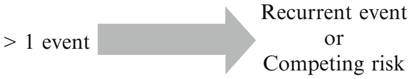
Start follow-up
TIME
Event

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



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 can be characterized as either a recurrent event or a **competing risk** problem, which are discussed in Chaps. 8 and 9, respectively.

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 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 until 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 rearrest." 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.

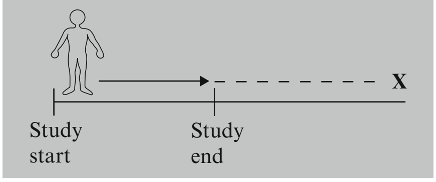
II. Censored Data

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:



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.

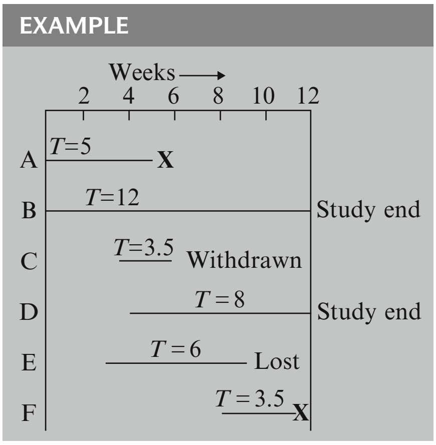
Why censor?

1. study ends – no event
2. lost to follow-up
3. withdraws

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 or other competing risk)

EXAMPLE



X ⟹ Event occurs

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 from 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 3rd week and is followed until he 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.

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

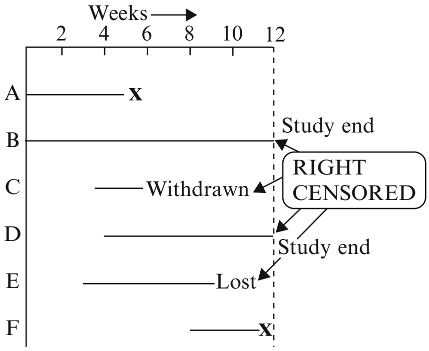
Event: A, F
Censored: B, C, D, E

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

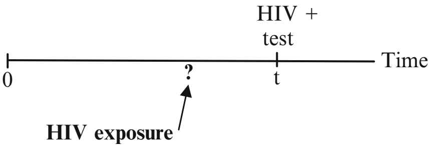
A table of the survival time data for the six persons in the graph is now presented. For each person, we have given the corresponding survival time up to the event's occurrence or up to censorship. We have indicated in the last column whether this time was censored or not (with 1 denoting failed and 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.

Right-censored: true survival time is equal to or greater than observed survival time



Notice in our example that for each of the four persons censored, we know that the person's true 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 of the observed survival time interval. Although data can also be **left-censored**, most survival data is right-censored.

Left-censored: true survival time is less than or equal to the observed survival time

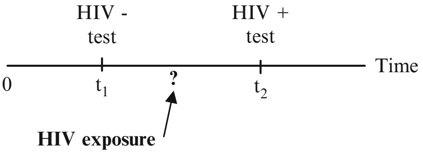


Left-censored: data can occur when a person's true survival time is less than or equal to that person's observed survival time. For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know the exact time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject's test is positive.

Event occurs between 0 and t
but
do not know the exact time.

In other words, if a person is left-censored at time t, we know they had an event between time 0 and t, but we do not know the exact time of event.

Interval-censored: true survival time is within a known time interval



Survival analysis data can also be **interval-censored**, which can occur if a subject's true (but unobserved) survival time is within a certain known specified time interval. As an example, again considering HIV surveillance, a subject may have had two HIV tests, where he/she was HIV negative at the time (say, t_1) of the first test and HIV positive at the time (t_2) of the second test. In such a case, the subject's true survival time occurred after time t_1 and before time t_2, i.e., the subject is interval-censored in the time interval (t_1, t_2).

Left censoring $\Rightarrow$ $t_1 = 0$, t_2 = upper bound

Right censoring $\Rightarrow$ t_1 = lower bound, $t_2 = \infty$

Interval-censoring actually incorporates both right-censoring and left-censoring as special cases. Left-censored data occur whenever the value of t_1 is 0 and t_2 is a known upper bound on the true survival time. In contrast, right-censored data occurs whenever the value of t_2 is infinity, and t_1 is a known lower bound on the true survival time.

Right-censored due to competing risk,
e.g., death from another cause
$\Downarrow$
$t_2 = \infty$
gives upper bound for true survival time
assuming that competing risk had not
occurred.

If an individual is right-censored due to a competing event (e.g., death from another cause), then in this context, we consider what the true survival time would have been if the competing event had not occurred. In other words, when we state that the value of the upper bound for the true survival time is infinity for right-censored data, we are considering what would have occurred in the absence of a competing risk. Competing risks are fully discussed in Chapter 9.

III. Terminology and Notation

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

random variable

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 = specific value for T

EXAMPLE

Survives > 5 years?
$T > t = 5$

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.

$d = (0, 1)$ random variable

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

- study ends
- lost to follow-up
- withdraws

Finally, we denote the **small letter *d*** to define a (0,1) random variable indicating either failure or censorship. That is, $d = 1$ for failure if the event occurs during the study period, or $d = 0$ 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, $d = 0$ 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.

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

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 $h(t)$.

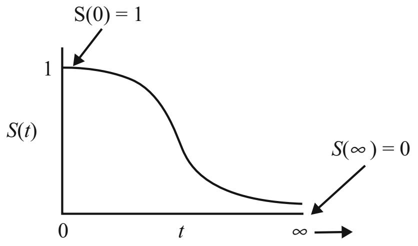
$S(t) = P(T > t)$

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

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.

Theoretical $S(t)$:

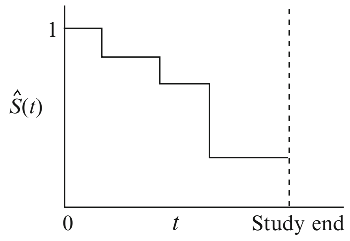


Theoretically, as t ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph, where t identifies the X-axis, all survivor functions have the following characteristics:

- they are nonincreasing; that is, they head downward as t 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.

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

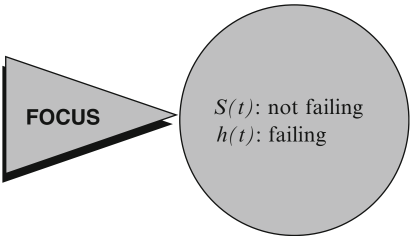


In practice, when using actual data, we usually obtain graphs that are **step functions**, as illustrated here, rather than smooth curves. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event.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

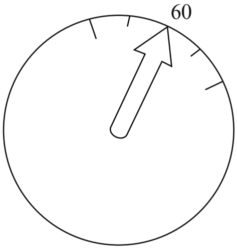
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

The hazard function, denoted by $\boldsymbol{h(t)}$, 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.

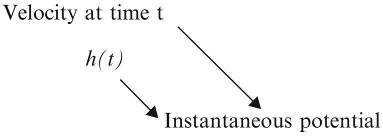
$h(t)$ = instantaneous potential



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 opposite side of the information given by the survivor function.



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.

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

Conditional probabilities: $P(A|B)$

$P(t \le T < t + \Delta t \mid T \ge t)$
$=$ P(individual fails in the interval $[t, t + \Delta t]$ | survival up to time t)

In mathematical terms, the given part of the formula for the hazard function is found in the probability statement in 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 a person's survival time, T, will lie in the time interval between t and $t + \Delta t$, given that the survival time is greater than or equal to t. Because of the given sign here, the hazard function is sometimes called a **conditional failure rate**.

Hazard function ≡ conditional failure **rate**

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

Probability per unit time

Rate: 0 to ∞

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.

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

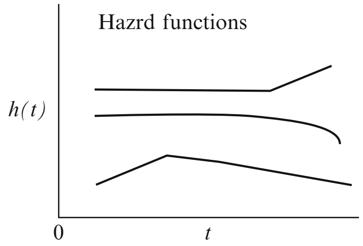
P	Δt	$P/\Delta t =$ rate
$\frac{1}{3}$	$\frac{1}{2}$day	$\frac{1/3}{1/2} = 0.67/\text{day}$
$\frac{1}{3}$	$\frac{1}{14}$week	$\frac{1/3}{1/14} = 4.67/\text{week}$

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**.

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

Gives instantaneous potential

When we take the limit of the right-side expression as the time interval approaches zero, we are essentially getting an expression for the instantaneous probability of failing at time t per unit time. Another way of saying this is that the conditional failure rate or hazard function $h(t)$ gives the instantaneous **potential** for failing at time t per unit time, given survival up to time t.



As with a survivor function, the hazard function $h(t)$ can be graphed as t ranges over various values. The graph at the left illustrates three different hazards. In contrast to a survivor function, the graph of $h(t)$ does not have to start at 1 and go down to zero, but rather can start anywhere and go up and down in any direction over time. In particular, for a specified value of t, the hazard function $h(t)$ has the following characteristics:

- $h(t) \geq 0$
- $h(t)$ has no upper bound

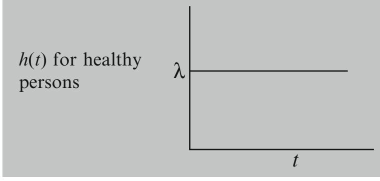
- it is always nonnegative, that is, equal to or greater than zero;
- it has no upper bound.

These two features follow from the ratio expression in the formula for $h(t)$, because both the probability in the numerator and the Δt in the denominator are nonnegative, and since Δt can range between 0 and ∞.

EXAMPLE

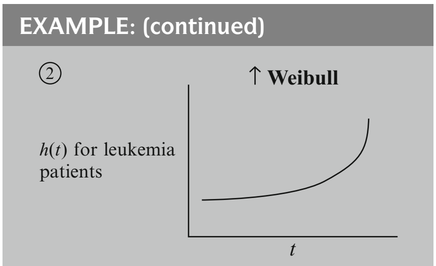
①

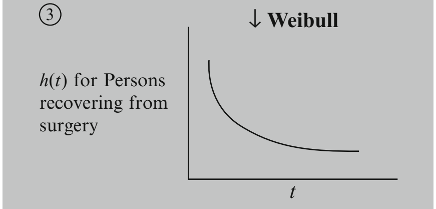
Constant hazard
(exponential model)

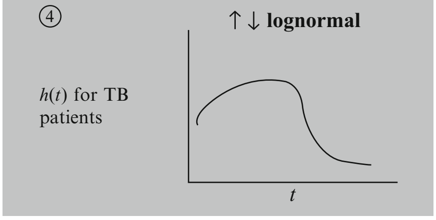


Now we show some graphs of different types of hazard functions. The first graph given shows a constant hazard for a study of healthy persons. In this graph, no matter what value of t is specified, $h(t)$ equals the same value—in this example, λ. Note that for a person who continues to be healthy throughout the study period, his/her instantaneous potential for becoming ill at any time during the period remains constant throughout the follow-up period. When the hazard function is constant, we say that the survival model is **exponential**. This term follows from the relationship between the survivor function and the hazard function. We will return to this relationship later.

EXAMPLE: (continued)







The second graph shows a hazard function that is increasing over time. An example of this kind of graph is called an **increasing Weibull** model. Such a graph might be expected for leukemia patients not responding to treatment, where the event of interest is death. As survival time increases for such a patient, and as the prognosis accordingly worsens, the patient's potential for dying of the disease also increases.

In the third graph, the hazard function is decreasing over time. An example of this kind of graph is called a **decreasing Weibull**. Such a graph might be expected when the event is death in persons who are recovering from surgery, because the potential for dying after surgery usually decreases as the time after surgery increases.

The fourth graph given shows a hazard function that is first increasing and then decreasing. An example of this type of graph is the **lognormal survival** model. We can expect such a graph for tuberculosis patients, since their potential for dying increases early in the disease and decreases later.

$S(t)$: directly describes survival

$h(t)$:
- a measure of instantaneous potential
- identify specific model form
- math model for survival analysis

Of the two functions we have considered, $S(t)$ and $h(t)$, the survivor function is more naturally appealing for analysis of survival data, simply because $S(t)$ directly describes the survival experience of a study cohort.

However, the hazard function is also of interest for the following reasons:

- it is a measure of instantaneous potential whereas a survival curve is a cumulative measure over time;
- it may be used to identify a specific model form, such as an exponential, a Weibull, or a lognormal curve that fits one's data;
- it is the vehicle by which mathematical modeling of survival data is carried out; that is, the survival model is usually written in terms of the hazard function.

Relationship of *S(t) and h(t)*:
If you know one, you can determine the other.

Regardless of which function $S(t)$ or $h(t)$ one prefers, **there is a clearly defined relationship between the two**. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa. For example, if the hazard function is constant, i.e., $h(t) = \lambda$, for some specific value λ, then it can be shown that the corresponding survival function is given by the following formula: $S(t)$ equals **e** to the power minus λ times t.

EXAMPLE

$h(t) = \lambda$ if and only if $S(t) = e^{-\lambda t}$

General formulae:

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

More generally, the relationship between $S(t)$ and $h(t)$ can be expressed equivalently in either of two calculus formulae shown here.

The first of these formulae describes how the survivor function $S(t)$ can be written in terms of an integral involving the hazard function. The formula says that $S(t)$ equals the exponential of the negative integral of the hazard function between integration limits of 0 and t.

The second formula describes how the hazard function $h(t)$ can be written in terms of a derivative involving the survivor function. This formula says that $h(t)$ equals minus the derivative of $S(t)$ with respect to t divided by $S(t)$.

$$S(t) \Longleftrightarrow h(t)$$

In any actual data analysis, a computer program can make the numerical transformation from $S(t)$ to $h(t)$, or vice versa, without the user ever having to use either formula. The point here is simply that if you know either $S(t)$ or $h(t)$, you can get the other directly.

SUMMARY

T = survival time random variable
t = specific value of T
d = (0.1) variable for failure/censorship
$S(t)$ = survivor function
$h(t)$ = hazard function

At this point, we have completed our discussion of key terminology and notation. **The key notation is *T* for the survival time variable, *t* for a specified value of *T*, and *d* for the dichotomous variable indicating event occurrence or censorship. The key terms are the survivor function *S(t)* and the hazard function *h(t)***, which are in essence opposed concepts, in that the survivor function focuses on surviving whereas the hazard function focuses on failing, given survival up to a certain time point.

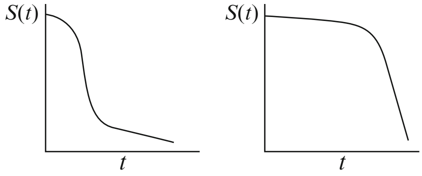
IV. Goals of Survival Analysis

We now state the basic goals of survival analysis.

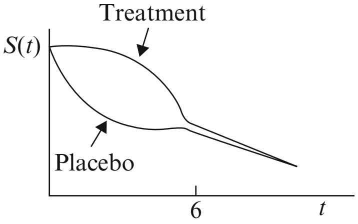
Goal 1: To estimate and interpret survivor and/or hazard functions from survival data.

Goal 2: To compare survivor and/or hazard functions.

Goal 3: To assess the relationship of explanatory variables to survival time.



Regarding the first goal, consider, for example, the two survivor functions pictured at the left, which give very different interpretations. The function farther on the left shows a quick drop in survival probabilities early in follow-up but a leveling off thereafter. The function on the right, in contrast, shows a very slow decrease in survival probabilities early in follow-up but a sharp decrease later on.



We compare survivor functions for a treatment group and a placebo group by graphing these functions on the same axis. Note that up to 6 weeks, the survivor function for the treatment group lies above that for the placebo group, but thereafter the two functions are at about the same level. This dual graph indicates that up to 6 weeks the treatment is **more effective** for survival than the placebo but has about the same effect thereafter.

Goal 3: Use math modeling, e.g., Cox proportional hazards

Goal 3 usually requires using some form of mathematical modeling, for example, the Cox proportional hazards approach, which will be the subject of subsequent chapters.

V. Basic Data Layout for Computer

Two types of data layouts:

- for computer use
- for understanding

We previously considered some examples of survival analysis problems and a simple data set involving six persons. We now consider the general data layout for a survival analysis. We will provide two types of data layouts, one giving the form appropriate for computer use, and the other giving the form that helps us understand how a survival analysis works.

For computer:

Indiv. #	t	d	X_1	X_2	• • •	X_p
1	t_1	d_1	X_{11}	X_{12}	• • •	X_{1p}
2	t_2	d_2	X_{21}	X_{22}	• • •	X_{2p}
•						•
•						•
•						•
5	$t_5 = 3$ got event					
•						•
•						•
•						•
8	$t_8 = 3$ consored					
•						•
•						•
•						•
n	t_n	d_n	X_{n1}	X_{n2}	• • •	X_{np}

We start by providing, in the table shown here, the basic data layout for the computer. Assume that we have a data set consisting of n persons. The first column of the table identifies each person from 1, starting at the top, to n, at the bottom.

The remaining columns after the first one provide survival time and other information for each person. The second column gives the survival time information, which is denoted t_1 for individual 1, t_2 for individual 2, and so on, up to t_n for individual n. Each of these t's gives the observed survival time regardless of whether the person got the event or is censored. For example, if person 5 got the event at 3 weeks of follow-up, then $t_5 = 3$; on the other hand, if person 8 was censored at 3 weeks, without getting the event, then $t_8 = 3$ also.

To distinguish persons who get the event from those who are censored, we turn to the third column, which gives the information for status (i.e., d) the dichotomous variable that indicates censorship status.

		Failure status ↓	Explanatory variables			
Indiv. #	t	d	X_1	X_2	• • •	X_p
1	t_1	d_1	X_{11}	X_{12}	• • •	X_{1p}
2	t_2	d_2	X_{21}	X_{22}	• • •	X_{2p}
•						•
•						•
•						•
5	$t_5 = 3$	$d_5 = 1$				
•						•
•		$\sum_1^n d_i =$ # failures				•
•						•
8	$t_8 = 3$	$d_8 = 0$				
•						•
•						•
•						•
n	t_n	d_n	X_{n1}	X_{n2}	• • •	X_{np}

Thus, d_1 is 1 if person 1 gets the event or is 0 if person 1 is censored; d_2 is 1 or 0 similarly, and so on, up through d_n. In the example just considered, person 5, who failed at 3 weeks, has a d of 1; that is, d_5 equals 1. In contrast, person 8, who was censored at 3 weeks, has a d of 0; that is, d_8 equals 0.

Note that if all of the d values in this column are added up, their sum will be the total number of failures in the data set. This total will be some number equal to or less than n, because not every one may fail.

The remainder of the information in the table gives values for explanatory variables of interest. An explanatory variable, X_l is any variable like age or exposure status, E, or a product term like age × race that the investigator wishes to consider to predict survival time. These variables are listed at the top of the table as X_1, X_2, and so on, up to X_p. Below each variable are the values observed for that variable on each person in the data set.

Columns

Rows

#	t	d	X_1	X_2	$\cdots$	X_p
1	t_1	d_1	X_{11}	X_{12}	$\cdots$	X_{1p}
2	t_2	d_2	X_{21}	X_{22}	$\cdots$	X_{2p}
$\vdots$						$\vdots$
i	t_i	d_i	X_{i1}	X_{i2}	$\cdots$	X_{ip}
$\vdots$						$\vdots$
n	t_n	d_n	X_{n1}	X_{n2}	$\cdots$	X_{np}

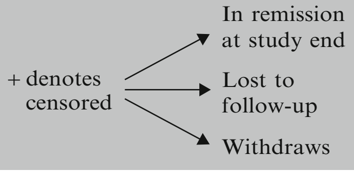
For example, in the column corresponding to X_1 are the values observed on this variable for all n persons. These values are denoted as X_{11}, X_{21}, and so on, up to X_{n1}; the first subscript indicates the person number, and the second subscript, a one in each case here, indicates the variable number. Similarly, the column corresponding to variable X_2 gives the values observed on X_2 for all n persons. This notation continues for the other X variables up through X_p.

We have thus described the basic data layout by columns. Alternatively, we can look at the table line by line, that is, by rows. For each line or row, we have the information obtained on a given individual. Thus, for individual i, the observed information is given by the values t_i, d_i, X_{i1}, X_{i2}, etc., up to X_{ip}. This is how the information is read into the computer, that is, line by line, until all persons are included for analysis.

EXAMPLE

The data: Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	



As an example of this data layout, consider the following set of data for two groups of leukemia patients: one group of 21 persons has received a certain treatment; the other group of 21 persons has received a placebo. The data come from Freireich et al., *Blood*, 1963.

As presented here, the data are not yet in tabular form for the computer, as we will see shortly. The values given for each group consist of time in weeks a patient is in remission, up to the point of the patient's either going out of remission or being censored. Here, going out of remission is a failure. A person is censored if he or she remains in remission until the end of the study, is lost to follow-up, or withdraws before the end of the study. The censored data here are denoted by a plus sign next to the survival time.

EXAMPLE: (continued)

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

	Indiv. #	t (weeks)	d (failed or censored)	X (Group)
	1	6	1	1
	2	6	1	1
	③	6	1	1
	4	7	1	1
	5	10	1	1
	6	13	1	1
	7	16	1	1
	8	22	1	1
GROUP	9	23	1	1
1	10	6	0	1
	11	9	0	1
	12	10	0	1
	13	11	0	1
	⑭	17	0	1
	15	19	0	1
	16	20	0	1
	17	25	0	1
	18	32	0	1
	19	32	0	1
	20	34	0	1
	21	35	0	1

Here are the data again:

Notice that the first three persons in group 1 went out of remission at 6 weeks; the next 6 persons also went out of remission, but at failure times ranging from 7 to 23. All of the remaining persons in group 1 with pluses next to their survival times are censored. For example, on line three the first person who has a plus sign next to a 6 is censored at 6 weeks. The remaining persons in group 1 are also censored, but at times ranging from 9 to 35 weeks.

Thus, of the 21 persons in group 1, nine failed during the study period, whereas the last 12 were censored. Notice also that none of the data in group 2 is censored; that is, all 21 persons in this group went out of remission during the study period.

We now put this data in tabular form for the computer, as shown at the left. The list starts with the 21 persons in group 1 (listed 1–21) and follows (on the next page) with the 21 persons in group 2 (listed 22–42). Our n for the composite group is 42.

The *second* column of the table gives the survival times in weeks for all 42 persons. The *third* column indicates failure or censorship for each person. Finally, the *fourth* column lists the values of the only explanatory variable we have considered so far, namely, group status, with 1 denoting treatment and 0 denoting placebo.

If we pick out any individual and read across the table, we obtain the line of data for that person that gets entered in the computer. For example, person #3 has a survival time of 6 weeks, and since $d = 1$, this person failed, that is, went out of remission. The X value is 1 because person #3 is in group 1. As a second example, person #14, who has an observed survival time of 17 weeks, was censored at this time because $d = 0$. The X value is again 1 because person #14 is also in group 1.

EXAMPLE: (continued)

	Indiv. #	t (weeks)	d (failed or censored)	X (Group)
	22	1	1	0
	23	1	1	0
	24	2	1	0
	25	2	1	0
	26	3	1	0
	27	4	1	0
GROUP	28	4	1	0
2	29	5	1	0
	30	5	1	0
	31	8	1	0
	32	8	1	0
	33	8	1	0
	34	8	1	0
	35	11	1	0
	36	11	1	0
	37	12	1	0
	38	12	1	0
	39	15	1	0
	40	17	1	0
	41	22	1	0
	42	23	1	0

As one more example, this time from group 2, person #32 survived 8 weeks and then failed, because $d = 1$; the X value is 0 because person #32 is in group 2.

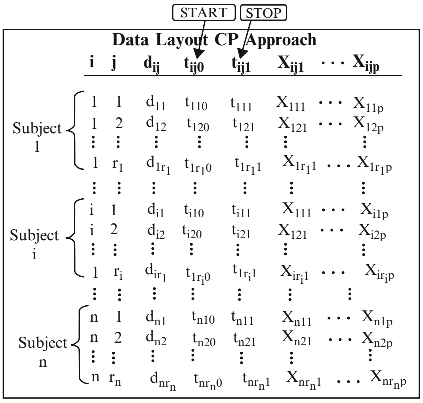
Alternative Data Layout: **Counting Process** (Start, Stop) Format

An alternative format for the computer is called the **Counting Process** (**CP**) format.

CP Format: applies to more complicated survival analysis

- Age-at follow-up is outcome
- Time-dependent variables
- Recurrent events
- Gaps in follow-up

The CP format is useful for more complicated survival analysis situations that we discuss in later chapters, in particular when age-at-follow-up time is used as the outcome variable instead of time of follow-up (Chap. 3), when there are time-dependent variables (Chap. 6), and when there are recurrent events and/or gaps in follow-up (Chap. 8).



Data Layout CP Approach

	i	j	d_{ij}	t_{ij0}	t_{ij1}	X_{ij1}	$\cdots$	X_{ijp}
Subject 1	1	1	d_{11}	t_{110}	t_{111}	X_{111}	$\cdots$	X_{11p}
	1	2	d_{12}	t_{120}	t_{121}	X_{121}	$\cdots$	X_{12p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	1	r_1	d_{1r_1}	t_{1r_10}	t_{1r_11}	X_{1r_11}	$\cdots$	X_{1r_1p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
Subject i	i	1	d_{i1}	t_{i10}	t_{i11}	X_{111}	$\cdots$	X_{i1p}
	i	2	d_{i2}	t_{i20}	t_{i21}	X_{121}	$\cdots$	X_{i2p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	1	r_i	d_{ir_1}	t_{1r_i0}	t_{1r_i1}	X_{ir_i1}	$\cdots$	X_{ir_ip}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Subject n	n	1	d_{n1}	t_{n10}	t_{n11}	X_{n11}	$\cdots$	X_{n1p}
	n	2	d_{n2}	t_{n20}	t_{n21}	X_{n21}	$\cdots$	X_{n2p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	n	r_n	d_{nr_n}	t_{nr_n0}	t_{nr_n1}	X_{nr_n1}	$\cdots$	X_{nr_np}

The general CP format is shown on the left. This format differs from the previously described "standard" data layout in two ways. First, the CP format allows **multiple lines of data for the same individual**; that is, each individual's total at-risk-follow-up time is subdivided into smaller time intervals to allow for recurrent events on the same individual. Second, there are **two time points specified for each individual**, labeled in the layout as t_{ij0} and t_{ij1}, and often referred to as **START** and **STOP times**, respectively.

The first two columns in this format are labeled **i** (for subject number) and **j** (for dataline number for the ith subject). As in the standard format, i ranges from 1 to n; also, in the CP format, j ranges from 1 to $\mathbf{r_i}$, where $\mathbf{r_i}$ denotes the number of datalines for the i-th subject.

The third column labeled $\mathbf{d_{ij}}$ denotes the failure status (1=failed, 0=censored) for the j-th dataline on the i-th subject.

The next two columns identify two time points required for each dataline, the **START time** ($\mathbf{t_{ij0}}$) and the **STOP time** ($\mathbf{t_{ij1}}$). These two columns are the primary distinguishing feature of the CP format.

Simplest CP format: 1 dataline/ subject

i	j	d_{ij}	t_{ij0}	t_{ij1}	X_{ij1}	⋯	X_{ijp}
1	1	d_{11}	0	t_1	X_{111}	⋯	X_{11p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
i	1	d_{i1}	0	t_i	X_{i11}	⋯	X_{i1p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
n	1	d_{n1}	0	t_n	X_{n11}	⋯	X_{n1p}

The simplest CP format occurs when the outcome is follow-up time since study entry and when there are no recurrent events or time-dependent covariates, as in our previously described Remission Time Dataset. In this situation, there is one dataline for each subject (i.e., r_i=1 for all i so that the only value that j takes is 1), the start time (t_{i10}) is 0 for each subject, and the stop time (t_{i11}) is the follow-up time (t) until either the event or censorship occurs.

CP Format for Group 1 of Remission Time Dataset

i	j	d_{ij}	start	stop	X(Group)
1	1	1	0	6	1
2	1	1	0	6	1
3	1	1	0	6	1
4	1	1	0	7	1
5	1	1	0	10	1
6	1	1	0	13	1
7	1	1	0	16	1
8	1	1	0	22	1
9	1	1	0	23	1
10	1	0	0	12	1
11	1	0	0	6	1
⋮	⋮	⋮	⋮	⋮	⋮
19	1	0	0	32	1
20	1	0	0	34	1
21	1	0	0	35	1

As an example, the CP format for Group 1 of the Remission Time Dataset is shown on the left. Note that the value of j is 1 throughout the table, the start times are all zero, and the stop times are the failure or censored survival times for each subject.

CP Format: First 15 Subjects-Bladder Canscer Study

i	j	d	start	stop	tx	num	size
1	1	0	0	0	0	1	1
2	1	0	0	1	0	1	3
3	1	0	0	4	0	2	1
4	1	0	0	7	0	1	1
5	1	0	0	10	0	5	1
6	1	1	0	6	0	4	1
6	2	0	6	10	0	4	1
7	1	0	0	14	0	1	1
8	1	0	0	18	0	1	1
9	1	1	0	5	0	1	3
9	2	0	5	18	0	1	3
10	1	1	0	12	0	1	1
10	2	1	12	16	0	1	1
10	3	0	16	18	0	1	1
11	1	0	0	23	0	3	3
12	1	1	0	10	0	1	3
12	2	1	10	15	0	1	3
12	3	0	15	23	0	1	3
13	1	1	0	3	0	1	1
13	2	1	3	16	0	1	1
13	3	1	16	23	0	1	1
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1
15	1	1	0	7	0	2	3
15	2	1	7	10	0	2	3
15	3	1	10	16	0	2	3
15	4	0	16	24	0	2	3

We now illustrate the CP format that allows for more than one dataline per subject as well as start times other than zero. We consider data on the first 15 subjects from a study of recurrent bladder cancer tumors (Byar, 1980; and Wei, Lin and Weissfeld, 1989). The entire dataset contained 86 patients, each followed for a variable amount of time up to 64 months. We describe how to analyze this dataset in Chapter 8 on Recurrent Event Survival Analysis. Here, we only describe how this data layout fits the CP format.

The event being analyzed is the recurrence of bladder cancer tumor after transurethral surgical excision. Each recurrence of new tumors was treated by removal at each examination.

The exposure variable of interest is drug treatment status (**tx**, 0=placebo, 1= treatment with thiotepa). Although each of 15 subjects shown here are in the placebo group (tx=0), several other subjects in the larger dataset are in the treatment group (tx=1).

The covariates listed here are initial number of tumors (**num**) and initial size of tumors (**size**) in centimeters. Both these variables have the same value for each subject (i.e., time-independent variables), although the general data layout also allows for time-dependent variables.

Notice that several subjects in this dataset, namely subjects 6, 9, 10, 12, 13, 14, and 15 have two or more datalines. Subject 6, for example, has two datalines (i.e., $r_6 = 2$), whereas subject 14 has 4 datalines (i.e., $r_{14} = 4$).

Bladder Canscer Study (cont'd)

i	j	d	start	stop	tx	num	size
6	1	1	0	6	0	4	1
6	2	0	6	10	0	4	1

The first of the two lines for subject 6 tells us that this subject had a (first) recurrent bladder cancer event (i.e., d=1) at 6 months (i.e., stop =6). This subject was then followed for another 4 months (from 6 to 10, as shown on the second dataline for this subject, where the start time is 6 and the stop time is 10. At 10 months, the subject is censored (d=0); in other words, this subject did not get a second recurrent event at 10 months, after which no further information is available.

i	j	d	start	stop	tx	num	size
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1

Subject 14 had three recurrent events, the first one at 3 months (i.e., stop =3), the second one at 9 months (i.e., stop =9), and the third one at 21 months (i.e., stop =21). This subject was followed for another 2 months (start =21 to stop =23 on dataline number j=4) without another event occurring (d=0).

CP format illustrated for other situations in later chapters.

As mentioned at the beginning of this section, the CP format is also applicable when age-at-follow-up time is used as the outcome variable instead of time of follow-up (Chapter 3), when there are time-dependent variables (Chapter 6), and when there are gaps in follow-up (Chapter 8). We will illustrate these latter situations within the later chapters just mentioned.

See Computer Appendix for computer code in CD format for SAS, STATA, and R.

In the Computer Appendix, we describe the computer code required by SAS, STATA, and R packages when the data is set up in CP format for the analysis of recurrent event survival data and when age is used as the time scale instead of time-on-study.

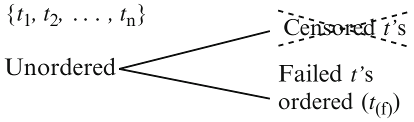
VI. Basic Data Layout for Understanding Analysis

We are now ready to look at another data layout, which is shown at the left. This layout helps provide some understanding of how a survival analysis actually works and, in particular, how survivor curves are derived.

For analysis:

Ordered failure times ($t_{(f)}$)	# of failures (m_f)	# censored in [$t_{(f}$, $t_{(f-1)}$) (q_f)	Risk set R ($t_{(f)}$)
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
·	·	·	·
·	·	·	·
·	·	·	·
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

The first column in this table gives ordered failure times. These are denoted by t's with subscripts within parentheses, starting with $t_{(0)}$, then $t_{(1)}$ and so on, up to $t_{(k)}$. Note that the parentheses surrounding the subscripts distinguish ordered failure times from the survival times previously given in the computer layout.



k = # of distinct times at whick subjects failed ($k \leq n$)

To get ordered failure times from survival times, we must first remove from the list of unordered survival times all those times that are censored; we are thus working only with those times at which people failed. We then order the remaining failure times from smallest to largest, and count ties only once. The value k gives the number of distinct times at which subjects failed.

EXAMPLE

Remission Data: Group 1
(n = 21, 9 failures, k = 7)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Remission Data: Group 2
(n = 21, 21 failures, k = 12)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 1$	2	0	21 persons survive ≥ 1 wks
$t_{(2)} = 2$	2	0	19 persons survive ≥ 2 wks
$t_{(3)} = 3$	1	0	17 persons survive ≥ 3 wks
$t_{(4)} = 4$	2	0	16 persons survive ≥ 4 wks
$t_{(5)} = 5$	2	0	14 persons survive ≥ 5 wks
$t_{(6)} = 8$	4	0	12 persons survive ≥ 8 wks
$t_{(7)} = 11$	2	0	8 persons survive ≥ 11 wks
$t_{(8)} = 12$	2	0	6 persons survive ≥ 12 wks
$t_{(9)} = 15$	1	0	4 persons survive ≥ 15 wks
$t_{(10)} = 17$	1	0	3 persons survive ≥ 17 wks
$t_{(11)} = 22$	1	0	2 persons survive ≥ 22 wks
$t_{(12)} = 23$	1	0	1 person survive ≥ 23 wks
Totals	21	0	

ties

For example, using the remission data for group 1, we find that 9 of the 21 persons failed, including 3 persons at 6 weeks and 1 person each at 7, 10, 13, 16, 22, and 23 weeks. These nine failures have $k = 7$ distinct survival times, because three persons had survival time 6 and we only count one of these 6's as distinct. The first ordered failure time for this group, denoted as $t_{(1)}$, is 6; the second ordered failure time $t_{(2)}$, is 7, and so on up to the seventh ordered failure time of 23.

Turning to group 2, we find that although all 21 persons in this group failed, there are several ties. For example, two persons had a survival time of 1 week; two more had a survival time of 2 weeks; and so on. In all, we find that there were $k = 12$ distinct survival times out of the 21 failures. These times are listed in the first column for group 2.

Note that for both groups we inserted a row of data giving information at time 0. We will explain this insertion when we get to the third column in the table.

The *second column* in the data layout gives frequency counts, denoted by m_f, of those persons who failed at each distinct failure time. When there are no ties at a certain failure time, then $m_f = 1$. Notice that in group 1, there were three ties at 6 weeks but no ties thereafter. In group 2, there were ties at 1, 2, 4, 5, 8, 11, and 12 weeks. In any case, the sum of all the m_f's in this column gives the total number of failures in the group tabulated. This sum is 9 for group 1 and 21 for group 2.

EXAMPLE: (continued)

q_j = censored in $[t_{(j)}, t_{(j+1)}]$

Remission Data: Group 1

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$ ties	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Remission Data: Group 1

#	t(weeks)	d	X(group)
1	6	1	1
2	6	1	1
3	6	1	1
4	7	1	1
5	10	1	1
6	13	1	1
7	16	1	1
8	22	1	1
9	23	1	1
10	6	0	1
11	9	0	1
12	10	0	1
13	11	0	1
14	17	0	1
15	19	0	1
16	20	0	1
17	25	0	1
18	32	0	1
19	32	0	1
20	34	0	1
21	35	0	1

The *third column* gives frequency counts, denoted by q_f, of those persons censored in the time interval starting with failure time $t_{(f)}$ up to the next failure time denoted $t_{(f+1)}$. Technically, because of the way we have defined this interval in the table, we include those persons censored at the beginning of the interval.

For example, the remission data, for group 1 includes 5 nonzero q's: $q_1 = 1$, $q_2 = 1$, $q_3 = 2$, $q_5 = 3$, $q_7 = 5$. Adding these values gives us the total number of censored observations for group 1, which is 12. Moreover, if we add the total number of q's (12) to the total number of m's (9), we get the total number of subjects in group 1, which is 21.

We now focus on group 1 to look a little closer at the q's. At the left, we list the unordered group 1 information followed (on the next page) by the ordered failure time information. We will go back and forth between these two tables (and pages) as we discuss the q's. Notice that in the table here, one person, listed as #10, was censored at week 6. Consequently, in the table at the top of the next page, we have $q_1 = 1$, which is listed on the second line corresponding to the ordered failure time $t_{(1)}$, which equals 6.

The next q is a little trickier, it is derived from the person who was listed as #11 in the table here and was censored at week 9. Correspondingly, in the table at the top of the next page, we have $q_2 = 1$ because this one person was censored within the time interval that starts at the second ordered failure time, 7 weeks, and ends just before the third ordered failure time, 10 weeks. We have *not* counted person #12 (who was censored at week 10) here because this person's censored time is exactly at the end of the interval. We count this person in the following interval.

EXAMPLE: (continued)

Group 1 using ordered failure times

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	$\boxed{1}$	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	①	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	②	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

We now consider, from the table of unordered failure times, person #12 who was censored at 10 weeks, and person #13, who was censored at 11 weeks. Turning to the table of ordered failure times, we see that these two times are within the third ordered time interval, which starts and includes the 10-week point and ends just before the 13th week. As for the remaining q's, we will let you figure them out for practice.

One last point about the q information. We inserted a row at the top of the data for each group corresponding to time 0. This insertion allows for the possibility that persons may be censored after the start of the study but before the first failure. In other words, it is possible that q_0 may be nonzero. For the two groups in this example, however, no one was censored before the first failure time.

EXAMPLE

Risk Set: $R(t_f)$ is the set of individual for whom

Remission Data: Group 1

$t_{(f)}$	$m_{(f)}$	$q_{(f)}$	$R(t_{(f)})$
$t_{(0)}$ = ⓪	0	0	21 persons survive ≥ 0 wks
$t_{(1)}$ = ⑥	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

The last column in the table gives the "**risk set**." The risk set is not a numerical value or count but rather a collection of individuals. By definition, the risk set $R(t_{(f)})$ is the collection of individuals who have survived at least to time $t_{(f)}$; that is, each person in $R(t_{(f)})$ has a survival time that is $t_{(f)}$ or longer, regardless of whether the person has failed or is censored.

For example, we see that at the start of the study everyone in group 1 survived at least 0 weeks, so the risk set at time 0 consists of the entire group of 21 persons. The risk set at 6 weeks for group 1 also consists of all 21 persons, because all 21 persons survived at least as long as 6 weeks. These 21 persons include the 3 persons who failed at 6 weeks, because they survived and were still at risk just up to this point.

EXAMPLE: (continued)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Now let's look at the risk set at 7 weeks. This set consists of 17 persons in group 1 that survived at least 7 weeks. We omit everyone in the X-ed area. Of the original 21 persons, we therefore have excluded the three persons who failed at 6 weeks and the one person who was censored at 6 weeks. These four persons did not survive at least 7 weeks. Although the censored person may have survived longer than 7 weeks, we must exclude him or her from the risk set at 7 weeks because we have information on this person only up to 6 weeks.

To derive the other risk sets, we must exclude all persons who either failed or were censored before the start of the time interval being considered. For example, to obtain the risk set at 13 weeks for group 1, we must exclude the five persons who failed before, but not including, 13 weeks and the four persons who were censored before, but not including, 13 weeks. Subtracting these 9 persons from 21, leaves 12 persons in group 1 still at risk for getting the event at 13 weeks. Thus, the risk set consists of these 12 persons.

How we work with censored data: Use all information up to time of censorship; don't throw away information.

The importance of the table of ordered failure times is that we can work with censored observations in analyzing survival data. Even though censored observations are incomplete, in that we don't know a person's survival time exactly, we can still make use of the information we have on a censored person up to the time we lose track of him or her. Rather than simply throw away the information on a censored person, we use all the information we have on such a person up until time of censorship. (Nevertheless, most survival analysis techniques require a key assumption that censoring is independent, i.e., censored subjects are not at increased risk for failure. See Chap. 9 on competing risks for further details.)

EXAMPLE

$t_{(f)}$	m_f	q_f		$R(t_{(f)})$
6	3	1	✓	21 persons
7	1	1	✓	17 persons
10	1	2	✓	15 persons
13	1	0	✓	12 persons
16	1	③	✓	11 persons
22	1	0		7 persons
23	1	5		6 persons

For example, for the three persons in group 1 who were censored between the 16th and 22nd weeks, there are at least 16 weeks of survival information on each that we don't want to lose. These three persons are contained in all risk sets up to the 16th week; that is, they are each at risk for getting the event up to 16 weeks. Any survival probabilities determined before, and including, 16 weeks should make use of data on these three persons as well as data on other persons at risk during the first 16 weeks.

Having introduced the basic terminology and data layouts to this point, we now consider some data analysis issues and some additional applications.

VII. Descriptive Measures of Survival Experience

EXAMPLE

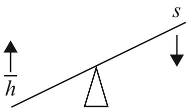
Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
$\overline{T}_1$ (ignoring + 's) = 17.1	$\overline{T}_2 = 8.6$
$\bar{h}_1 = \frac{9}{359} = .025$	$\bar{h}_2 = \frac{21}{182} = .115$

$$\text{Average hazard rate } (\bar{h}) = \frac{\#\text{ failures}}{\sum_{i=1}^{n} t_i}$$

We first return to the remission data, again shown in untabulated form. Inspecting the survival times given for each group, we can see that most of the treatment group's times are longer than most of the placebo group's times. If we ignore the plus signs denoting censorship and simply average all 21 survival times for each group we get an average, denoted by ***T*** "bar," of **17.1** weeks survival for the treatment group and **8.6** weeks for the placebo group. Because several of the treatment group's times are censored, this means that group 1's true average is even larger than what we have calculated. Thus, it appears from the data (without our doing any mathematical analysis) that, regarding survival, the treatment is more effective than the placebo.

As an alternative to the simple averages that we have computed for each group, another descriptive measure of each group is the **average hazard rate**, denoted as ***h*** "bar." This rate is defined by dividing the total number of failures by the sum of the observed survival times. For group 1, *h* "bar" is 9/359, which equals **.025**. For group 2, *h* "bar" is 21/182, which equals. **115**.



As previously described, the hazard rate indicates failure potential rather than survival probability. Thus, the higher the average hazard rate, the lower is the group's probability of surviving.

In our example, the average hazard for the treatment group is smaller than the average hazard for the placebo group.

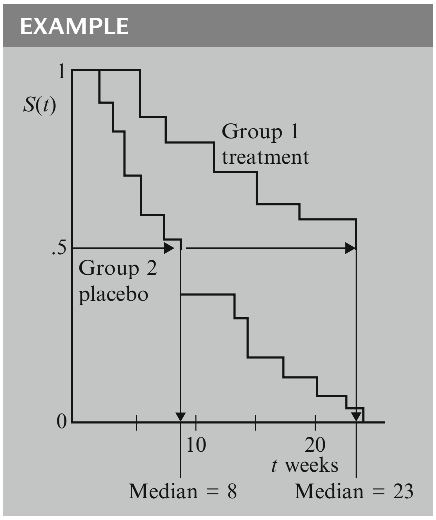
Placebo hazard > treatment hazard: suggests that treatment is more effective than placebo

Thus, using average hazard rates, we again see that the treatment group appears to be doing better overall than the placebo group; that is, the treatment group is less prone to fail than the placebo group.

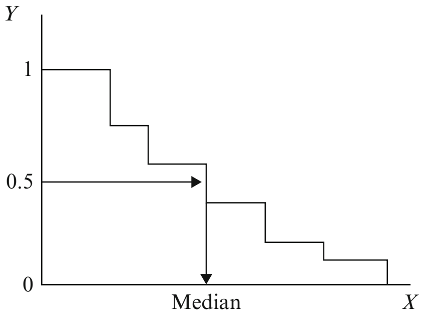
Descriptive measures ($\overline{T}$ and $\overline{h}$) give **overall** comparison; they do not give comparison over time.

The descriptive measures we have used so far—the ordinary average and the hazard rate average—provide overall comparisons of the treatment group with the placebo group. These measures don't compare the two groups at different points in time of follow-up. Such a comparison is provided by a graph of survivor curves.

EXAMPLE



Here we present the **estimated survivor curves** for the treatment and placebo groups. The method used to get these curves is called the Kaplan–Meier method, which is described in Chap. 2. When estimated, these curves are actually step functions that allow us to compare the treatment and placebo groups over time. The graph shows that the survivor function for the treatment group consistently lies above that for the placebo group; this difference indicates that the treatment appears effective at all points of follow-up. Notice, however, that the two functions are somewhat closer together in the first few weeks of follow-up, but thereafter are quite spread apart. This widening gap suggests that the treatment is more effective later during follow-up than it is early on.



Median (treatment) = 23 weeks
Median (placebo) = 8 weeks

Also notice from the graph that one can obtain estimates of the median survival time, the time at which the survival probability is.5 for each group. Graphically, the median is obtained by proceeding horizontally from the 0.5 point on the *Y*-axis until the survivor curve is reached, as marked by an arrow, and then proceeding vertically downward until the *X*-axis is crossed at the median survival time.

For the treatment group, the median is 23 weeks; for the placebo group, the median is 8 weeks. Comparison of the two medians reinforces our previous observation that the treatment is more effective overall than the placebo.

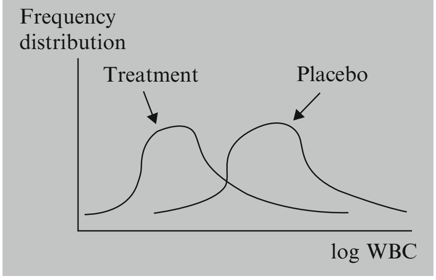
VIII. Example: Extended Remission Data

Group 1		Group 2	
t (weeks)	log WBC	*t* (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

Before proceeding to another data set, we consider the remission example data (Freireich et al., *Blood*, 1963) in an **extended form**. The table at the left gives the remission survival times for the two groups with additional information about white blood cell count for each person studied. In particular, each person's log white blood cell count is given next to that person's survival time. The epidemiologic reason for adding log WBC to the data set is that this variable is usually considered an important predictor of survival in leukemia patients; the higher the WBC, the worse the prognosis. Thus, any comparison of the effects of two treatment groups needs to consider the possible **confounding effect** of such a variable.

EXAMPLE: CONFOUNDING

Treatment group: $\overline{\log \text{WBC}} = 1.8$
Placebo group: $\overline{\log \text{WBC}} = 4.1$
Indicates **confounding** of treatment effect by log WBC



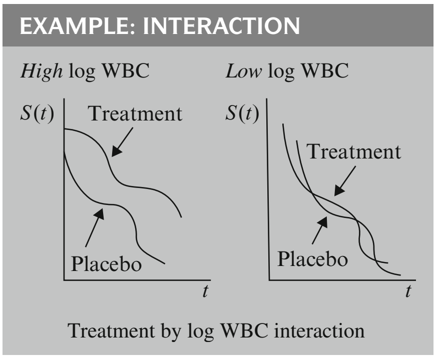
Although a full exposition of the nature of confounding is not intended here, we provide a simple scenario to give you the basic idea. Suppose all of the subjects in the treatment group had very low log WBC, with an average, for example, of 1.8, whereas all of the subjects in the placebo group had very high log WBC, with an average of 4.1. We would have to conclude that the results we've seen so far that compare treatment with placebo groups may be misleading.

The additional information on log WBC would suggest that the treatment group is surviving longer simply because of their low WBC and not because of the efficacy of the treatment itself. In this case, we would say that **the treatment effect is confounded by the effect of log WBC**.

Need to adjust for imbalance in the distribution of log WBC

More typically, the distribution of log WBC may be quite different in the treatment group than in the control group. We have illustrated one extreme in the graph at the left. Even though such an extreme is not likely, and is not true for the data given here, the point is that some attempt needs to be made to adjust for whatever imbalance there is in the distribution of log WBC. However, if high log WBC count was a consequence of the treatment, then white blood cell count should not be controlled for in the analysis.

EXAMPLE: INTERACTION



Treatment by log WBC interaction

Another issue to consider regarding the effect of log WBC is **interaction**. What we mean by interaction is that the effect of the treatment may be different, depending on the level of log WBC. For example, suppose that for persons with high log WBC, survival probabilities for the treatment are consistently higher over time than for the placebo. This circumstance is illustrated by the first graph at the left. In contrast, the second graph, which considers only persons with low log WBC, shows no difference in treatment and placebo effect over time. In such a situation, we would say that **there is strong treatment by log WBC interaction**, and we would have to qualify the effect of the treatment as depending on the level of log WBC.

Need to consider:

- interaction;
- confounding.

The example of interaction we just gave is but one way interaction can occur; on the other hand, interaction may not occur at all. As with confounding, it is beyond our scope to provide a thorough discussion of interaction. In any case, the assessment of interaction is something to consider in one's analysis in addition to confounding that involves explanatory variables.

The problem:
Compare two groups after adjusting for confounding and interaction.

Thus, with our extended data example, the basic **problem** can be described as follows: to compare the survival experience of the two groups after adjusting for the possible confounding and/or interaction effects of log WBC.

EXAMPLE

	Individual #	t (weeks)	d	X_1 (Group)	X_2 (log WBC)
Group 1	1	6	1	1	2.31
	2	6	1	1	4.06
	3	6	1	1	3.28
	4	7	1	1	4.43
	5	10	1	1	2.96
	6	13	1	1	2.88
	7	16	1	1	3.60
	8	22	1	1	2.32
	9	23	1	1	2.57
	10	6	0	1	3.20
	11	9	0	1	2.80
	12	10	0	1	2.70
	13	11	0	1	2.60
	14	17	0	1	2.16
	15	19	0	1	2.05
	16	20	0	1	2.01
	17	25	0	1	1.78
	18	32	0	1	2.20
	19	32	0	1	2.53
	20	34	0	1	1.47
	21	35	0	1	1.45
Group 2	22	1	1	0	2.80
	23	1	1	0	5.00
	24	2	1	0	4.91
	25	2	1	0	4.48
	26	3	1	0	4.01
	27	4	1	0	4.36
	28	4	1	0	2.42
	29	5	1	0	3.49
	30	5	1	0	3.97
	31	8	1	0	3.52
	32	8	1	0	3.05
	33	8	1	0	2.32
	34	8	1	0	3.26
	35	11	1	0	3.49
	36	11	1	0	2.12
	37	12	1	0	1.50
	38	12	1	0	3.06
	39	15	1	0	2.30
	40	17	1	0	2.95
	41	22	1	0	2.73
	42	23	1	0	1.97

The problem statement tells us that we are now considering two explanatory variables in our extended example, whereas we previously considered a single variable, group status. The data layout for the computer needs to reflect the addition of the second variable, log WBC. The extended table in computer layout form is given at the left. Notice that we have labeled the two explanatory variables X_1 (for group status) and X_2 (for log WBC). The variable X_1 is our primary study or exposure variable of interest here, and the variable X_2 is an extraneous variable that we are interested in accounting for because of either confounding or interaction.

Analysis alternatives:

- stratify on log WBC;
- use math modeling, e.g., proportional hazards model.

As implied by our extended example, which considers the possible confounding or interaction effect of log WBC, we need to consider methods for adjusting for log WBC and/or assessing its effect in addition to assessing the effect of treatment group. The two most popular alternatives for analysis are the following:

- to stratify on log WBC and compare survival curves for different strata; or
- to use mathematical modeling procedures such as the proportional hazards or other survival models; such methods will be described in subsequent chapters.

IX. Multivariable Example

- Describes general multivariable survival problem.
- Gives analogy to regression problems.

We now consider one other example. Our purpose here is to describe a more general type of multivariable survival analysis problem. The reader may see the analogy of this example to multiple regression or even logistic regression data problems.

EXAMPLE

13-year follow-up of fixed cohort from Evans County, Georgia

$n = 170$ white males (60+)

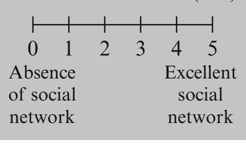
$T =$ years until death

Event = death

Explanatory variables:
- exposure variable
- confounders
- interaction variables

Exposure:

Social Network Index (**SNI**)



We consider a data set developed from a 13-year follow up study of a fixed cohort of persons in Evans County Georgia, during the period 1967–1980 (Schoenbach et al., *Amer. J. Epid.*, 1986). From this data set, we focus on a portion containing $n = 170$ white males who are age 60 or older at the start of follow-up in 1967.

For this data set, the outcome variable is T, time in years until death from start of follow-up, so the event of interest is **death**. Several explanatory variables are measured, one of which is considered the primary exposure variable; the other variables are considered as potential confounders and/or interaction variables.

The primary exposure variable is a measure called Social Network Index (SNI). This is an ordinal variable derived from questionnaire measurement and is designed to assess the extent to which a study subject has social contacts of various types. With the questionnaire, a scale is used with values ranging from 0 (absence of any social network) to 5 (excellent social network).

EXAMPLE: (continued)

Study goal: to determine whether **SNI** is protective against death, i.e., **SNI** ↗ ⇒ S(*t*) ↗.

Explanatory variables:

$$\begin{array}{ll}\textbf{SNI} & \text{Exposure variable}\\ \left.\begin{array}{l}\textbf{AGE}\\ \textbf{SBP}\\ \textbf{CHR}\\ \textbf{QUET}\\ \textbf{SOCL}\end{array}\right\} & \begin{array}{l}\text{Potential confounders/}\\ \text{interaction variables}\end{array}\end{array}$$

Note: $\text{QUET} = \dfrac{\text{Weight}}{(\text{height})} \times 100$

The study's goal is to determine whether one's social network, as measured by SNI, is protective against death. If this study hypothesis is correct, then the higher the social network score, the longer will be one's survival time.

In evaluating this problem, several explanatory variables, in addition to SNI, are measured at the start of follow-up. These include AGE, systolic blood pressure (SBP), an indicator of the presence or absence of some chronic disease (CHR), body size as measured by Quetelet's index (QUET = weight over height squared times 100), and social class (SOCL).

These five additional variables are of interest because they are thought to have their own special or collective influence on how long a person will survive. Consequently, these variables are viewed as potential confounders and/or interaction variables in evaluating the effect of social network on time to death.

The problem:

To describe the relationship between **SNI** and time to death, after controlling for **AGE, SBP, CHR, QUET**, and **SOCL**.

We can now clearly state the problem being addressed by this study: To describe the relationship between SNI and time to death, controlling for AGE, SBP, CHR, QUET, and SOCL.

Goals:

- Measure of effect (adjusted)
- Survivor curves for different SNI categories (adjusted)
- Decide on variables to be adjusted; determine method of adjustment

Our goals in using survival analysis to solve this problem are as follows:

- to obtain some measure of effect that will describe the relationship between SNI and time until death, after adjusting for the other variables we have identified;
- to develop survival curves that describe the probability of survival over time for different categories of social networks; in particular, we wish to compare the survival of persons with excellent networks to the survival of persons with poor networks. Such survival curves need to be adjusted for the effects of other variables.
- to achieve these goals, two intermediary goals are to decide which of the additional variables being considered need to be adjusted and to determine an appropriate method of adjustment.

The computer data layout for this problem is given below. The first column lists the 170 individuals in the data set. The second column lists the survival times, and the third column lists failure or censored status. The remainder of the columns list the 6 explanatory variables of interest, starting with the exposure variable SNI and continuing with the variables to be accounted for in the analysis.

Computer layout: 13-year follow-up study (1967–1980) of a fixed cohort of $n = 170$ white males (60+) from Evans County, Georgia

#	t	d	SNI	AGE	SBP	CHR	QUET	SOCL
1	t_1	d_1	SNI_1	AGE_1	SBP_1	CHR_1	QUET_1	SOCL_1
2	t_2	d_2	SNI_2	AGE_2	SBP_2	CHR_2	QUET_2	SOCL_2
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
170	t_{170}	d_{170}	SNI_{170}	AGE_{170}	SBP_{170}	CHR_{170}	QUET_{170}	SOCL_{170}

X. Math Models in Survival Analysis

General framework

Controlling for $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots \boldsymbol{C}_p$.

SNI study:

$E = \text{SNI} \Rightarrow D = \text{survival time}$

Controlling for **AGE, SBP, CHR, QUET**, and **SOCL**

It is beyond the scope of this presentation to provide specific details of the survival analysis of these data. Nevertheless, the problem addressed by these data is closely analogous to the typical multivariable problem addressed by linear and logistic regression modeling. Regardless of which modeling approach is chosen, the typical problem concerns describing the relationship between an exposure variable (e.g., E) and an outcome variable (e.g., D) after controlling for the possible confounding and interaction effects of additional variables (e.g., C_1, C_2, and so on up to C_p). In our survival analysis example, E is the social network variable SNI, D is the survival time variable, and there are $p = 5$ C variables, namely, AGE, SBP, CHR, QUET, and SOCL.

	Model	Outcome
	Survival analysis	Time to event (with censoring)
follow-up time info not used	Linear regression	Continuous (SBP)
	Logistic regression	Dichotomous (CHD yes/no)

Nevertheless, an important distinction among modeling methods is the type of outcome variable being used. In survival analysis, the outcome variable is "time to an event," and there may be censored data. In linear regression modeling, the outcome variable is generally a continuous variable, like blood pressure. In logistic modeling, the outcome variable is a dichotomous variable, like CHD status, yes or no. And with linear or logistic modeling, we usually do not have information on follow-up time available.

As with linear and logistic modeling, one statistical goal of a survival analysis is to obtain some measure of effect that describes the exposure–outcome relationship adjusted for relevant extraneous variables.

Measure of effect:

Linear regression:
regression coefficient β

In linear regression modeling, the measure of effect is usually some regression coefficient β.

Logistic regression:
odds ratio e^{β}

In logistic modeling, the measure of effect is an odds ratio expressed in terms of an exponential of one or more regression coefficients in the model, for example, *e* to the β.

Survival analysis:
hazard ratio e^{β}

In survival analysis, the measure of effect typically obtained is called a **hazard ratio**; as with the logistic model, this hazard ratio is expressed in terms of an exponential of one or more regression coefficients in the model.

EXAMPLE

SNI study: hazard ratio (HR) describes relationship between SNI and *T*, after controlling for covariates.

Thus, from the example of survival analysis modeling of the social network data, one may obtain a hazard ratio that describes the relationship between SNI and survival time (*T*), after controlling for the appropriate covariates.

Interpretation of HR (like OR):

HR = 1 ⇒ no relationship

HR = 10 ⇒ exposed hazard 10 times unexposed

HR = 1/10 ⇒ exposed hazard 1/10 times unexposed

The hazard ratio, although a different measure from an odds ratio, nevertheless has a similar interpretation of the strength of the effect. A hazard ratio of 1, like an odds ratio of 1, means that there is no effect; that is, 1 is the null value for the exposure–outcome relationship. A hazard ratio of 10, on the other hand, is interpreted like an odds ratio of 10; that is, the exposed group has ten times the hazard of the unexposed group. Similarly, a hazard ratio of 1/10 implies that the exposed group has one-tenth the hazard of the unexposed group.

XI. Censoring Assumptions

Three assumptions about censoring:

Independent (vs. non-independent) censoring

Random (vs. non-random) censoring

Non-informative (vs. informative) censoring

There are three assumptions about censoring often considered for survival data: **independent censoring**, **random censoring**, and **non-informative censoring**. Although these assumptions have similarities, they are also somewhat different and are often confused in the textbook and published literature as being interchangeable synonyms.

Mathematic definitions have been provided elsewhere.

Mathematical definitions of independent (vs. nonindependent), random (vs. nonrandom), and non-informative (vs. informative) censoring have been given elsewhere (Kalbfleisch and Prentice, 1980; Klein and Moeschberger, 2003). Here, however, we prefer to provide more intuitive definitions and examples.

Independent (vs. non-independent) censoring

- most useful
- affects validity

Random (vs. non-random) censoring

- more restrictive than independent, i.e., random ⇒ indep, whereas indep ⇏ random.

The assumption of independent censoring is the most useful of the three types for drawing correct inferences that compare the survival experience of two or more groups (e.g., treatment vs. placebo). In particular, the presence of non-independent censoring typically affects the validity of one's estimated effect. Random censoring is a stronger assumption and more restrictive than independent censoring.

Random Censoring:

	Failure rate	
Censored		Not censored
$h_{Ce}(t)$	=	$h_{NCe}(t)$

To be more specific, **random censoring** essentially means that *subjects who are censored at time t should be representative of all the study subjects who remained at risk at time t with respect to their survival experience.* In other words, the failure rate for subjects who are censored is assumed to be equal to the failure rate for subjects who remained in the risk set who are not censored.

Independent censoring:

		Failure rate	
Subgrp	Censored		Not censored
A	$h_{A,Ce}(t)$	=	$h_{A,NCe}(t)$
B	$h_{B,Ce}(t)$	=	$h_{B,NCe}(t)$

Independent censoring essentially means that *within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience.* In other words, censoring is independent provided that it is random within any subgroup of interest. We illustrate these ideas with an example.

EXAMPLE

Group A

Time	# at risk	# events	# survived
0–3 yrs	100	20	80

3-yr risk = 20/100 = 0.20
3-yr survival = 80/100 =0.80

Suppose that we are interested in estimating the 3-year survival (for some disease) among those in Group A. We follow 100 individuals intially disease free for 3 years. Over the 3-year period, 20 contract disease. We estimate the 3-year risk of disease for those in Group A to be 0.20 and the 3-year survival to be 0.80 (since 80 of 100 survived).

Time	# at risk	# events	# survived
0–3	100	20	80
	40 leave study		
3–5	40	5	35

5-year survival?

Now suppose we wish to continue the study for another 2 years in order to estimate the 5-year survival for Group A. We want to continue the following for the 80 individuals who participated in the study and survived for the first 3 years. However, half or 40 of those 80 individuals refused to continue in the study and were therefore lost to follow-up (censored). Of the other 40 individuals who remained in the study, 5 individuals contract the disease. With this information, what is the estimate of the 5-year survival for Group A and under what assumptions?

EXAMPLE: (continued)

What happened to 40 individuals who were censored at 3 years? *Don't know*

Assuming indep and random censoring:

40 at risk at time 5
similar to
40 censored at time 3

i.e.,
expect 5 events from 40 censored at time 3 since 5 events from 40 at risk

If we know what happened to the 40 individuals who were censored at the 3-year mark, then we could sum the total number of events and the total number of individuals who survived (out of the original 100 at risk). **Under an assumption of independent and random censoring, we assume that the 40 individuals who were censored were similar to the 40 who remained at risk with respect to their survival experience.** Since 5 of the 40 who remained in the study after 3 years contracted disease over the next 2 years, we estimate that 5 of the 40 who were censored also contracted the disease over the same time period, even though their disease experience was unobserved.

Estimated # of cases over 5 years:

$$\underbrace{20}_{\text{first 3 years}} + \underbrace{5}_{\text{next 2 years}} + \underbrace{5}_{\text{censored cases}}$$

= 30 estimated cases from original 100 over 5 years

Estimated 5-year survival = 70/100 = 0.70

So over the course of the 5 years: 20 contracted disease in the first 3 years, 5 were observed to get disease after 3 years, and 5 of the censored individuals were estimated to have contracted disease. This yields 20 + 5 + 5 = 30 who are estimated to have contract disease leaving 70 of the original 100 who have survived over the 5-year period. The estimated 5-year survival among Group A is 0.70 under the assumptions of random and independent censoring.

The idea:
Assume survival experience of subjects censored at t is *as expected* if randomly selected from subjects who are at risk at t.

The idea behind independent and random censoring is that it is *as if* the subjects censored at time t were randomly selected to be censored from the group of subjects who were in the risk set at time t. Even though the censored subjects were not randomly selected, their survival experience would be expected to be the same as if they had been randomly selected from the risk set at time t.

So far, there is no distinction between independent and random censoring.

Reason: Only considering one group

In this example, there is no distinction made between independent censoring and random censoring. The reason is because we are only considering one group of individuals (i.e., there are no predictor variables considered). The distinction comes if we consider more than one group for comparison. We illustrate this distinction by continuing the example.

EXAMPLE: (continued)

Group B

Time	# at risk	# events	# survived
0–3	100	40	60
	10 leave study		
3–5	50	10	40

Failure risk from 3 to 5 yrs =10/50 = **0.20**.

We extend the previous example to include 100 subjects from Group B who are disease free at the start of follow-up. The goal is to estimate their 5-year survival and compare it to the 5-year survival for Group A. Suppose over the first 3-year period, 40 of the 100 individuals contract disease. Then, of the 60 who survive the first 3 years, 10 refuse to continue in the study and are therefore censored. For the 50 who remain in the study, 10 individuals contract disease by the 5th year (10 of 50 = 20%).

Under independent censoring, we estimate that 20% or 2 of the 10 censored subjects contract disease by the 5th year.

Assuming independent censoring:
expect **0.20**×10 =2 cases
from 10 censored at time 3

Estimated # of cases over 5 years:

$$\underset{\text{first 3 years}}{40} + \underset{\text{next 2 years}}{40} + \underset{\text{censored cases}}{2}$$

= 52 estimated cases from original 100 over 5 years

Estimated 5-year survival = 48/100 = 0.48

So, over the course of the 5 years among the original 100 in Group B: 40 contracted disease in the first 3 years, 10 were observed to get disease after 3 years, and 2 of the censored individuals were estimated to the contracted disease. This yields 40 + 10 + 2 = 52 who are estimated to have contracted disease, leaving 48 of the original 100 who survived over the 5-year period. The estimated 5-year survival among Group B is 0.48 under the assumptions of independent censoring.

Groups A and B combined

Time	# at risk			# events			# survived		
	A	B	total	A	B	total	A	B	total
0–3	100	100	**200**	20	40	**60**	80	60	**140**
40 from **A** and **10** from **B** leave study									
3–5	40	50	90	5	10	15	35	40	75

$$\left\{\begin{array}{l} \mathrm{p_A(censored)} = 40/80 = 0.50 \text{ or } 50\% \\ \mathrm{p_B(censored)} = 10/60 = 0.17 \text{ or } 17\% \\ \mathrm{p_A(censored)} \gg \mathrm{p_B(censored)} \end{array}\right.$$

	Group A	**Group B**
5-yr survival	0.70	0.48

⇓

Censoring not random

Over all, combining both groups, there were 200 originally at risk of whom 60 contracted disease within the first 3 years (20 from Group A and 40 from Group A) leaving 140 who have survived for the first 3 years (80 from Group A and 60 from Group B). At the 3-year mark, 50 subjects were censored (40 in Group A and 10 in Group B).

A much **higher proportion of censoring occurred in Group A** (i.e., 40/80 = 0.50) **than in Group B** (i.e., 10/60 = 0.17).

Moreover, subjects in **Group A** had a higher survival probability than those in **Group B**.

Therefore, the censoring was not random.

EXAMPLE (continued)

Random censoring
within **Group A** *and* within **Group B**
⇓
Independent censoring
(i.e., random censoring conditional on covariates)

Nevertheless,
(overall) random censoring not met

However, conditional on each level of covariates (conditional on group status in this example), the censoring was random. Therefore, the censoring was independent because **independent censoring is random censoring conditional on each level of covariates**.

Nevertheless, the random censoring assumption was not met overall because the censored individuals were not representative of all who remained in the risk set at time t with respect to the rate of failure.

ALTERNATIVE EXAMPLE

Time	# at risk			# events			# survived		
	A	B	total	A	B	total	A	B	total
0–3	100	100	200	20	40	60	80	60	140
40 from **A** and **30** from **B** leave study									
3–5	40	30	70	5	10	15	35	20	55

$p_{\mathbf{A}}(\text{censored}) = 40/80 = 0.50$ or 50%
$p_{\mathbf{B}}(\text{censored}) = 30/60 = 0.50$ or 50%

$p_{\mathbf{A}}(\text{censored}) = p_{\mathbf{B}}(\text{censored})$
⇓
Random censoring (overall)

If instead in the previous example, suppose **40** of 80 from Group A and **30** of 60 from Group B were censored at the 3-year mark, as shown in the table on the left.

Then the censoring would have been random because an equal proportion of Group A and Group B would have been censored from the risk set and those censored would be representative of those who remained at risk.

Non-informative censoring depends on

- distribution of time-to-event
- distribution of time-to-censorship

We next consider the assumption of **non-informative censoring**. Whether censoring is non-informative or informative depends on two distributions: (1) the distribution of the time-to-event random variable and (2) the distribution of time-to-censorship random variable.

Time-to-event random variable (T): Distribution of survival times assuming:

- no loss-to-follow-up
- study continues until all subjects get event

We can conceptualize the distribution for the time-to-event random variable by considering the distribution of survival times if there was no loss to follow-up and the study did not end until all subjects got the event.

Time-to-censorship random variable (C):
Distribution of censoring times assuming:

- study ends before all subjects get event
- censored subjects do not get event prior to the end of study

Similarly, we can conceptualize a time-to-censorship random variable by considering the distribution of censoring times for those subjects who would not have gotten the event if the study ended before all subjects got the event.

Non-informative censoring:

T distribution $\overset{\text{no information}}{\nLeftrightarrow}$ *C distribution*

Note: must still need to know which subjects are censored or not censored.

Non-informative censoring occurs if the distribution of survival times (T) provides no information about the distribution of censorship times (C), and vice versa. Otherwise, the censoring is **informative**. Note, however, that the data must still identify which subjects are or are not censored.

Non-informative
Independent
Random

- Often all justifiable together
- Not all equivalent

The assumption of non-informative censoring is often justifiable when censoring is independent and/or random; nevertheless, these three assumptions are not equivalent.

EXAMPLE: Independent and random but informative censoring

Subject A gets event
⇓
Subject B (randomly selected) gets event, e.g., family member of Subject A leaves study

Assume: censored subjects represent subjects at risk at any time

Then

- independent and random censoring
- **informative censoring** since T ⇒ C. (i.e., T distribution specifies C distribution)

To illustrate how independent censoring could be different from non-informative censoring, we describe an artificial example where the censoring is informative but also random and independent.

Suppose every time an individual gets an event, another individual in the study is randomly selected to leave the study, e.g., after an event, a family member decides to leave the study. If those censored were representative of those who remained in the risk set, then the censoring would be random and independent. However, the censoring would also be informative as **the censoring mechanism would be related to the time-to-event distribution** (since events cause censorships). In fact, if this was the only mechanism in which individuals were censored, the distribution of survival times would completely specify the distribution of censoring times (highly informative).

EXAMPLE: Not independent censoring

- Drug side effect causes censoring
- Censored subjects not representative of subjects still at risk
- Censored subjects more vulnerable than subjects still at risk

⇓

Assuming independent censoring would overestimate survival

To see how bias can occur if the censoring is not independent, consider a drug study in which some individuals are censored from the study due to the occurrence of some side effects. It may be that the unobserved survival experience among those who are censored due to a drug side effect is not representative of those who remained in the study. If those with a side effect are more vulnerable to the health outcome, then we would likely overestimate their survival with an assumption of independent censoring.

Independent censoring most relevant: affects validity

Many of the analytic techniques discussed in the chapters that follow Kaplan–Meier survival estimation, the log rank test, and the Cox model, rely on an assumption of independent censoring for valid inference in the presence of right-censored data.

Chapters

✓ 1. Introduction
2. Kaplan–Meier Survival Curves and the Log–Rank Test

This presentation is now complete. We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

In Chap. 2 we describe how to estimate and graph survival curves using the Kaplan–Meier (KM) method. We also describe how to test whether two or more survival curves are estimating a common curve. The most popular such test is called the log–rank test.

Detailed Outline

I. **What is survival analysis**? (pages 4–5)

A. Type of problem addressed: outcome variable is **time until an event occurs**.

B. Assume one event of interest; more than one type of event implies **a competing risk** problem.

C. Terminology: time = survival time; event = failure.

D. Examples of survival analysis:

i. leukemia patients/time in remission
ii. disease-free cohort/time until heart disease
iii. elderly population/time until death
iv. parolees/time until rearrest (recidivism)
v. heart transplants/time until death

II. **Censored data** (pages 5–8)

A. Definition: don't know survival time exactly.

B. Typical reasons: study ends, loss to follow-up, withdrawal from study.

C. Example that illustrates (right-) censoring.

D. Right-censoring: true survival time is equal to or greater than observed survival times.

E. Left-censoring: true survival time is less than or equal to observed survival time

F. Interval-censoring: true survival time is within a known time interval (t1, t2)

G. Interval-censoring incorporates right- and left-censoring as special cases, i.e.,

right-censoring ⇒ t_1 =lower bound, $t_2=\infty$;
left-censoring ⇒ $t_1=0$, t_2=upper bound.

III. **Terminology and notation** (pages 9–15)

A. Notation: T = survival time random variable:

t = specific value for T
d = (0–1) variable for failure/censorship status

B. Terminology: $S(t)$ = survivor function
$h(t)$ = hazard function

C. Properties of survivor function:

- theoretically, graph is smooth curve, decreasing from $S(t) = 1$ at time $t = 0$ to $S(t) = 0$ at $t = \infty$;
- in practice, graph is step function that may not go all the way to zero at end of study if not everyone studied gets the event.

D. Hazard function formula:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

E. Hazard function properties:

- $h(t)$ gives instantaneous potential for event to occur given survival up to time t;
- instantaneous potential idea is illustrated by velocity;
- hazard function also called "conditional failure rate";
- $h(t) \geq 0$; has no upper bound; not a probability; depends on time units.

F. Examples of hazard curves:

i. exponential
ii. increasing Weibull
iii. decreasing Weibull
iv. log normal

G. Uses of hazard function:

- gives insight about conditional failure rates;
- identifies specific model form;
- math model for survival analysis is usually written in terms of hazard function.

H. Relationship of $S(t)$ to $h(t)$: if you know one, you can determine the other:

- example: $h(t) = \lambda$ if and only if $S(t) = e^{-\lambda t}$
- general formulae:

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

IV. **Goals of survival analysis** (page 16)

A. Estimate and interpret survivor and/or hazard functions.

B. Compare survivor and/or hazard functions.

C. Assess the relationship of explanatory variables to survival time.

V. Basic data layout for computer (pages 16–23)

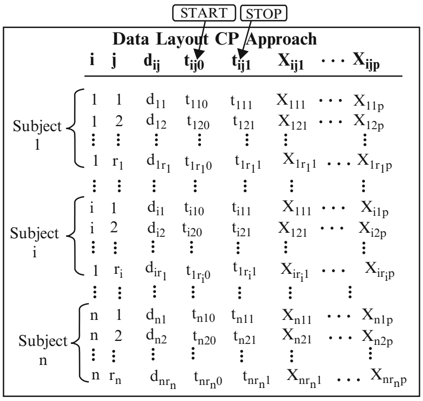
A. General layout:

#	t	d	X_1	$X_2 \ldots X_p$
1	t_1	d_1	X_{11}	$X_{12} \ldots X_{1p}$
2	t_2	d_2	X_{21}	$X_{22} \ldots X_{2p}$
.	.	.		.
.	.	.		.
.	.	.		.
i	t_i	d_i	X_{i1}	$X_{i2} \ldots X_{ip}$
.	.	.		.
.	.	.		.
.	.	.		.
n	t_n	d_n	X_{n1}	$X_{n2} \ldots X_{np}$

B. Example: Remission time data

C. Alternative Data Layout for Computer: Couuting Process (Start, Stop) Format

- Useful for more complicated survival analysis:
  - i. Age-at-follow-up as time scale (Chapter 3)
  - ii. Time-dependent variables (Chapter 6)
  - iii. Recurrent events (Chapter 7)
- CP data layout



START STOP

Data Layout CP Approach

	i	**j**	$\mathbf{d_{ij}}$	$\mathbf{t_{ij0}}$	$\mathbf{t_{ij1}}$	$\mathbf{X_{ij1}}$	···	$\mathbf{X_{ijp}}$
Subject 1	1	1	d_{11}	t_{110}	t_{111}	X_{111}	···	X_{11p}
	1	2	d_{12}	t_{120}	t_{121}	X_{121}	···	X_{12p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	1	r_1	d_{1r_1}	t_{1r_10}	t_{1r_11}	X_{1r_11}	...	X_{1r_1p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
Subject i	i	1	d_{i1}	t_{i10}	t_{i11}	X_{111}	···	X_{i1p}
	i	2	d_{i2}	t_{i20}	t_{i21}	X_{121}	···	X_{i2p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	1	r_i	d_{ir_1}	t_{1r_i0}	t_{1r_i1}	X_{ir_i1}	···	X_{ir_ip}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Subject n	n	1	d_{n1}	t_{n10}	t_{n11}	X_{n11}	···	X_{n1p}
	n	2	d_{n2}	t_{n20}	t_{n21}	X_{n21}	···	X_{n2p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	n	r_n	d_{nr_n}	t_{nr_n0}	t_{nr_n1}	X_{nr_n1}	...	X_{nr_np}

- Simplest CP format: 1 dataline subject

$\mathbf{i}$	$\mathbf{j}$	$\mathbf{d_{ij}}$	$\mathbf{t_{ij0}}$	$\mathbf{t_{ij1}}$	$\mathbf{X_{ij1}}$	$\cdots$	$\mathbf{X_{ijp}}$
1	1	d_{11}	0	t_1	X_{111}	$\cdots$	X_{11p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
i	1	d_{i1}	0	t_i	X_{i11}	$\cdots$	X_{i1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	d_{n1}	0	t_n	X_{n11}	$\cdots$	X_{n1p}

- Example from Remission Time Dataset
- Example from Study of Recurrent Bladder Cancer Tumors (Byar, 1980; Wei, Lin, and Weisfeld, 1989)
- Computer Appendix gives programming code

VI. Basic data layout for understanding analysis (pages 23–28)

A. General layout:

Ordered failure times ($t_{(f)}$)	# of failures (m_f)	# censored in $[t_{(f)}, t_{(f+1)})$ (q_{f})	Risk set $R(t_{(f)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

Note: k = # of distinct times at which subjects failed; n = # of subjects ($k \leq n$); $R(t_{(f)})$, the risk set, is the set of individuals whose survival times are at least $t_{(f)}$ or larger.

B. Example: Remission time data

Group 1 ($n = 21$, 9 failures, $k = 7$);

Group 2 ($n = 21$, 21 failures, $k = 12$)

C. How to work with censored data:

Use all information up to the time of censorship; don't throw away information.

VII. Descriptive measures of survival experience (pages 28–30)

A. Average survival time (ignoring censorship status):

$$\overline{T} = \frac{\sum_{i=1}^{n} t_i}{n}$$

$\overline{\mathrm{T}}$ underestimates the true average survival time, because censored times are included in the formula.

B. Average hazard rate:

$$\bar{h} = \frac{\#\text{ failures}}{\sum_{i=1}^{n} t_i}$$

C. Descriptive measures $\overline{T}$ and $\bar{h}$ give overall comparison; estimated survivor curves give comparison over time.

D. Estimated survivor curves are step function graphs.

E. Median survival time: graphically, proceed horizontally from 0.5 on the Y-axis until reaching graph, then vertically downward until reaching the X-axis.

VIII. **Example: Extended remission data** (pages 30–33)

A. Extended data adds log WBC to previous remission data.

B. Need to consider **confounding** and **interaction**.

C. Extended data problem: compare survival experience of two groups, after adjusting for confounding and interaction effects of log WBC.

D. Analysis alternatives:

i. stratify on log WBC and compare survival curves for different strata;

ii. use math modeling, e.g., proportional hazards model.

IX. **Multivariable example** (pages 33–35)

A. The problem: to describe the relationship between social network index (**SNI**) and time until death, controlling for **AGE**, systolic blood pressure (**SBP**), presence or absence of chronic disease (**CHR**), Quetelet's index (**QUET** – a measure of body size), and social class (**SOCL**).

B. Goals:

- to obtain an adjusted measure of effect;
- to obtain adjusted survivor curves for different SNI categories;
- to decide on variables to be adjusted.

C. The data: 13-year follow-up study (1967–1980) of a fixed cohort of $n = 170$ white males (60+) from Evans County, Georgia.

X. **Math models in survival analysis** (pages 35–37)

A. Survival analysis problem is analogous to typical multivariable problem addressed by linear and/or logistic regression modeling: describe relationship of exposure to outcome, after accounting for possible confounding and interaction.

B. Outcome variable (time to event) for survival analysis is different from linear (continuous) or logistic (dichotomous) modeling.

C. Measure of effect typically used in survival analysis: hazard ratio (**HR**).

D. Interpretation of HR: like OR. SNI study: **HR** describes relationship between SNI and *T*, after controlling for covariates.

XI. **Censoring assumptions** (pages 37–43)

A. Three different assumptions about censoring:

i. Independent (vs. Non-independent) censoring

a. most useful- concerns validity of estimated effect

ii. Random (vs. Non-random) censoring

a. more restrictive than independent censoring

iii. Non-informative (vs. Informative) censoring

a. typically affects efficiency of estimated effect

B. Examples.

Practice Exercises

True or False (Circle T or F):

T F 1. In a survival analysis, the outcome variable is dichotomous.

T F 2. In a survival analysis, the event is usually described by a (0, 1) variable.

T F 3. If the study ends before an individual has gotten the event, then his or her survival time is censored.

T F 4. If, for a given individual, the event occurs **before** the person is lost to follow-up or withdraws from the study, then this person's survival time is censored.

T F 5. $S(t) = P(T > t)$ is called the hazard function.

T F 6. The hazard function is a probability.

T F 7. Theoretically, the graph of a survivor function is a smooth curve that decreases from $S(t) = 1$ at $t = 0$ to $S(t) = 0$ at $t = \infty$.

T F 8. The survivor function at time t gives the instantaneous potential per unit time for a failure to occur, given survival up to time t.

T F 9. The formula for a hazard function involves a conditional probability as one of its components.

T F 10. The hazard function theoretically has no upper bound.

T F 11. Mathematical models for survival analysis are frequently written in terms of a hazard function.

T F 12. One goal of a survival analysis is to compare survivor and/or hazard functions.

T F 13. Ordered failure times are censored data.

T F 14. Censored data are used in the analysis of survival data up to the time interval of censorship.

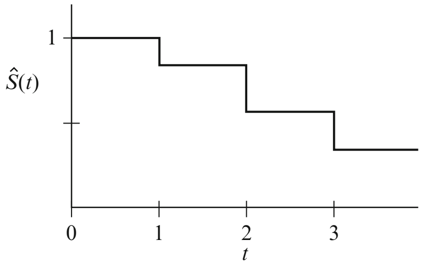
T F 15. A typical goal of a survival analysis involving several explanatory variables is to obtain an adjusted measure of effect.

16. Given the following survival time data (in weeks) for n = 15 subjects,
    1, 1, 1+, 1+, 1+, 2, 2, 2, 2+, 2+, 3, 3, 3+, 4+, 5+
    where + denotes censored data, complete the following table:

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
0	0	0	15 persons survive $\geq$ 0 weeks
1			
2			
3			

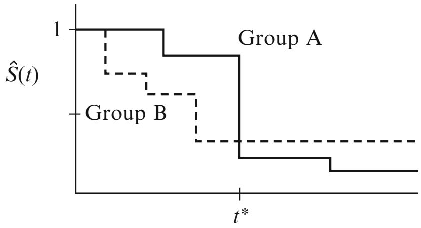
Also, compute the average survival time ($\overline{T}$) and the average hazard rate ($\overline{h}$) using the raw data (ignoring + signs for $\overline{T}$).

17. Suppose that the estimated survivor curve for the above table is given by the following graph:



What is the median survival time for this cohort?

Questions 18–20 consider the comparison of the following two survivor curves:



18. Which group has a better survival prognosis **before** time t^*?
19. Which group has a better survival prognosis **after** time t^*?
20. Which group has a longer median survival time?

Test

True or False (Circle T or F):

T F 1. Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable is **time until an event occurs**.

T F 2. In survival analysis, the term "event" is synonymous with "failure."

T F 3. If a given individual is lost to follow-up or withdraws from the study before the end of the study without the event occurring, then the survival time for this individual is said to be "censored."

T F 4. In practice, the survivor function is usually graphed as a smooth curve.

T F 5. The survivor function ranges between 0 and ∞.

T F 6. The concept of instantaneous potential is illustrated by velocity.

T F 7. A hazard rate of one per day is equivalent to seven per week.

T F 8. If you know the form of a hazard function, then you can determine the corresponding survivor curve, and vice versa.

T F 9. One use of a hazard function is to gain insight about conditional failure rates.

T F 10. If the survival curve for group 1 lies completely above the survival curve for group 2, then the median survival time for group 2 is longer than that for group 1.

T F 11. The risk set at 6 weeks is the set of individuals whose survival times are less than or equal to 6 weeks.

T F 12. If the risk set at 6 weeks consists of 22 persons, and 4 persons fail and 3 persons are censored by the 7th week, then the risk set at 7 weeks consists of 18 persons.

T F 13. The measure of effect used in survival analysis is an odds ratio.

T F 14. If a hazard ratio comparing group 1 relative to group 2 equals 10, then the potential for failure is ten times higher in group 1 than in group 2.

T F 15. The outcome variable used in a survival analysis is different from that used in linear or logistic modeling.

16. State two properties of a hazard function.

17. State three reasons why hazard functions are used.

18. State three goals of a survival analysis.

19. The following data are a sample from the 1967–1980 Evans County study. Survival times (in years) are given for two study groups, each with 25 participants. Group 1 has no history of chronic disease (CHR = 0), and group 2 has a positive history of chronic disease (CHR = 1):

Group 1 (CHR = 0):	12.3+, 5.4, 8.2, 12.2+, 11.7, 10.0, 5.7, 9.8, 2.6, 11.0, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3.0
Group 2 (CHR = 1):	5.8, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 1.8, 3.1, 11.4, 2.4, 1.4, 5.9, 1.6, 2.8, 4.9, 3.5, 6.5, 9.9, 3.6, 5.2, 8.8, 7.8, 4.7, 3.9

For group 1, complete the following table involving ordered failure times:

	$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
Group 1:	0.0	0	0	25 persons survived $\geq$ 0 years
	1.8	1	0	25 persons survived $\geq$ 1.8 years
	2.2			
	2.5			
	2.6			
	3.0			
	3.5			
	3.8			
	5.3			
	5.4			
	5.7			
	6.6			
	8.2			
	8.7			
	9.2			
	9.8			
	10.0			
	10.2			
	10.7			
	11.0			
	11.1			
	11.7			

20. For the data of Problem 19, the average survival time ($\overline{T}$) and the average hazard rate ($\overline{h}$) for each group are given as follows:

	$\overline{T}$	$\overline{h}$
Group 1:	7.5	.1165
Group 2:	5.3	.1894

a. Based on the above information, which group has a better survival prognosis? Explain briefly.

b. How would a comparison of survivor curves provide additional information to what is provided in the above table?

Answers to Practice Exercises

1. F: the outcome is continuous; time until an event occurs.
2. T
3. T
4. F: the person fails, i.e., is not censored.
5. F: $S(t)$ is the survivor function.
6. F: the hazard is a rate, not a probability.
7. T
8. F: the hazard function gives instantaneous potential.
9. T
10. T
11. T
12. T
13. F: ordered failure times are data for persons who are failures.
14. T
15. T
16.

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
0	0	0	15 persons survive ≥ 0 weeks
1	2	3	15 persons survive ≥ 1 weeks
2	3	2	10 persons survive ≥ 2 weeks
3	2	3	5 persons survive ≥ 3 weeks

$$\overline{T} = \frac{33}{15} = 2.2, ; \quad \overline{h} = \frac{7}{33} = 0.22$$

17. Median = 3 weeks
18. Group A
19. Group B
20. Group A