

SURVIVAL ANALYSIS

Chapter 1. Introduction to Survival Analysis

Instructor: Seokho Lee

2012Fall at HUFS

What is Survival Analysis?

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



Event: death
disease
relapse
recovery

Assume 1 event



Time \equiv survival time

Event \equiv failure

- 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.
- 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.
- 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.
- 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 events or a **competing risk** problem, which are discussed in later chapter, if time is allowed for us.
- 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. 2

What is Survival Analysis?

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)

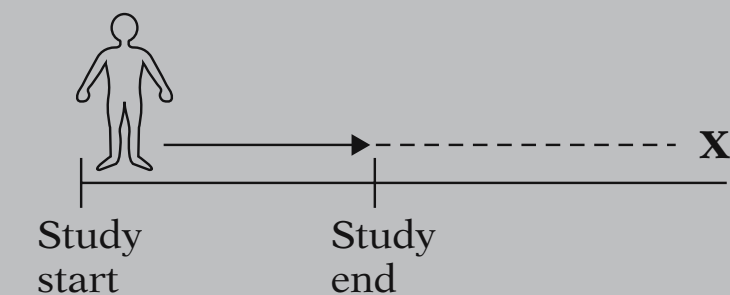
- Five examples of survival analysis problems are briefly mentioned here. The first is a study follows leukemia patients in remission over several weeks to see how long they stay in remission. The second example follows a disease-free cohort of individuals over several years to see who develops heart 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 outcome becoming “time until death (in months from receiving a transplant).”

Censored Data

Censoring: don't know survival time exactly

EXAMPLE

Leukemia patients in remission:

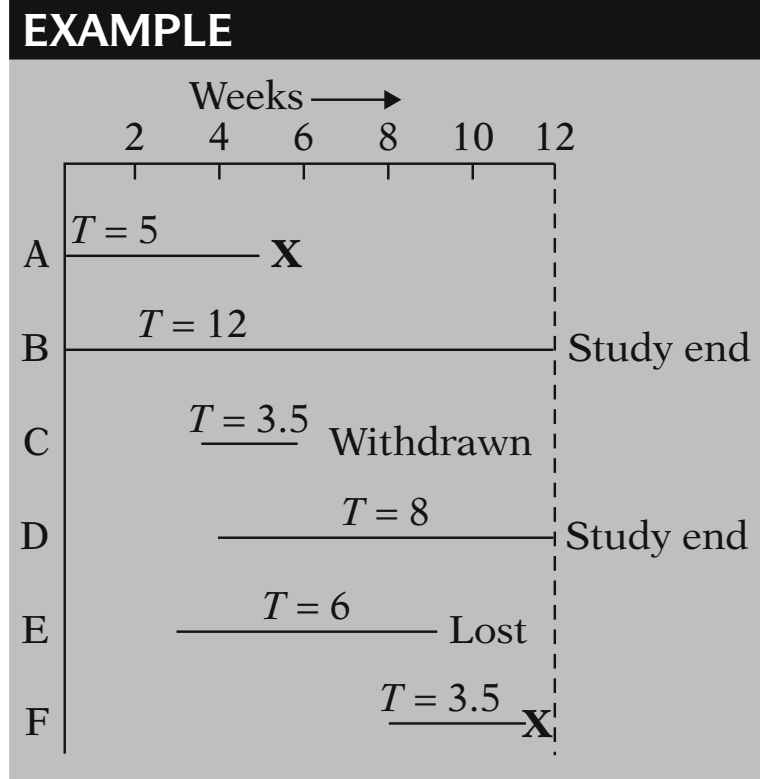


Why censor?

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

- 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**.
- As a simple example of censoring, consider leukemia patients followed until they go out of remission, shown here as **X**. If for a given patient, the study ends while the patient is still in remission (i.e., doesn't get the event), then that patient's survival time is considered censored. We know that, for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission after the study ends, we do not know the complete survival time.
- There are generally three reasons why censoring may occur:
 - (1) a person does not experience the event before **the study ends**;
 - (2) a person is **lost to follow-up** during the study period;
 - (3) a person **withdraws from the study** because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk)

Censored Data



- 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 third 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

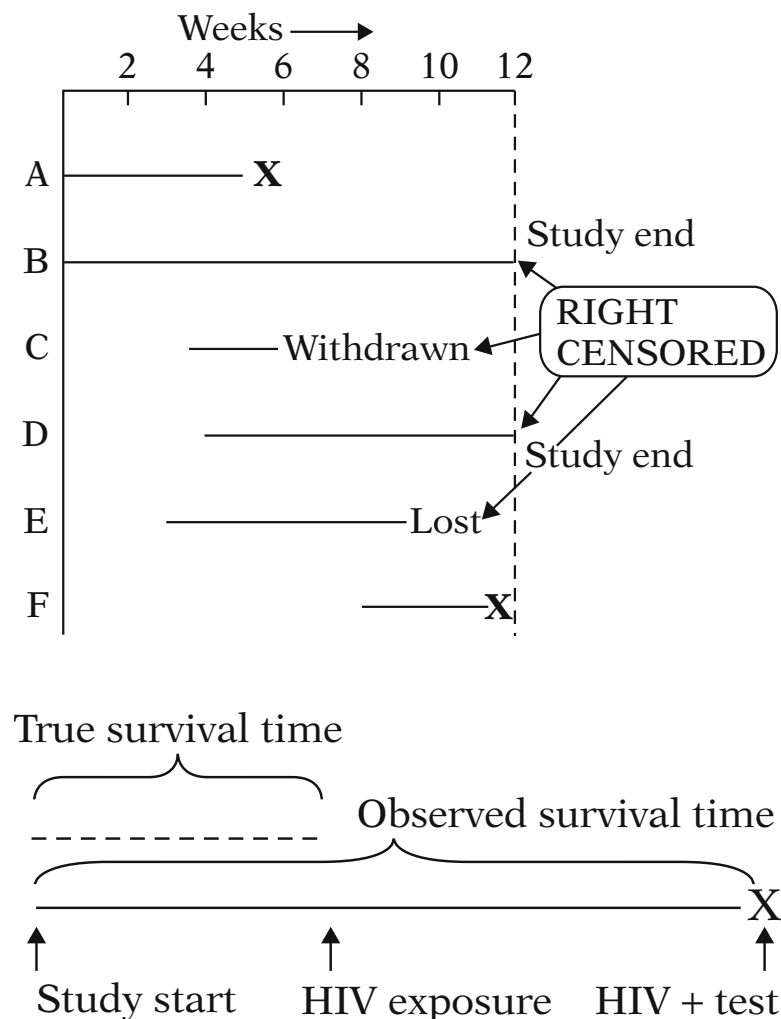
Censored: B, C, D, E

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

Censored Data

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



Terminology and Notation

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

← random variable

t = specific value for T

EXAMPLE

Survives > 5 years?

$T > t = 5$

$\delta = (0, 1)$ random variable

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

- study ends
- lost to follow-up
- withdraws

$S(t)$ = survivor function

$h(t)$ = hazard function

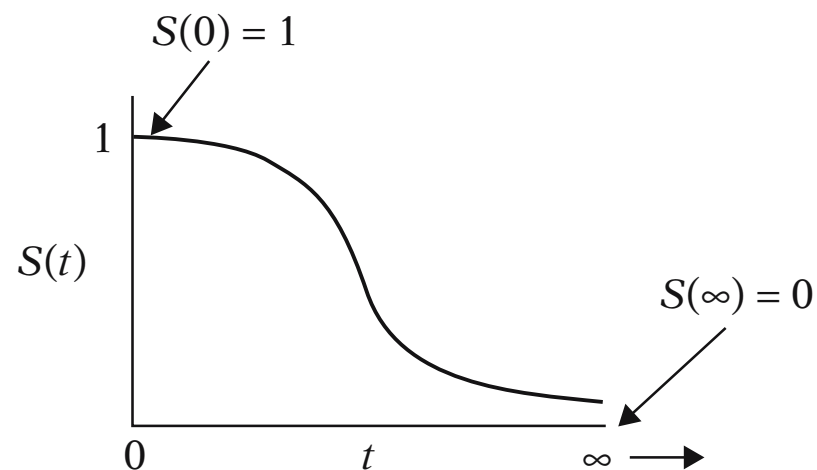
- We are now ready to introduce basic mathematical terminology and notation for survival analysis. First, we denote by a **capital T** the random variable for a person's survival time. Since T denotes time, its possible values include all nonnegative numbers; that is, T can be any number equal to or greater than zero.
- 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.
- Finally, we let the Greek letter delta (δ) denote a (0,1) random variable indicating either failure or censorship. That is, $\delta = 1$ for failure if the event occurs during the study period, or $\delta = 0$ if the survival time is censored by the end of the study period. Note that if a person does not fail, that is, does not get the event during the study period, censorship is the **only** remaining possibility for that person's survival time. That is, $\delta = 0$ if and only if one of the following happens: a person survives until the study ends, a person is lost to follow-up, or a person withdraws during the study period.
- 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 $h(t)$.

Terminology and Notation

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

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

Theoretical $S(t)$:

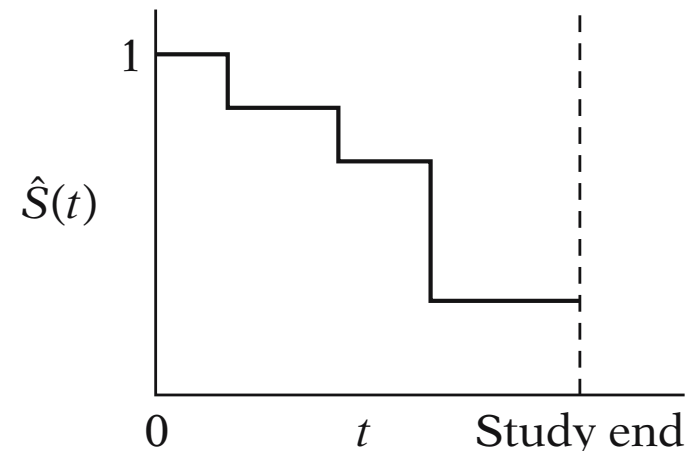



- The survivor function $S(t)$ gives the probability that a person survives longer than some specified time t : that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t .
- The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values of t provides crucial summary information from survival data.
- Theoretically, as t ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph, 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 that survivor curve must eventually fall to zero.

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

Terminology and Notation

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



Given 

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

Conditional probabilities: $P(A|B)$

$$P(t \leq T < t + \Delta t \mid T \geq t)$$

$$= \text{P(individual fails in the interval } [t, t + \Delta t] \mid \text{survival up to time } t)$$

- 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.
- The hazard function, denoted by $h(t)$, is given by the formula: $h(t)$ equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt , where Δt denotes a small interval of time. This mathematical formula is difficult to explain in practical terms.
- In mathematical terms, the given part of the formula for the hazard function is found in the probability statement-the numerator to the right of the limit sign. This statement is a conditional probability because it is of the form, “ P of A , given B ,” where the P denotes probability and where the long vertical line separating A from B denotes “given.” In the hazard formula, the conditional probability gives the probability that 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**.

Terminology and Notation

Hazard function \equiv conditional failure **rate**

$$\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Probability per unit time

Rate: 0 to ∞

$$P = P(t \leq T < t + \Delta t \mid T \geq t)$$

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

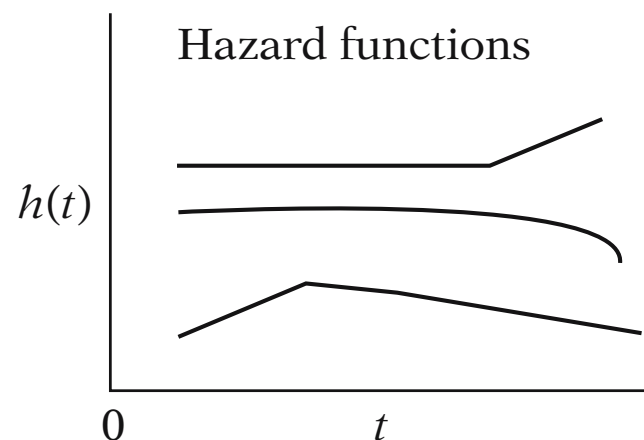
- We now explain why the hazard is a **rate** rather than a probability. Note that in the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities. The numerator is the conditional probability we just discussed. the denominator is Δt , which denotes a small 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.

- 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.**

Terminology and Notation

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

Gives instantaneous potential



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

- 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 specific value of t , the hazard function $h(t)$ has the following characteristics:
 - 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 ∞ .

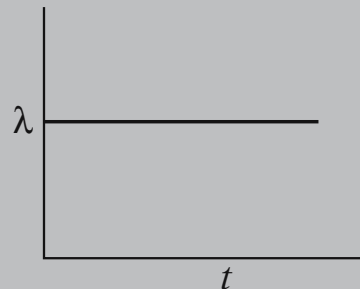
Terminology and Notation

EXAMPLE

①

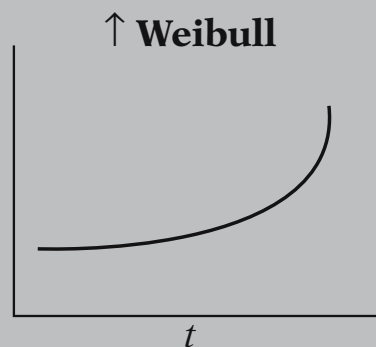
Constant hazard
(**exponential model**)

$h(t)$ for healthy
persons



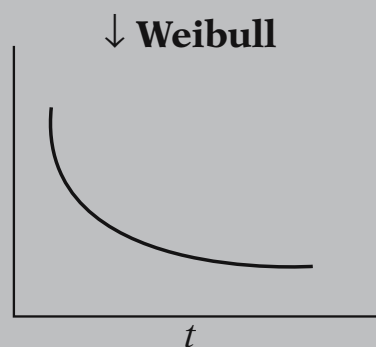
②

$h(t)$ for leukemia
patients



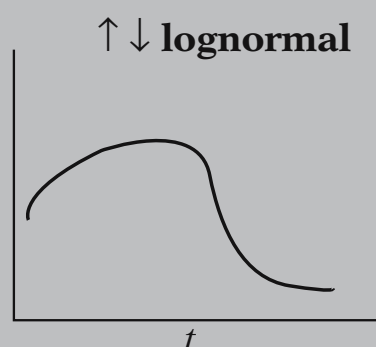
③

$h(t)$ for persons
recovering from
surgery



④

$h(t)$ for TB
patients



- Now we show some of different types of hazard functions. The first graph shows a constant hazard for a study of healthy persons. In this graphs, 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.
- 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 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.

Terminology and Notation

$S(t)$: directly describes survival

$h(t)$: • a measure of instantaneous potential

- identify specific model form
- math model for survival analysis

Relationship of $S(t)$ and $h(t)$:

If you know one, you can determine the other.

EXAMPLE

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

- 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.
- 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 .

Terminology and Notation

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 $S(t)$.
- 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.

Terminology and Notation

SUMMARY

T = survival time random variable

t = specific value of T

δ = (0, 1) variable for failure/censorship

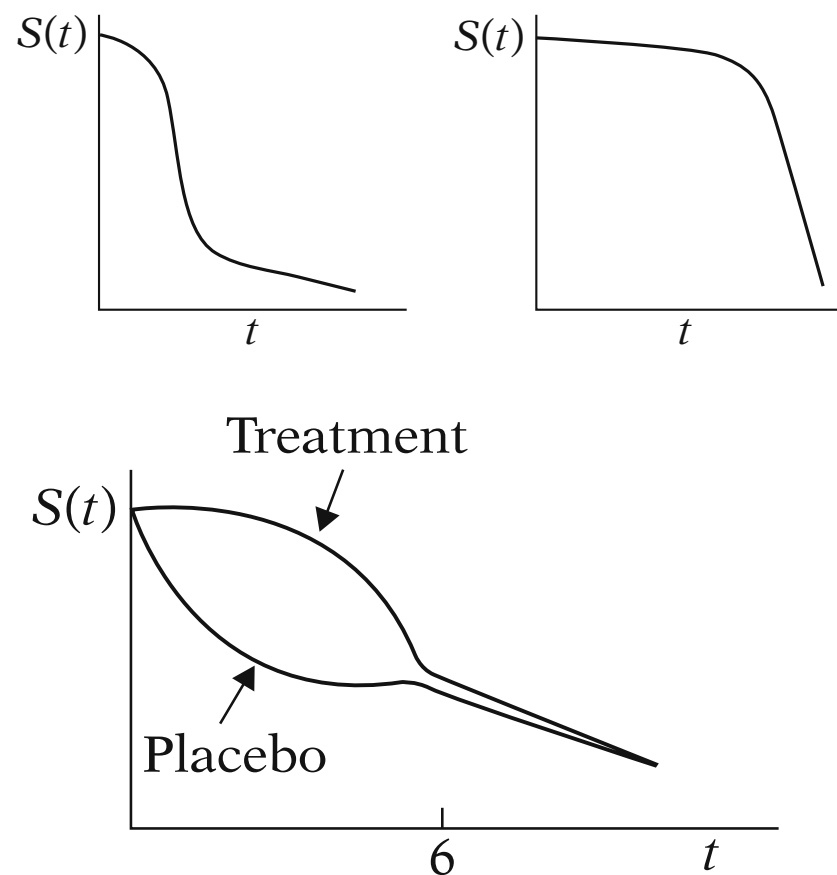
$S(t)$ = survivor function

$h(t)$ = 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 δ 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.**

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.



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

- 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 usually requires using some form of mathematical modeling, for example, the Cox proportional hazard approach, which will be the subject of subsequent chapter.

Basic Data Layout for Computer

Two types of data layouts:

- for computer use
- for understanding

Indiv. #	t	δ	X_1	X_2	\cdots	X_p
1	t_1	δ_1	X_{11}	X_{12}	\cdots	X_{1p}
2	t_2	δ_2	X_{21}	X_{22}	\cdots	X_{2p}
•						•
•						•
•						•
5	$t_5 = 3$ got event					
•						•
•						•
•						•
8	$t_8 = 3$ censored					
•						•
•						•
•						•
n	t_n	δ_n	X_{n1}	X_{n2}	\cdots	X_{np}

- 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.
- 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.

Basic Data Layout for Computer

Indiv. #	t	Failure status	Explanatory variables			
		δ	X_1	X_2	\cdots	X_p
1	t_1	δ_1	X_{11}	X_{12}	\cdots	X_{1p}
2	t_2	δ_2	X_{21}	X_{22}	\cdots	X_{2p}
.						.
.						.
.						.
5	$t_5 = 3$	$\delta_5 = 1$				
.						.
.						.
.						.
8	$t_8 = 3$	$\delta_8 = 0$				
.						.
.						.
.						.
n	t_n	δ_n	X_{n1}	X_{n2}	\cdots	X_{np}

$X_i = \text{Age}, E, \text{ or } \text{Age} \times \text{Race}$

- 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. δ) the dichotomous variable that indicates censorship status.
- Thus, δ_1 is 1 if person 1 gets the event or is 0 if person 1 is censored; δ_2 is 1 or 0 similarly, and so on, up through δ_n . In the example just considered, person 5, who failed at 3 weeks, has a δ of 1; that is, δ_5 equals 1. In contrast, person 8, who was censored at 3 weeks, has a δ of 0; that is, δ_8 equals 0.
- Note that if all of the δ_j 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_i , is any variable like age or exposure status, E , or a product term like $\text{age} \times \text{race}$ 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.

Basic Data Layout for Computer

		Columns					
		#	t	δ	X_1	X_2	$\cdots X_p$
Rows	1	t_1	δ_1	X_{11}	X_{12}	\cdots	X_{1p}
	2	t_2	δ_2	X_{21}	X_{22}	\cdots	X_{2p}
	\vdots						
	\vdots						
	\vdots						
	j	t_j	δ_j	X_{j1}	X_{j2}	\cdots	X_{jp}
	\vdots						
	n	t_n	δ_n	X_{n1}	X_{n2}	\cdots	X_{np}

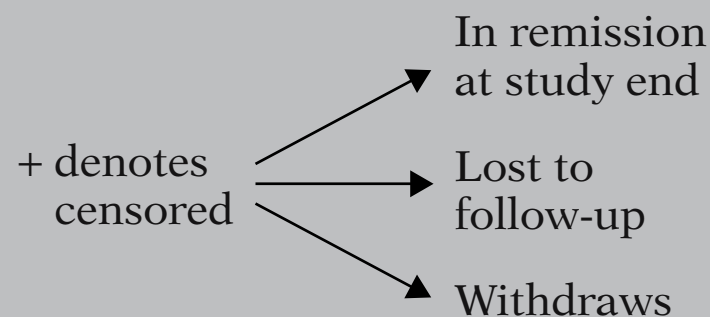
- For example, in the column corresponding to X_1 are the values observed on this variable for all n persons. These values are denoted as X_{11} , X_{21} , and so on, up to X_{n1} ; the first subscript indicates the person number, and the second subscript, a one in each case here, indicates the variable 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 j , the observed information is given by the values t_j , δ_j , X_{j1} , X_{j2} , etc., up to X_{jp} . This is how the information is read into the computer, that is, line by line, until all persons are included for analysis.

Basic Data Layout for Computer

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 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,	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+	

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

	Indiv. #	t (weeks)	δ (failed or censored)	X (Group)
GROUP 1	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

Layout for Computer

- Here are the data again: notice that the first three persons in group 1 went out of remission at 6 weeks; the next six 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 six weeks. The remaining persons in group one 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 put this data in tabular form for the computer, as shown at the left. The list starts with 21 persons in group 1 (listed 1-21) and follow (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 $\delta=1$, this person failed, that is, went out of remission. The X value is 1 because person #3 is in group1. As a second example, person #14, who has an observed survival time of 17 weeks, was censored at this time because $\delta=0$. The X value is again 1 because person #14 is also in group 1.

Basic Data Layout for Computer

EXAMPLE (continued)

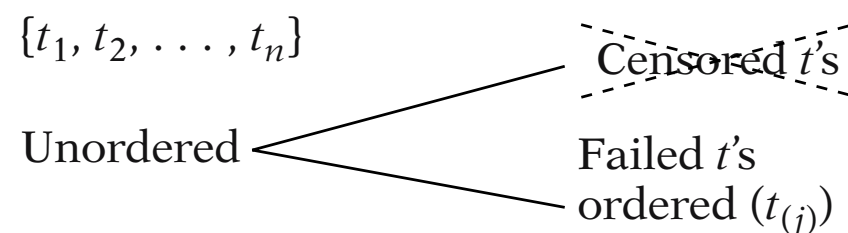
	Indiv. #	t (weeks)	δ (failed or censored)	X (Group)
	22	1	1	0
	23	1	1	0
	24	2	1	0
	25	2	1	0
	26	3	1	0
	27	4	1	0
GROUP	28	4	1	0
2	29	5	1	0
	30	5	1	0
	31	8	1	0
	32	8	1	0
	33	8	1	0
	34	8	1	0
	35	11	1	0
	36	11	1	0
	37	12	1	0
	38	12	1	0
	39	15	1	0
	40	17	1	0
	41	22	1	0
	42	23	1	0

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

Basic Data Layout for Understanding Analysis

For analysis:

Ordered failure times ($t_{(j)}$)	# of failures (m_j)	# censored in $[t_{(j)}, t_{(j+1)})$ (q_j)	Risk set $R(t_{(j)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
\bullet	\bullet	\bullet	\bullet
\bullet	\bullet	\bullet	\bullet
\bullet	\bullet	\bullet	\bullet
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$



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

- 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.
- The first column in this table gives ordered failure times. These are denoted by t 's with subscripts within parentheses, starting $t_{(0)}$, 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.
- 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.

Basic Data Layout for Understanding Analysis

EXAMPLE

Remission Data: Group 1

($n = 21$, 9 failures, $k = 7$)

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	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, $k = 12$)

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 1$	2	0	21 persons survive ≥ 1 wk
$t_{(2)} = 2$	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	

- For example, using the remission data for group 1, we find that nine of the 21 persons failed, including three persons at 6 weeks and one 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_j , of those persons who failed at each distinct failure time. When there are no ties at a certain failure time, then $m_j=1$. 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_j '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

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	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 1

#	$t(\text{weeks})$	δ	$X(\text{group})$
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

LAYOUT for Understanding Analysis

- The *third column* gives frequency counts, denoted by q_j , of those persons censored in the time interval starting with failure time $t_{(j)}$ up to the next failure time denoted by $t_{(j+1)}$. 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_j '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 here person #12, who was censored at week 10, because this person's censored time is exactly at the end of the interval. We count this person in the following interval.

Basic Data Layout for Understanding Analysis

EXAMPLE (continued)

Group 1 using ordered failure times

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	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	

- 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 two 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.
- The last column in the table gives the “**risk set.**” The risk set is not a numerical or count but rather a collection of individuals. By definition, the risk set $R(t_{(j)})$ is the collection of individuals who have survived at least to time $t_{(j)}$; that is, each person in $R(t_{(j)})$ has a survival time that is $t_{(j)}$ 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

Risk Set: $R(t_{(j)})$ is the set of individuals for whom $T \geq t_{(j)}$.

Remission Data: Group 1

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = \textcircled{0}$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = \textcircled{6}$	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	

Basic Data Layout for Understanding Analysis

EXAMPLE (continued)

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = \textcircled{7}$	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	

$t_{(0)} = 0$	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
$t_{(4)} = \textcircled{13}$	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 seventeen 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 nine persons from 21, leaves twelve persons in group 1 still at risk for getting the event at 13 weeks. Thus, the risk set consists of these twelve persons.

Basic Data Layout for Understanding Analysis

How we work with censored data:

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

EXAMPLE			
$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
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

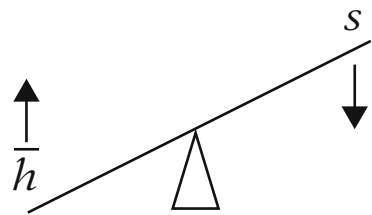
- The importance of the table of ordered failure times is that we can work with censored observations in analyzing survival data. Even though censored observations are incomplete, in that we don't know a person's survival time exactly, we can still make use of the information we have on a censored person 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 until time of censorship. (Nevertheless, most survival analysis techniques require a key assumption that censoring is **non-informative**—censored subjects are not at increased risk for failure.)
- 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.

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
\bar{T}_1 (ignoring + 's) = 17.1	$\bar{T}_2 = 8.6$
$\bar{h}_1 = \frac{9}{359} = .025$	$\bar{h}_2 = \frac{21}{182} = .115$
<div> Average hazard rate $(\bar{h}) = \frac{\text{\# failures}}{\sum_{i=1}^n t_i}$ </div>	

- 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.

Descriptive Measures of Survival Experience

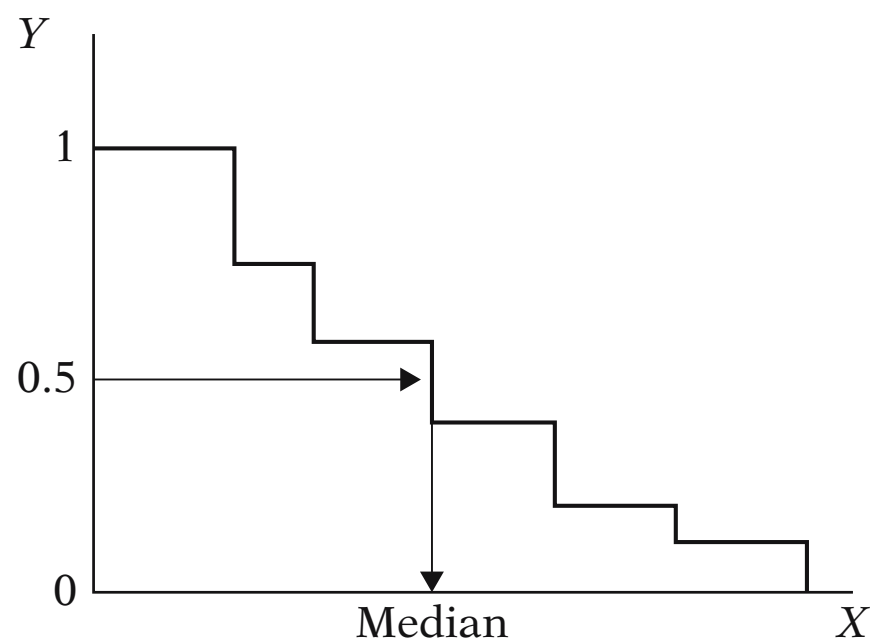
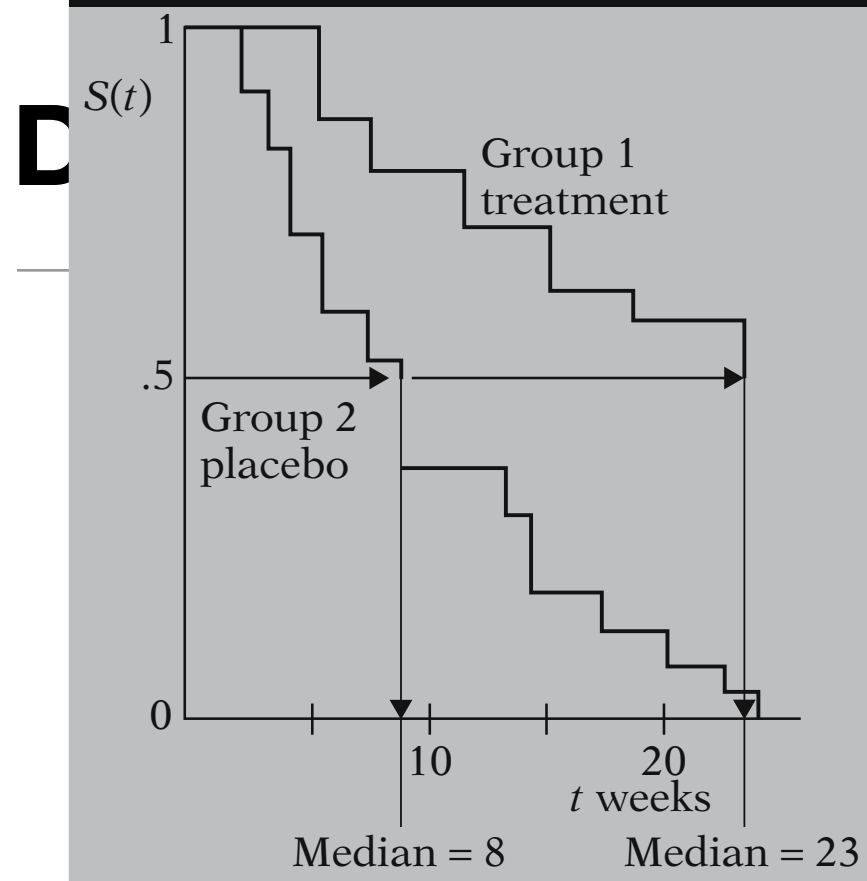


Placebo hazard > treatment hazard:
suggests that treatment is more
effective than placebo

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

- 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.
- 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.
- 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



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

asures of Survival Experience

- 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 Chapter 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.
- 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.

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 (WBC) for each person studied. In particular, each person's log white blood cell count (log WBC) 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: Extended Remission Data

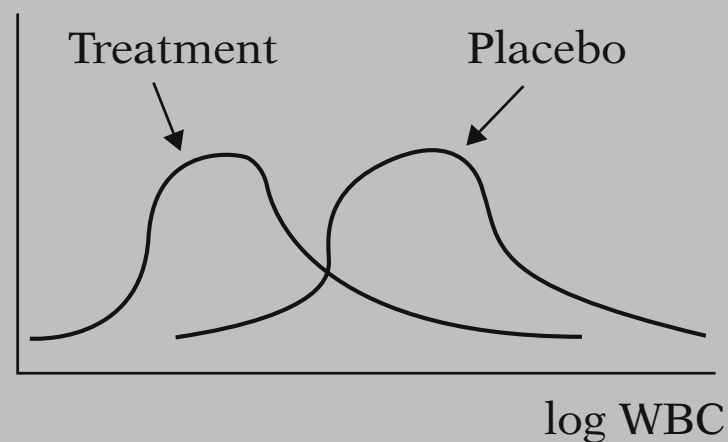
EXAMPLE: CONFOUNDING

Treatment group: $\overline{\log \text{WBC}} = 1.8$

Placebo group: $\overline{\log \text{WBC}} = 4.1$

Indicates **confounding** of treatment effect by log WBC

Frequency distribution

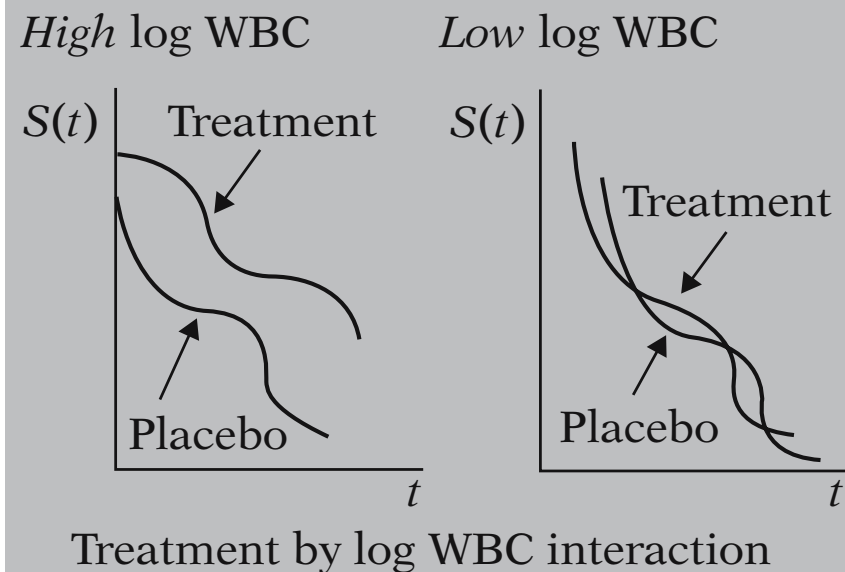


Need to adjust for imbalance in the distribution of log WBC

- 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**.
- 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: Extended Remission Data

EXAMPLE: 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 problem:

Compare two groups after adjusting for confounding and interaction.

- 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.
- 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)	δ	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

Extended Remission Data

- 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.
- 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.

Multivariable Example

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

EXAMPLE

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

$n = 170$ white males (60+)

T = years until death

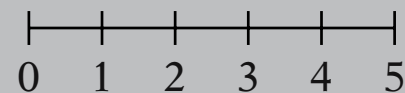
Event = death

Explanatory variables:

- exposure variable
- confounders
- interaction variables

Exposure:

Social Network Index (**SNI**)



Absence
of social
network

Excellent
social
network

- 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.
- 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).

Multivariable Example

EXAMPLE (continued)

Study goal: to determine whether **SNI** is protective against death, i.e., **SNI** $\nearrow \Rightarrow S(t) \nearrow$

Explanatory variables:

SNI	}	Exposure variable
AGE		
SBP	}	Potential confounders/ interaction variables
CHR		
QUET		
SOCL		

Note : $QUET = \frac{\text{weight}}{(\text{height})^2} \times 100$

The problem:

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

- 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.
- 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.

Multivariable Example

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.

Multivariable Example

- 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	δ	SNI	AGE	SBP	CHR	QUET	SOCL
1	t_1	δ_1	SNI_1	AGE_1	SBP_1	CHR_1	QUET_1	SOCL_1
2	t_2	δ_2	SNI_2	AGE_2	SBP_2	CHR_2	QUET_2	SOCL_2
.
.
.
170	t_{170}	δ_{170}	SNI_{170}	AGE_{170}	SBP_{170}	CHR_{170}	QUET_{170}	SOCL_{170}

Math Models in Survival Analysis

General framework



Controlling for $C_1, C_2, \dots C_p$.

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.

follow-up time info not used {	Model	Outcome
	Survival analysis	Time to event (with censoring)
	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.

Math Models in Survival Analysis

Measure of effect:

Linear regression:

regression coefficient β

Logistic regression:

odds ratio e^β

Survival analysis:

hazard ratio e^β

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

Math Models in Survival Analysis

EXAMPLE

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

Interpretation of HR (like OR):

$HR = 1 \Rightarrow$ no relationship

$HR = 10 \Rightarrow$ exposed hazard
10 times unexposed

$HR = 1/10 \Rightarrow$ exposed hazard
1/10 times unexposed

- 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.
- 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.

R Implementation for Data Layout

```
install.packages('survival') # install survival package
```

```
library(survival) # loading survival package
```

```
# survival time
```

```
t<-c(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)
```

```
# censorship index
```

```
delta<-c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
```

```
# independent variable: group information
```

```
x1<-rep(c(0,1),c(21,21))
```

```
# covariate: log(WBC)
```

```
x2<-c(2.31, 4.06, 3.28, 4.43, 2.96, 2.88, 3.60, 2.32, 2.57, 3.20, 2.80, 2.70, 2.60, 2.16, 2.06,  
2.01, 1.78, 2.20, 2.53, 1.47, 1.45, 2.80, 5.00, 4.91, 4.48, 4.01, 4.36, 2.42, 3.49, 3.97, 3.52,  
3.05, 2.32, 3.26, 3.49, 2.12, 1.50, 3.06, 2.30, 2.95, 2.73, 1.97)
```

```
# create 'survival' object using the function Surv()
```

```
remission<-Surv(t,delta)
```

```
remission
```

```
remission[x1==0] # for group 1
```

```
remission[x1==1] # for group 2
```


R Implementation for Data Layout

```
> remission<-Surv(t,delta)
> remission
 [1]  6   6   6   7  10  13  16  22  23   6+  9+ 10+ 11+ 17+ 19+ 20+ 25+
[18] 32+ 32+ 34+ 35+  1   1   2   2   3   4   4   5   5   8   8   8   8
[35] 11  11  12  12  15  17  22  23
> remission[x1==0] # for treatment group
 [1]  6   6   6   7  10  13  16  22  23   6+  9+ 10+ 11+ 17+ 19+ 20+ 25+
[18] 32+ 32+ 34+ 35+
> remission[x1==1] # for placebo group
 [1]  1   1   2   2   3   4   4   5   5   8   8   8   8  11  11  12  12
[18] 15  17  22  23
```

Math Models in Survival Analysis

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 Chapter 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.