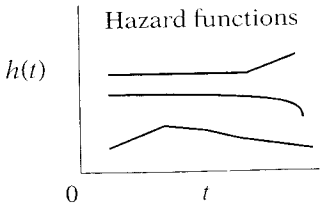
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Gives instantaneous potential



- $h(t) \ge 0$
- $h(t)$ 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.

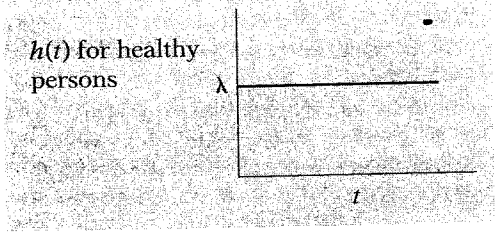
As with a survivor function, the hazard function $h(t)$ can be graphed as t ranges over various values. The graph at the left illustrates three different hazards. In contrast to a survivor function, the graph of $h(t)$ does not have to start at 1 and go down to zero, but rather can start anywhere and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

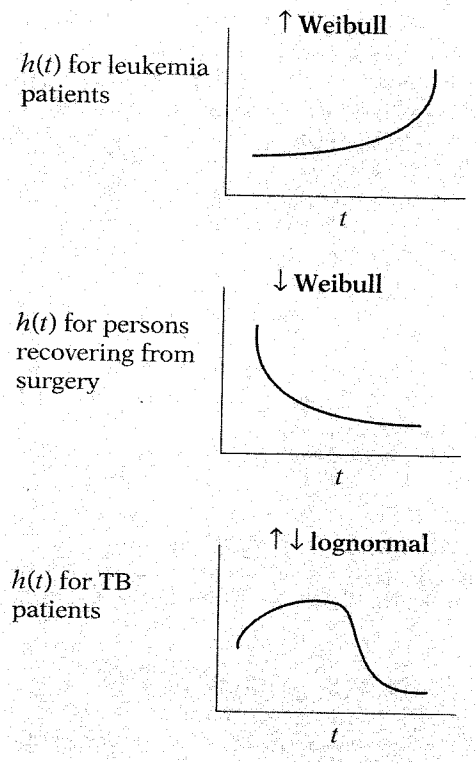
EXAMPLE

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 no matter what time is picked. 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 hazard function illustrated shows a graph that is increasing. 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.

The third hazard function illustrated shows a graph that is decreasing. 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 hazard function given shows a graph that is first increasing and then decreasing. This type of graph is called a **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

$h(t)$:
- insight about conditional failure rates
- 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 provides insight about conditional failure rates;
- 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.

EXAMPLE

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

General formulae:

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

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

Formulae not important:

$S(t)$ [redacted] $h(t)$

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.

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 $S(t)$.

The actual formulae are not important, because 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
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.

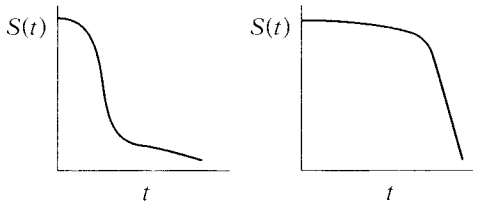
IV. Goals of Survival Analysis

We now state the basic goals of survival analyses.

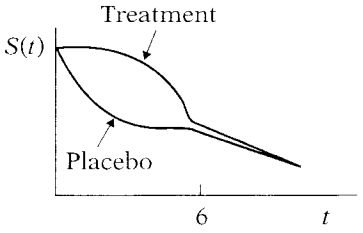
Goal 1: To estimate and interpret survivor and/or hazard functions from survival data.

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 graph for the treatment group lies above that for the placebo group, but thereafter the two graphs are at about the same level. This dual graph indicates that up to 6 weeks the treatment is more effective 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 modules.

V. Basic Data Layout for Computer

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	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}	$X_{n2} \cdots X_{np}$

		Failure status	Explanatory variables	
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	$t_5 = 3$	$\delta_5 = 1$		
•				
•		$\sum_1^n \delta_i$ = # failures		
•				
8	$t_8 = 3$	$\delta_8 = 0$		
•				
•				
•				
n	t_n	δ_n	X_{n1}	$X_{n2} \cdots X_{np}$

X_i = Age, E, or Age × Race

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.

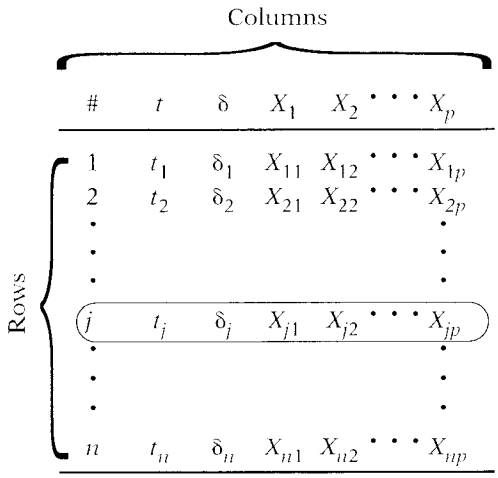
The remaining columns after the first one provide survival time and other information for each person. The second column gives the survival time information, which is denoted t_1 for individual 1, t_2 for individual 2, and so on, up to t_n for individual n. Each of these t's gives the observed survival 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 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_i 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 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.



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.

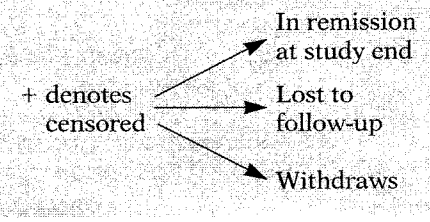
We have thus described the basic data layout by columns. Alternatively, we can look at the table line by line, that is, by rows. For each line or row, we have the information obtained on a given individual. Thus, for individual j, the observed information is given by the values t_j, δ_j, X_{j1}, X_{j2}, 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.

EXAMPLE

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.

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

Group 1 (Treatment) $n = 21$	Group 2 (Treatment) $n = 21$
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 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 (Treatment) $n = 21$
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)
	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

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 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 because $\delta = 1$, which means that 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 a 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.

EXAMPLE (continued)

	Indiv. #	t (weeks)	δ (failed or censored)	X (Group)
	20	34	0	1
	21	35	0	1
	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

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.

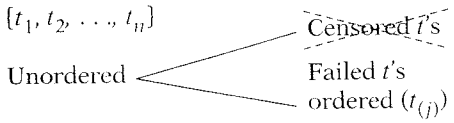
VI. Basic Data Layout for Understanding Analysis

For analysis:

Ordered failure $(t_{(j)})$	# 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	$R(t_{(2)})$
•	•	•	•
•	•	•	•
•	•	•	•
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

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, t_1, and so on, up to t_k by $t_{(0)}$, $t_{(1)}$ 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 which 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)

$t_{(j)}$	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	

Ties

For example, using the remission data for group 1, we find that nine of the 21 persons failed, including three persons each 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, shown at the left, 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, shown at the bottom left, 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
$t_{(2)} = 7$ (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)	δ	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_j, of those persons censored in the time interval starting with failure time $t_{(j)}$ up to the next failure time denoted $t_{(j+1)}$. Technically, because of the way we have defined this interval in the table, we include those persons censored at the beginning of the interval but not at its end.

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 in parentheses, 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 at 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.