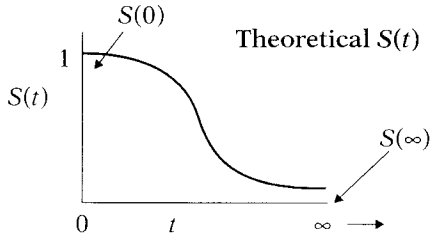
δ = (0,1) random variable
$$= \begin{cases} 1 \text{ if failure} \\ 0 \text{ if censored} \end{cases}$$

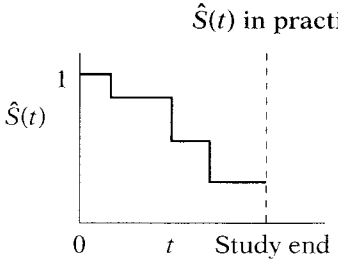
We let the Greek letter delta (δ) denote a (0,1) random variable indicating either censorship or failure. A person who does not fail, that is, does not get the event during the study period, must have been censored either before or at the end of the study.

$S(t)$ = survivor function
= PR $(T > t)$

The survivor function, denoted by ***S(t),*** gives the probability that the random variable T exceeds the specified time t.



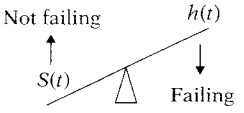
Theoretically, as t ranges from 0 up to infinity, the survivor function is graphed as a decreasing smooth curve, which begins at $S(t) = 1$ at $t = 0$ and heads downward toward zero as t increases toward infinity.



In practice, using data, we usually obtain survivor function graphs which are **step functions,** as illustrated here, rather than smooth curves.

$h(t)$ = hazard functon
= instantaneous potential given survival up to time t

The hazard function, denoted by ***h(t),*** gives the **instantaneous potential** per unit time for the event to occur given that the individual has survived up to time t.



In contrast to the survivor function, which focuses on **not** failing, the hazard function focuses on failing; in other words, as $S(t)$ goes up, $h(t)$ goes down, and vice versa. The hazard is a **rate,** rather than a probability. Thus, the values of the hazard function range between zero and infinity.

$h(t)$ is a rate: 0 to ∞

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.

Data Layout:

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}	$X_{22} \cdots X_{2p}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	t_n	δ_n	X_{n1}	$X_{n2} \cdots X_{np}$

The general data layout for a survival analysis is given by the table shown here. The first column of the table identifies the study subjects. The second column gives the observed survival time information. The third column gives the information for δ, the dichotomous variable that indicates censorship status. The remainder of the information in the table gives values for explanatory variables of interest.

Alternative (ordered) data layout:

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	$R(t_{(2)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

An alternative data layout is shown here. This layout is the basis upon which **Kaplan–Meier** survival curves are derived. The first column in the table gives ordered survival times from smallest to largest. The second column gives frequency counts of failures at each distinct failure time. 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 but not including the next failure time denoted $t_{(j+1)}$. The last column gives the **risk set,** which denotes the collection of individuals who have survived at least to time $t_{(j)}$.

Table of ordered failures:

- Uses all information up to time of censorship;
- $S(t)$ is derived from $R(t)$.

To compute the survival probability at a given time, we make use of the risk set at that time to include the information we have on a censored person up to the time of censorship, rather than simply throw away all the information on a censored person.

Survival probability:

Use **Kaplan–Meier (KM)** method.

The actual computation of such a survival probability can be carried out using the Kaplan–Meier (KM) method. We introduce the KM method in the next section by way of an example.

II. An Example of Kaplan–Meier Curves

EXAMPLE

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

Group 1 ($n = 21$) treatment	Group 2 ($n = 21$) placebo
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+	

Note: + denotes censored

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

Descriptive statistics:

$\bar{t}_1$ (ignoring +'s) = 17.1, $\bar{t}_2$ = 8.6

$\bar{h}_1 = .025$, $\bar{h}_2 = .115$, $\frac{\bar{h}_2}{\bar{h}_1} = 4.6$

The data for this example derive from a study of the remission times in weeks for two groups of leukemia patients, with 21 patients in each group. Group 1 is the treatment group and group 2 is the placebo group. The basic question of interest concerns comparing the survival experience of the two groups.

Of the 21 persons in group 1, 9 failed during the study period and 12 were censored. In contrast, none of the data in group 2 are censored; that is, all 21 persons in the placebo group went out of remission during the study period.

In Chapter 1, we observed for this data set that group 1 appears to have better survival prognosis than group 2, indicating that the treatment is effective. This conclusion was supported by descriptive statistics for the average survival time and average hazard rate shown here. Note, however, that descriptive statistics provide overall comparisons but do not compare the two groups at different times of follow-up.

EXAMPLE (continued)

Ordered failure times:

Group 1 (treatment)

j	n_j	m_j	q_j
0	21	0	0
5	21	3	1
7	17	1	1
10	15	1	2
13	12	1	0
16	11	1	3
22	7	1	0
23	6	1	5
	—	—	—

Group 2 (placebo)

j	n_j	m_j	q_j
0	21	0	0
1	21	2	0
2	19	2	0
3	17	1	0
4	16	2	0
5	14	2	0
6	12	4	0
11	8	2	0
12	6	2	0
15	4	1	0
17	3	1	0
22	2	1	0
23	1	1	0

Group 2: no censored subjects

A table of ordered failure times is shown here for each group. These tables provide the basic information for the computation of KM curves.

Each table begins with a survival time of zero, even though no subject actually failed at the start of follow-up. The reason for the zero is to allow for the possibility that some subjects might have been censored before the earliest failure time.

Also, each table contains a column denoted as n_j that gives the number of subjects in the risk set at the start of the interval. It is assumed that n_j includes those persons failing at time $t_{(j)}$; in other words, n_j counts those subjects at risk for failing instantaneously prior to time $t_{(j)}$.

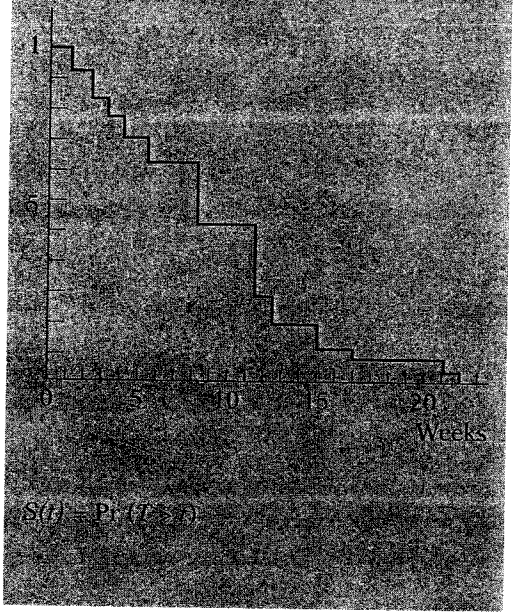
We now describe how to compute the KM curve for the table for group 2. The computations for group 2 are quite straightforward because there are no censored subjects for this group.

EXAMPLE (continued)

Group 2 placebo

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
1	21	2	0	19/21 = .90
2	19	2	0	17/21 = .81
3	17	1	0	16/21 = .76
4	16	2	0	14/21 = .67
5	14	2	0	12/21 = .57
8	12	4	0	8/21 = .38
11	8	2	0	6/21 = .29
12	6	2	0	4/21 = .19
15	4	1	3	3/21 = .14
17	3	1	0	2/21 = .10
22	2	1	0	1/21 = .05
23	1	1	0	0/21 = .00

KM Curve for Group 2 (Placebo)



The table of ordered failure times for group 2 is presented here again with the addition of another column that contains survival probability estimates. These estimates are the KM survival probabilities for this group. We will discuss the computations of these probabilities shortly.

A plot of the KM survival probabilities corresponding to each ordered failure time is shown here for group 2. Empirical plots such as this one are typically plotted as a step function that starts with a horizontal line at a survival probability of 1 and then steps down to the other survival probabilities as we move from one ordered failure time to another.

We now describe how the survival probabilities for the group 2 data are computed. Recall that a survival probability gives the probability that a study subject survives past a specified time.