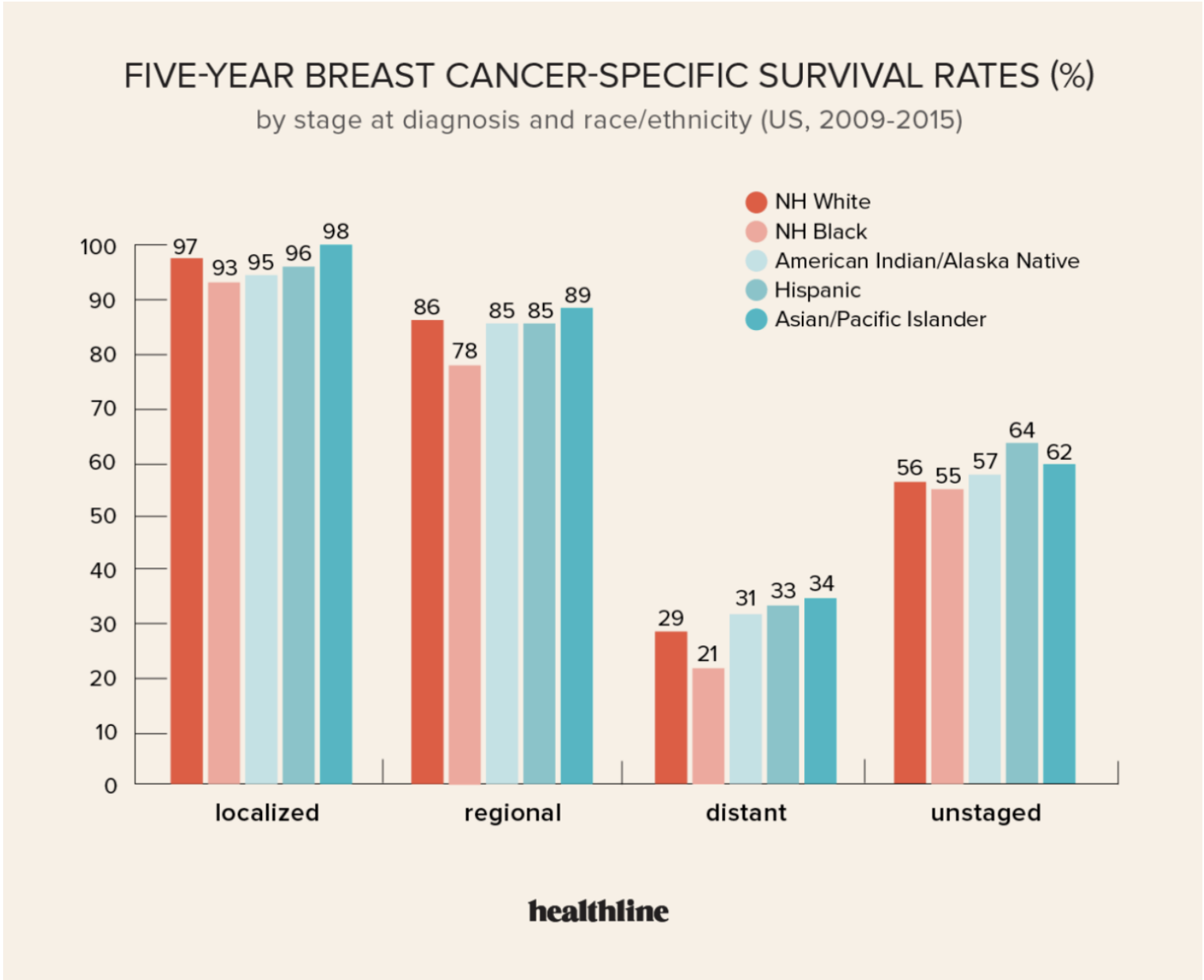
BIST P8110: Applied Regression II

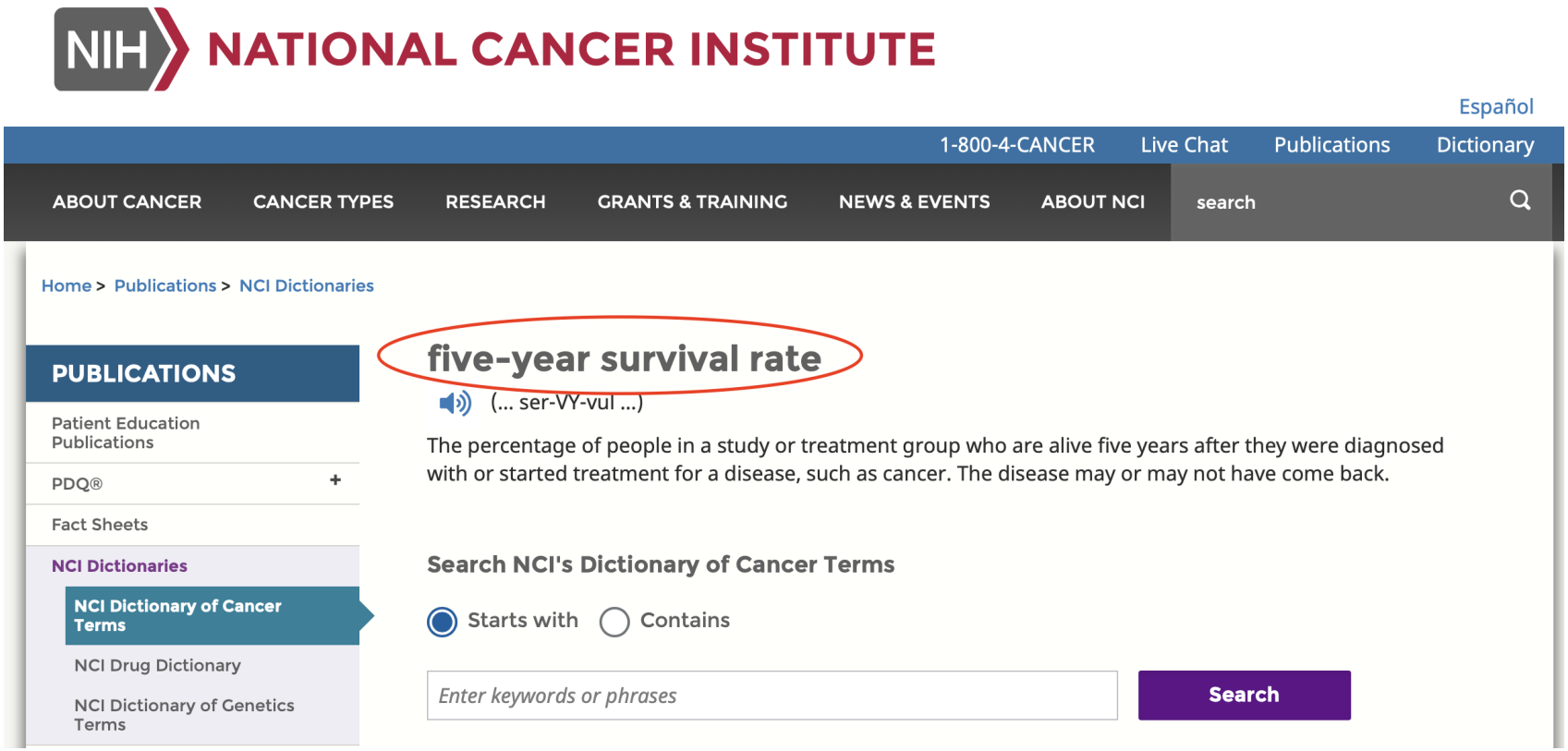
2. Kaplan-Meier Estimator of Survival Function

Qixuan Chen

Department of Biostatistics
Columbia University



Source: American Cancer Society. Bar chart created by Yaja' Mulcare


NIH NATIONAL CANCER INSTITUTE
Español
1-800-4-CANCER
Live Chat
Publications
Dictionary
ABOUT CANCER
CANCER TYPES
RESEARCH
GRANTS & TRAINING
NEWS & EVENTS
ABOUT NCI
search
Home > Publications > NCI Dictionaries
PUBLICATIONS
Patient Education Publications
PDQ®
Fact Sheets
NCI Dictionaries
NCI Dictionary of Cancer Terms
NCI Drug Dictionary
NCI Dictionary of Genetics Terms
five-year survival rate
(... ser-VY-vul ...)
The percentage of people in a study or treatment group who are alive five years after they were diagnosed with or started treatment for a disease, such as cancer. The disease may or may not have come back.
Search NCI's Dictionary of Cancer Terms
Starts with
Contains
Enter keywords or phrases
Search

This lecture's big ideas

- What is survival function?
- What is Kaplan-Meier estimator of survival function?

Survival function

- T denotes response variable, $T \geq 0$
- The survival function is

$$S(t) = Pr(T > t) = 1 - F(t)$$

  where $F(t)$ is called cumulative distribution function (CDF)
  - 5-year survival rate: $S(t = 5\text{yrs}) = Pr(T > 5\text{yrs})$
- How to estimate $S(t)$?
  - Survival function without censoring
  - Survival function with censoring

Estimating $S(t)$ without censoring

- We have a sample of n subjects and the time T_i at which the event has occurred for the ith subject
- Such data is called completely uncensored because every subject has been observed to have an event
- Suppose we pick a point in time t at which we want to know $S(t)$
- Because we have an event time for every subject, $S(t)$ is easy to estimate:

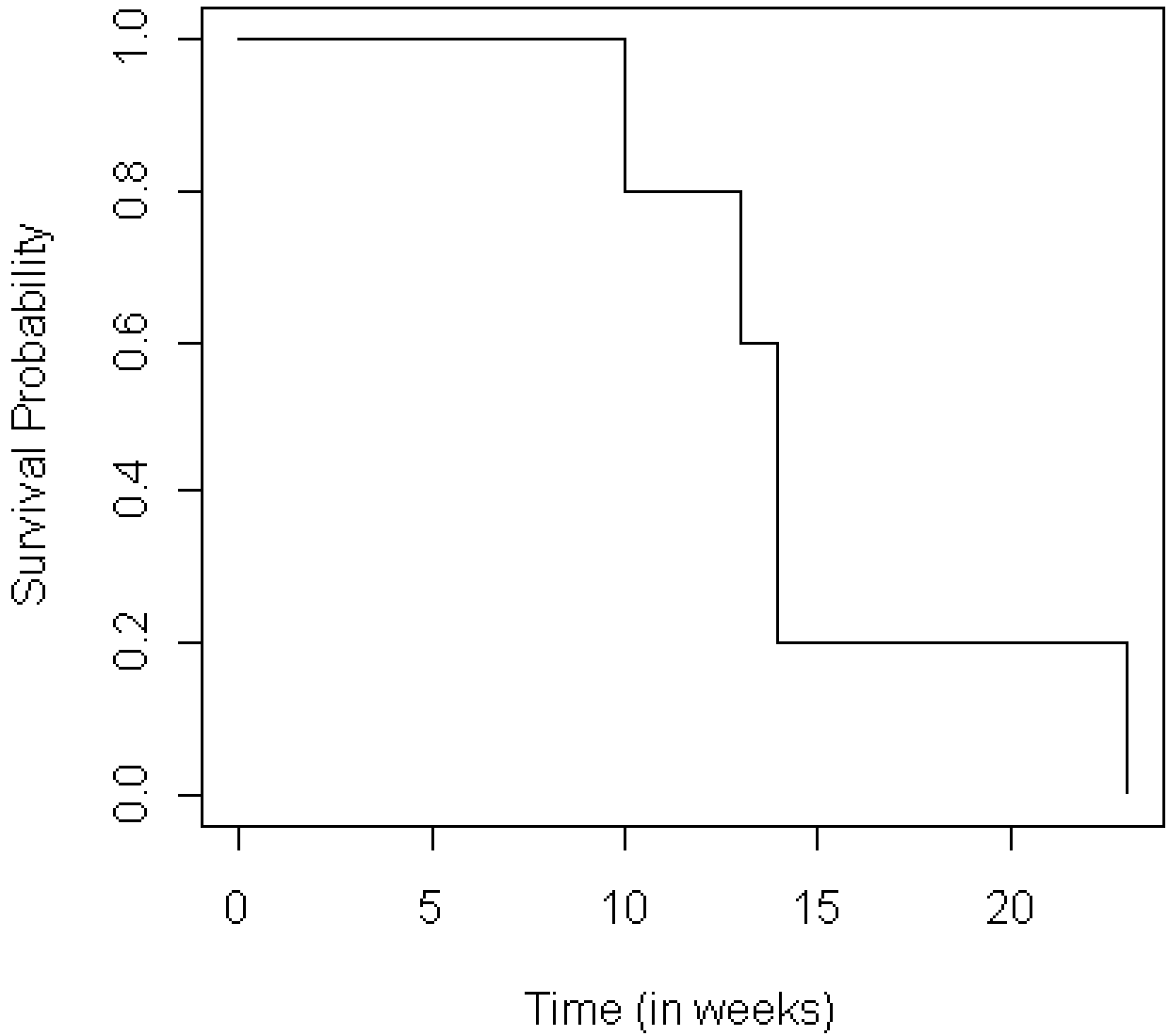
$$S(t) = Pr(T > t) = \frac{\#\text{ of subjects with } T_i > t}{n}$$

Estimating $S(t)$ without censoring

- Suppose we have 5 subjects whose cancer returned at 10, 13, 14, 14, and 23 weeks, respectively, after treatment
- Then we can compute $S(t)$ for any value of t:

$$\hat{S}(t) = \begin{cases} 5/5 = 1.0 & 0 \leq t < 10 \\ 4/5 = 0.8 & 10 \leq t < 13 \\ 3/5 = 0.6 & 13 \leq t < 14 \\ 1/5 = 0.2 & 14 \leq t < 23 \\ 0/5 = 0.0 & t \geq 23 \end{cases}$$

Displaying $\hat{S}(t)$ graphically



Estimating $S(t)$ with censoring

- When right censoring occurs, our previous method of computing $\hat{S}(t)$ breaks down
- Why?

Estimating $S(t)$ with censoring

- Returning to our previous example, suppose one of the subjects recorded at 14 weeks actually left the study at 14 weeks and was cancer-free
- The data are now recorded as 10, 13, 14, 14^+, and 23 weeks
  - using 14^+ to denote right-censoring at 14 weeks
- Questions:
  - Can we estimate $\hat{S}(t)$ for any value of $t < 14$?
  - Can we estimate $\hat{S}(t)$ for any value of $t \geq 14$?

Estimating $S(t)$ with censoring

- We can still compute $\hat{S}(t)$ for any $t < 14$ using the formula below because we know the censored subject was cancer-free for at least 14 weeks.

$$S(t) = Pr(T > t) = \frac{\#\text{ of subjects with } T_i > t}{n}$$

- We can NOT compute $\hat{S}(t)$ for any $t \geq 14$ because we do not know whether the censored subject had cancer recurrence after leaving the study.

Estimating $S(t)$ with censoring

With the event and censored information, we have:

$$\hat{S}(t) = \begin{cases} 5/5 = 1.0 & 0 \leq t < 10 \\ 4/5 = 0.8 & 10 \leq t < 13 \\ 3/5 = 0.6 & 13 \leq t < 14 \\ ??? & 14 \leq t < 23 \\ ??? & t \geq 23 \end{cases}$$

- We can fill in the ??? if we make some assumption about the future of the censored subject.

Estimating $S(t)$ with censoring

If we assume that the censored subject had cancer recurrence immediately after leaving, we have:

$$\hat{S}(t) = \begin{cases} 5/5 = 1.0 & 0 \leq t < 10 \\ 4/5 = 0.8 & 10 \leq t < 13 \\ 3/5 = 0.6 & 13 \leq t < 14 \\ 1/5 = 0.2 & 14 \leq t < 23 \\ 0/5 = 0.0 & t \geq 23 \end{cases}$$

- Such an approach is too conservative and makes $S(t)$ too low (negative biased) for $t \geq 14$

Estimating $S(t)$ with censoring

If we assume that the censored subject lived longer than 23 weeks, we have:

$$\hat{S}(t) = \begin{cases} 5/5 = 1.0 & 0 \leq t < 10 \\ 4/5 = 0.8 & 10 \leq t < 13 \\ 3/5 = 0.6 & 13 \leq t < 14 \\ 2/5 = 0.4 & 14 \leq t < 23 \\ 1/5 = 0.2 & t \geq 23 \end{cases}$$

- Such an approach is too liberal and makes $S(t)$ too large (positive biased) for $t \geq 14$

Estimating $S(t)$ with censoring

The "right" answer lies somewhere in-between the two biased approaches:

$$\hat{S}(t) = \begin{cases} 1.0 & 0 \leq t < 10 \\ 0.8 & 10 \leq t < 13 \\ 0.6 & 13 \leq t < 14 \\ [0.2, 0.4] & 14 \leq t < 23 \\ [0.0, 0.2] & t \geq 23 \end{cases}$$

- The "right" answer was developed by Kaplan & Meier (1958), leading to the so-called **Kaplan-Meier estimator** of $S(t)$.

Kaplan-Meier estimate of $S(t)$

- The first step is to identify and order the **unique** survival time points among the subjects **with an event**
  - For our example with data of 10, 13, 14, 14^+, and 23, there are $J = 4$ values: $t_1 = 10$, $t_2 = 13$, $t_3 = 14$, and $t_4 = 23$
  - These time points define $J + 1 = 5$ non-overlapping intervals: $[0, 10), [10, 13), [13, 14), [14, 23), [23, \infty)$

Kaplan-Meier estimate of $S(t)$

- For each interval, we compute p_j, the **conditional survival probability**
  - the probability of surviving through the entire interval, given being in the study (at risk) at the beginning of the interval.
- For the j^{th} interval, $j = 1, 2, ..., J$, we have:

$$\begin{aligned} p_j &= 1 - \frac{\#\text{ of events at } t_j}{\#\text{of subjects at risk just prior to } t_j} \\ &= 1 - \frac{d_j}{n_j} \\ &= \frac{n_j - d_j}{n_j} \end{aligned}$$

Kaplan-Meier estimate of $S(t)$

- For the cancer recurrence example, we can obtain:

j	t_j	Interval	n_j	d_j	$p_j = \frac{n_j - d_j}{n_j}$
0	0	[0,10)	5	0	
1	10	[10,13)	5	1	
2	13	[13,14)	4	1	
3	14	[14,23)	3	1	
4	23	[23,∞)	1	1	0

Kaplan-Meier estimate of $S(t)$

- Thus the probability of surviving to the end of interval j^* is simply the product of the conditional survival probability of all intervals prior to and including interval j^*.
  - For example:

$$\begin{aligned}\text{Prob(surviving to the end of third interval)} = \\ \text{Prob(surviving the first interval)} \times \\ \text{Prob(surviving the second interval)} \times \\ \text{Prob(surviving the third interval)}\end{aligned}$$

Kaplan-Meier estimate of $S(t)$

- This concept defines the K-M estimate of $S(t)$

$$\hat{S}(t) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j}) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}$$

- In words, the Kaplan-Meier estimate of the survival probability to time t is the product of surviving each interval that occurs before or including t
- For our example, the estimated survival probability to $t = 20$ would be the product of surviving the intervals $[0, 10), [10, 13), [13, 14)$, and $[14, 23)$

Kaplan-Meier estimate of $S(t)$

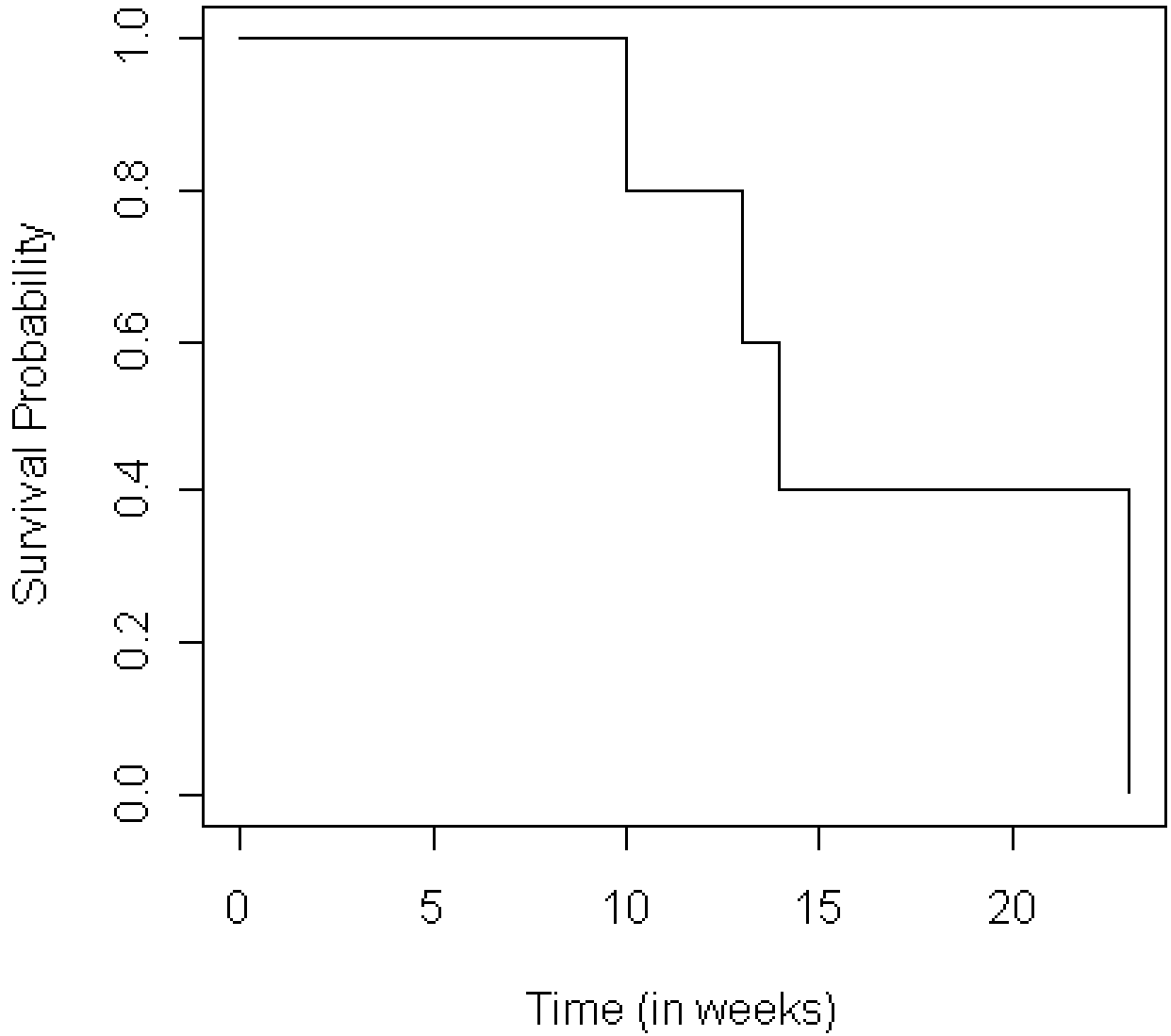
- For the cancer recurrence example, we have:

j	t_j	n_j	d_j	$p_j = \frac{n_j - d_j}{n_j}$
0	0	5	0	1
1	10	5	1	4/5
2	13	4	1	3/4
3	14	3	1	2/3
4	23	1	1	0

- Using the above table, we can compute the K-M estimates:

For t in	$\hat{S}(t)$
$[0, 10)$	1
$[10, 13)$	$1 \times 4/5 = 0.8$
$[13, 14)$	$1 \times 4/5 \times 3/4 = 0.6$
$[14, 23)$	$1 \times 4/5 \times 3/4 \times 2/3 = 0.4$
$[23, \infty)$	$4/5 \times 3/4 \times 2/3 \times 0/1 = 0.0$

Kaplan-Meier survival curve



More on Kaplan-Meier Estimator

- Kaplan-Meier estimator is also called the product limit estimator.
  - Subjects who die contribute to the number at risk until their time of death, at which point they also contribute to the number of deaths. Subjects who are censored contribute to the number at risk until they are censored.
  - If the last observed time point corresponds to a censored observation, then the estimate of the survival function does not go to zero and is undefined after the last time point.
- Life-table estimator of survival function is an alternative to the Kaplan-Meier estimator (Page 24 Hosmer).

Independence of censoring

- In using KM estimator, we make a crucial assumption that **censoring is independent of survival time**.
  - Put another way, censoring is non-informative, i.e. not due to causes related to when a subject would have an event
  - such an assumption is valid with administrative censoring (i.e. study ends)

Informative censoring

- Examples of informative censoring:
  - Suppose we follow a cohort of new graduate students to see what factors affect how long it takes them to get a Ph.D.
  - Some students drop out before completing the degree - these observations are right censored.
  - There is a good reason to suspect that those who drop out are among those who would take more time to finish if they stayed until completion.
- Informative censoring can lead to severe bias, but it is difficult in most situations to estimate the magnitude or direction of the bias