

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

# When Time Is of Interest: The Case for Survival Analysis

**John McGready**  
**Johns Hopkins University**

# Lecture Topics

- ◆ Why another set of methods?
- ◆ Event times versus censoring times
- ◆ Estimating the survival curve—Kaplan Meier and life table methods
- ◆ Comparing survival curves



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

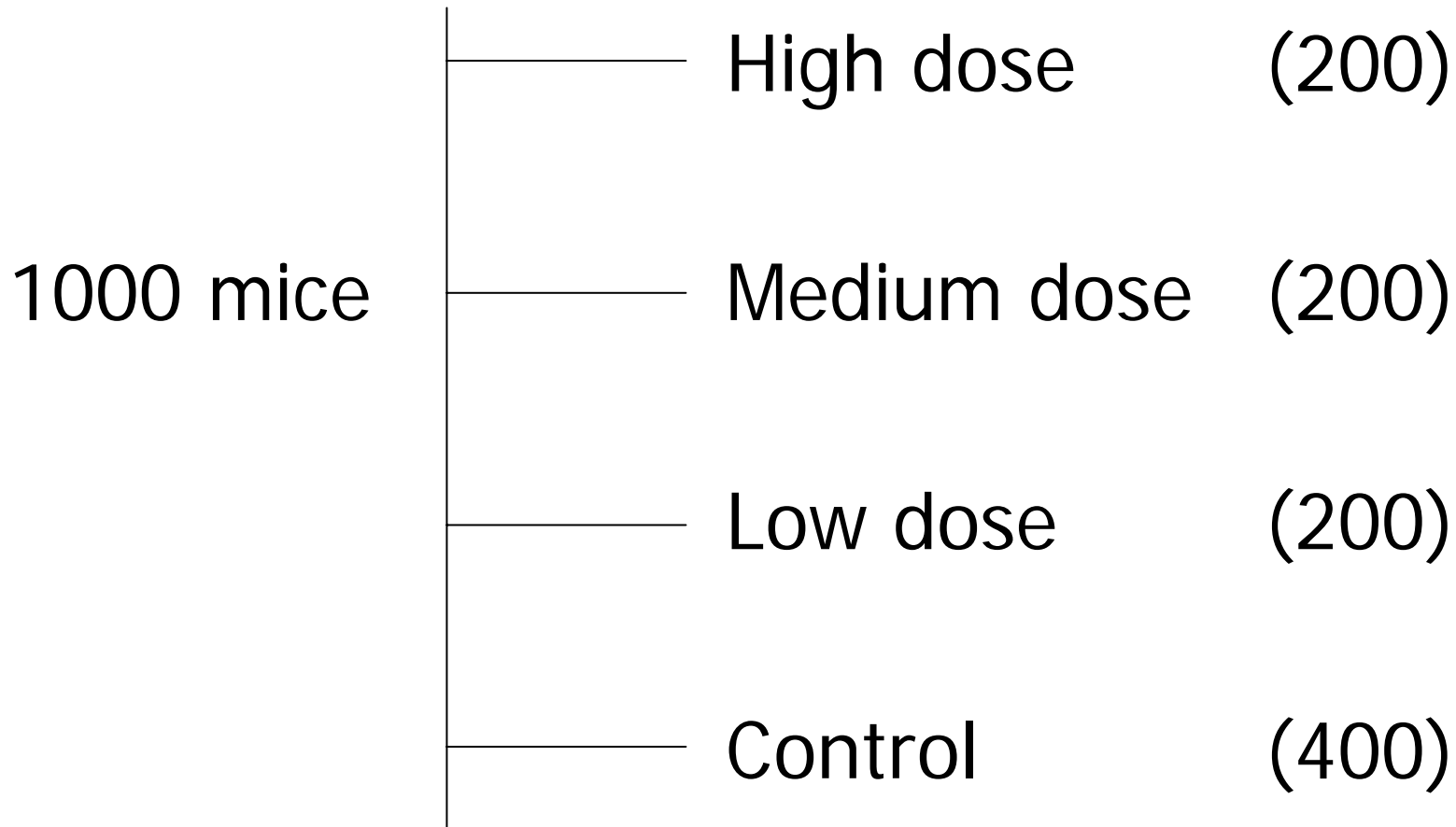
# Section A

*Motivating the Need*

# Survival Analysis

- ◆ Statistical methods for the study of time to an event
- ◆ Accounts for . . .
  - Time that events occur
  - Different follow-up times

# Artificial Sweetener and Cancer



# Artificial Sweetener

- ◆ Rats were followed for one year to see if tumors developed
- ◆ Proportion of rats with tumors the same in the unexposed and exposed

# Artificial Sweetener

- ◆ However, evidence of “acceleration”
- ◆ Tumors happened sooner in those exposed
- ◆ How can we incorporate information not only about whether tumors occur, but also how long it takes them to occur?

# Survival Analysis

- ◆ Survival analysis methods allow us to incorporate information about both frequency of event occurrence and time to event information
- ◆ Subjects are followed until they have an “event,” or the study ends

# Endpoint

- ◆ The endpoint doesn't have to be "death"; it can be any well-defined event
  - Death
  - Disease onset
  - Menopause
  - Pregnancy
  - Relapse

# Time Scale

- ◆ When do you start the clock?
  - Time from diagnosis of disease to death
  - Time from HIV infection to AIDS

# Time Scale

- ◆ When do you start the clock?
  - Time from birth (chronological age)
  - Time from randomization in clinical trial

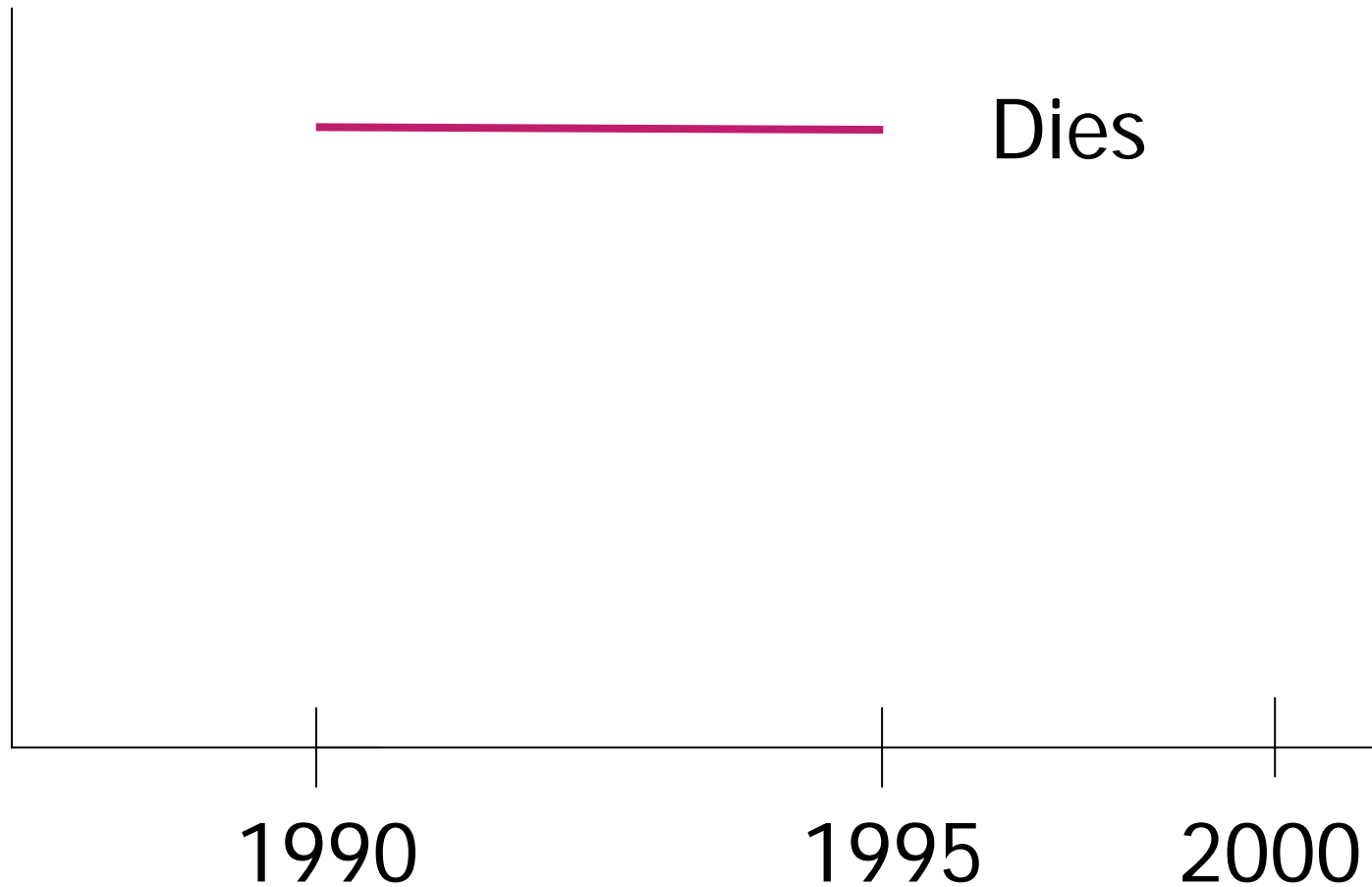
# Why Is Survival Analysis Tricky?

- ◆ Suppose we have designed a study to estimate survival after chemotherapy treatment for patients with a certain cancer
- ◆ Patients received chemotherapy between 1990 and 1994 and were followed until death or the year 2000, whichever occurred first

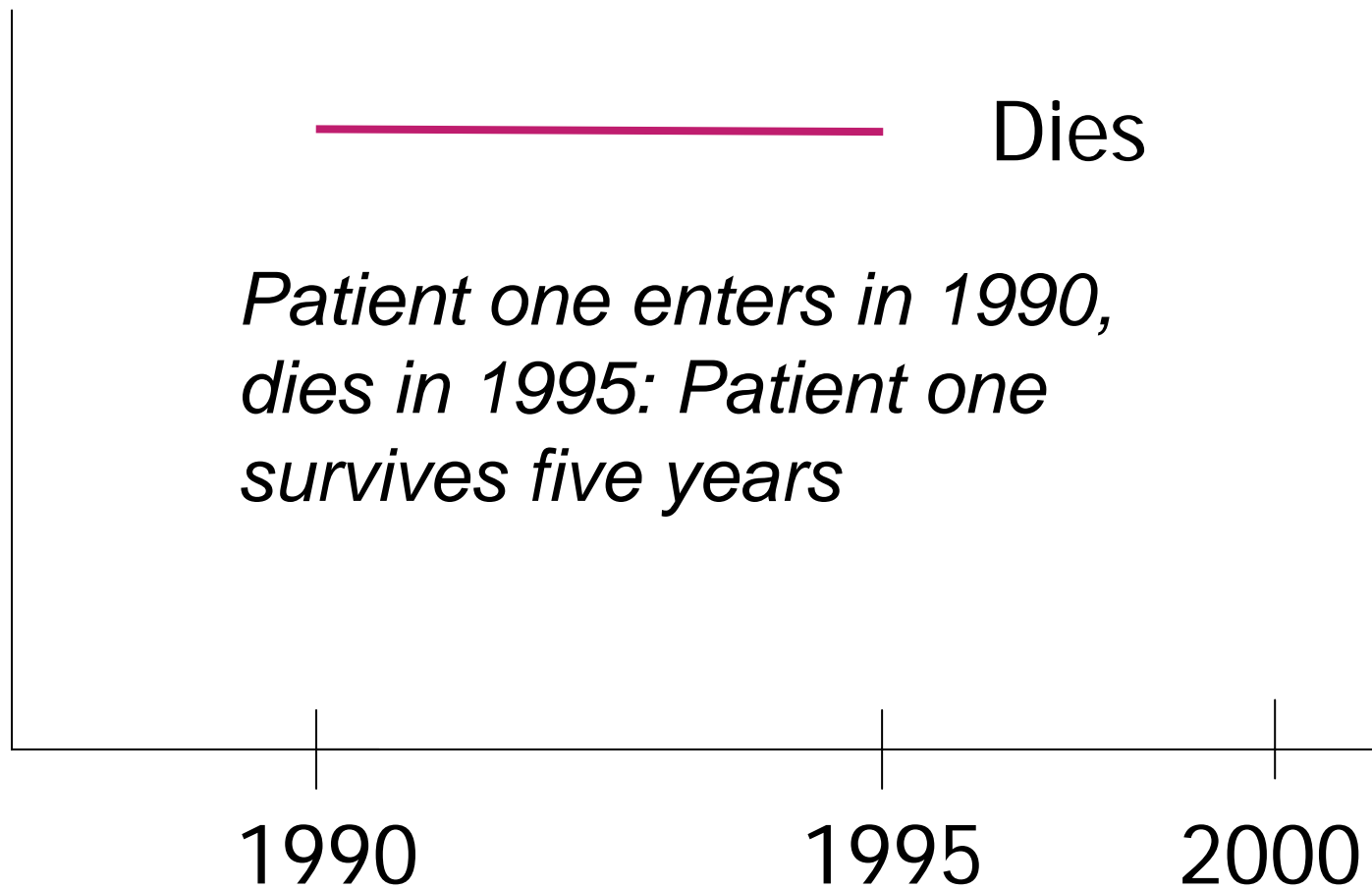
# Why Is Survival Analysis Tricky?

- ◆ In this study the event of interest is death
- ◆ The time clock starts as soon as the subject finishes his/her chemotherapy treatments

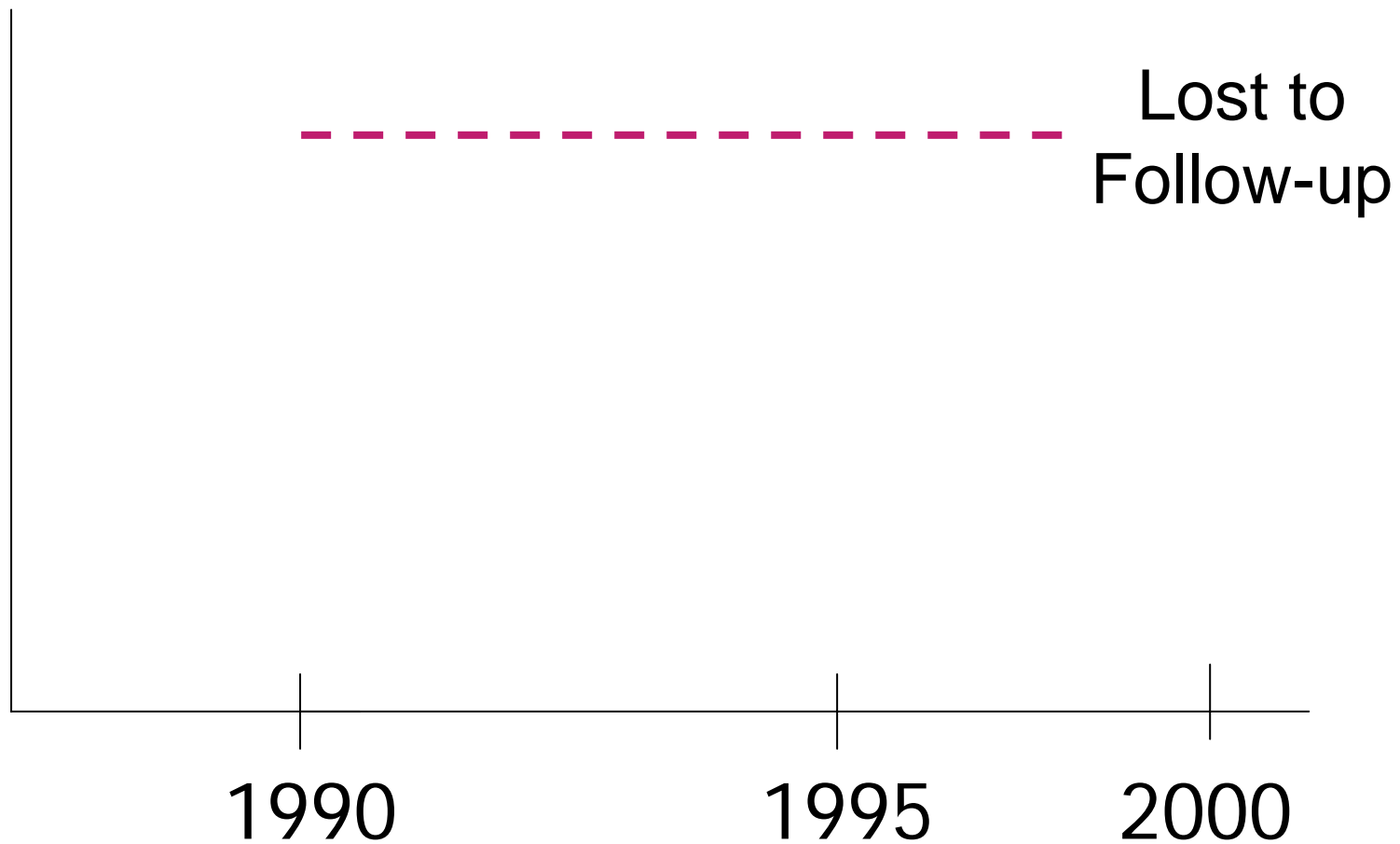
# Why Is Survival Analysis Tricky?



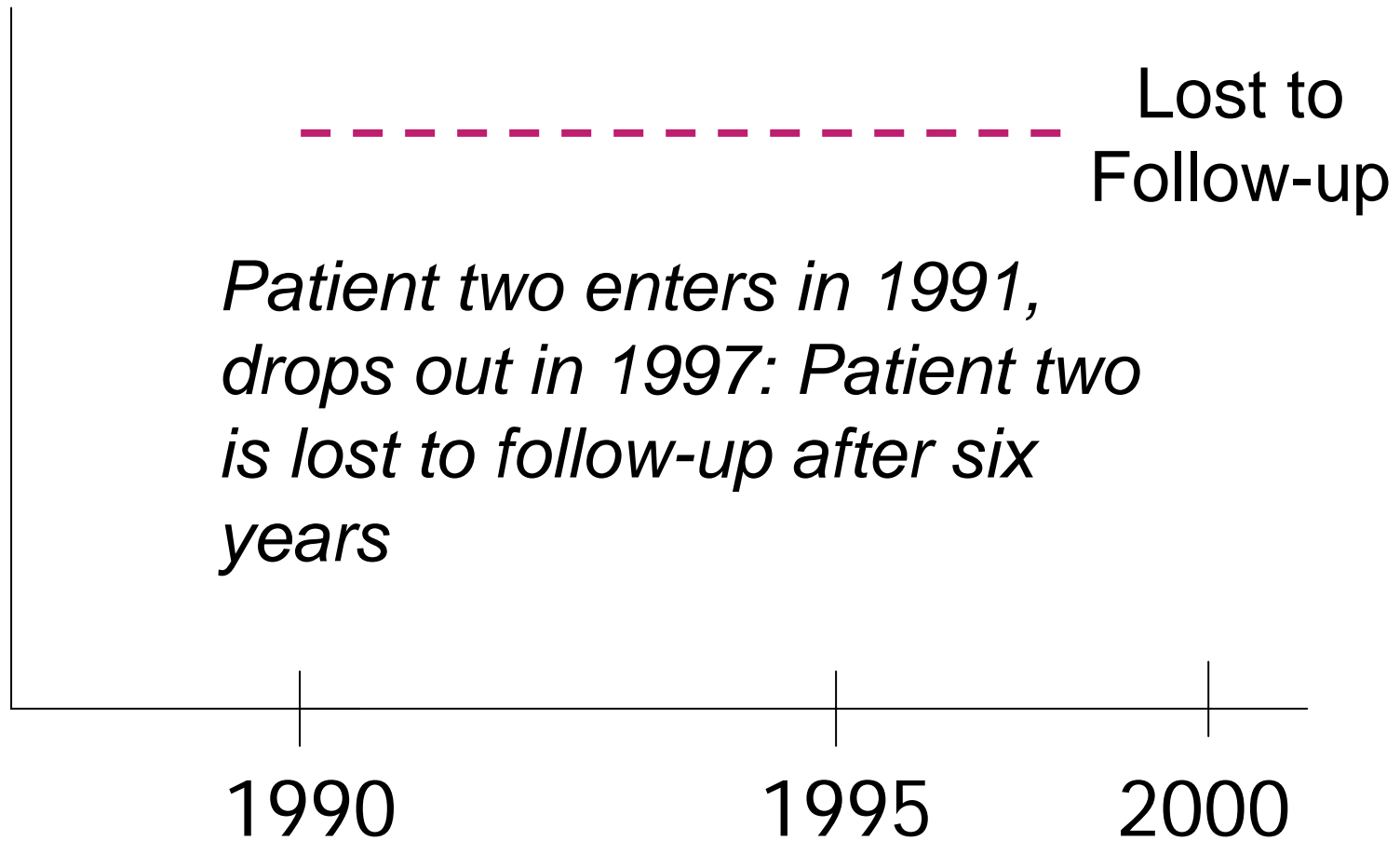
# Why Is Survival Analysis Tricky?



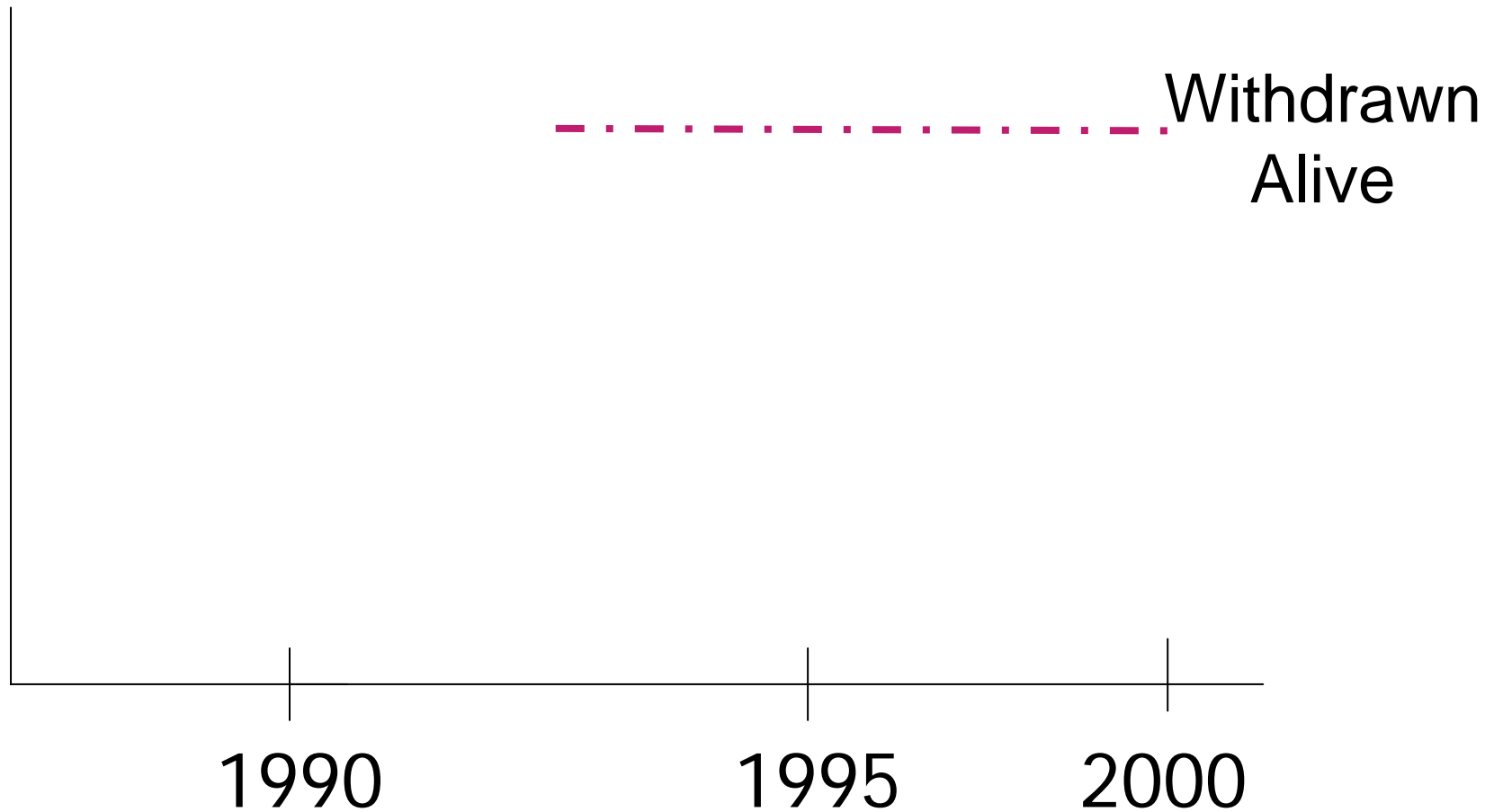
# Why Is Survival Analysis Tricky?



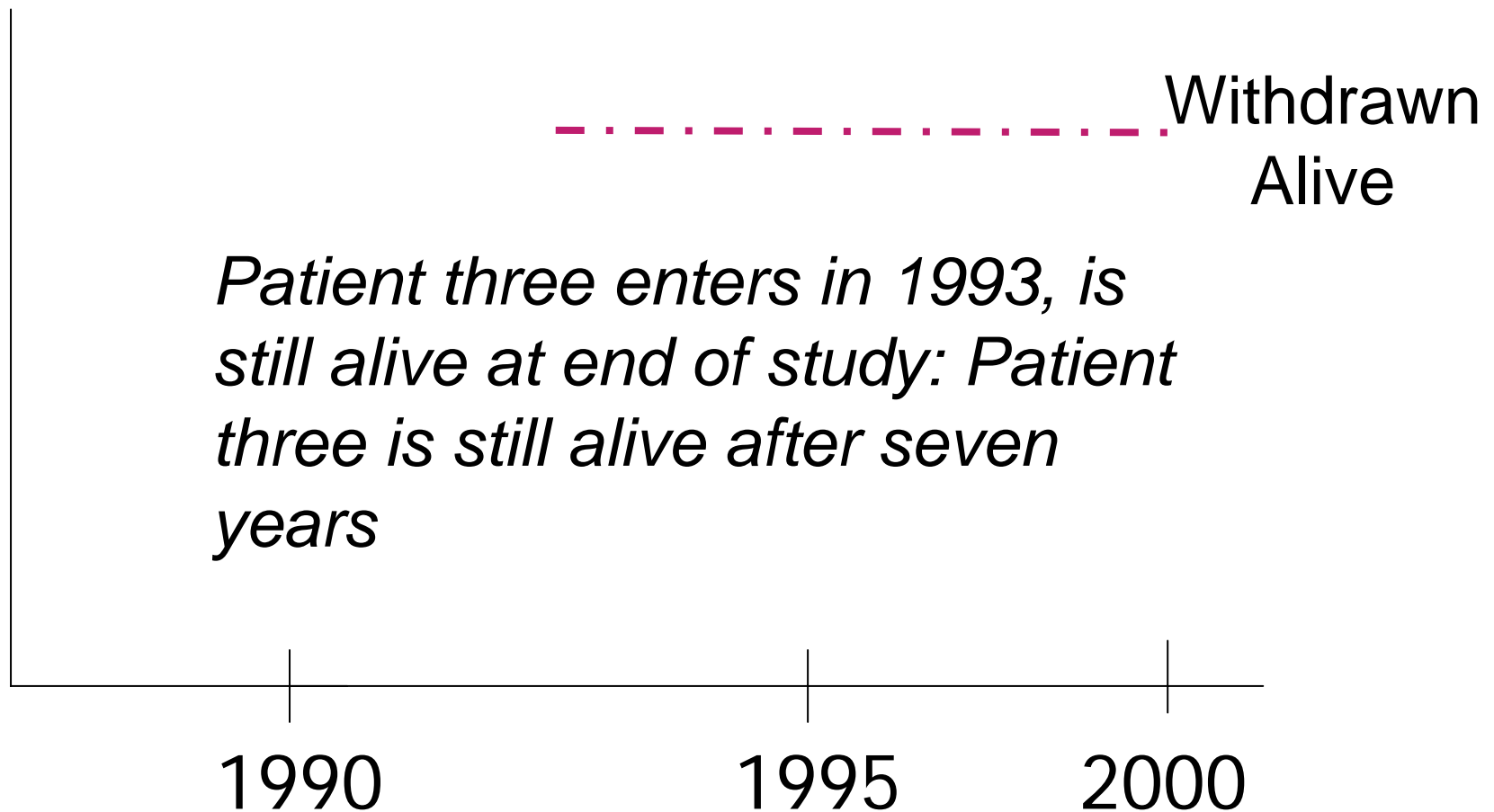
# Why Is Survival Analysis Tricky?



# Why Is Survival Analysis Tricky?



# Why Is Survival Analysis Tricky?



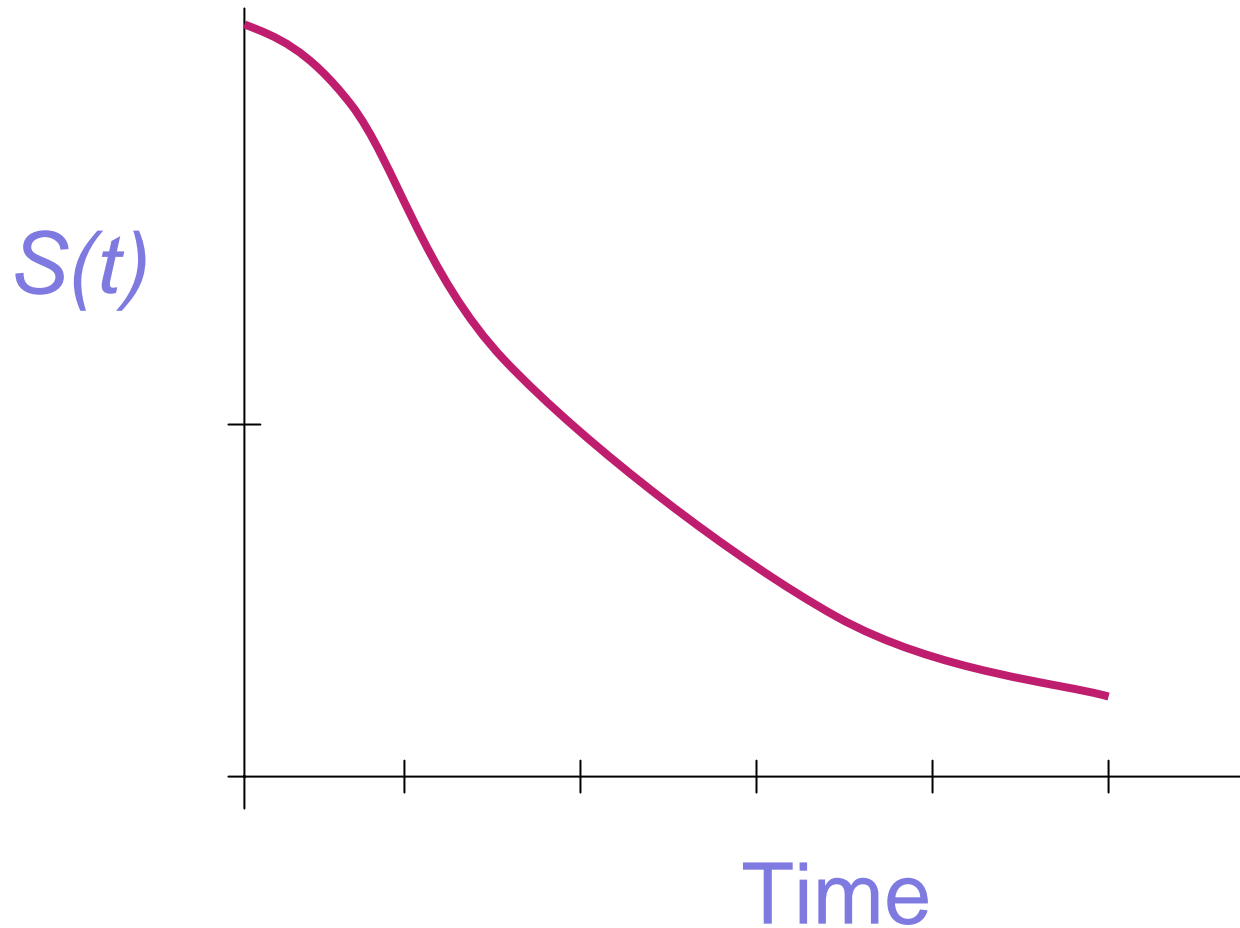
# Why Is Survival Analysis Tricky?

- ◆ Patient:
  - 1: 1990 → 1995 5 years
  - 2: 1991 → 1997 6+ years
  - 3: 1993 → 2000 7+ years
- ◆ Patients two and three are called censored observations

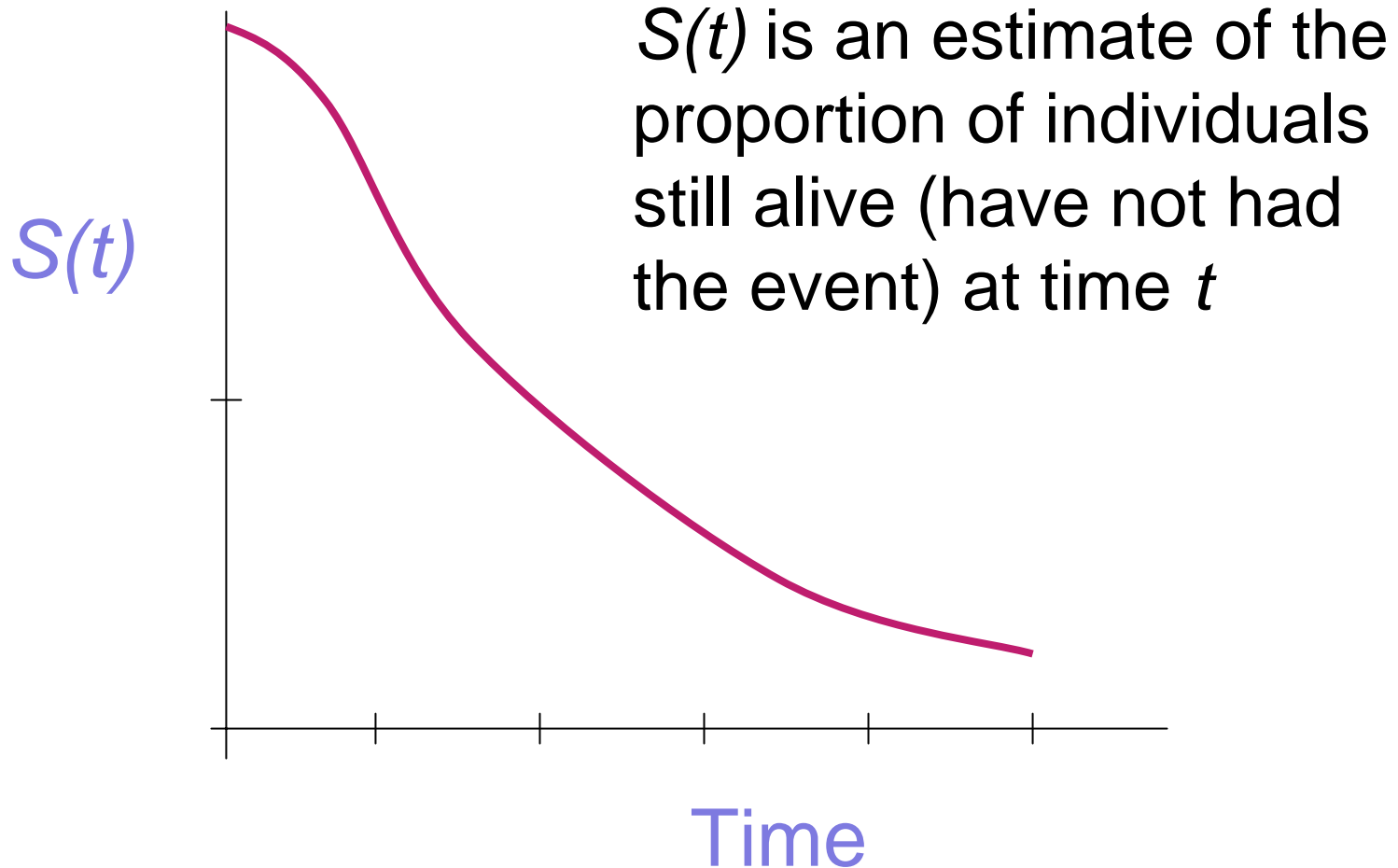
# Why Is Survival Analysis Tricky?

- ◆ We need a method which can incorporate information about censored data into an analysis

# The Survival Curve



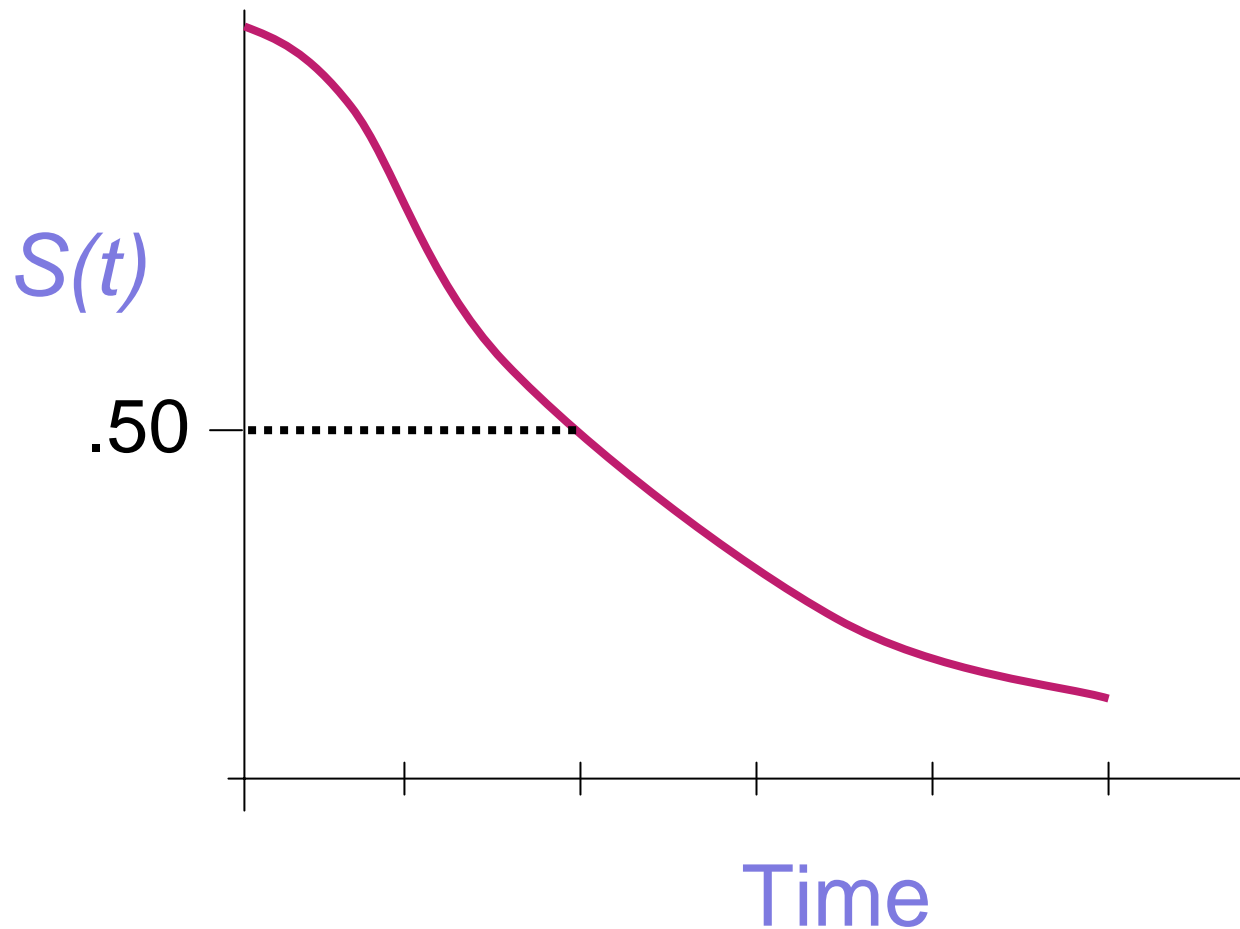
# The Survival Curve



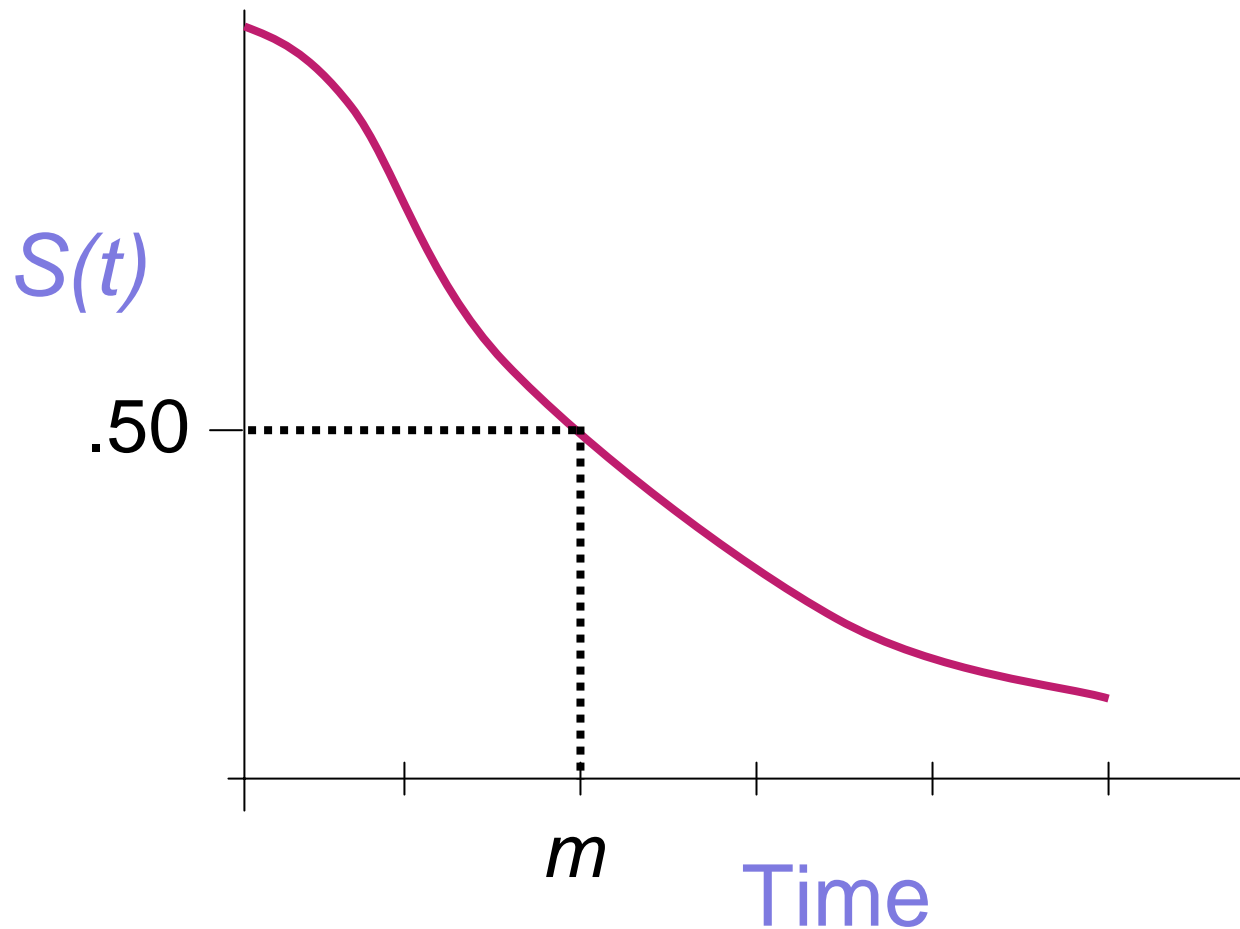
# Summary Statistics

- ◆ We can estimate summary statistics from estimated survival curve
  - Median survival time
  - One, two year survival rates

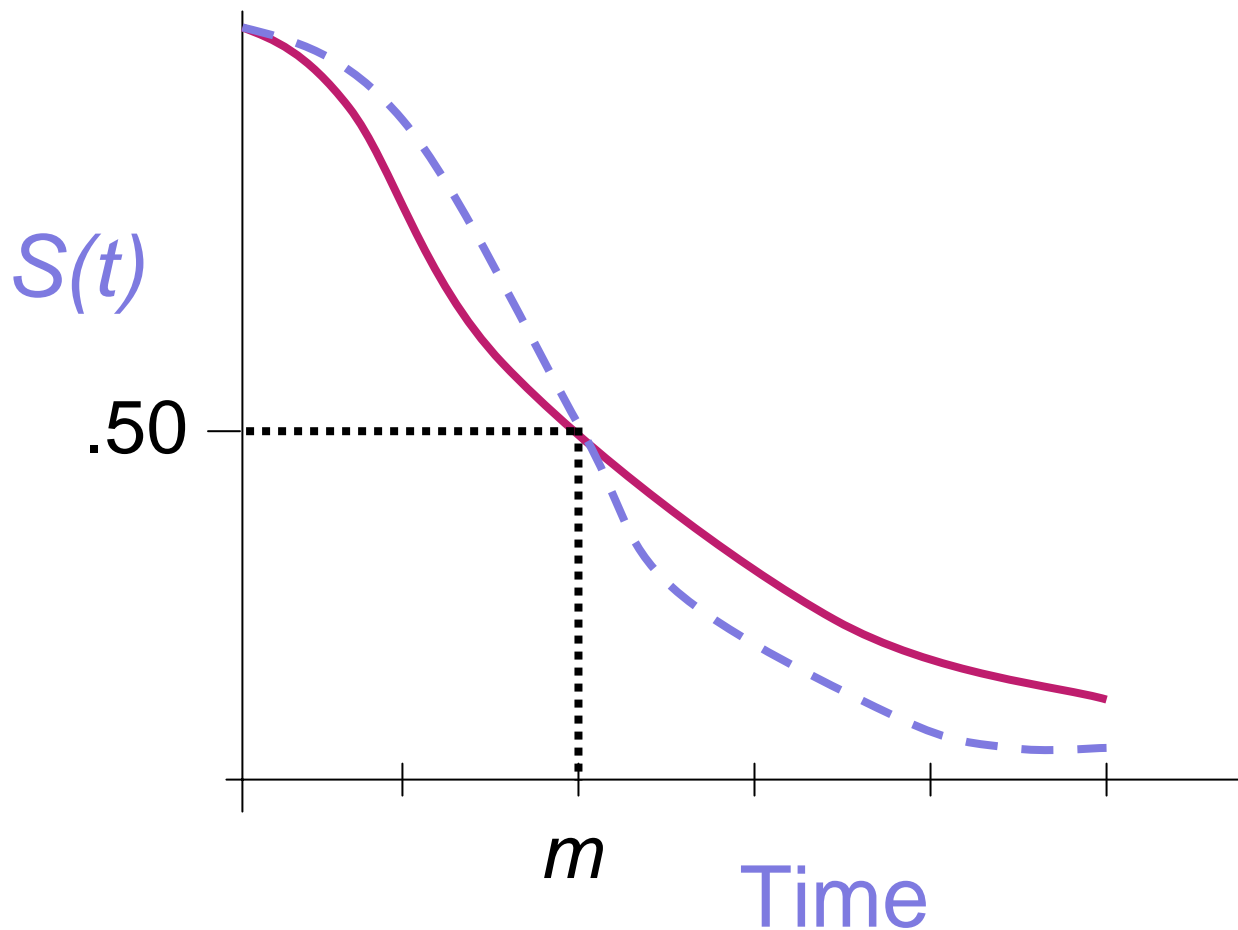
# Estimating Median Survival



# Estimating Median Survival



# Caveat—Medians Do Not Describe Whole Curve





JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

# Section A

## *Practice Problems*

# Practice Problems

1. Explain the difference between a censored observation versus an observation in which an actual event time is observed. For a single subject, how will his/her censoring time compare to his/her actual, unobserved event time?

# Practice Problems

2. What is the “survival function” or “survival curve?” What information about the survival experience of a cohort of subjects is described by the cohort’s “survival curve?”



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

# Section A

*Practice Problem Solutions*

# Question

1. Explain the difference between a censored observation versus an observation in which an actual event time is observed. For a single censored subject, how will his/her censoring time compare to his/her actual, unobserved event time?

# Answer

- ◆ A censored observation is a time measure on a subject who does not have the outcome/event under study. A subject may be censored because he/she drops out of the study before having the event, or makes it to the end of the study without having the event.

# Answer

- ◆ Censoring times will always be less than actual event times for those who are censored—so the censoring time is an underestimate of the event time.

# Question

2. What is the “survival function” or “survival curve?” What information about the survival experience of a cohort of subjects is described by the cohort’s “survival curve?”

# Answer

- ◆ The survival function tracks the estimated percentage of individuals in a cohort who are still “event free” as a function of time—the survivor function gives an estimate of these percentages for a given time between the beginning of the time interval and the end of the study.



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

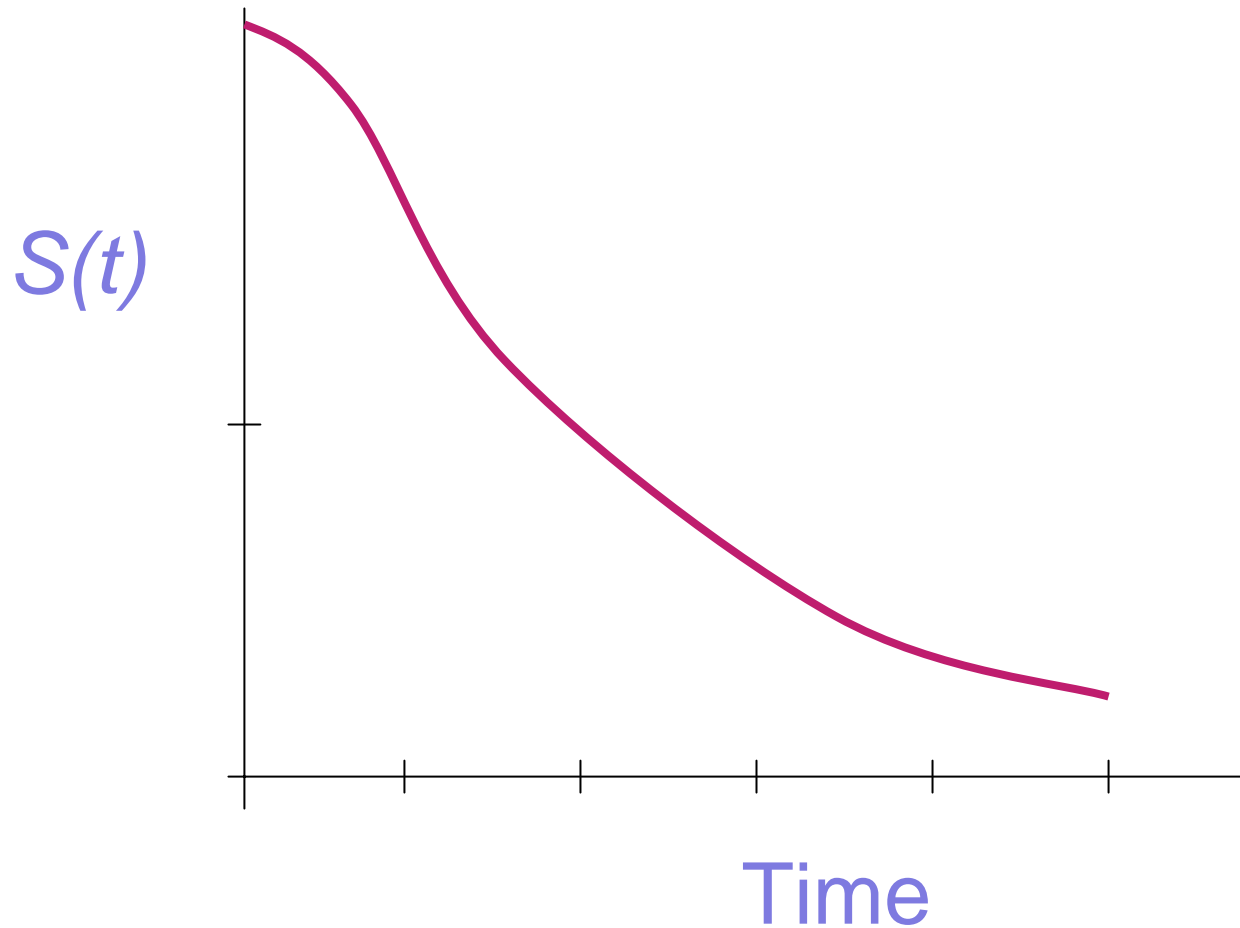
## Section B

*Estimating the Survival Curve:  
The Kaplan Meier and Life Table  
Approaches*

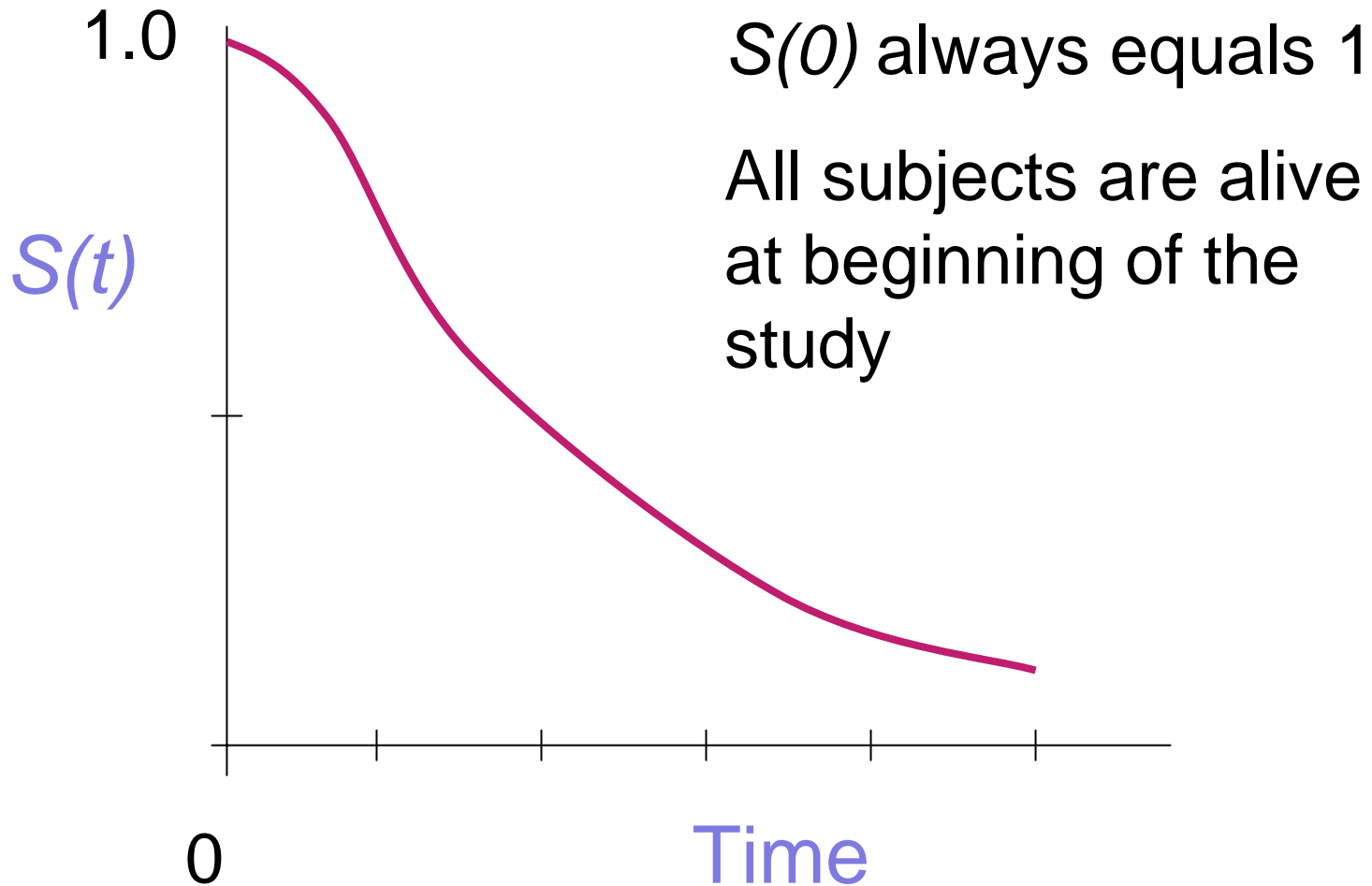
# Central Problem

- ◆ Estimation of the survival curve
- ◆  $S(t)$  = Proportion surviving at least to time  $t$  or beyond

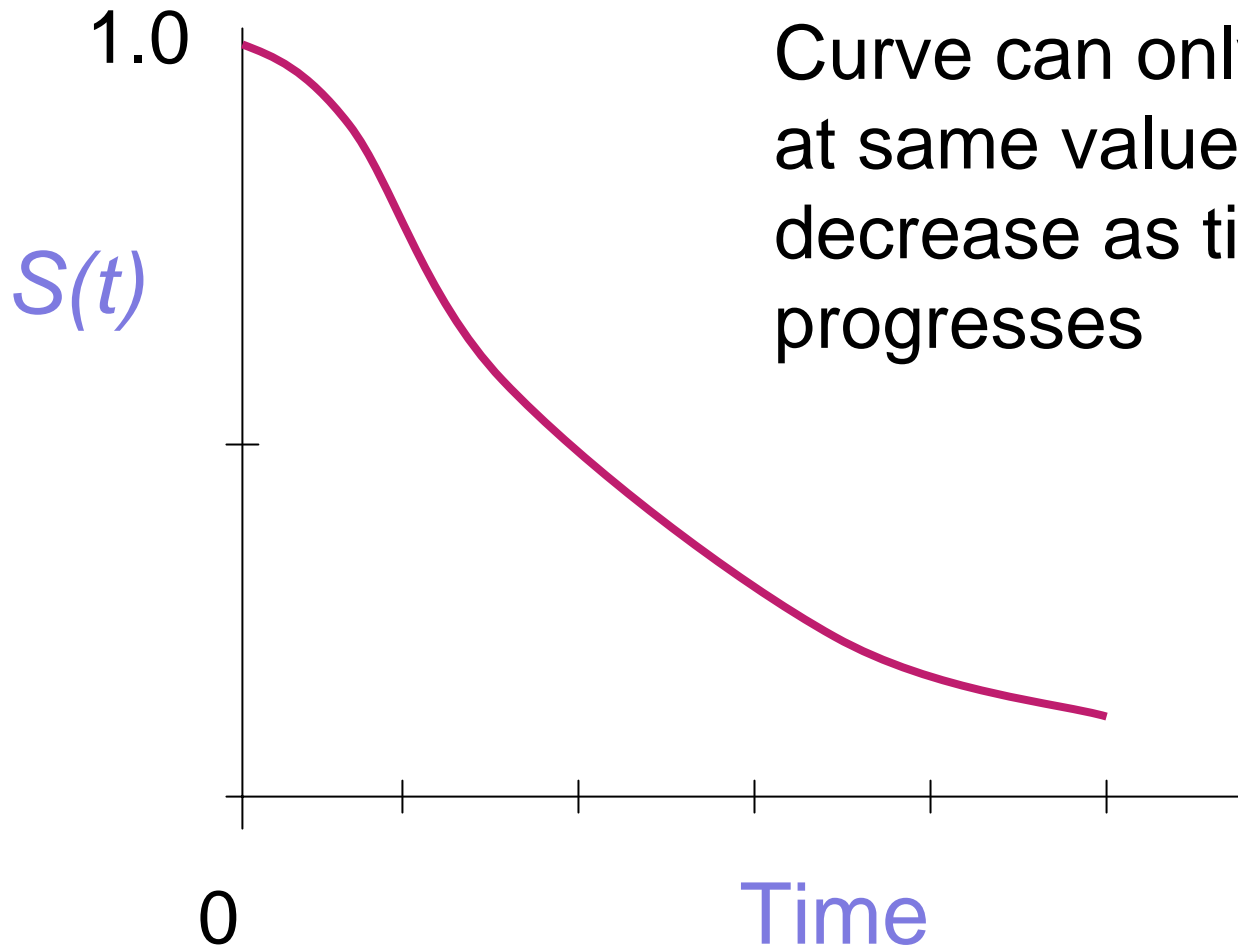
# The Survival Curve



# The Survival Curve

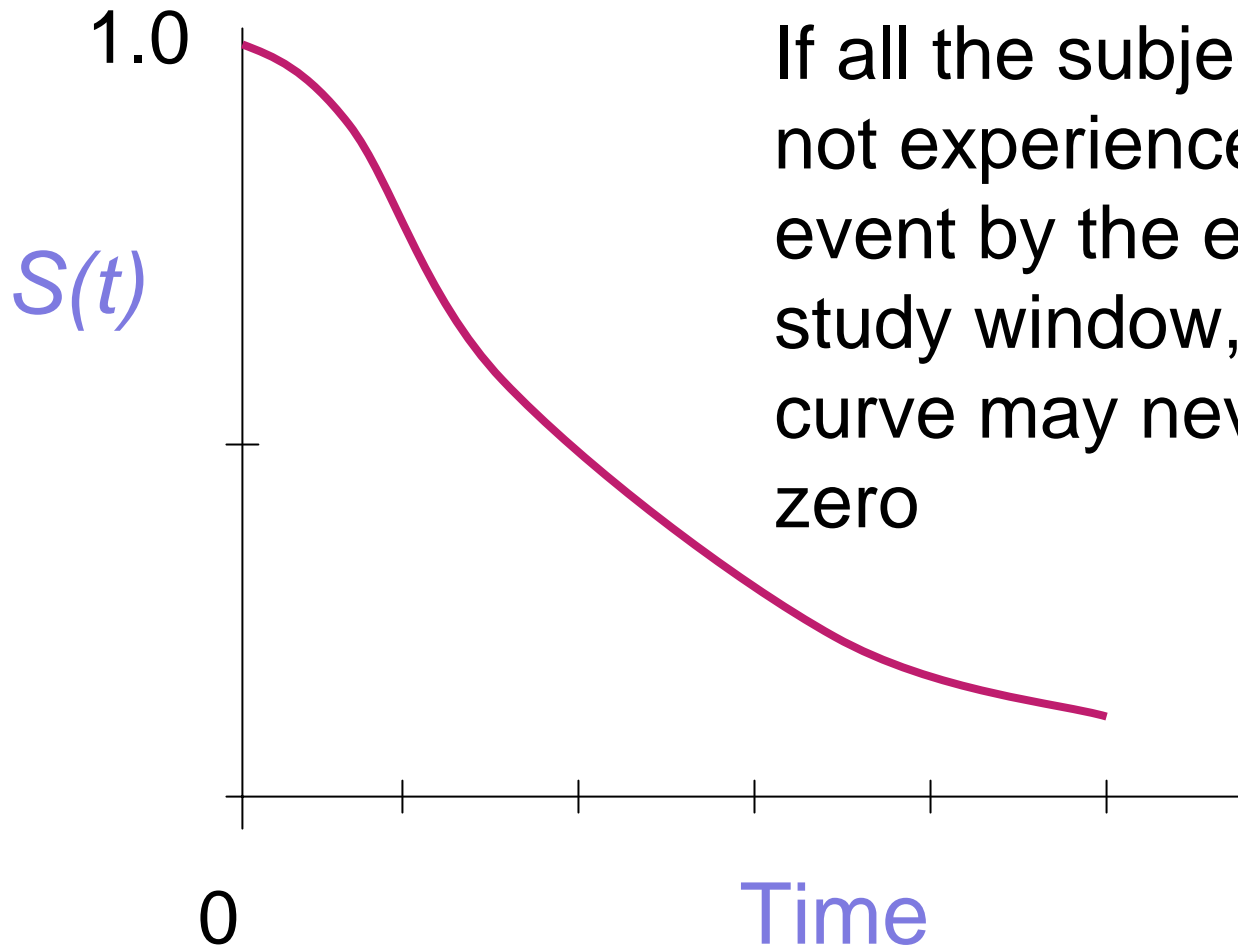


# The Survival Curve



Curve can only remain at same value or decrease as time progresses

# The Survival Curve



If all the subjects do not experience the event by the end of the study window, the curve may never reach zero

# Approaches

- ◆ Life table method
  - Grouped in intervals
- ◆ Kaplan-Meier (1958)
  - Ungrouped data
  - Small samples

# Kaplan-Meier Estimate

- ◆ Example

- Survival times ( $N = 10$ )

- 8+    4    5+    6    12    3    10+    17    6

# Kaplan-Meier Estimate

- ◆ Example

- Order of survival times

- 3 4 5+ 6 6 8+ 10+ 12 15 17

# Kaplan-Meier Estimate

- ◆ Curve can be estimated at each event, but not at censoring times
- ◆  $S(t)$  = proportion of individuals surviving beyond time  $t$

# Kaplan-Meier Estimate

- ◆ Curve can be estimated at each event, but not at censoring times

$$S(t) = \left( \frac{N(t) - E(t)}{N(t)} \right) \times S(\text{Previous\_Event\_Time})$$

- ◆  $E(t)$  = # events at time  $t$
- ◆  $N(t)$  = # subjects at risk for event at time  $t$

# Kaplan-Meier Estimate

- ◆ Curve can be estimated at each event, but not at censoring times

$$S(t) = \left( \frac{N(t) - E(t)}{N(t)} \right) \times S(\text{Previous\_Event\_Time})$$



Proportion of original sample  
making it to time  $t$

# Kaplan-Meier Estimate

- ◆ Curve can be estimated at each event, but not at censoring times

$$S(t) = \left( \frac{N(t) - E(t)}{N(t)} \right) \times S(\text{Previous\_Event\_Time})$$



Proportion surviving to time  $t$   
who survive beyond time  $t$

# Kaplan-Meier Estimate

- ◆ Start estimate at first event time

– 3 4 5+ 6 6 8+ 10+ 12 15 17

$$S(3) = \left( \frac{N(3) - E(3)}{N(3)} \right) = \frac{10 - 1}{10} = \frac{9}{10} = .9$$

# Kaplan-Meier Estimate

- ◆ Can estimate  $S(t)$  at each event time
  - 3 4 5+ 6 6 8+ 10+ 12 15 17

$$S(4) = \left( \frac{N(4) - E(4)}{N(4)} \right) \times S(3) = \left( \frac{9 - 1}{9} \right) \times (.9)$$
$$= \frac{8}{9} \times .9 = .8$$

# Kaplan-Meier Estimate

- ◆ Skip over censoring times—remove from number at risk for next event time

– 3 4 5+ 6 6 8+ 10+ 12 15 17

$$S(6) = \left( \frac{N(6) - E(6)}{N(6)} \right) \times S(4) = \left( \frac{7 - 2}{7} \right) \times (.8)$$

$$= \frac{5}{7} \times .8 = .57$$

# Kaplan-Meier Estimate

- ◆ Continue through final event time

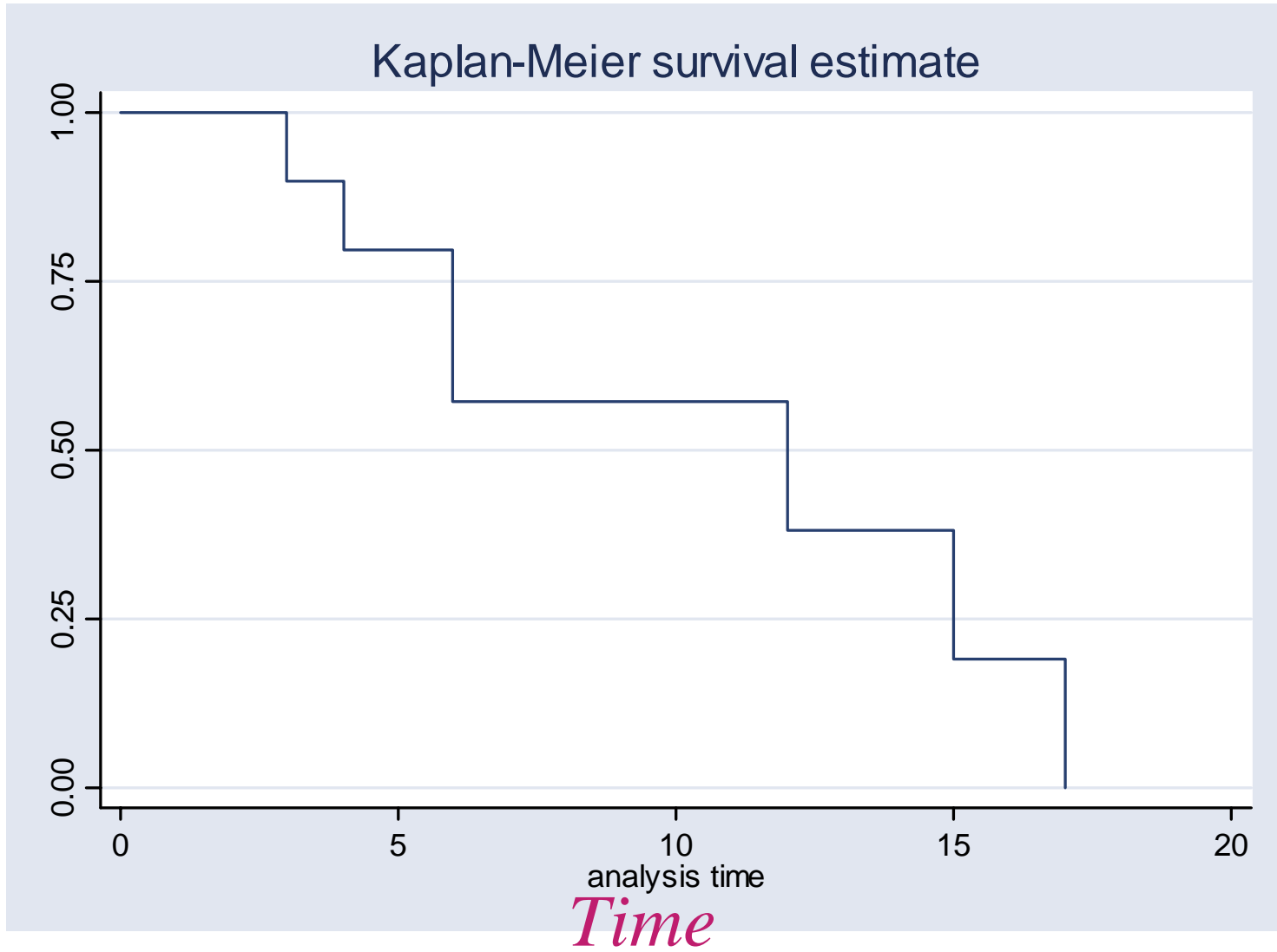
$t$	$S(t)$
3	0.9
4	0.8
6	0.57
12	0.38
15	0.19
17	0

# Kaplan-Meier Estimate

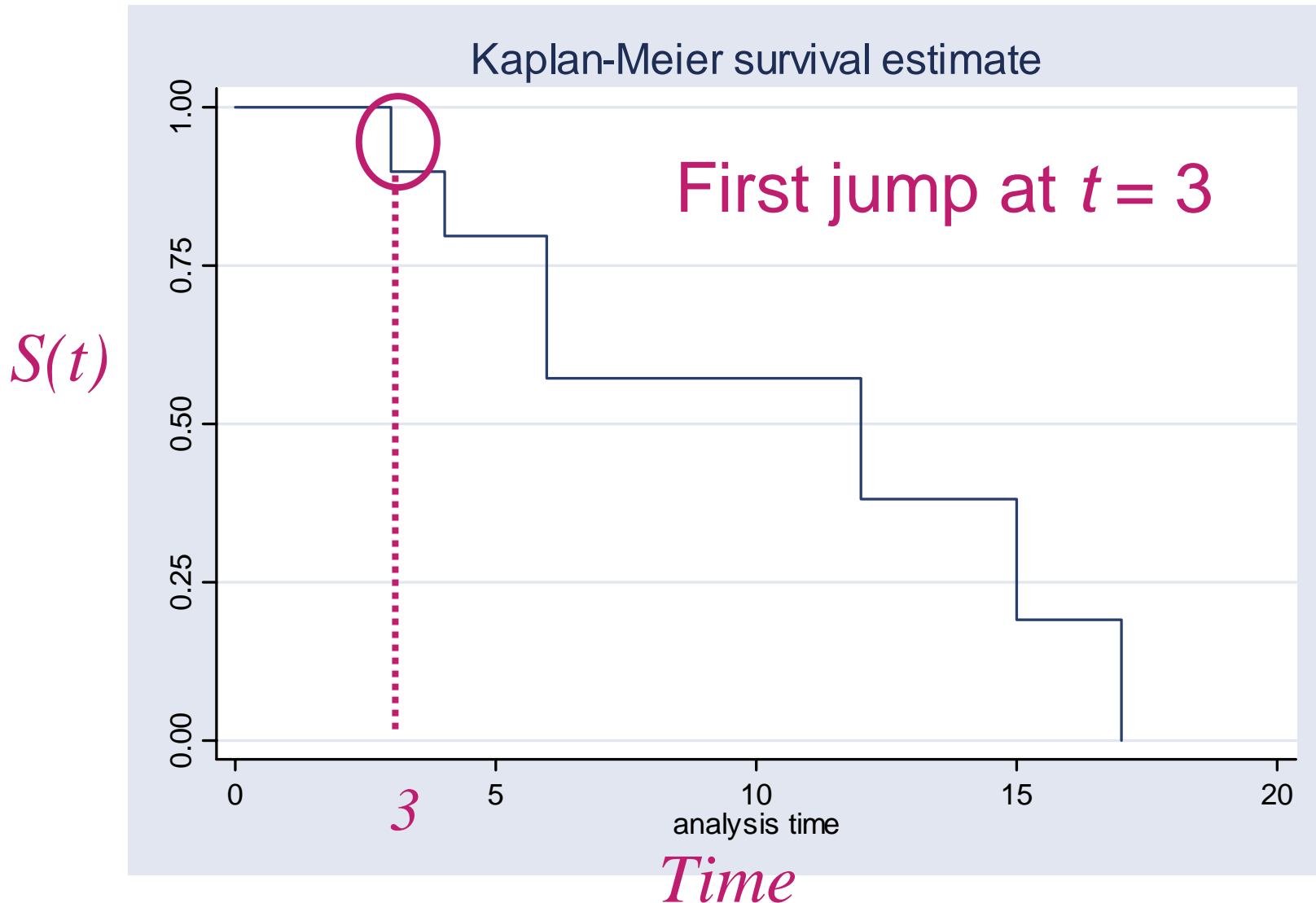
- ◆ Graph is a step function
- ◆ “Jumps” at each observed event time
- ◆ Nothing is assumed about curved shape between each observed event time

# Kaplan-Meier Estimate

$S(t)$

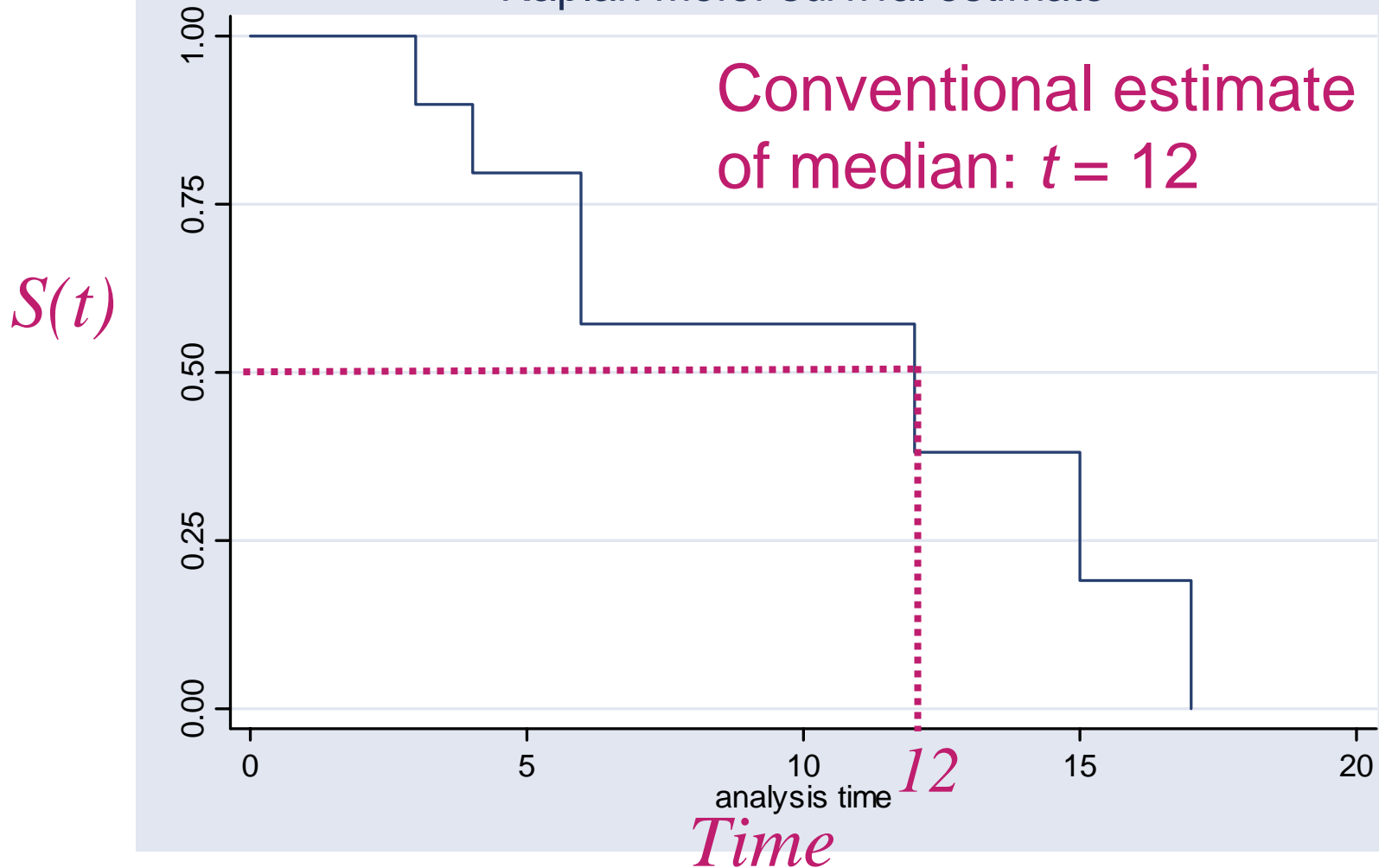


# Kaplan-Meier Estimate



# Kaplan-Meier Estimate

Kaplan-Meier survival estimate



# Kaplan-Meier Estimate

- ◆ Product limit estimate
  - Order survival times
  - Computed at observed events
  - Multiplying conditional probabilities

# Life Table Estimate

- ◆ Data is grouped into intervals of time
- ◆  $S(t)$  is an estimate of proportion surviving beyond the end of the interval

# Life Table Estimate

- ◆ Three key quantities needed in order to estimate via life table method:
  - $N_i$  = number at risk at start of interval  $i$
  - $d_i$  = number of deaths in interval  $i$
  - $c_i$  = number censored in interval  $i$

# Life Table Estimate

- ◆ From these three key quantities, we can calculate three more:
  - $N_i^*$  = effective # at risk =  $N_i - 1/2 c_i$   
in  $i^{\text{th}}$  interval
  - $Q_i$  = proportion who die in  $i^{\text{th}}$  interval  
=  $d_i/N_i^*$   
(interval specific death rate)

# Life Table Estimate


- ◆ From these three key quantities, we can calculate three more:
  - $P_i = 1 - Q_i =$  proportion of those alive at beginning of  $i^{\text{th}}$  interval who make it through the end

# Clinical Life Table

Year	$N_i$ #At risk	$d_i$ #Deaths	$C_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8			
[1-2)	170	38	12			
[2-3)	120	48	25			
[3-4)	47	19	10			
[4-5)	18	4	8			
[5-6)	6	1	3			
[6-	2					

# Clinical Life Table

Year	$N_i$ #At risk	$d_i$ #Deaths	$c_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8	196		
[1-2)						
[2-3)						
[3-4)						
[4-5)						
[5-6)						
[6-						

$200 - 1/2^*(8) = 196$ 


# Clinical Life Table


Year	$N_i$ #At risk	$d_i$ #Deaths	$C_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8	196	.11	
[1-2)						
[2-3)						
[3-4)						
[4-5)						
[5-6)						
[6-						

$$\frac{22}{196} = .11$$


# Clinical Life Table

Year	$N_i$ #At risk	$d_i$ #Deaths	$C_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8	196	.11	.89
[1-2)						
[2-3)						
[3-4)					$1 - .11$	$= .89$
[4-5)						
[5-6)						
[6-						



# Clinical Life Table

Year	$N_i$ #At risk	$d_i$ #Deaths	$C_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8	196	.11	.89
[1-2)	170					
[2-3)						
[3-4)						
[4-5)						
[5-6)						
[6-						

$200 - (22+8) = 170$ 


# Clinical Life Table

Year	$N_i$ #At risk	$d_i$ #Deaths	$C_i$ #Censored	$N_i^*$	$Q_i$	$P_i$
[0-1)	200	22	8	196	.11	.89
[1-2)	170	38	12	164	.23	.77
[2-3)	120	48	25	107.5	.45	.55
[3-4)	47	19	10	42	.45	.55
[4-5)	18	4	8	14	.29	.71
[5-6)	6	1	3	4.5	.22	.78
[6-	2					

# Clinical Life Table

- ◆ To estimate survival curve, start at  $S(0) = 1$  by convention
- ◆ Estimates of survival curve done at end of each interval:
  - Product of  $P_i$  for the interval, and  $S(t)$  for previous interval

# Clinical Life Table

<b>Year</b>	<b>P<sub>i</sub></b>	<b>S<sub>i</sub></b>
[0-1)	.89	$.89 * S(0) = .89$
[1-2)	.77	$.77 * S(1) =$ $.77 * .89 = .68$
[2-3)	.55	$.55 * S(2) = .38$
[3-4)	.55	.21
[4-5)	.71	.15
[5-6)	.78	.12

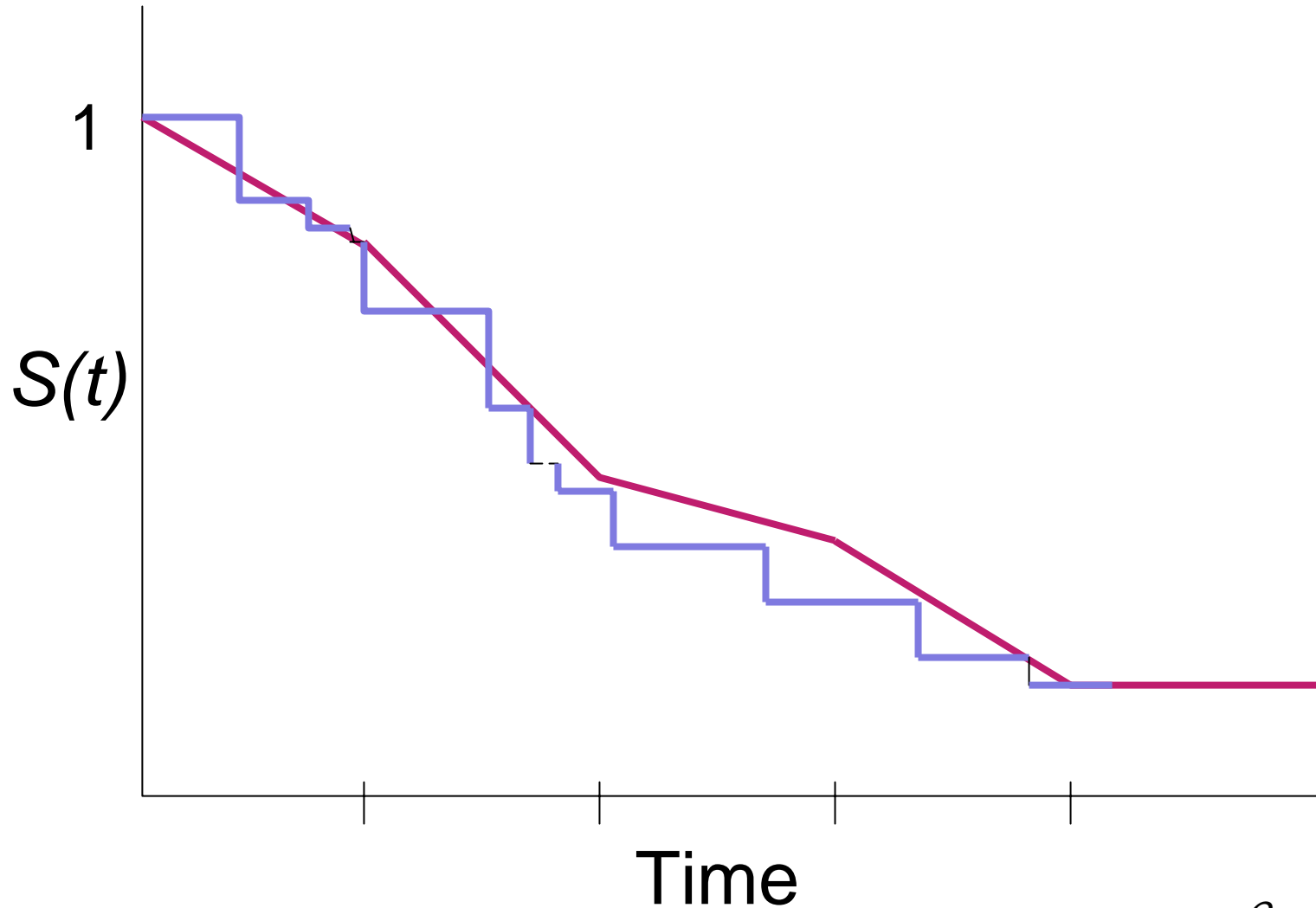
# 95% CI for the Four Year Survival Rate

- ◆  $.21 \pm 2 \times \text{standard error}$
- ◆  $.21 \pm 2 \times (.072)$   
(0.07, 0.35)

# 95% CI for the Four Year Survival Rate

- ◆ The standard error (.072) comes from Greenwood's formula
- ◆ We estimate that about 21% of individuals will survive at least four years
- ◆ The 95% CI is 7%–35%

# Life Table vs. Kaplan-Meier



# Life Table vs. Kaplan-Meier

- ◆ The difference between the Kaplan-Meier and the life-table is small if . . .
  - N is large
  - Intervals are small

# Big Assumption

- ◆ Independence of censoring and survival
- ◆ Those censored at time  $t$  have the same prognosis as those not censored at  $t$

# Big Assumption

- ◆ Examples of possible violations
  - Time to tumor—animal
  - Occupation—loss to follow up



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

# Section B

## *Practice Problems*

# Practice Problems

1. How does the Kaplan-Meier method use censored observations in its estimation process for construction of the “survival function?”

# Practice Problems

2. How can the median survival time for a cohort of subjects be estimated with the Kaplan-Meier curve?

# Practice Problems

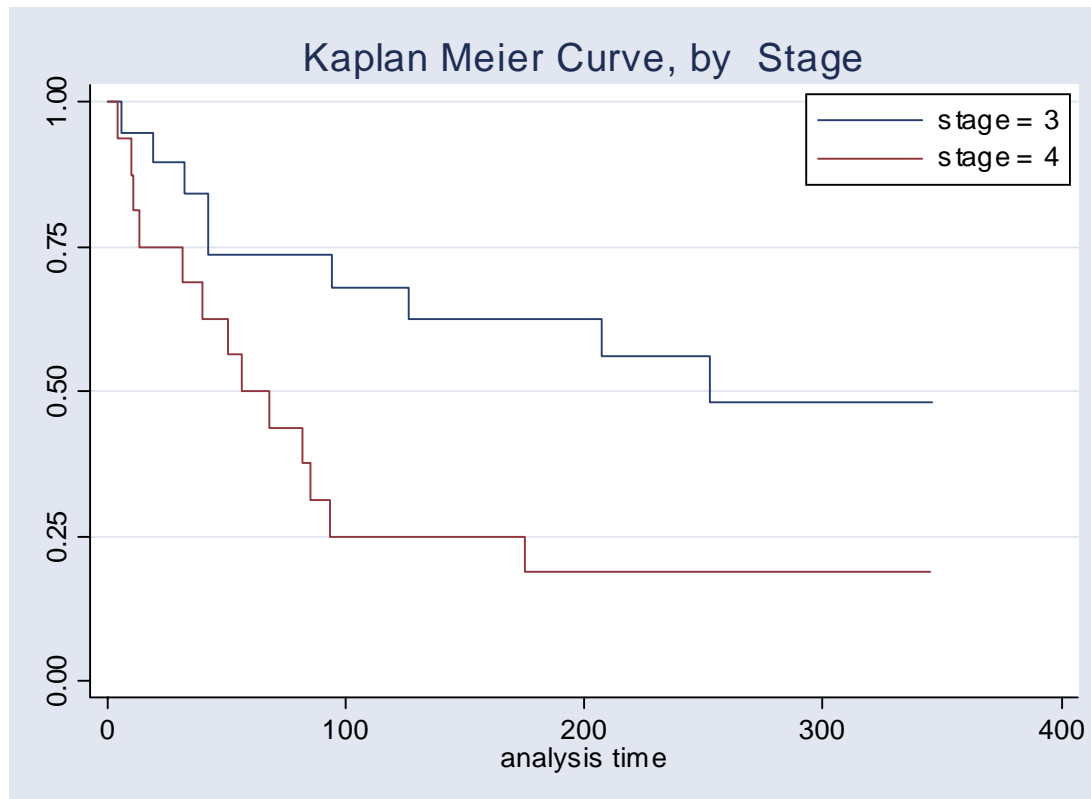
3. On the next slide, you will see two Kaplan-Meier curves graphed on the same set of axes. The curves are estimates of the survival functions for two groups of patients with diffuse histiocytic lymphoma:

- 19 patients with stage three cancer
- 16 patients with stage four cancer

Patients were followed for a year after their diagnosis until death or censoring.

# Practice Problems

3. Here are the survival curves.



# Practice Problems

3. a) Which group has the better one-year survival prognosis based on the curves?
- b) Estimate the median survival time for each of the groups?
- c) Based only on the curves, can you ascertain whether the patient's with the largest times in each of the groups were censored?

# Practice Problems

4. What is the primary difference between the Kaplan-Meier approach and the life table approach to estimating the survival function?



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

# Section B

*Practice Problem Solutions*

# Question

1. How does the Kaplan-Meier method use censored observations in its estimation process for construction the “survival function?”

# Answer

- ◆ Subjects that are censored are considered to be “at risk” for having the event of interest until censoring occurs—these observations help “inform” the survival curve via their usage in the denominators for time specific survival percentage estimates.

# Question

2. How can the median survival time for a cohort of subjects be estimated with the Kaplan-Meier curve?

# Answer

- ◆ The median survival time is estimated by the time at which 50% of the cohort being studied are still event free.
  - If the Kaplan-Meier curve does not hit 50% exactly, the convention is to use the first event time where the curve drops below 50%.

# Question

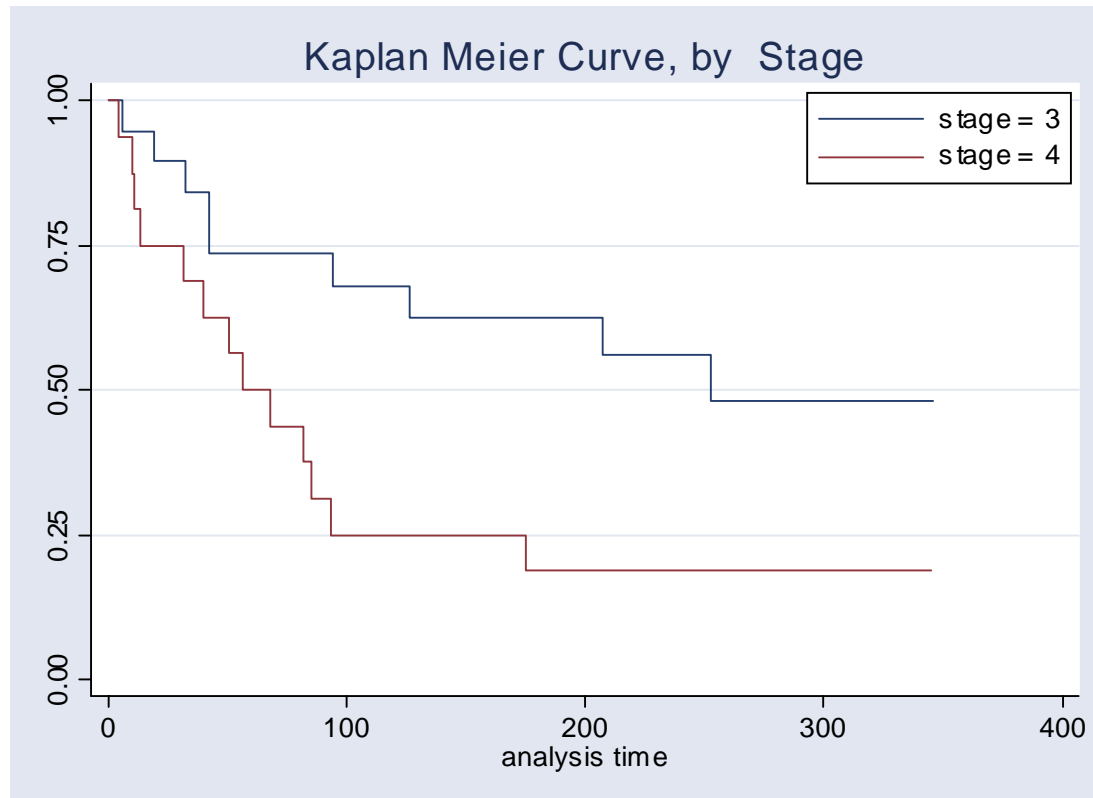
3. On the next slide, you will see two Kaplan-Meier curves graphed on the same set of axes. The curves are estimates of the survival functions for two groups of patients with diffuse histiocytic lymphoma:

- 19 patients with stage three cancer
- 16 patients with stage four cancer

Patients were followed for a year after their diagnosis until death or censoring.

# Question

3. Here are the survival curves.

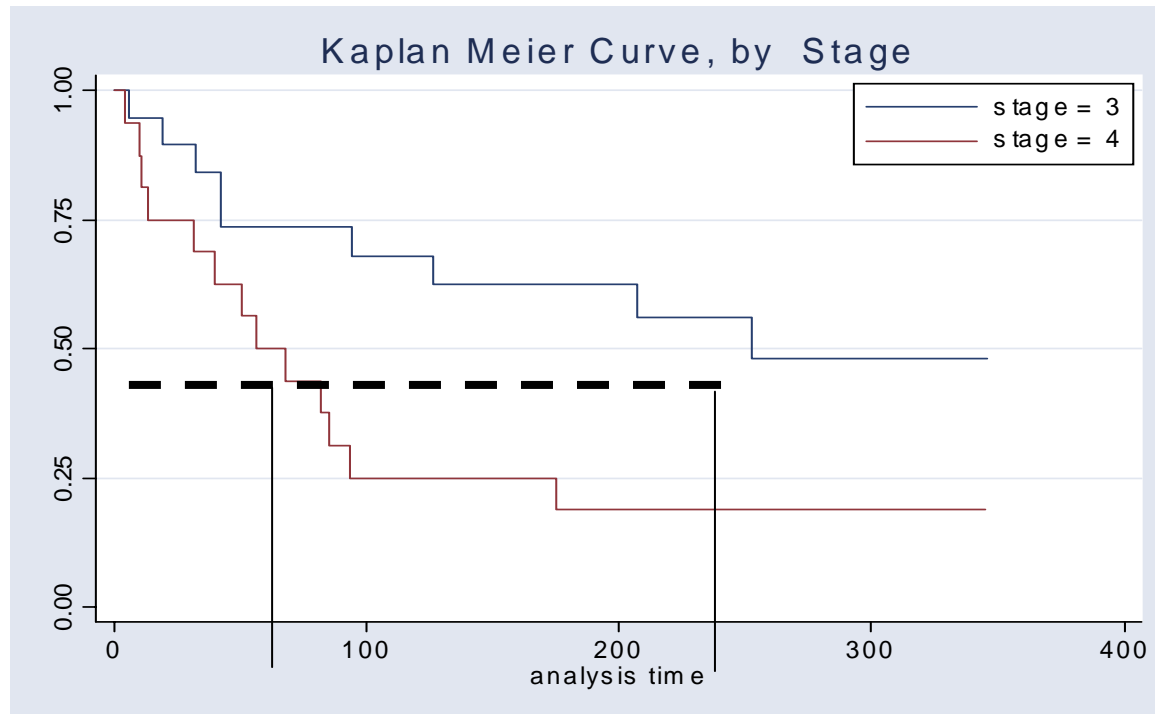


# Answer

3. a.) Which group has the better one-year survival prognosis based on the curves?
  - ◆ As the curve for the stage three group is “higher” than the curve for the stage four group across the entire time interval, stage three patients in this sample have a better survival prognosis.

# Question

3. b) Estimate the median survival time for each of the groups.



# Answer

3. b) Estimate the median survival time for each of the groups.
  - ◆ By drawing lines on the graph to show the time at which curve “hits” 50% or first drops below 50%, the estimated median survival times are 75 days for stage four patients, and 250 days for stage three patients.

# Answer

3. c) Based only on the curves, can you ascertain whether the patient's with the largest times in each of the groups were censored?
- ◆ As neither curve drops to 0% by the end of the time interval (study time frame of 365 days), this indicates that the patients with the largest times were censored observations in both groups.

# Answer

4. What is the primary difference between the Kaplan-Meier approach and the life table approach to estimating the survival function?
  - ◆ The Kaplan-Meier curve estimates a change in the survival function at every observed event time—the life table estimates changes at specific intervals of time.



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

## Section C

### *Statistically Comparing Survival Curves*

# Comparing Survival Curves

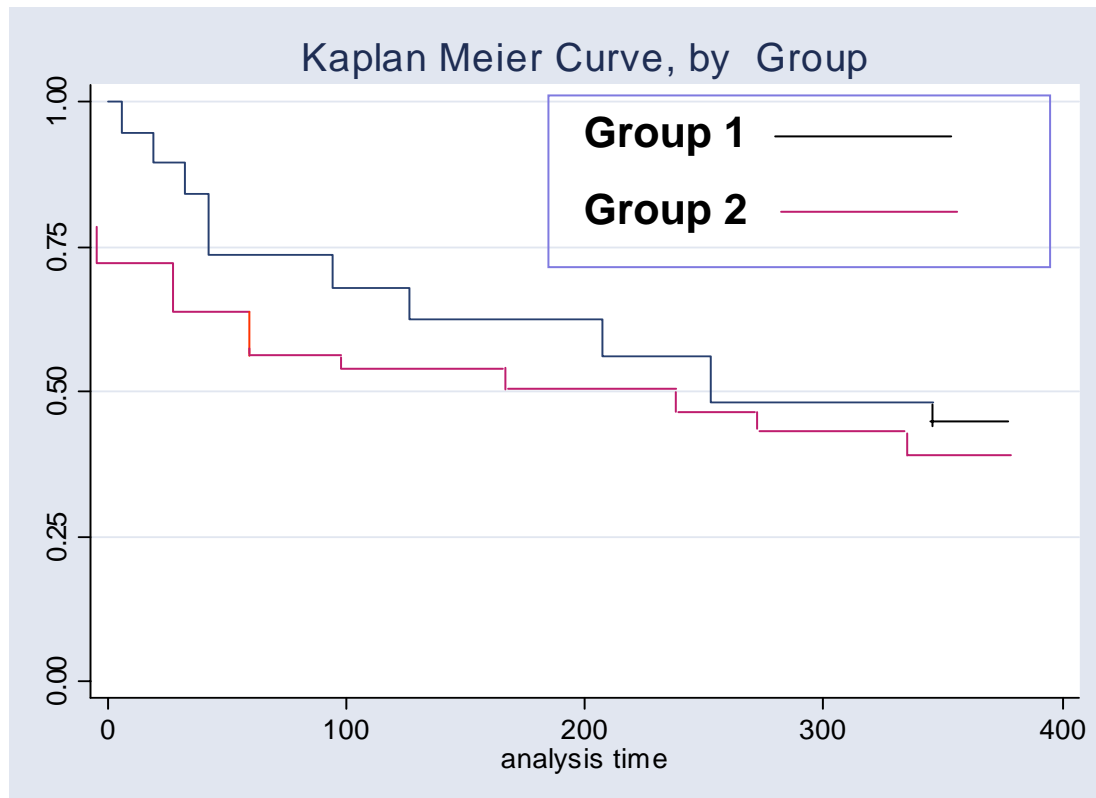
- ◆ Common statistical tests
  - Generalized Wilcoxon (Breslow, Gehan)
  - Logrank

# Comparing Survival Curves

- ◆ Both compare survival curves across multiple time points to answer the question—“is overall survival different between any of the groups?”
  - $H_0$ : No difference in  $S(t)$
  - $H_a$ : Difference in  $S(t)$

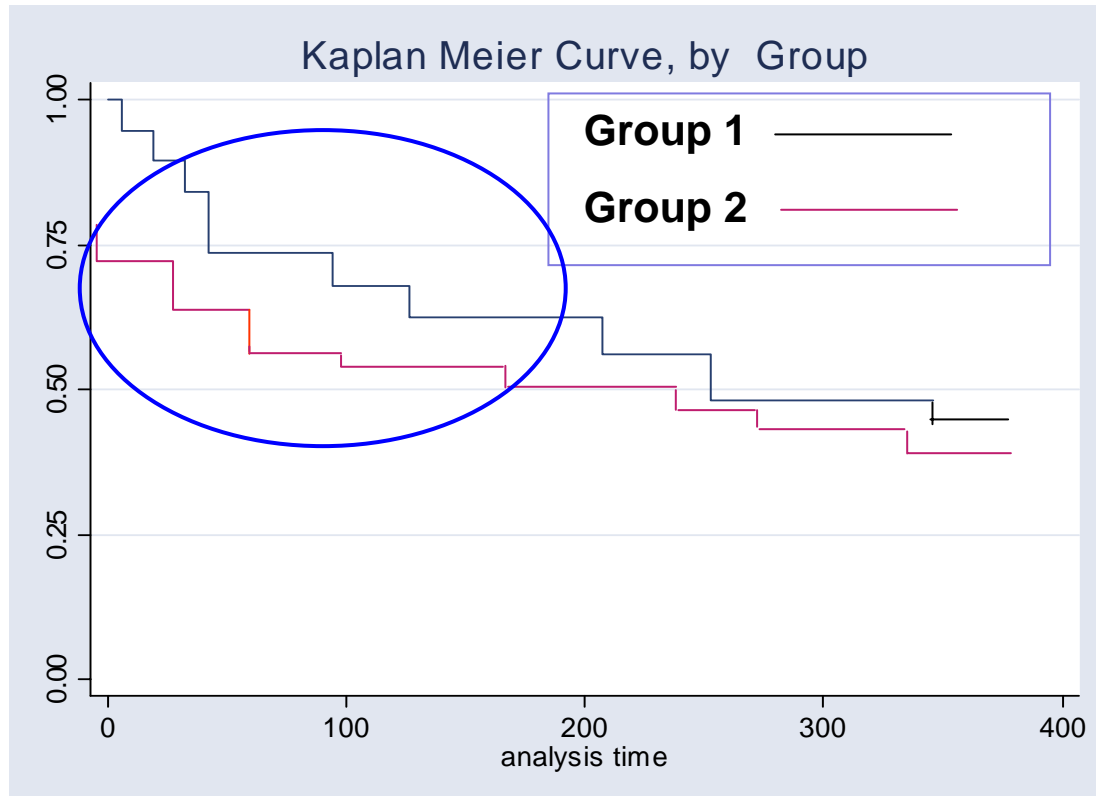
# Comparing Survival Curves

- ◆ Wilcoxon (Breslow, Gehan) more sensitive to early survival differences



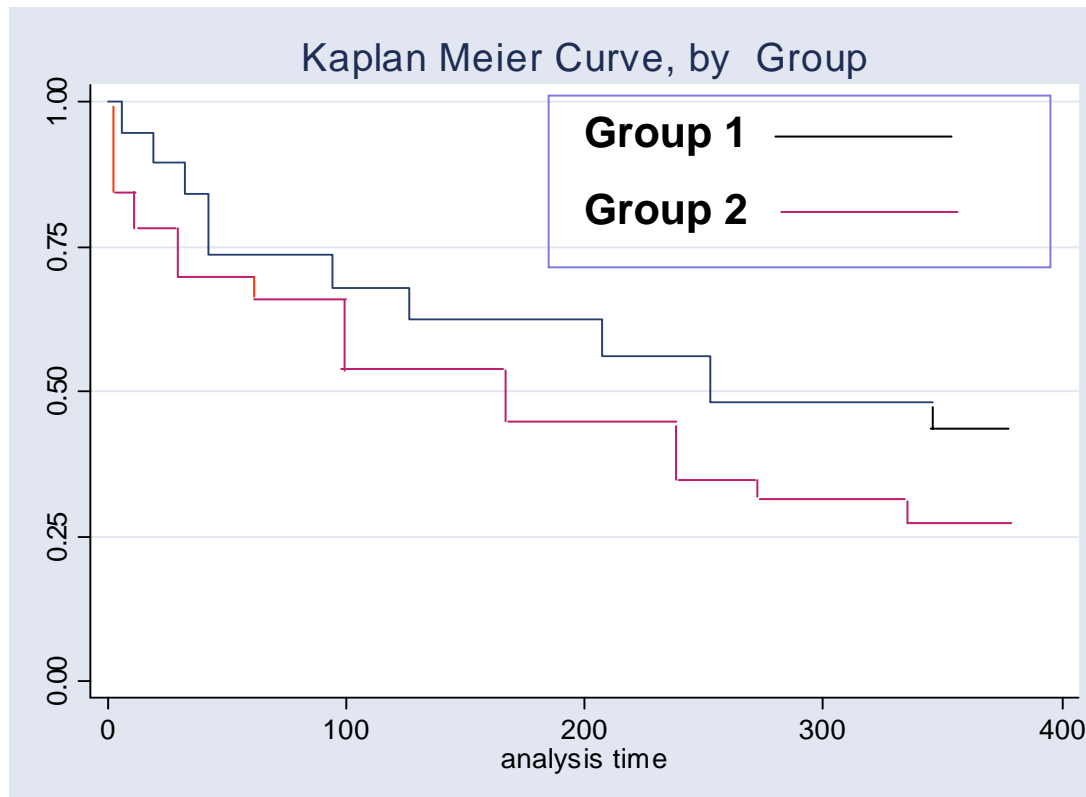
# Comparing Survival Curves

- ◆ Wilcoxon (Breslow, Gehan) more sensitive to early survival differences



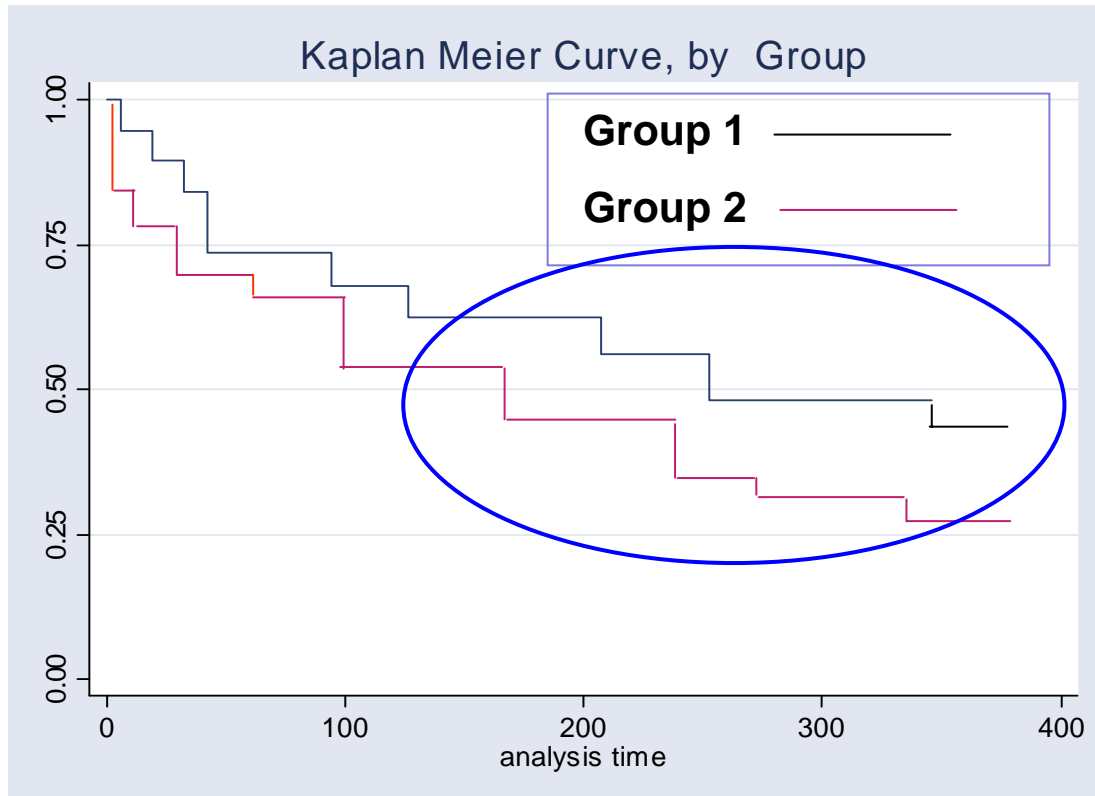
# Comparing Survival Curves

- ◆ Logrank more sensitive to later survival differences



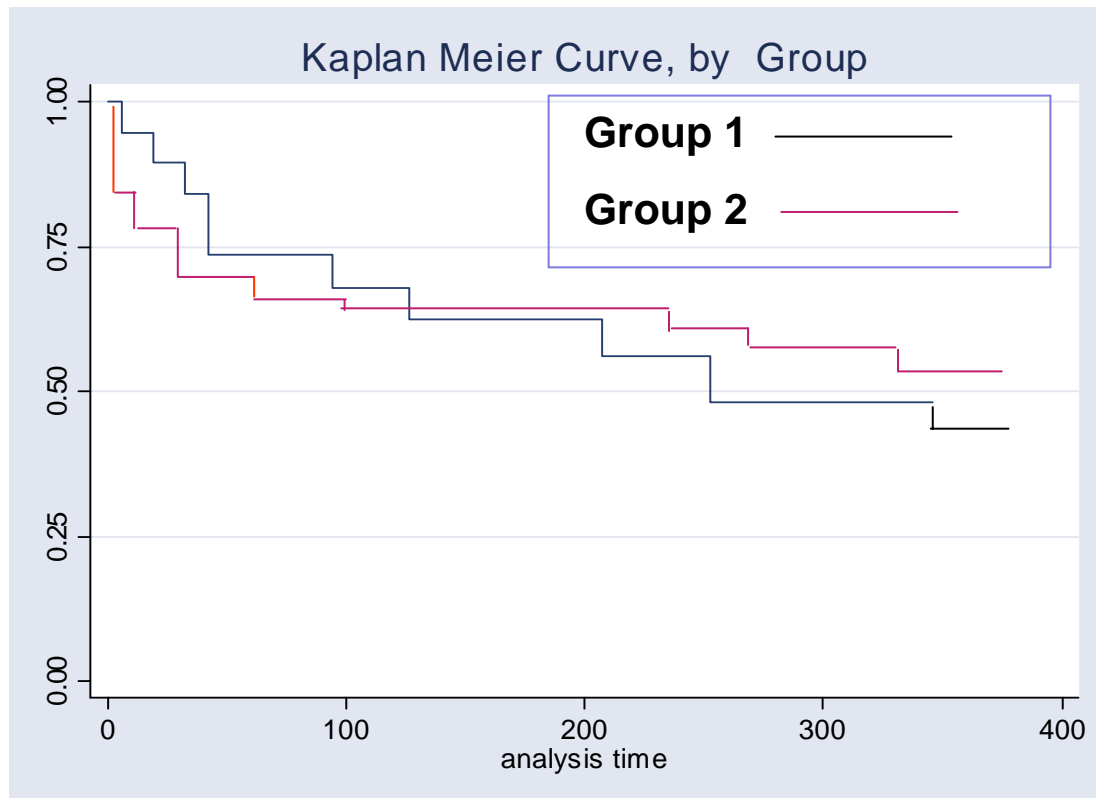
# Comparing Survival Curves

- ◆ Logrank more sensitive to later survival differences



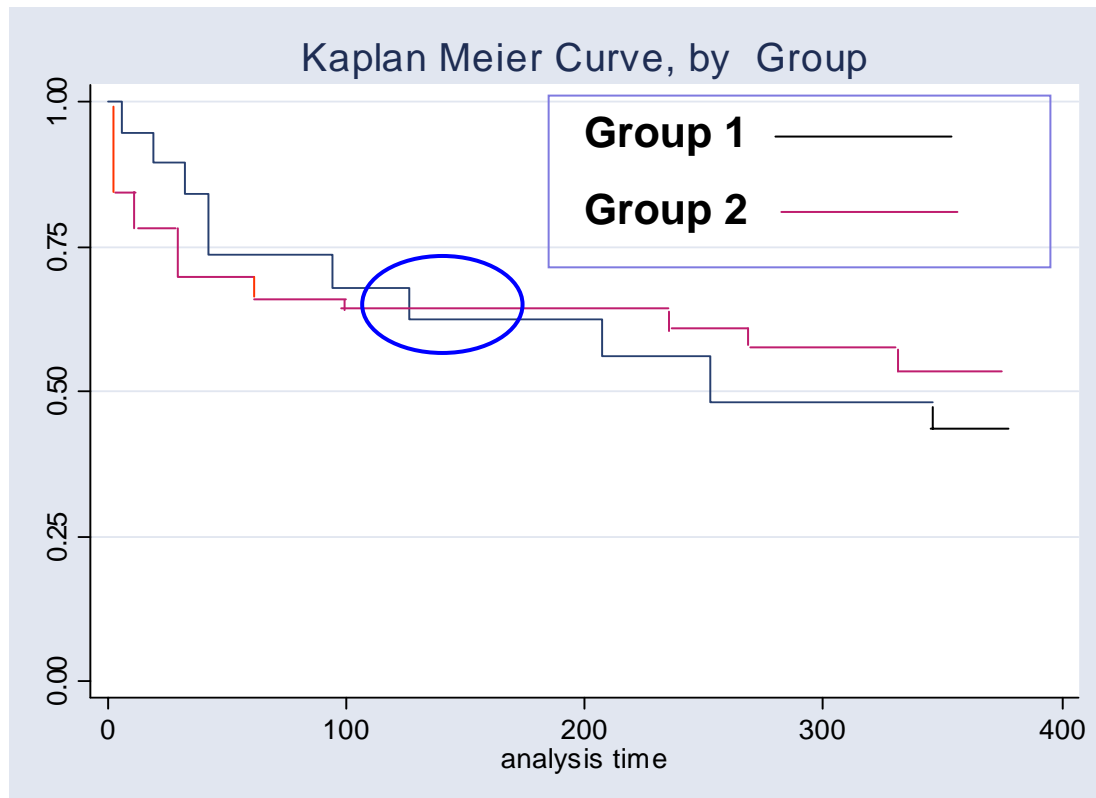
# Comparing Survival Curves

- ◆ Neither test very good if curves “crossover”



# Comparing Survival Curves

- ◆ Neither test very good if curves “crossover”



# Examples of Logrank and Breslow-Gehan Test

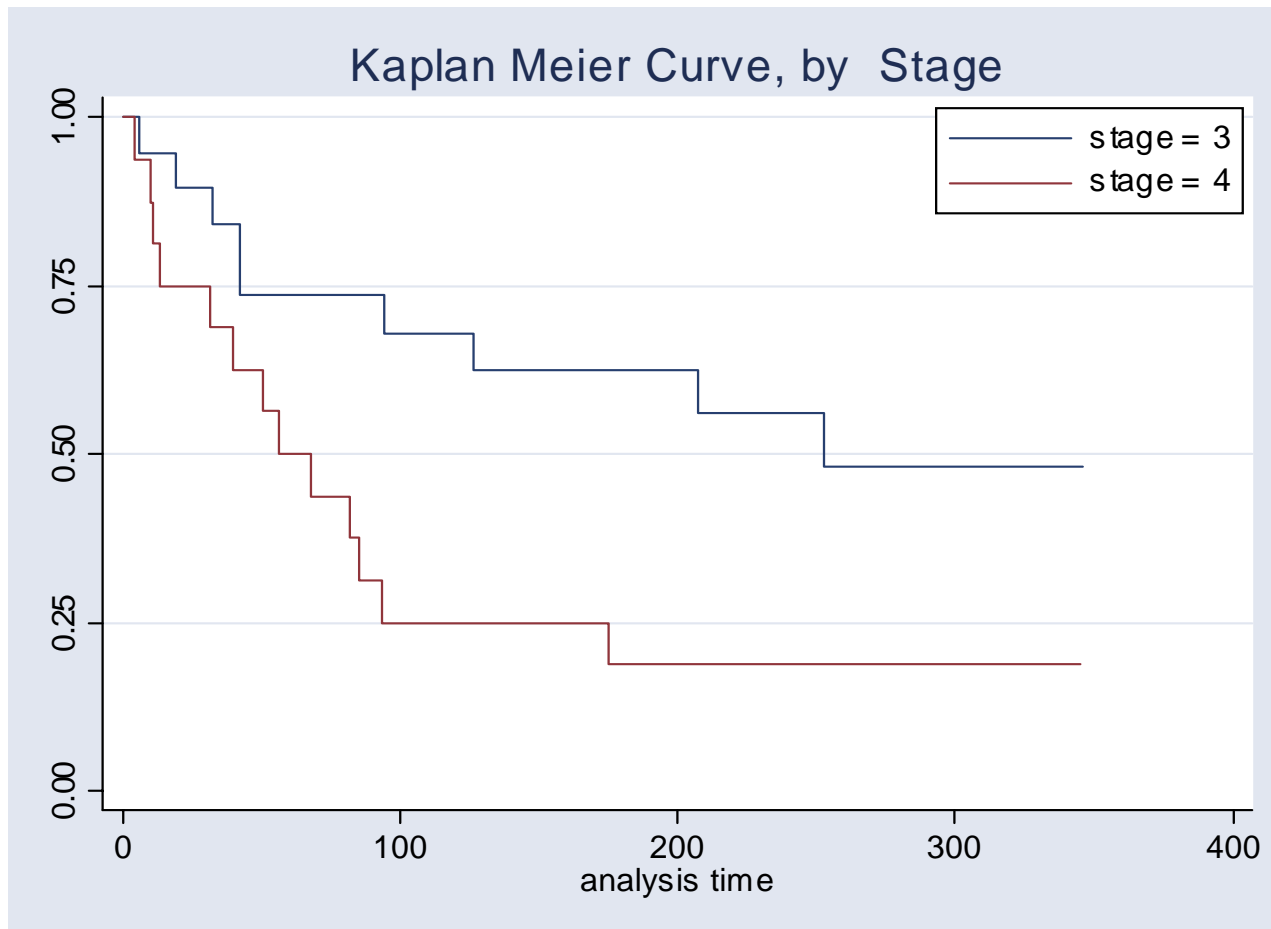
- ◆ Survival times (measured in days) for two groups of patients with diffuse histiocytic lymphoma
  - Group A ( $n = 19$ ) has stage-3 cancer
  - Group B ( $n = 16$ ) has stage-4 cancer

# Examples of Logrank and Breslow-Gehan Test

- ◆ Nine of nineteen observations censored in group A
- ◆ Three of sixteen observations censored in group B
- ◆ We want to know if stage of cancer is associated with survival

# Kaplan Meier Curves

- ◆ Survival (event = death) by stage of cancer



# Testing Stage/Survival Relationship

- ◆ Log-rank results:
  - $p = .02$
- ◆ Breslow/Wilcoxon/Gehan results:
  - $p = .03$