PBIO 504
Experimental Design & Analysis

Survival Analysis

Nonparametric Analysis: Survival data analysis

- In many biomedical research problems, the outcome is an event: Death, re-infection, or a particular clinical condition.
- The time until the event occurs is referred to as "survival" time.
- Survival time can be viewed as a random variable whose distribution may differ between treatment groups.

- Examples:
  - Time from diagnosis of cancer until death
  - Time from HIV infection to development of AIDS

Sometimes, the event is a positive outcome, e.g.,

- Time from treatment with antibiotics to being uninfected

Why not treat the event as a binary variable?

Whether the event occurred within a specified period of time can be viewed as a binary random variable

- For example, suppose there is interest in the survival rate after diagnosis with lung cancer. Death within 3 year can be viewed as a binary random variable.

Why not treat the event as a binary variable?

- However, there are two reasons why survival analysis is advantageous over this approach

1. The follow-up time of the study subjects may vary due to loss to follow-up or later recruitment. (*e.g.*, in the above example, not all patients will be observed for the entire three years).

2. Treating survival as binary entails a loss of information (*e.g.*, the binary summary tells you whether someone died within 3 years, but does not tell you when in that period they died.)

Example of how to estimate survival functions from uncensored data

Example data: Survival times post treatment for terminal lung cancer (in months):

1,3,4,4,5,5,5,5,6,6,6,6,6,6,7,7,7,7,8,8

Month	Numb. in risk set at start of month	Numb. dying during month	Proportion (of all people) dying during that month	Proportion (of all those starting the month) who died in that month	Proportion who survived beyond that month
1	20	1	1/20 = .05	1/20 = .05	19/20 = .95
2	19	0	0/20 = 0	0/19 = 0	19/20 = .95
3					
4					
5					
6					
7					
8					

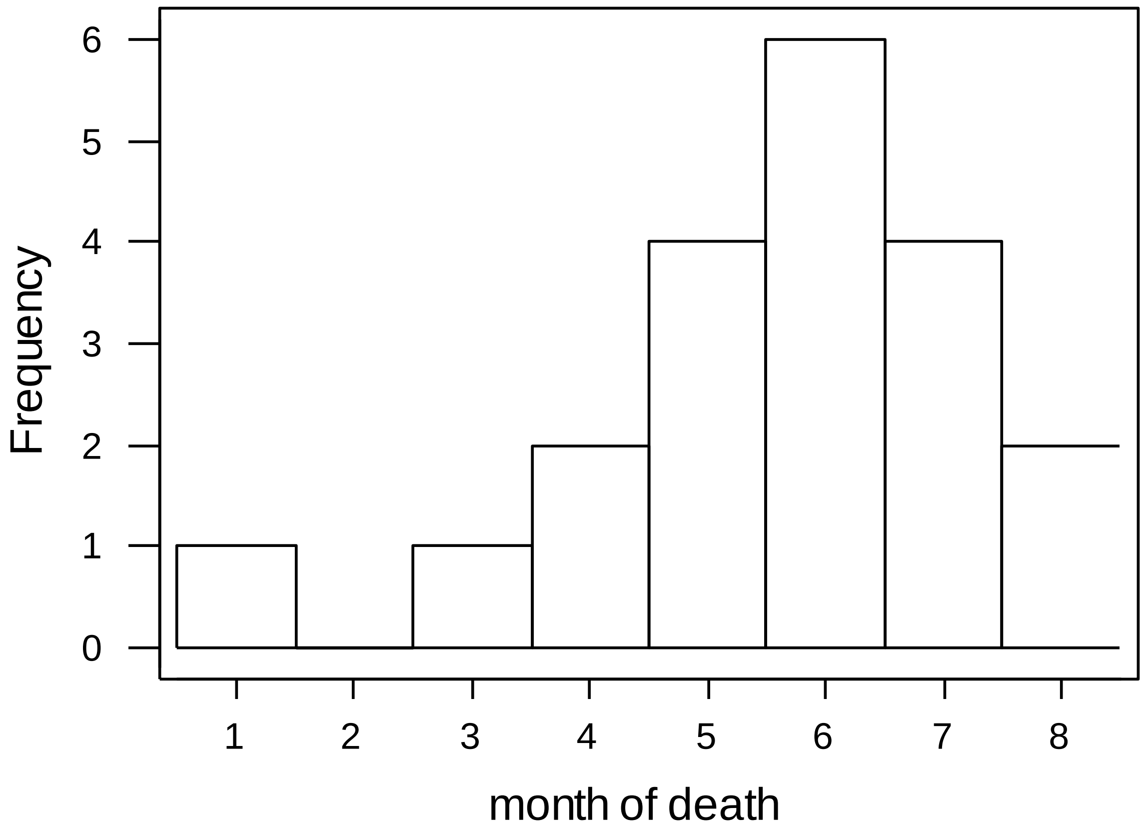
Example of how to estimate survival functions from uncensored data

Example data: Survival times post treatment for terminal lung cancer (in months):

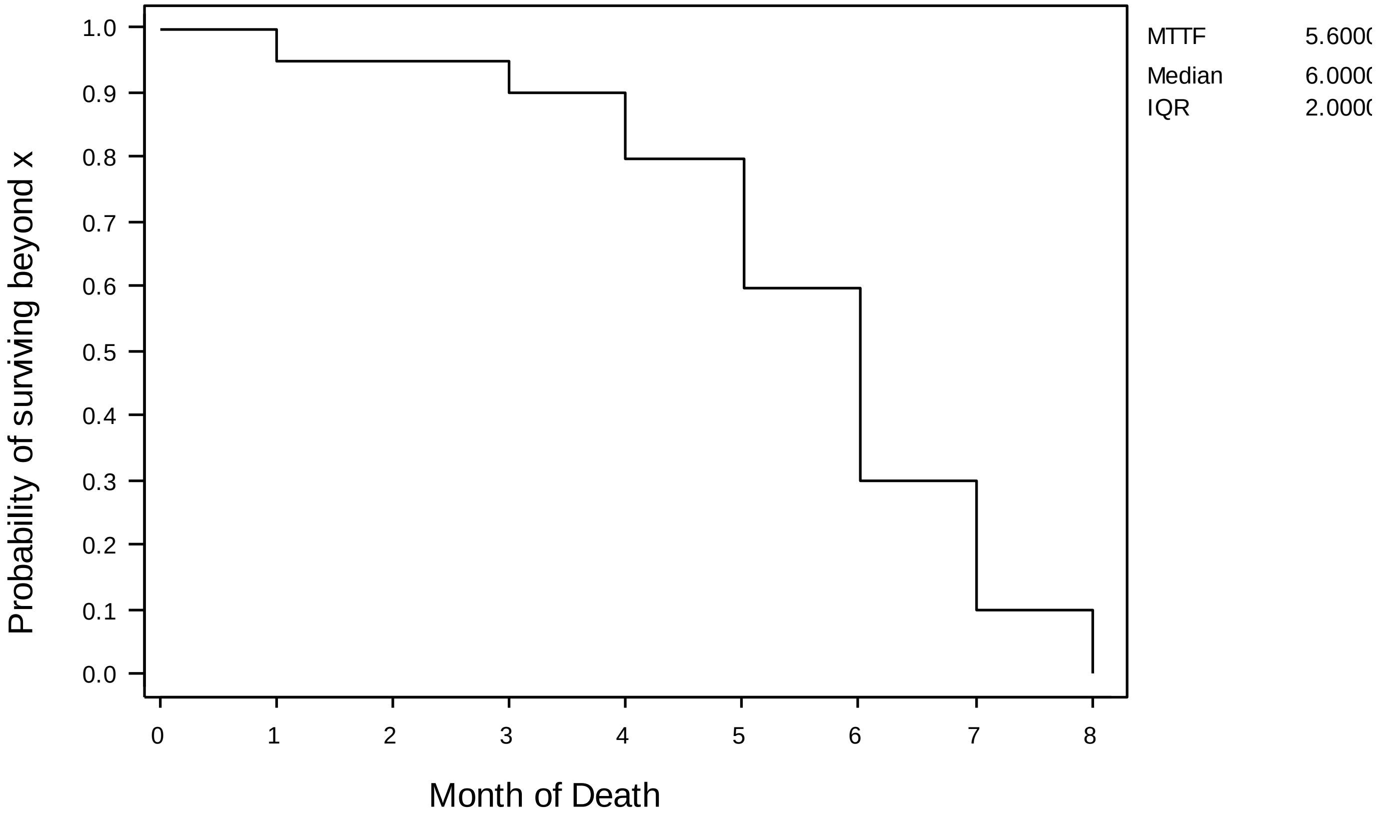
1,3,4,4,5,5,5,5,6,6,6,6,6,6,7,7,7,7,8,8

Month	Numb. at risk at start of month	Numb. dying during month	Proportion (of all people) dying during that month	Proportion (of all those starting the month) who died in that month	Proportion who survived beyond that month
1	20	1	1/20 = .05	1/20 = .05	19/20 = .95
2	19	0	0/20 = 0	0/19 = 0	19/20 = .95
3	19	1	1/20 = .05	1/19 = .053	18/20 = .90
4	18	2	2/20 = .10	2/18 = .11	16/20 = .80
5	16	4	4/20 = .20	4/16 = .25	12/20 = .60
6	12	6	6/20 = .30	6/12 = .50	6/20 = .30
7	6	4	4/20 = .20	4/6 = .67	2/20 = .10
8	2	2	2/20 = .10	2/2 = 1.00	0/20 = 0

Based on the table, the PDF can be estimated as :



The Survival Function can be estimated as:



Common case: censored survival data

- However, in reality, when the study ended, someone had been followed for 8 months, and could be still alive.
- In this case, we do not know the exact time until death for that subject, but we know the time until death must exceed 8 months.
- This is referred to as *right censored* data and this introduces some complications in the analysis.
- Often many of the study subjects in our research do not have the event of interest (e.g., death), before the study is over.

Example of right censored data:

Survival times post treatment for advanced lung cancer (in months): 1,2,2*,3,3,3*,4,4*, 5*

(where an asterisk means that the patient did not die but was censored at that point)

In other words, when the study ended, one person had been followed for 2 months, and was still alive at that time, but we do not know when he died

Given these data, what would be a good estimate of the probability of surviving beyond 4 months?

Kaplan-Meier Approach

- Kaplan and Meier developed an approach based on the following insights:

  Even though there is no direct way to estimate the survival function:

1. We can estimate the hazard function by only using those at risk for dying each month, and

2. We can estimate the survival function from our estimates of hazard function.

Specifically:

Prob of surviving beyond month t =

(Prob of surviving to start of month t) X

(1-Hazard for month t)

Hazard: if survival to time t, the prob to die at the next moment

Example

Data: 1,2,2*,3,3,3*,3*,4,4*,5*

As of the time of analysis

Table:

Month	Numb. at risk at start of month	Numb. dying during month	Proportion (of all those starting the month) who died in that month (Hazard)	Proportion who survived beyond that month
1	10	1	1/10	9/10 = .90
2	9	1	1/9	(.90)(1-1/9) = .80
3	7	2	2/7	(.80)(1-2/7) = .64
4	3	1	1/3	(.64)(1-1/3)=.43
5				

Can you fill in the rest?!

Example

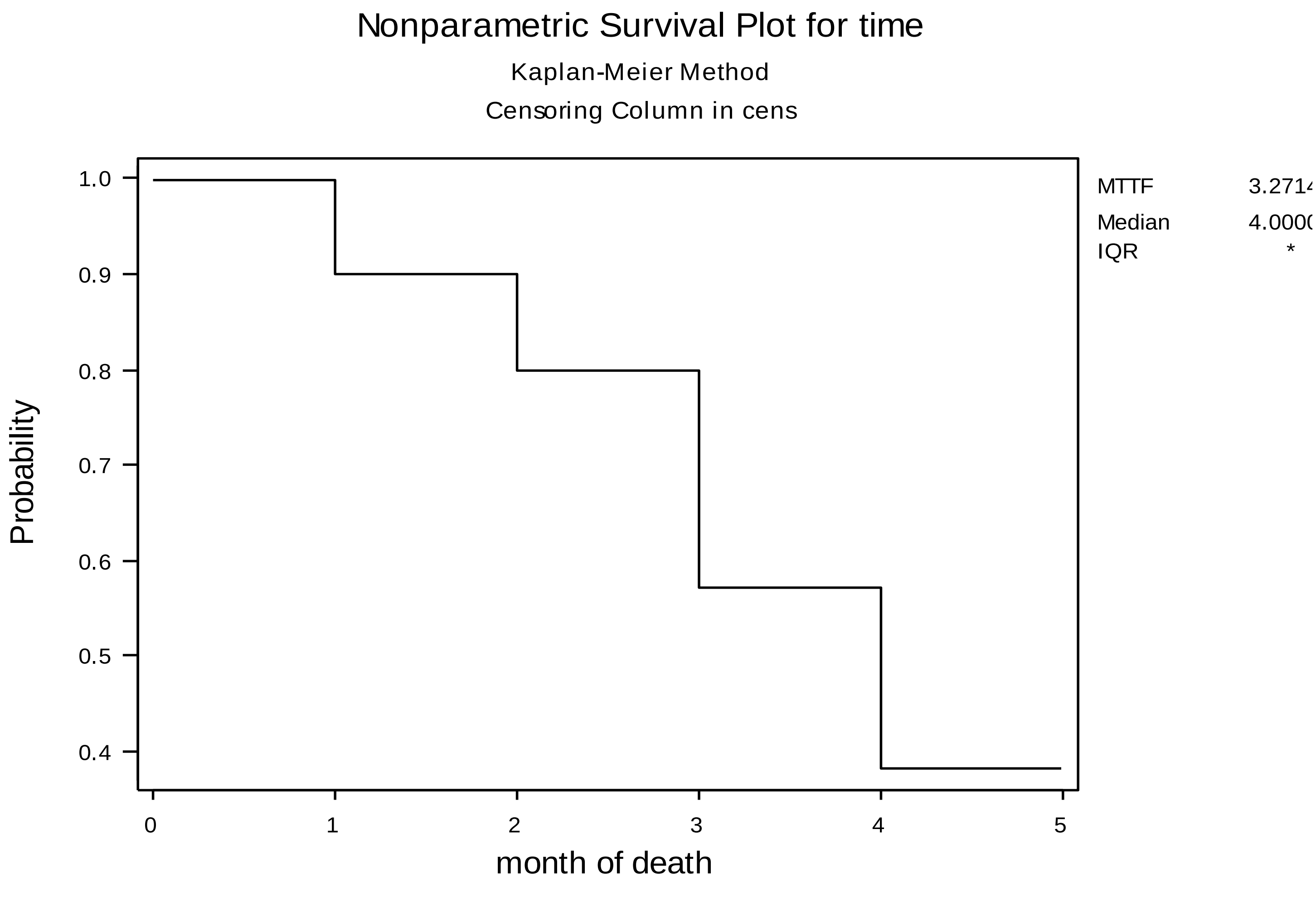
Data: 1,2,2*,3,3,3*,3*,4,4*,5*

Observe the times and list them in order

Table:

Month	Numb. in risk set at start of month	Numb. dying during month	Proportion (of all those starting the month) who died in that month (Hazard)	Proportion who survived beyond that month
1	10	1	1/10	9/10 = .90
2	9	1	1/9	(.90)(1-1/9) = .80
3	7	2	2/7	(.80)(1-2/7) = .64
4	3	1	1/3	(.64)(1-1/3)=.43
5	1	0	0/1	(0.43)(1-0)=0.43

Resulting (KM or Product Limit) Estimate of the Survival Distribution (Kaplan-Meier plot):



Statistical Inference about Survival Functions. P-values

- Suppose you have two treatments, and you want to compare the survival functions for each treatment.
- Example:
- Survival times for treatment A: 1,2,2*,3,3,3*,3*,4,4*,5*

  Survival times for treatment B: 2, 4, 4*, 5,5,5*,7,7
- Is the survival significantly better with treatment B?

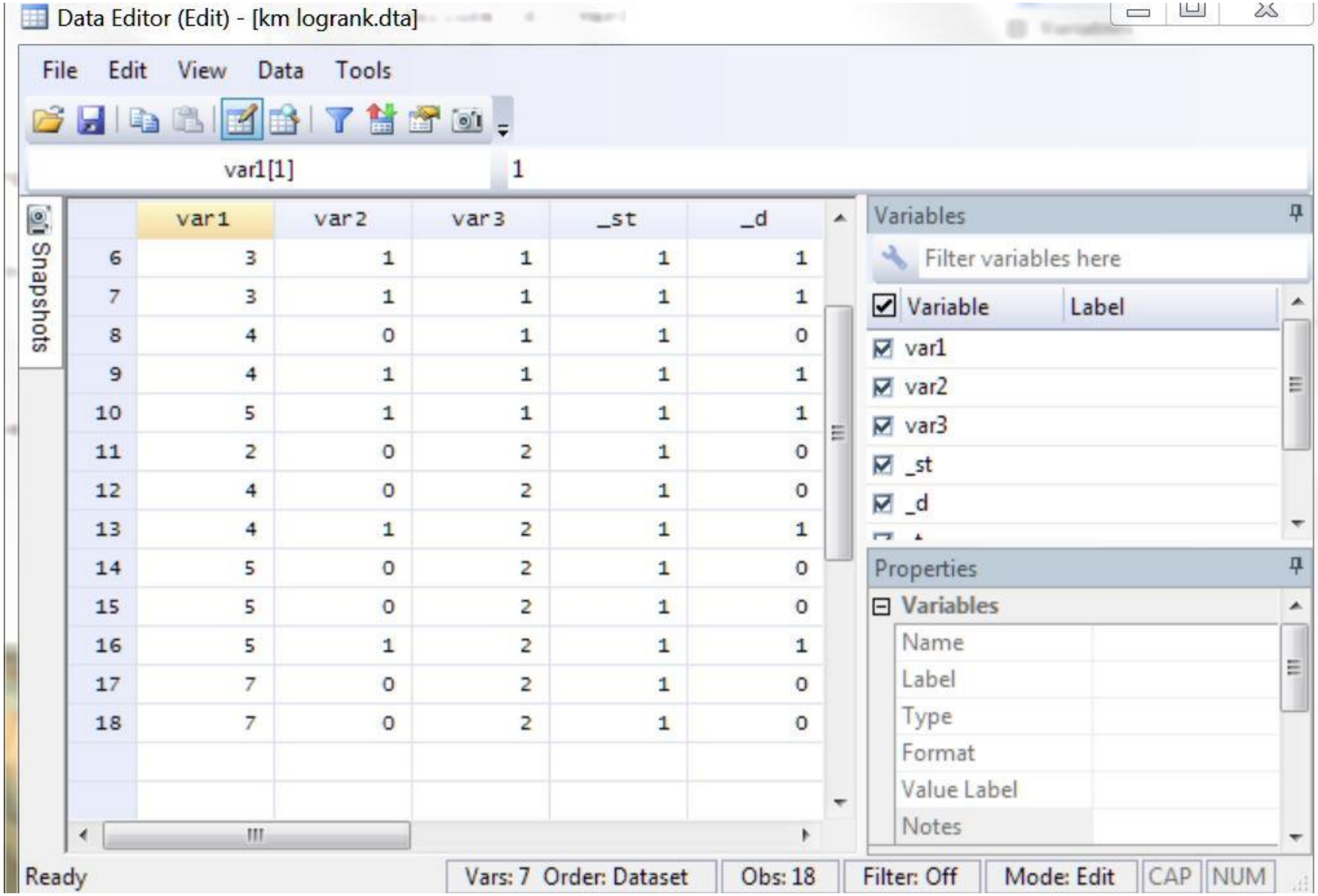
P-value

- **Null Hypothesis: The true survival functions are equal**
  - **Ho: Survival Function (A) = Survival Function (B)**
- **Most widely used statistical test:**
  - **Log Rank Test (p=.026 for above data)**

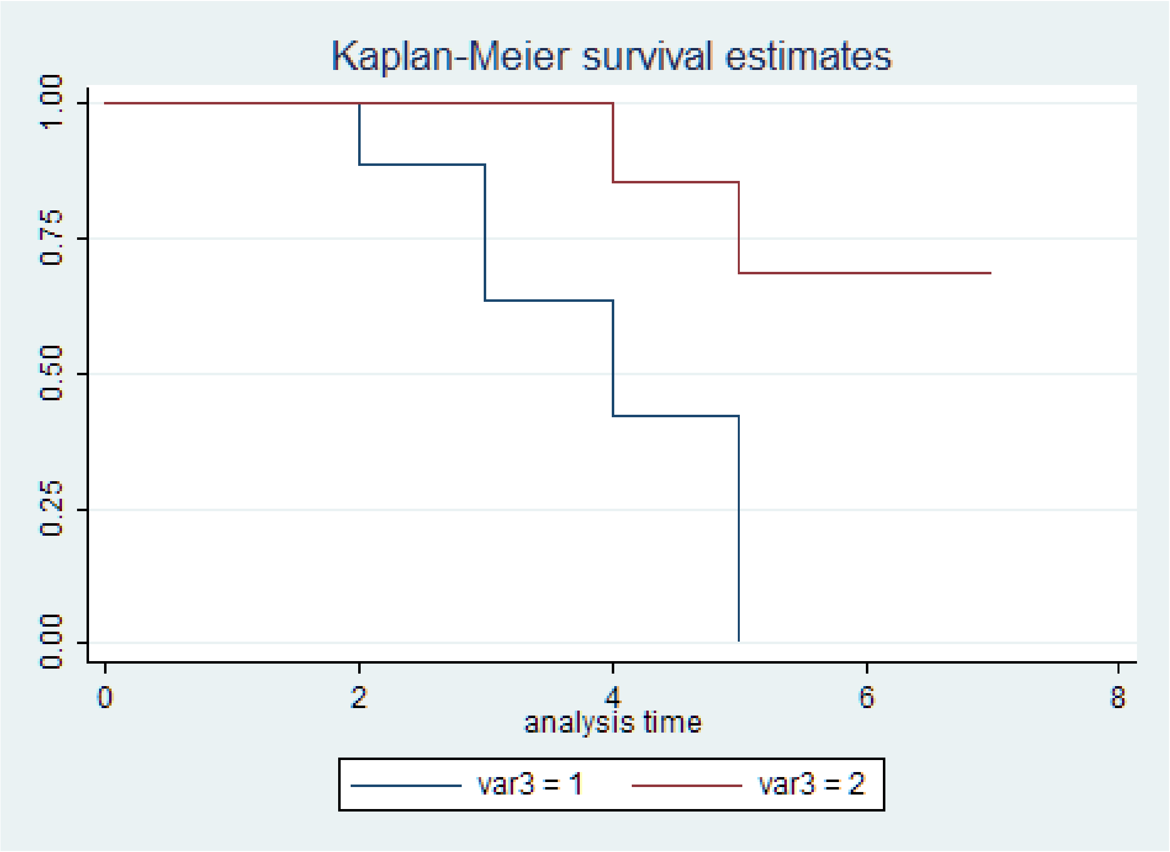
Import the data into STATA using data editor

Then go to Survival Analysis – Setup and Utilities –

- Declare Data to be Survival Time data

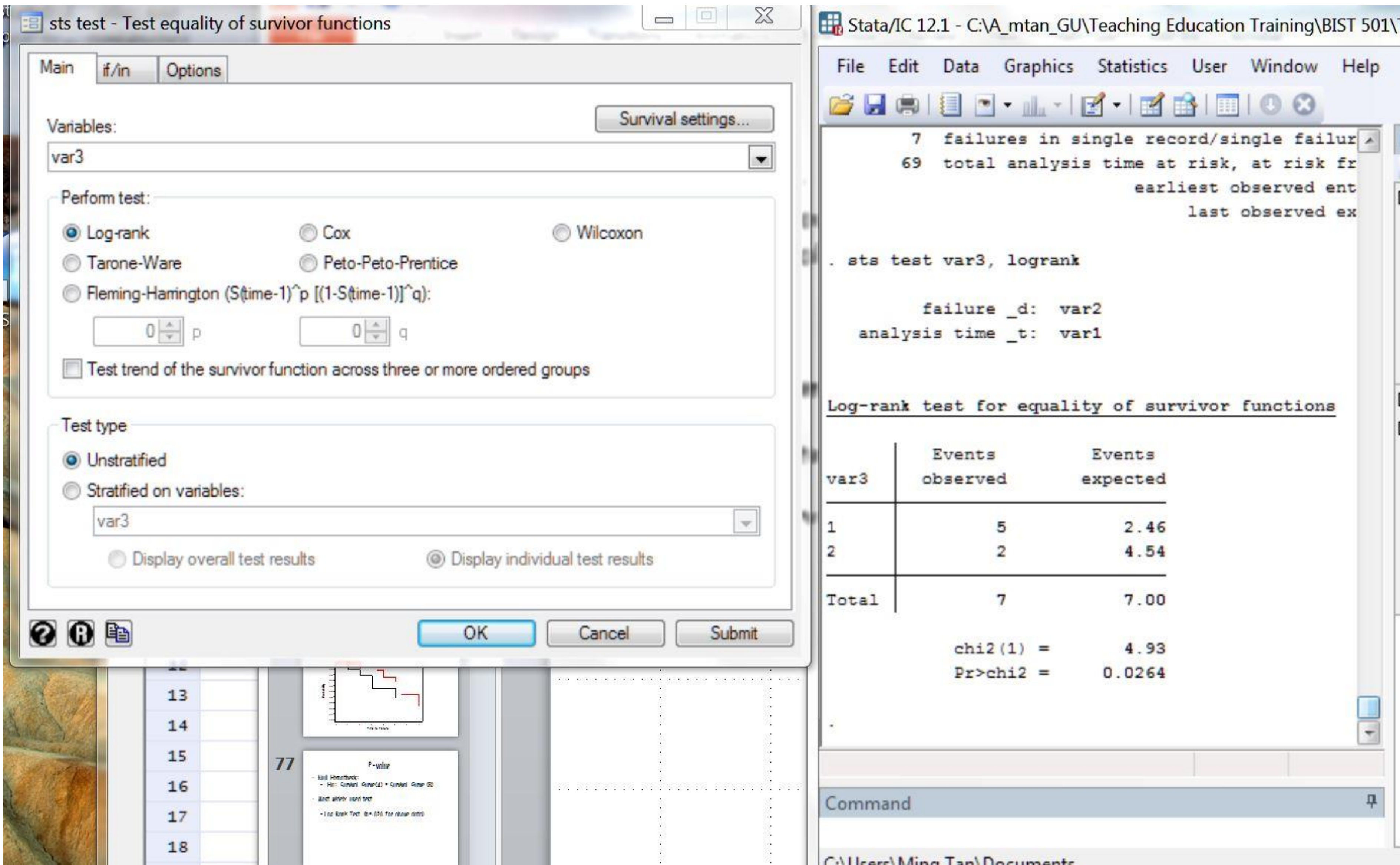


Statistical Inference about Survival Functions using the Logrank test



Logrank test to compare two survival functions

- Ho: Survival Function (A) = Survival Function (B)



Practice Example

- The following are survival times in months since diagnosis for 10 AIDS patients suffering from concomitant esophageal candidiasis. Censored observations are denoted by (*).

  Data: 1, 1, 1, 1*, 2, 5*, 8*, 9, 10*, 12*

- How many deaths were observed for these 10 patients?

- Plot the Kaplan Meier Survival Curve using the table on the next slide. This table has to be continued up to 12 months (12 rows).

Practice Example

Table:

Month	Numb. in risk set at start of month	Numb. dying during month	Proportion (of all those starting the month) who died in that month (Hazard)	Proportion who survived beyond that month
1	10	3	3/10	7/10 = .70
2	6	1	1/6	(.70)(1-1/6) = .58
3	5	0	0/5	(.58)(1-0/5) = .58