Biostatistics Short Course
Introduction to Survival Analysis

Zhangsheng Yu

Division of Biostatistics
Department of Medicine
Indiana University School of Medicine

Outline

Course objectives

- Know why special methods for the analysis of survival data are needed.
- Understand the basics of the Kaplan-Meier technique.
- Learn how to compare the survival time between two groups (graphically and statistically).
- Learn the basics of the Cox proportional hazards model.

What is "survival analysis"?

Survival analysis is also known as *time to event* analysis:

- time to death
- time until recurrence in a cancer study after surgery
- time to disease progression
- time until first sex transmitted infection

Survival analysis vs. logistic regression

We want to predict 1-year survival rate or probability using patient characteristics such as patient demographics, donor's characteristics, blood type, etc. Is logistic regression sufficient?
Yes, if:

- The 1-year survival rate is the only interest (i.e. not the distribution of time to relapse).
- The binary outcome (death or alive) is available for all subjects.

Survival analysis vs. logistic regression

No, because:

- What if interest becomes 2-year survival rate? For example, you may want to compare with another study which predicts 2-year survival.
- Some patients may drop out of study or die from other causes before 1-year follow-up. Say a patient drops out at 0.9 years before death, then he/she might can quite likely to be 2-year survival. Can we at least use this partial information.
- A patient with death at 1.5 years are quite different from a patient dies at 5 years. (In logistic regression using 1-year death status, their outcomes are treated the same!)

Survival analysis vs. logistic regression

No, because:

- What if interest becomes 2-year survival rate? For example, you may want to compare with another study which predicts 2-year survival.
- Some patients may drop out of study or die from other causes before 1-year follow-up. Say a patient drops out at 0.9 years before death, then he/she might can quite likely to be 2-year survival. Can we at least use this partial information.
- A patient with death at 1.5 years are quite different from a patient dies at 5 years. (In logistic regression using 1-year death status, their outcomes are treated the same!)

Survival analysis vs. logistic regression

No, because:

- What if interest becomes 2-year survival rate? For example, you may want to compare with another study which predicts 2-year survival.
- Some patients may drop out of study or die from other causes before 1-year follow-up. Say a patient drops out at 0.9 years before death, then he/she might can quite likely to be 2-year survival. Can we at least use this partial information.
- A patient with death at 1.5 years are quite different from a patient dies at 5 years. (In logistic regression using 1-year death status, their outcomes are treated the same!)

Why are special methods necessary?

Special methods for analysis of survival data are necessary for reasons such as follows:

1. To allow analysis before all events have been observed; namely presence of *censored observations*.
2. To accommodate for *staggered entry of patients*. Usually not all patients are enrolled into the study at the same time. When patients enter at different times during the study and some have not experienced the event at the time of analysis.
3. To utilize detail survival time information. Survival analysis methods are more powerful than logistic regression in general.

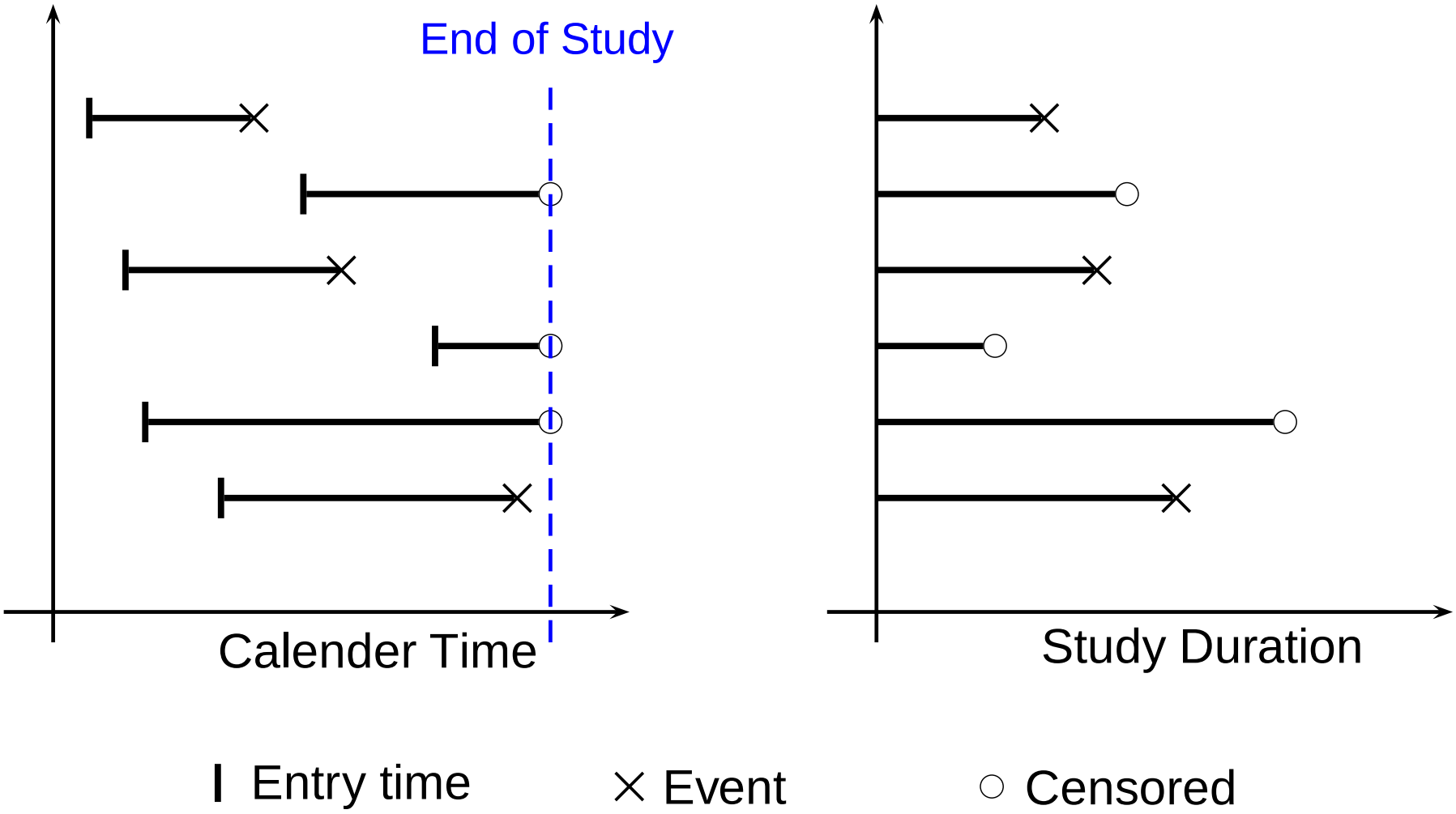
Censoring

1. Right censoring: the event time is larger than the censoring time:
   - The study is closed (administrative censoring).
   - The subject is lost from follow-up.
2. Left censoring: the event time is smaller than the censoring time.

```
Q: When did you first use marijuana?%
A: I have used it but can not recall just
   when the first time was.
```

3. Interval censoring: the event time is only known to fall in an interval. Frequently happen when we have periodic follow-up.

Example of survival data



Data on 42 children with acute leukemia

Pair	Base[1]	T_P[2]	T_{6MP}[3]		Pair	Base[1]	T_P[2]	T_{6MP}[3]
1	1	1	10		12	1	5	20^+
2	2	22	7		13	2	4	19^+
3	2	3	32^+		14	2	15	6
4	2	12	23		15	2	8	17^+
5	2	8	22		16	1	23	35^+
6	1	17	6		17	1	5	6
7	2	2	16		18	2	11	13
8	2	11	34^+		19	2	4	9^+
9	2	8	32^+		20	2	1	6^+
10	2	12	25^+		21	2	8	10^+
11	2	2	11^+					

[1] Remission status at randomization (1=partial, 2=complete)

[2] Time to relapse for placebo patients, months

[3] Time to relapse for 6-MP patients, months; +: censored

Some common survival estimates

How can the survival experience be summarized?

1. **Mean follow-up**
   For the Placebo group, this is $\frac{1}{21}(1 + 22 + 3... + 8) = 8.7$ months.
   For the 6-MP group, this is $\frac{1}{21}(10 + 7 + 32 + ... + 10) = 17.1$ months.

2. **Mean survival**
   We can also say the 8.7 is the mean survival time for the Placebo group. However due to the presence of censoring for the 6-MP group, 17.1 is less than the true mean survival time.

3. **Median survival**
   This is the length of time when 50% of the group under study die.

Empirical survival estimation without censoring

When no observation is censored (e.g. in the Placebo group) :

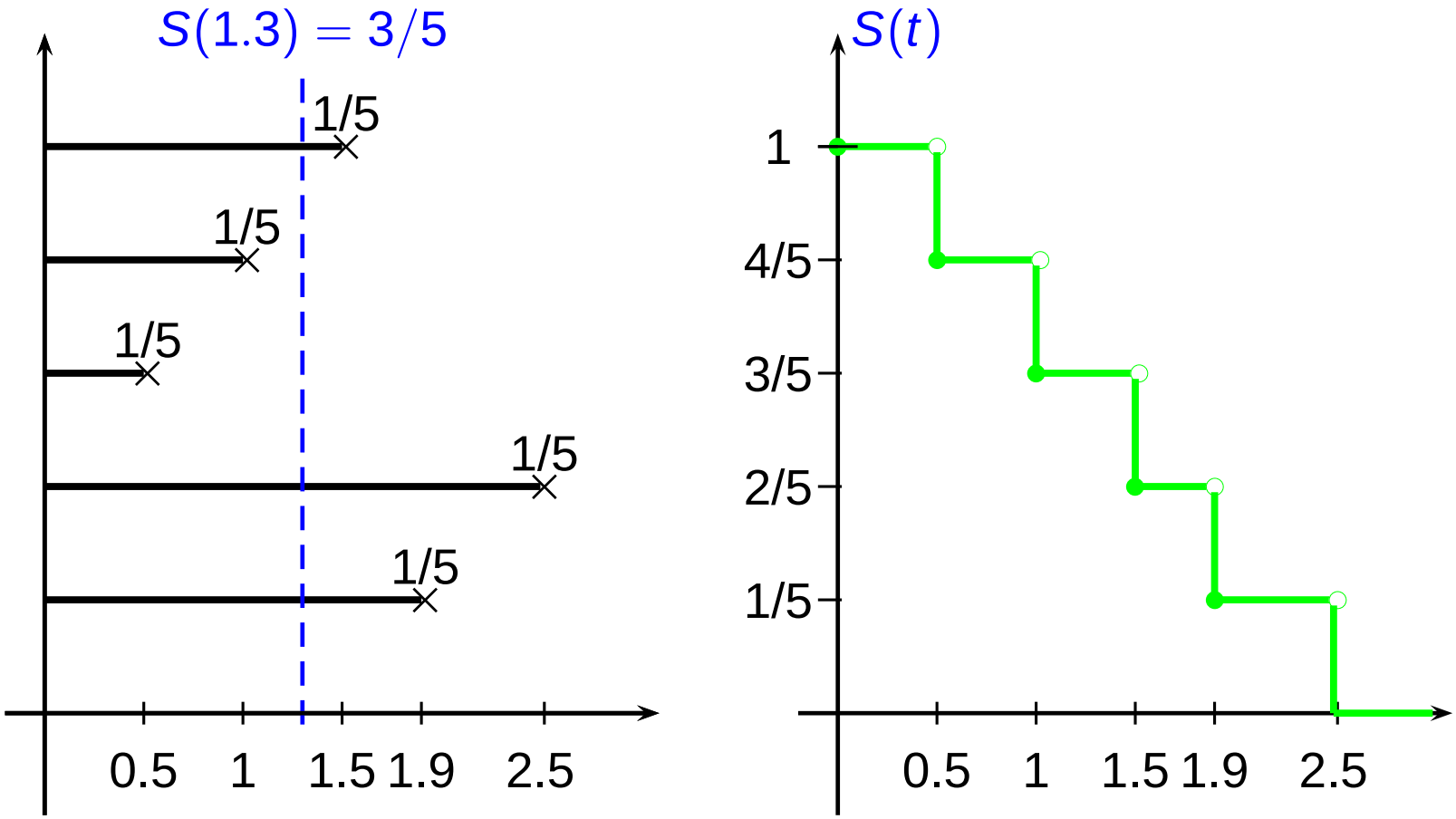
$$S(t) = Prob\left\{T_p > t\right\}$$

it is estimated using *the average number of patients who survive time t*. For example,

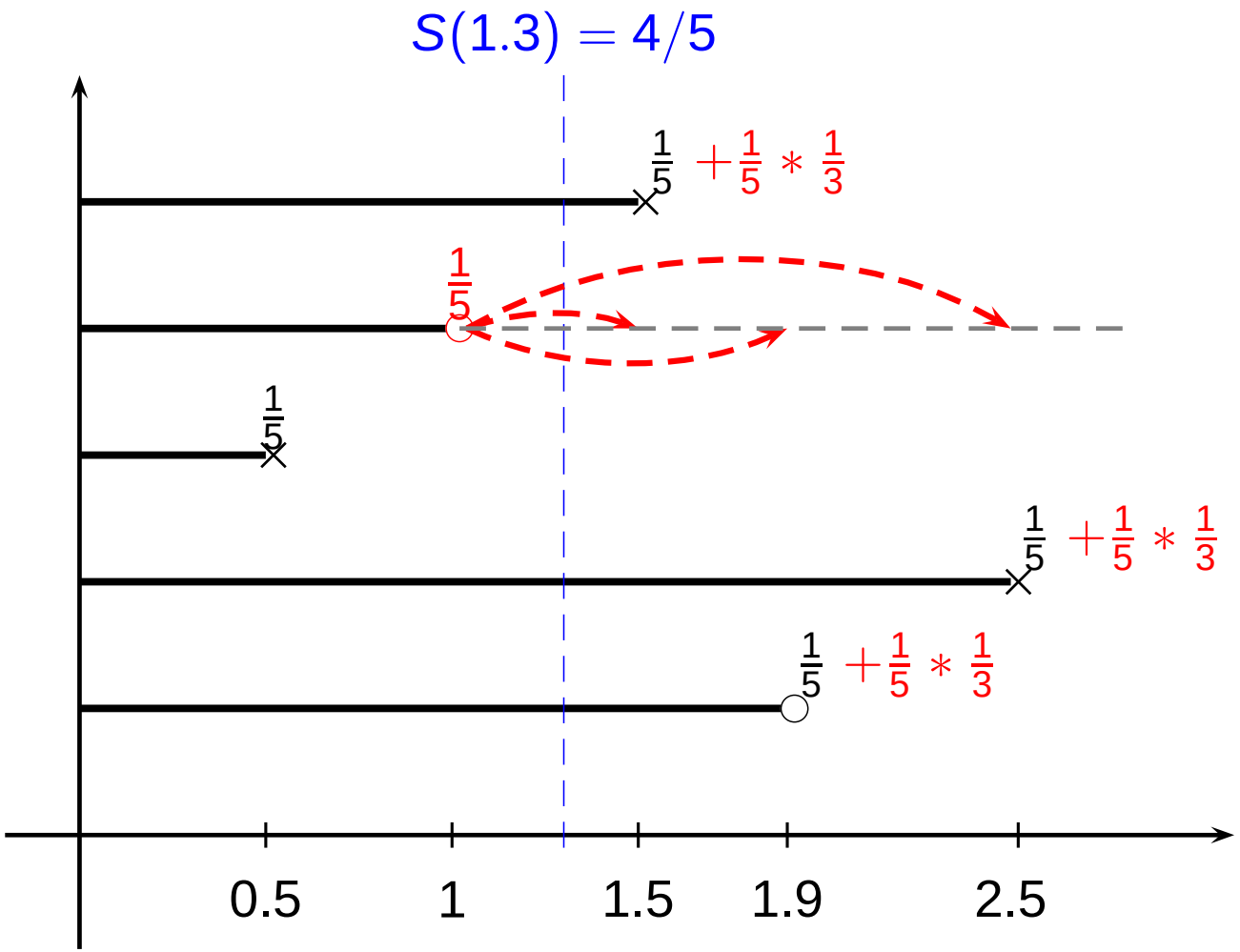
$$\hat{S}(12) = \frac{1}{21} * 4 = 0.19$$

this is the same as put a mass of $1/21$ on each failure time and count the total mass after 12 months.

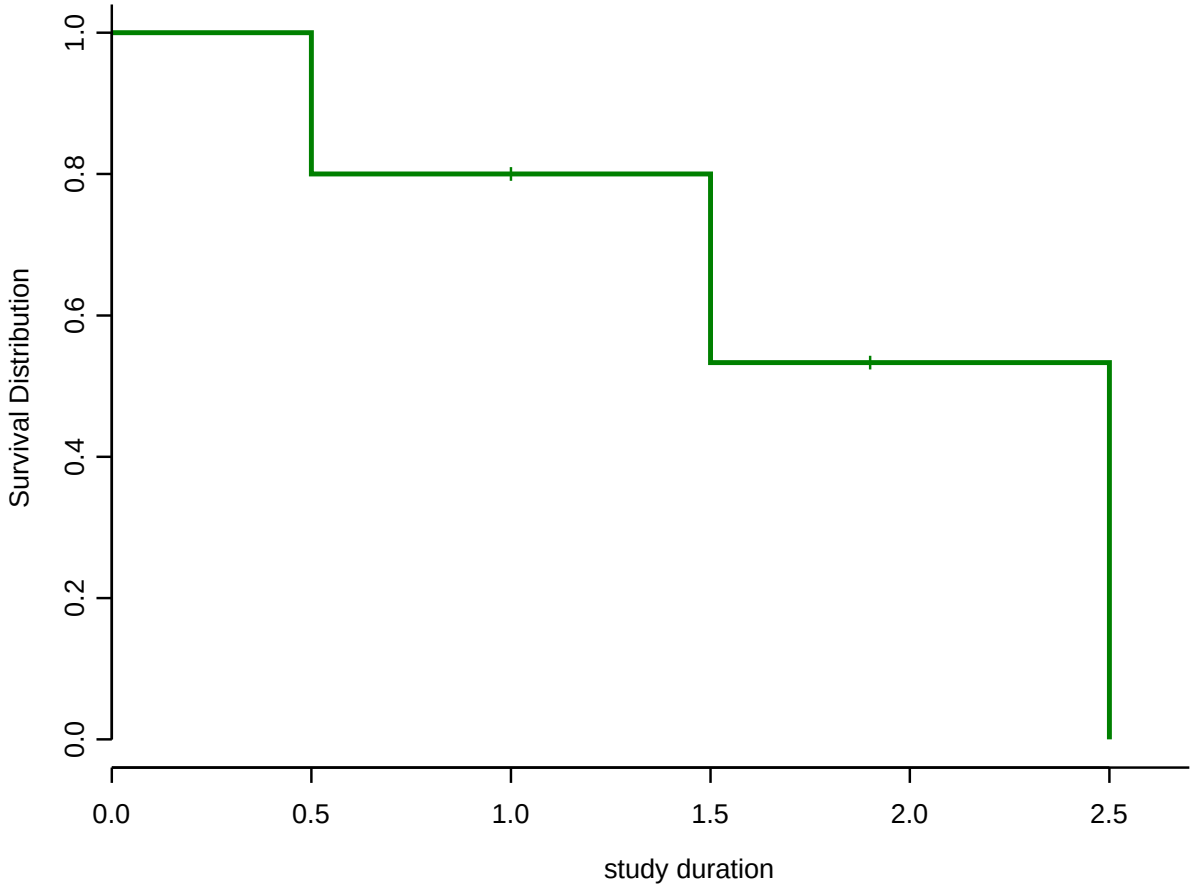
Empirical estimation of distribution



Redistribution of weights and Kaplan-Meier estimates



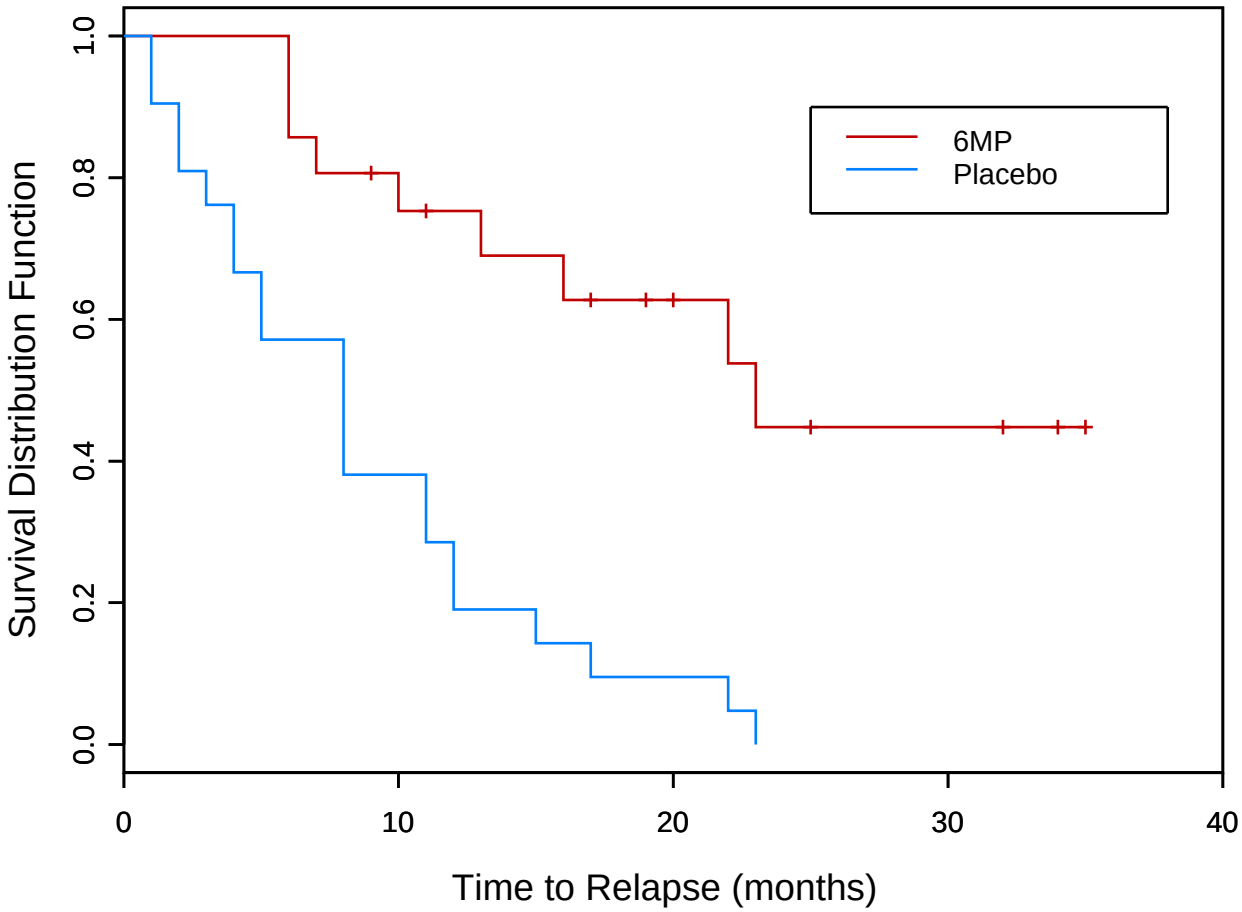
The Kaplan-Meier curve for the mocking data



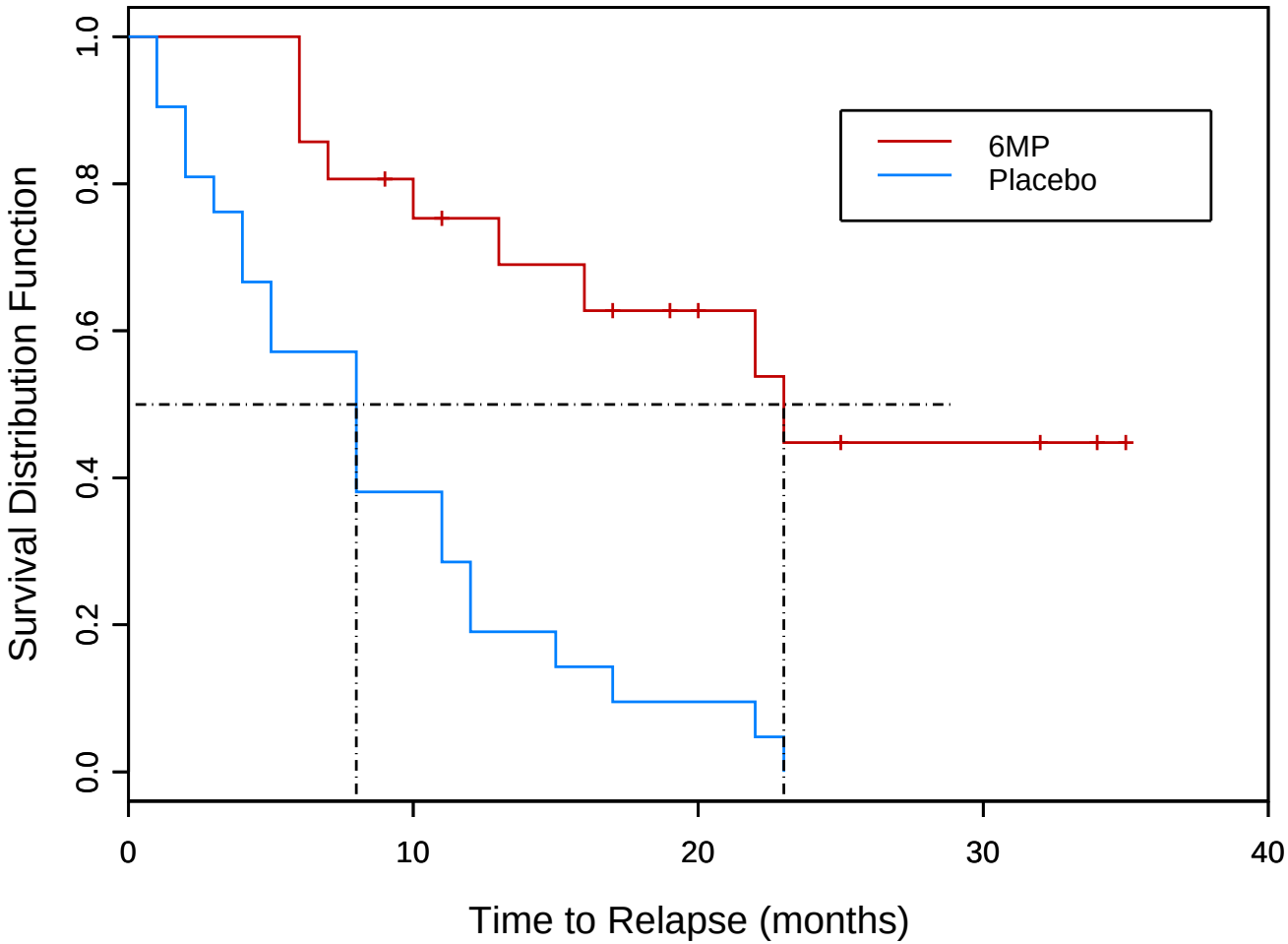
Some facts about the Kaplan-Meier curve

- The KM method is non-parametric; namely the survival curve is step-wise, not smooth. Any jumping point is a failure time point.
- If the largest observed study time t_{max} corresponds to a death time, then the estimated KM survival curve is 0 beyond t_{max}. If t_{max} is censored, then survival curve is not 0 beyond t_{max}.
- The Kaplan-Meier estimator is also known as the *Product-Limit Estimator* of survival due to the formula.

KM curves for the placebo and 6-MP groups



Extract information from the KM curve



Output of the KM estimates of the survival distribution for 6-MP group

time	n.risk	n.event	survival	std.err	l. 95% CI	u. 95% CI
6	21	3	0.857	0.0764	0.720	1.000
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

Comparison of survival between two groups

Eyeballing the KM curves for the Placebo and 6-MP groups, we see that

1. Median survival time is 22.5 months for 6-MP and 8 for placebo.
   $\Longrightarrow$ 14.5 month difference.
2. The Kaplan-Meier curve for 6-MP group lies above that for the Placebo group and there is a big gap between the two curves
   $\Longrightarrow$ the survival of 6-MP seems to be superior.
3. The gap seems to become bigger as time progresses.

Statistical comparison between two survival curves

Main idea:

If survival is unrelated to group assignment, then, at each time point, roughly the same proportion in each group will fail. Statistical tests are based on chi-square-type of statistics that compare the *expected* with the *observed* survival rates.

Test

H_0: no difference between the survival curves of treatment A and B

H_1: there is difference.

Computer calculation of the log-rank test

Using a computer we obtain the following results:

```
              N   Observed Expected (O-E)^2/E (O-E)^2/V
trt=Placebo  21         21     10.7      9.77      16.8
trt=6-MP     21          9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 0.0000417
```

The p value of the test is $p < 0.001$, which implies a significant difference in the survival of the two groups.

Methods for analysis of multiple variables

Although log-rank test can be extended to test differences in more than 2 groups, The method fall short however in the following situations:

- Single-variable analysis with a continuous factor.
- Multi-variable analysis with any combination of categorical and continuous factors.
- Quantify the differences.

The Crook study of prostate cancer (*Cancer*, 1997)

Variable	Explanation	Coding
age	patient age	
anyfail	any failure	0 = no 1 = yes
months	time to any failure	
prerx_psa_group	pretreatment psa classification	1 = 1-5 2 = 5-10 3 = 10-15 4 = 15-20 5 = 20-50 6 = $>$ 50
tumor_stage	stage of tumor	1 = T1b-c 3 = T2a 4 = T2b-c 6 = T3-T4

Research questions

An example of the type of questions that may be asked in a survival analysis is as follows:

- What is the effect of age (a continuous factor) on survival?
- What is the effect of tumor stage?
- What is the effect of tumor stage *adjusted for* the effect of age?

The Cox proportional hazards model

It addresses survival through modelling the hazard $\Rightarrow$ larger hazards are directly related to shorter survival.

By *hazard* we mean the propensity for failure for an individual at each time point. It is the instantaneous risk of failure.

The general Cox-type model is as follows:

$$h(t) = h_0(t) \times \exp\{\beta_1 X_1\}$$

where $h_0(t)$ is some unspecified baseline hazard at time t and X_1 is a covariate.

Behavior of the Cox model

If two individuals have covariates X_{11} and X_{12}, then the *hazard ratio, or risk ratio* $h_{12}(t) = \frac{h_1(t)}{h_2(t)}$ is

$$h_{12}(t) = \frac{h_0(t)\exp\{\beta_1 X_{11}\}}{h_o(t)\exp\{\beta_1 X_{12}\}} = \frac{e^{\beta_1 x_{11}}}{e^{\beta_1 x_{12}}} = e^{\beta_1(x_{11}-x_{12})} = r_{12}$$

Note that, by taking ratios, we do not have to specify the baseline hazard $h_o(t)$.
If $ratio_{12} > 1$, subjects with $X = X_{11}$ have a larger hazard than those with $X = X12$.

Behavior of the Cox model

If $X_{11} = 1$ and $X_{12} = 0$ which represents different groups two patients belong to, then the *hazard ratio, or risk ratio* of patient 1 and patient 2 is

$$h_{12}(t) = e^{\beta_1(x_{11}-x_{12})} = e^{\beta_1}$$

and $\beta_1 = \log[h_{12}(t)]$ is the log hazard ratio.

If by X_1 is continuous (e.g., PSA levels) then the *hazard ratio, or risk ratio* of two patients with PSA levels that differ by one unit (i.e., $X_{11} = X_{12} + 1$) is

$$h_{12}(t) = e^{\beta_1(x_{11}-x_{12})} = e^{\beta_1}$$

Hence $\beta_1 = \log[h_{12}(t)]$ is the log hazard ratio between two patients differing by a single unit in their measurements of PSA levels.

Effect of a factor with more than two groups

A categorical factor X_3 with more than two groups is coded by creating dummy variables.

There are four tumor stages which can be coded as:

	Tumor	Coding		
	stage (X_3)	Z_1	Z_2	Z_3
reference category $\Rightarrow$	T1b-2	0	0	0
	T2a	1	0	0
	T2b-c	0	1	0
	T3-4	0	0	1

The β associated with each dummy variable is the log hazard ratio of belonging in that category versus the reference category.

Analysis of the Crook data

The Cox PH analysis of prostate-cancer survival with respect to *age* and *tumor stage*.
The output for regression coefficient estimates and P-values:

```
                                                          95% CI
     coef  exp(coef) se(coef)    z    p-value  lower  upper
AGE -0.0105   0.990    0.016  -0.645  0.5200   0.96   1.02
 Z1 -0.0238   0.977    0.708  -0.033  0.9700   0.24   3.91
 Z2  1.1924   3.295    0.537   2.221  0.0260   1.15   9.43
 Z3  1.8972   6.667    0.533   3.560  0.0004   2.35  18.95

Rsquare= 0.135   (max possible= 0.957 )
Likelihood ratio test= 29.9  on 4 df,   p=0.000005
Wald test            = 24.4  on 4 df,   p=0.000066
Score (logrank) test = 29.5  on 4 df,   p=0.000006
```

Output interpretation: individual factors

- **Age**
  The log hazard ratio $\beta_1 = -0.011$ and the hazard ratio is $e^{\beta_1} = 0.99$.

  $\Rightarrow$ for each increase in age by one year, the risk of death is slightly decreasing by about 1%. Age is non-significant as a predictor of survival (p=0.52).

- **Tumor stage**
  Z_1, Z_2, and Z_3 compares tumor stage T2a, T2b-c and T3-4 with T1b-2. T2b-c and T3-4 are significantly different from T1b-2 (p=0.026 and 0.00037). The hazard ratios are 3.295 and 6.667.

  $\Rightarrow$ the risks of death are about 3 and 6.7 times higher compared with T1b-2.

Acknowledgement

Slides courtesy of Dr. Menggang Yu.