Introduction to Survival Analysis

Jianan Hui

2/19/2015

Outline

1. Introduction
2. Kaplan-Meier Survival Curves
3. The Log-Rank Test
4. Cox Proportional Hazards Model

Introduction

- Survival analysis:
  - method for analyzing timing of events;
  - data analytic approach to estimate the time until an event occurs.
- Historically, survival time refers to the time that an individual "survives" over some period until the event of death occurs.
- Event is also named failure.

Areas of application

- Survival analysis is used as a tool in many different settings:
  - proving or disproving the value of medical treatments for diseases;
  - evaluating reliability of technical equipment;
  - monitoring social phenomena like divorce and unemployment.

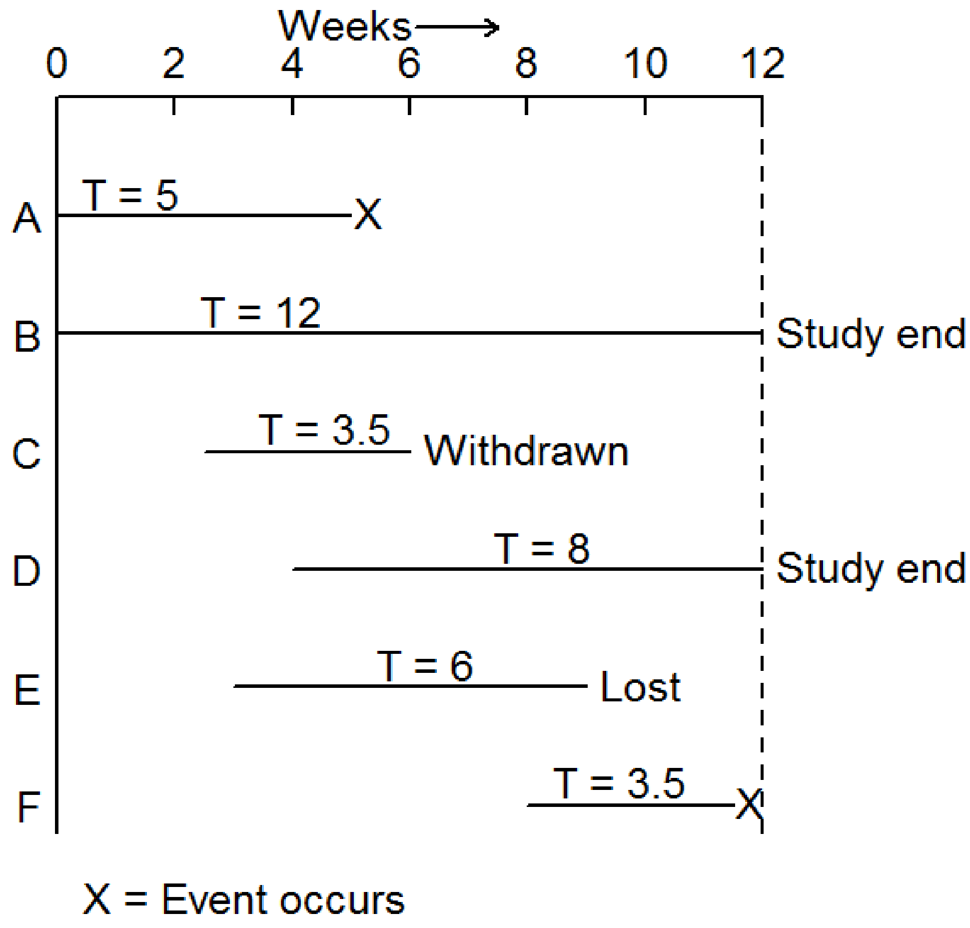
Examples

- Time from…
  - **marriage to divorce;**
  - **birth to cancer diagnosis;**
  - **entry to a study to relapse.**

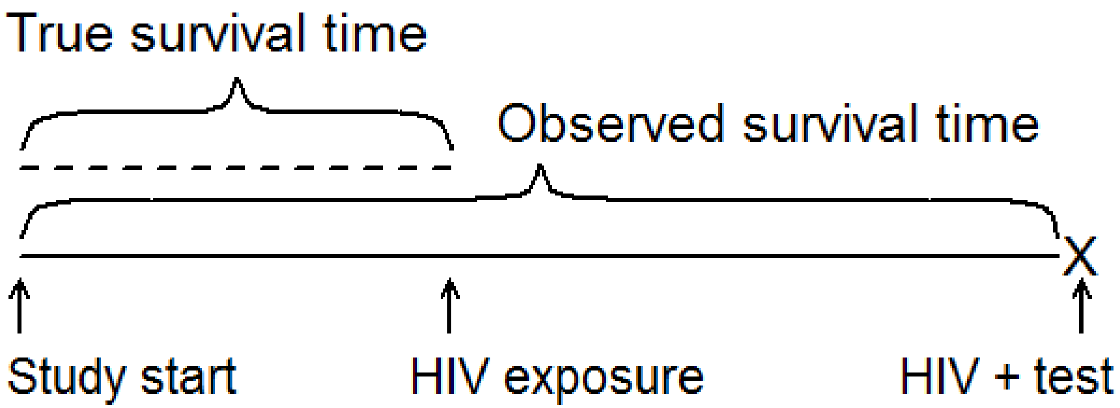
Censoring

- The survival time is not known exactly! This may occur due to the following reasons:
  - a person does not experience the event before the study ends;
  - a person is lost to follow-up during the study period;
  - a person withdraws from the study because of some other reason.

Right Censored



Left censored



Outcome variable

- Time until an event occurs
- T = survival time (T>0)
- T is a random variable
- t = specific value of interest for T
- Ask whether T > t if we are interested in the question whether an individual survives longer than t .

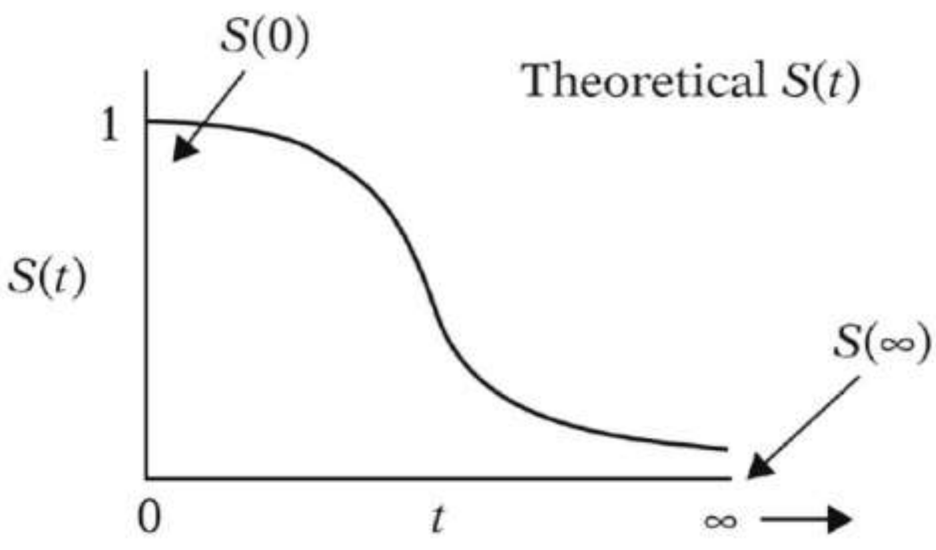
Survivor function

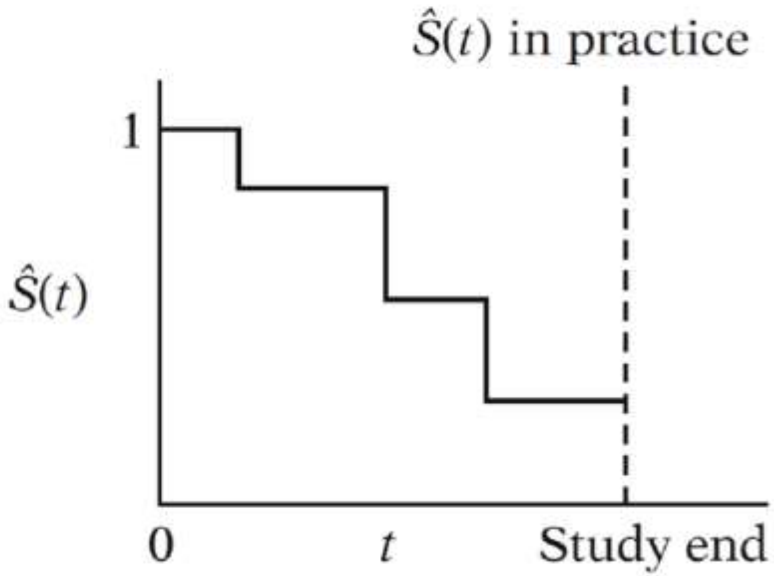
- $S(t)=P(T>t)$
- Probability that random variable T exceeds specified time t
- Fundamental to survival analysis

t	S(t)
1	S(1) = P(T > 1)
2	S(2) = P(T > 2)
3	S(3) = P(T > 3)
.	.
.	.
.	.

Survivor function

$$S(t) = \Pr(T > t)$$





Hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t}$$

- Often called: Conditional failure rate
- h(t) has no upper bounds
- Depends on whether time is measured in days, weeks, months, or years, etc. (Example next page)

Example: Hazard function

Assume having a huge follow-up study on heart attacks:

- 600 heart attacks (events) per year;
- 50 events per month;
- 11.5 events per week;
- 0.0011 events per minute.

$h(t)$ = rate of events occurring per time unit

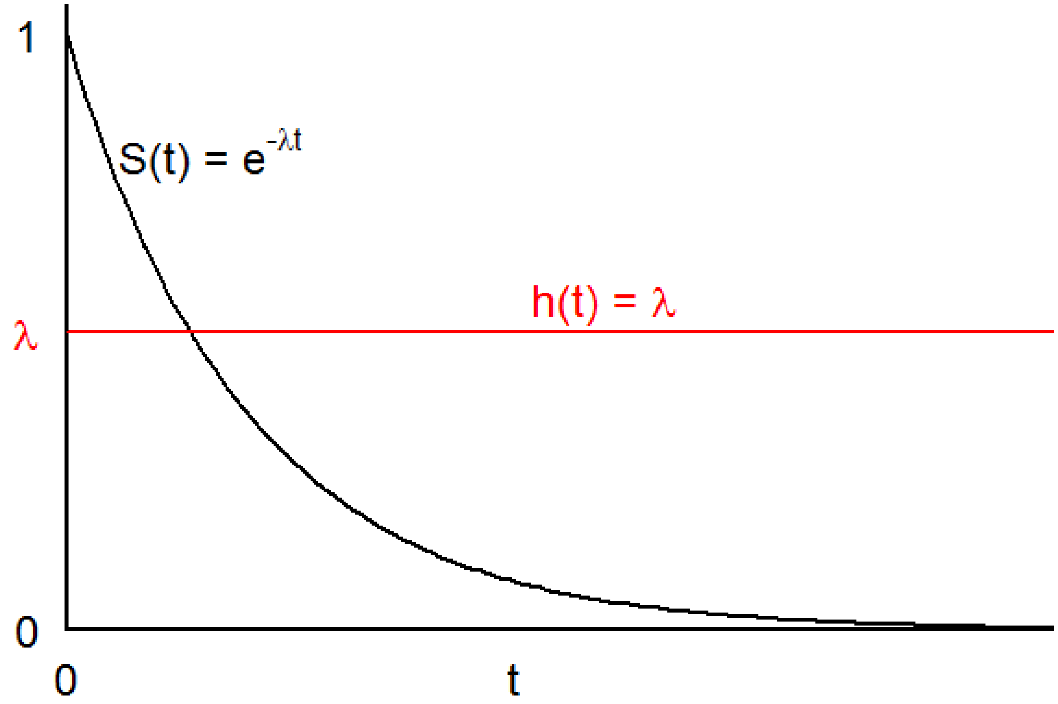
Relation between S(t) and h(t)

- If T continuous:

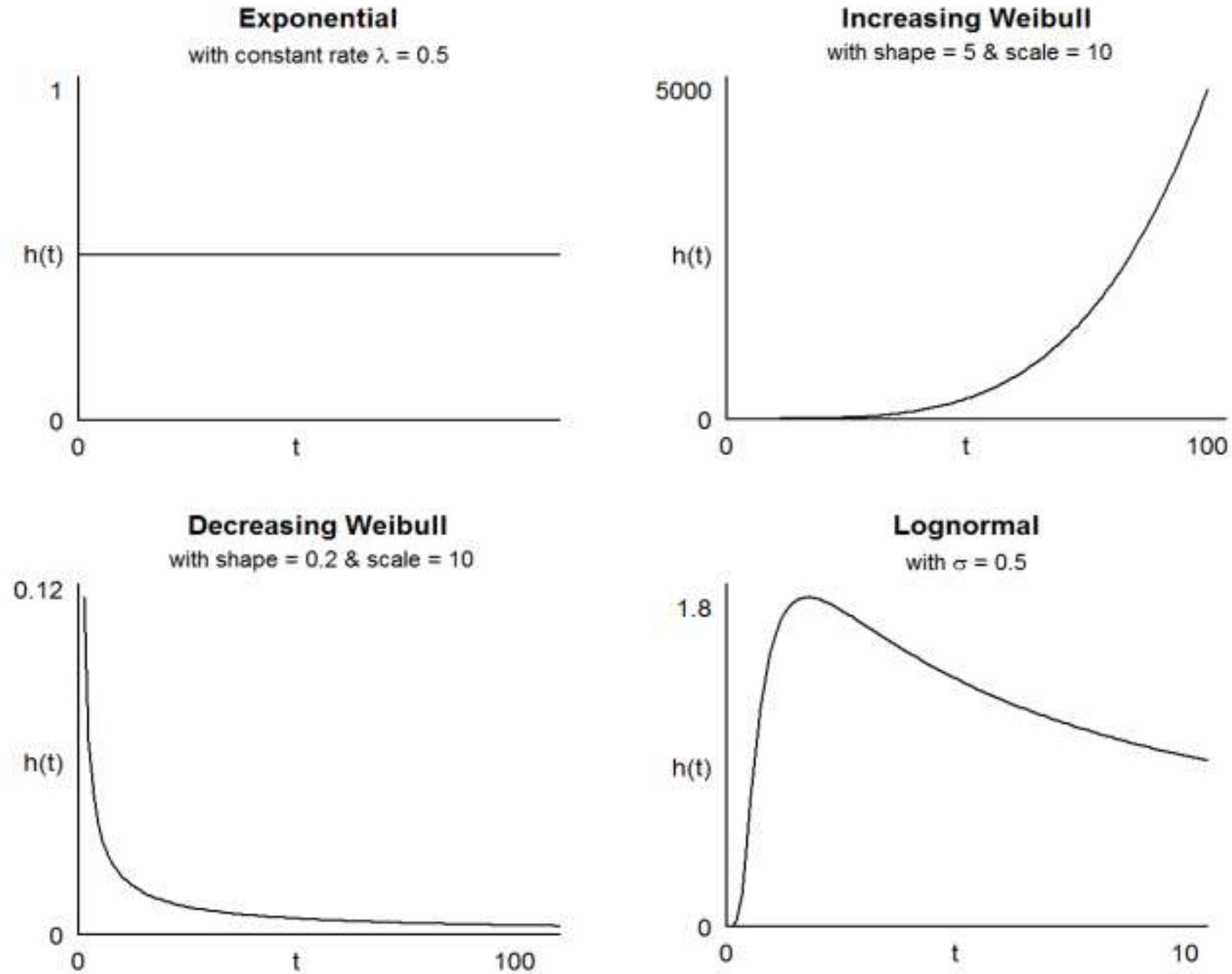
$$S(t) = \exp[-\int_0^t h(u)du]$$

$$h(t) = -\frac{S'(t)}{S(t)}$$

Example: Relationship



Types of hazard functions



Goals (of survival analysis)

- to estimate and interpret survivor and/or hazard function;
- to compare survivors and/or hazard functions;
- to assess the relationship of explanatory variables to survival times -> we need mathematical modelling (Cox model).

Computer layout

individual	t (in weeks)	δ (failed or censored)
1	5	1
2	12	0
3	3.5	0
4	8	0
5	6	0
6	3.5	1

Notation & terminology

- Ordered failures: **unordered** < **~~censored t's~~** / **failed t's ordered ($t_{(i)}$)**
- Frequency counts:
  - m_i = # individuals who failed at $t_{(i)}$
  - q_i = # ind. censored in $[t_{(i)}, t_{(i+1)})$
- Risk set $R(t_{(i)})$: Collection of individuals who have survived at least until time $t_{(i)}$

Manual analysis layout

Ordered failure times	# of failures m_i	# censored in $[t_{(i)}, t_{(i+1)})$	Risk set $R(t_{(i)})$
$t_{(0)}=0$	m_i	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
....	...	...	...
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

Manual analysis layout

Ordered failure times	# of failures m_i	# censored in $[t_{(i)}, t_{(i+1)})$	Risk set $R(t_{(i)})$
$t_{(0)} = 0$	0	0	6 persons survive ≥ 0 weeks
$t_{(1)} = 3.5$	1	1	6 persons survive ≥ 3.5 weeks
$t_{(2)} = 5$	1	3	4 persons survive ≥ 5 weeks

2 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, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 25+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

+ denotes censored

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

Descriptive statistic:

$$\overline{T_1}(ignoring\ +'s) = 17.1,\quad \overline{T_2} = 8.6$$

- Table of ordered failure times

Group 1 (treatment)

$t_{(j)}$	n_j	m_j	q_j
0	21	0	0
6	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
>23	-	-	-

Group 1 (treatment)	Group 2 (placebo)
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 25+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

+ denotes censored

Group 2 (placebo)

$t_{(j)}$	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
8	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

$\longrightarrow$ Remark: no censorship in group 2

Computation of KM-curve for group 2 (no censoring)

$t_{(j)}$	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	0	3/21 = .14
17	3	1	0	2/21 = .10
22	2	1	0	1/21 = .05
23	1	1	0	0/21 = .00

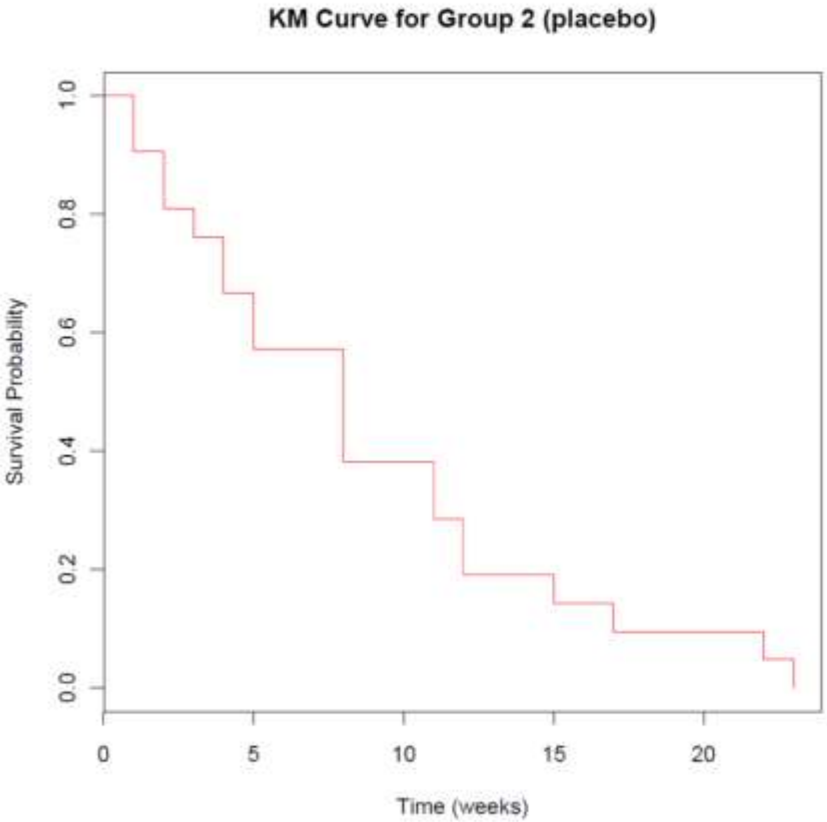
$$\hat{S}(t_{(j)}) = \frac{\#\ surviving\ past\ t_{(j)}}{21}$$

KM Curve for Group 2 (Placebo)

```
> time2 <-
c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,
22,23)
> status2 <-
c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

> fit2 <- survfit(Surv(time2, status2) ~ 1)

> plot(fit2,conf.int=0, col = 'red', xlab =
'Time (weeks)', ylab = 'Survival Probability')
> title(main='KM Curve for Group 2 (placebo)')
```



General KM formula

- Alternative way to calculate the survival probabilities
- KM formula = product limit formula

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \hat{P}r(T > t_{(i)} \,|\, T \geq t_{(i)})$$

$$= \hat{S}(t_{(j-1)}) \times \hat{P}r(T > t_{(j)} \,|\, T \geq t_{(j)})$$

Computation of KM-curve for group 1 (treatment)

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	1×18/21=.8571
7	17	1	1	.8571×16/17=.8067
10	15	1	2	
13	12	1	0	
16	11	1	3	
22	7	1	0	
23	6	1	5	

Fraction at $t_{(j)}$:
$\Pr(T > t_{(j)} \mid T \geq t_{(j)})$

$= \dfrac{n_j - m_j}{n_j}$

Computation of KM-curve for group 1 (treatment)

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	1×18/21=.8571
7	17	1	1	.8571×16/17=.8067
10	15	1	2	.8067×14/15=.7529
13	12	1	0	.7529×11/12=.6902
16	11	1	3	.6902×10/11=.6275
22	7	1	0	.6275×6/7=.5378
23	6	1	5	.5378×5/6=.4482

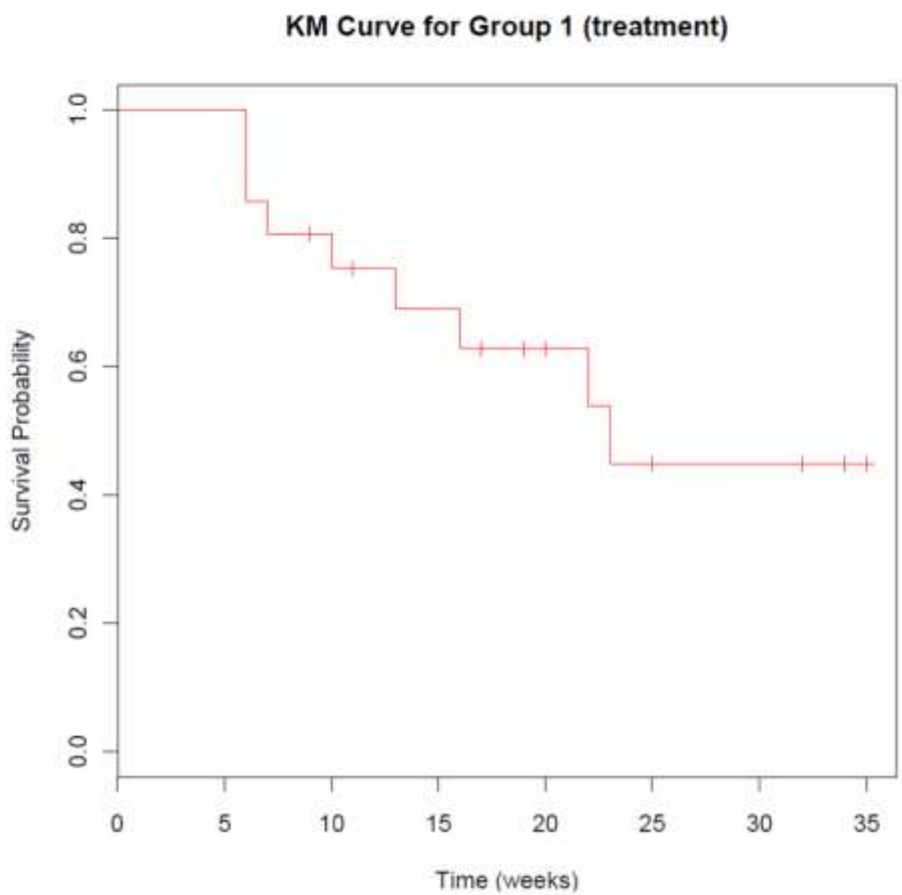
Fraction at $t_{(j)}$:
$\Pr(T > t_{(j)} \mid T \geq t_{(j)})$

KM-curve for group 1 (treatment)

```
> time1 <-
c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,
25,32,32,34,35)
> status1 <-
c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

> fit1 <- survfit(Surv(time1, status1) ~ 1)

> plot(fit1,conf.int=0, col = 'red', xlab =
'Time (weeks)', ylab = 'Survival
Probability')
> title(main='KM Curve for Group 1
(treatment)')
```



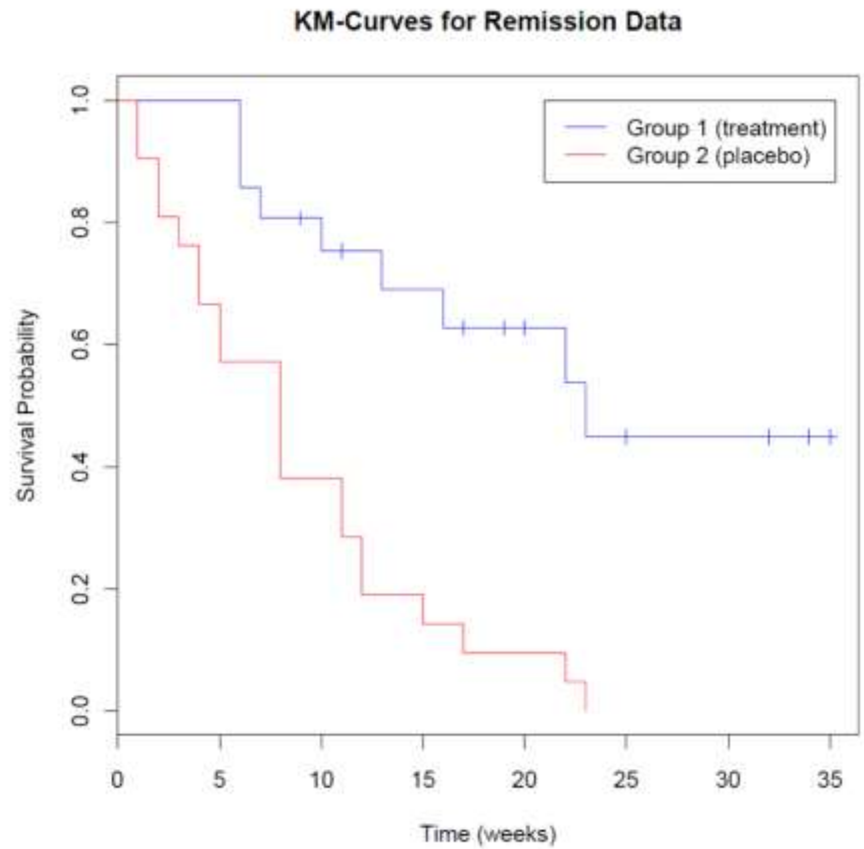
Comparison of KM Plots for Remission Data

```
> time1 <-
c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25
,32,32,34,35)
> status1 <-
c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

> time2 <-
c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,
22,23)
> status2 <-
c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

> fit1 <- survfit(Surv(time1, status1) ~ 1)
> fit2 <- survfit(Surv(time2, status2) ~ 1)

> plot(fit1,conf.int=0, col ='blue', xlab =
'Time (weeks)', ylab = 'Survival Probability')
> lines(fit2, col = 'red')
> legend(21,1,c('Group 1 (treatment)', 'Group
2 (placebo)'), col = c('blue','red'), lty = 1)
> title(main='KM-Curves for Remission Data')
```



→ Question: Do we have any reason to claim that group 1 (treatment) has better survival prognosis than group 2?