Survival Analysis
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-11-09 15:08

Outline

- Intro to survival data/analysis
- KM estimator, SE, CI
- Log-rank test

Survival Analysis

- Chapter 16 text; BIOS 680/780
- Survival analysis: response is time to event
- Measure time from beginning of follow-up until an event such as incident disease, death, or relapse
- Examples:

  time from kidney transplant until death

  time from leukemia treatment to remission

  time from release from jail to rearrest

Survival Analysis: Notation

- Let T^* denote the survival time; assume $T^* > 0$
- Define the survival function

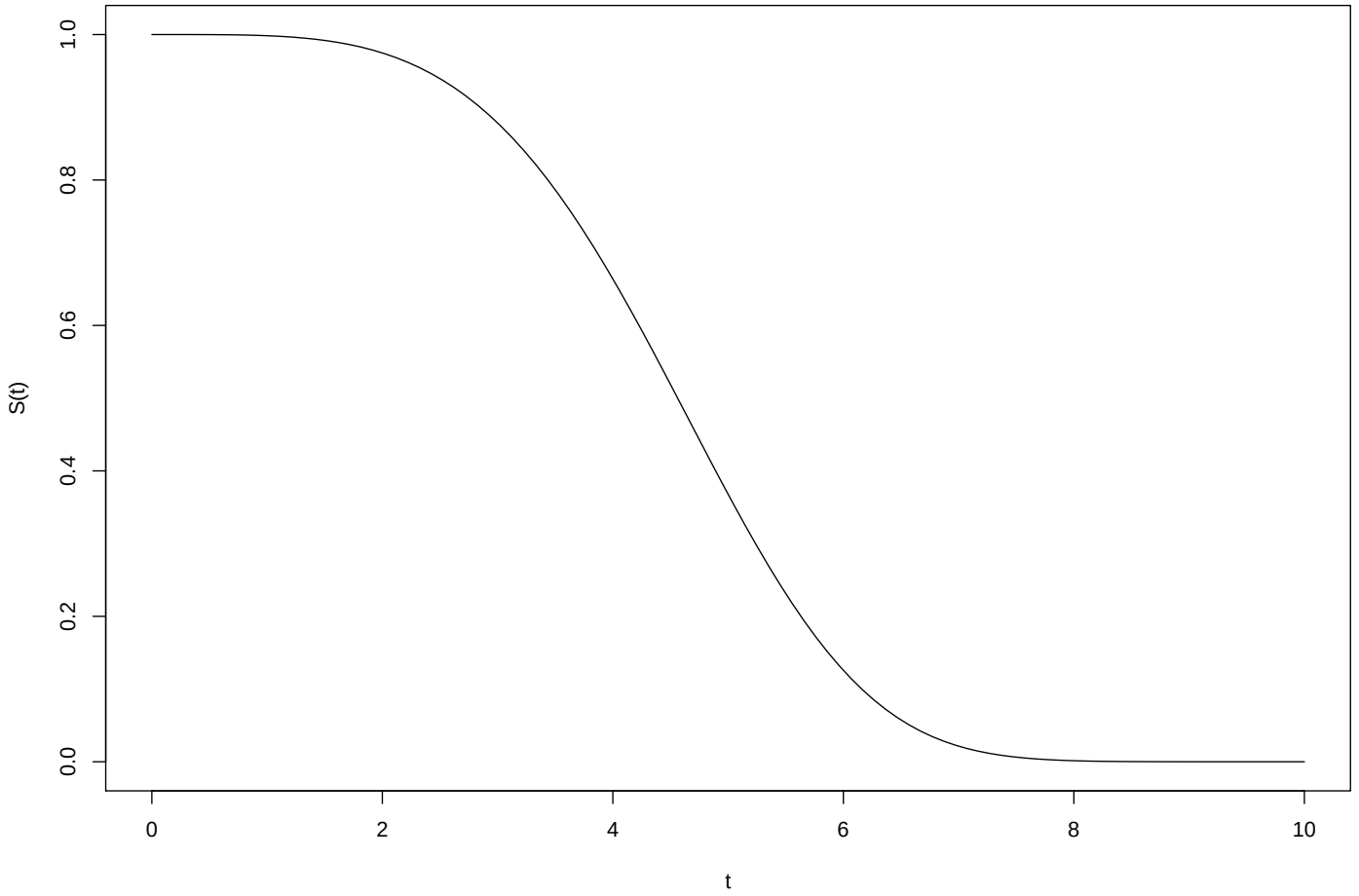
$$S(t) = \Pr[T^* > t] = 1 - \Pr[T^* \leq t] = 1 - F(t)$$

  where $F(t)$ is CDF of T^*
- Properties

$$S(0) = 1; S(\infty) = 0$$

$$\text{If } t_1 \leq t_2, \text{ then } S(t_1) \geq S(t_2)$$

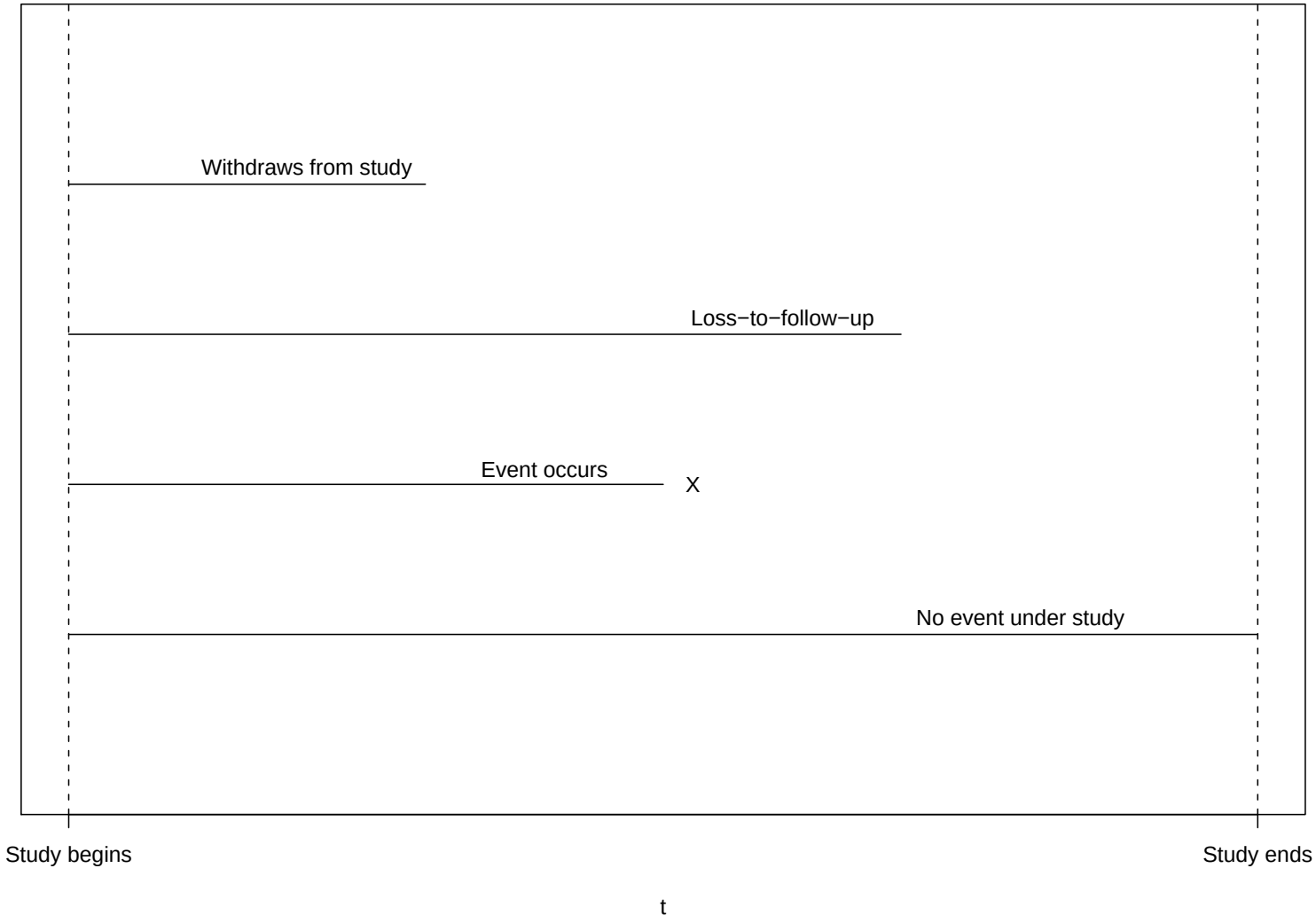
Example Survival Curve/Function



Censoring

- Often do not know the exact time of failure
- Reasons for **right** censoring:
  - subject does not experience event before the end of the study
  - subject is lost to follow-up during the study (eg withdraws from study, moves, death from something other than event of interest)
- Failure times can also be left or interval censored

Right Censoring



Survival Data

- Let T_i^* and C_i denote the survival and right censoring times for the i^{th} individual
- Observe $T_i = \min\{T_i^*, C_i\}$
- Censoring indicator

$$\delta_i = \begin{cases} 1 \text{ if failure i.e. } T_i = T_i^* \\ 0 \text{ if right censored i.e. } T_i = C_i \end{cases}$$

- Observe (T_i, δ_i) for $i = 1, 2, \ldots, N$

Example

- Remission time in weeks for leukemia patients ($N = 21$)

(T_i, δ_i)	(T_i, δ_i)	(T_i, δ_i)
(6,1)	(6,1)	(6,1)
(6,0)	(7,1)	(9,0)
(10,1)	(10,0)	(11,0)
(13,1)	(16,1)	(17,0)
(19,0)	(20,0)	(22,1)
(23,1)	(25,0)	(32,0)
(32,0)	(34,0)	(35,0)

Estimation

- How do we estimate $S(t)$ w/ minimal assumptions?
- Answer 1: In the absence of censoring, use 1-EDF
- Answer 2: Otherwise, use Kaplan-Meier estimator

Tabular Summary of Data

- Let $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$ be the distinct ordered failure times

Ordered failures $t_{(j)}$	No. of failures m_j	No. censored in $[t_{(j)}, t_{(j+1)})$ q_j	Risk set $R(t_{(j)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)}) = N$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

- $R(t_{(j)}) = R(t_{(j-1)}) - m_{j-1} - q_{j-1}$

Leukemia Example

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
0	0	0	21
6	3	1	21
7	1	1	17
10	1	2	15
13	1	0	12
16	1	3	11
22	1	0	7
23	1	5	6

Kaplan-Meier estimator of $S(t)$

- For $t \in [0, t_{(1)})$

$$\hat{S}(t) = 1$$

- For $t \in [t_{(j)}, t_{(j+1)})$

$$\hat{S}(t) = \hat{S}(t_{(j-1)})\widehat{\Pr}[T > t_{(j)} | T \geq t_{(j)}]$$

$$= \hat{S}(t_{(j-1)}) \left\{ \frac{R(t_{(j)}) - m_j}{R(t_{(j)})} \right\}$$

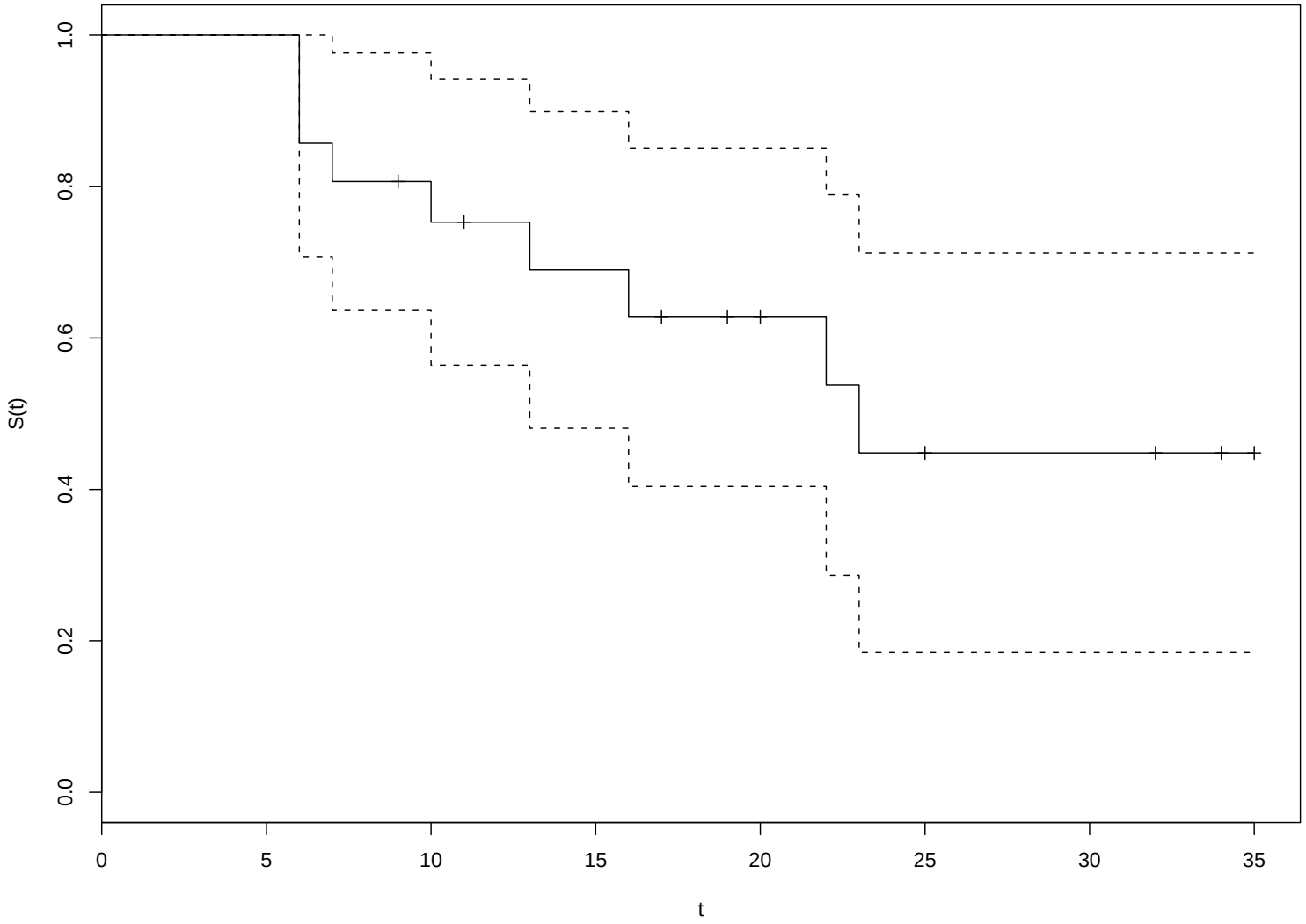
Kaplan-Meier

- KM is a NPMLE (nonparametric maximum likelihood estimator)
- Assumes independent censoring
- Aka *product limit estimator*
- If no censoring, KM equals 1-EDF
- Alternative: *life-table or actuarial method*

Leukemia Example

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$	$\hat{S}(t_{(j)})$
0	0	0	21	1
6	3	1	21	18/21=.857
7	1	1	17	$.857(16/17) = .807$
10	1	2	15	$.807(14/15) = .753$
13	1	0	12	$.753(11/12) = .690$
16	1	3	11	$.690(10/11) = .627$
22	1	0	7	$.627(6/7) = .538$
23	1	5	6	$.538(5/6) = .448$

Kaplan-Meier Estimate for Leukemia Example



Kaplan-Meier Estimate: R code

```
> library("survival")

> fit <- survfit(Surv(t, delta),conf.type="plain")

> plot(fit,xlab="t",ylab="S(t)")

> summary(fit)
Call: survfit(formula = Surv(t, delta))

 time n.risk n.event survival std.err
    6     21       3    0.857  0.0764
    7     17       1    0.807  0.0869
   10     15       1    0.753  0.0963
   13     12       1    0.690  0.1068
   16     11       1    0.627  0.1141
   22      7       1    0.538  0.1282
   23      6       1    0.448  0.1346
```

Kaplan-Meier Estimate: SAS code

```
proc lifetest;  time t*delta(0);
```

The LIFETEST Procedure

Product-Limit Survival Estimates

t	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
6.0000	.	.	.	1	20
6.0000	.	.	.	2	19
6.0000	0.8571	0.1429	0.0764	3	18
6.0000*	.	.	.	3	17
7.0000	0.8067	0.1933	0.0869	4	16
9.0000*	.	.	.	4	15
10.0000	0.7529	0.2471	0.0963	5	14
10.0000*	.	.	.	5	13
11.0000*	.	.	.	5	12

Derivation of Greenwood SE/CI of KM

- Let $n_j = R(t_{(j)})$
- Write KM as

$$\hat{S}(t) = \prod_{j=1}^{i} \hat{p}_j \text{ for } t \in [t_{(j)}, t_{(j+1)}),$$

where $\hat{p}_j = (n_j - m_j)/n_j$ is the estimated probability of surviving interval $[t_{(j)}, t_{(j+1)})$ conditional on survival up to $t_{(j)}$

Derivation of Greenwood

- Take logs

$$\log \hat{S}(t) = \sum_{j=1}^{i} \log \hat{p}_j$$

  such that

$$V\{\log \hat{S}(t)\} = \sum_{j=1}^{i} V\{\log \hat{p}_j\}$$

- Binomial argument

$$\hat{V}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$$

Derivation of Greenwood

- Taylor series approximation

$$\hat{V}\{g(X)\} \approx \{g'(\mu)\}^2 \hat{V}(X)$$

  implies

$$\hat{V}(\log \hat{p}_j) \approx \frac{1-\hat{p}_j}{n_j \hat{p}_j} = \frac{m_j}{n_j(n_j - m_j)}$$

- Thus

$$\hat{V}\{\log \hat{S}(t)\} \approx \sum_{j=1}^{i} \frac{m_j}{n_j(n_j - m_j)}$$

Greenwood SE/CI

- Additional application of Taylor series approximation

$$\hat{V}\{\log \hat{S}(t)\} \approx \{\hat{S}(t)\}^{-2}\hat{V}\{\hat{S}(t)\}$$

  implying

$$\hat{V}\{\hat{S}(t)\} \approx \{\hat{S}(t)\}^{2}\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}$$

- Thus

$$\widehat{SE}(\hat{S}(t)) \approx \hat{S}(t)\sqrt{\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}}$$

  for $t_{(i)} \leq t < t_{(i+1)}$

Greenwood SE/CI

- For Leukemia example,

$$\widehat{SE}(\hat{S}(6)) = .8571\sqrt{\frac{3}{21 * 18}} = 0.0764$$

$$\widehat{SE}(\hat{S}(7)) = .8067\sqrt{\frac{3}{21 * 18} + \frac{1}{17 * 16}} = 0.0869$$

- Approx $(1 - \alpha)$ x 100% CI

$$\hat{S}(t) \pm z_{1-\alpha/2}\widehat{SE}(\hat{S}(t))$$

Greenwood CIs

- Greenwood based CIs are symmetric
- Problematic when survivor function near 0 or 1; can have CI lie outside (0,1)
- Pragmatic solution: set equal to 0 or 1 in this case
- Many other methods exist to estimate standard error and obtain confidence intervals
- All have pointwise interpretation; different methods exist to obtain *confidence bands*

Testing

- How do we test if two survival functions are different under minimal assumptions?
- For example: leukemia patients are randomized to treatment or placebo. Are the survival functions the same between the two groups?
- Without censoring, use a rank test (e.g., Wilcoxon rank sum)
- In the presence of right censoring, use logrank test

Logrank test

- Data from two samples

$$(T_{ij}, \delta_{ij})$$

  for $i = 1, 2$ and $j = 1, 2, \ldots, n_i$

- Want to test

$$H_0 : S_1(t) = S_2(t)$$

  where

$$S_j(t) = \Pr[T_j^* > t] \text{ for } j = 1, 2$$

Logrank test

- Let $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$ be the distinct ordered failure times in the two groups combined
- At each time $t_{(j)}$, construct the table:

Group	Event	Survive	At risk
1	m_{1j}	$R_1(t_{(j)}) - m_{1j}$	$R_1(t_{(j)})$
2	m_{2j}	$R_2(t_{(j)}) - m_{2j}$	$R_2(t_{(j)})$
	m_j	$R(t_{(j)}) - m_j$	$R(t_{(j)})$

Logrank Test

- Under H_0, the expected number of deaths in group 1 is

$$E_{1j} = R_1(t_{(j)})\frac{m_j}{R(t_{(j)})}$$

- The hypergeometric variance is

$$V_{1j} = \frac{R_1(t_{(j)})R_2(t_{(j)})m_j\{R(t_{(j)}) - m_j\}}{R(t_{(j)})^2\{R(t_{(j)}) - 1\}}$$

Logrank Test

- The logrank (Mantel-Haenszel) statistic:

$$E_1 = \sum_{j=1}^{k} E_{1j},\ O_1 = \sum_{j=1}^{k} m_{1j},\ V_1 = \sum_{j=1}^{k} V_{1j}$$

- Under $H_0 : S_1(t) = S_2(t)$ for all t,

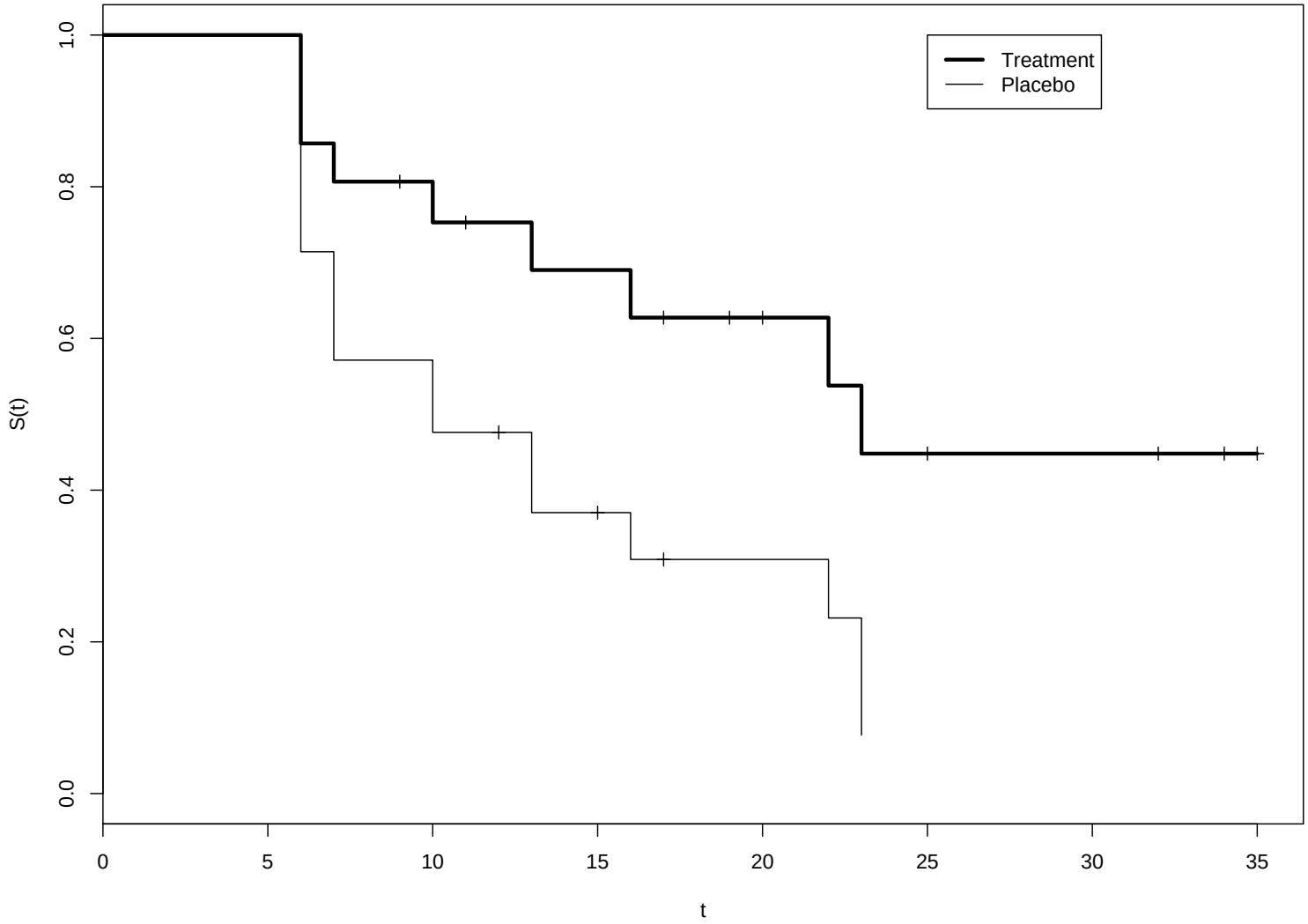
$$X = \frac{(O_1 - E_1)^2}{V_1} \sim \chi^2_1$$

Logrank Test

- Leukemia Example

Treatment ($n = 21$)	Placebo ($n = 21$)
6, 6, 6, 6+	6, 6, 6, 6
7, 9+, 10, 10+	6, 6, 7, 7
11+, 13, 16, 17+	7, 10, 10, 12+, 13
19+, 20+, 22, 23	13, 15+, 16, 17+
25+, 32+, 32+ 34+, 35+	22, 23, 23, 23+

Logrank Test: Leukemia Example



Code for Plotting Kaplan-Meier Curves

- R

```
library("survival")
fit <- survfit(Surv(t, delta)~rx,conf.type="none")
pdf("surv_leuk1.pdf",width=11,height=8.5)
plot(fit,xlab="t",ylab="S(t)",lwd=c(1,3))
legend(25,1,c("Treatment","Placebo"),lwd=c(3,1))
dev.off()
```

- SAS

```
proc lifetest plot=(s) graphics;
```

Logrank test "by hand": Leukemia Example

$t_{(j)}$	m_{1j}	$R_1(t_{(j)})$	m_{2j}	$R_2(t_{(j)})$	m_j	$R(t_{(j)})$	E_{1j}	V_{1j}
6	3	21	6	21	9	42	4.50	1.81
7	1	17	3	15	4	32	2.13	0.90
10	1	15	2	12	3	27	1.67	0.68
13	1	12	2	9	3	21	1.71	0.66
16	1	11	1	6	2	17	1.29	0.43
22	1	7	1	4	2	11	1.27	0.42
23	1	6	2	3	3	9	2.00	0.50
	9						14.57	5.4

Logrank test: Leukemia Example

- Therefore

$$X = \frac{(9 - 14.57)^2}{5.4} = 5.75$$

$$\Pr[\chi^2_1 > 5.75] = 0.0165$$

- R code:

```
Call:
survdiff(formula = Surv(t, delta) ~ rx)

      N Observed Expected (O-E)^2/E (O-E)^2/V
rx=p 21       17     11.4      2.72      5.75
rx=t 21        9     14.6      2.13      5.75

 Chisq= 5.8  on 1 degrees of freedom, p= 0.0165
```

Logrank test: Leukemia Example

- SAS code

```
proc lifetest;
  time t*delta(0);
  strata trt;
```

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	5.7507	1	0.0165
Wilcoxon	4.3357	1	0.0373
-2Log(LR)	6.0441	1	0.0140

Logrank test: SAS

```
data;
  input time group death wt;
cards;
6 1 1 3
6 1 0 18
6 2 1 6
6 2 0 15
7 1 1 1
7 1 0 16
7 2 1 3
7 2 0 12
.
.
.

proc freq order=data;
  tables time*group*death/chisq cmh;
  weight wt;
```

Logrank test: SAS

The FREQ Procedure

Summary Statistics for group by death
Controlling for time

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	5.7507	0.0165
2	Row Mean Scores Differ	1	5.7507	0.0165
3	General Association	1	5.7507	0.0165