BIOS 662 Fall 2016

Survival Analysis

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Introduction to survival data/analysis
- Kaplan-Meier estimator, standard error and CI
- Log-rank test
- (Cox / proportional hazards model)

Survival Analysis

- Chapter 16 of the text; BIOS 680/780
- Survival analysis: Response is time to an event
- Measure time from beginning of follow-up until an event such as incident disease, death, or relapse
- In a clinical trial, the beginning of follow-up is almost always the time of randomization
- In an epidemiology study, beginning of follow-up is usually the time of (initial) exposure assessment
- Examples:
  - time from kidney transplant until death
  - time from leukemia treatment to remission

Survival Analysis: Notation

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

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

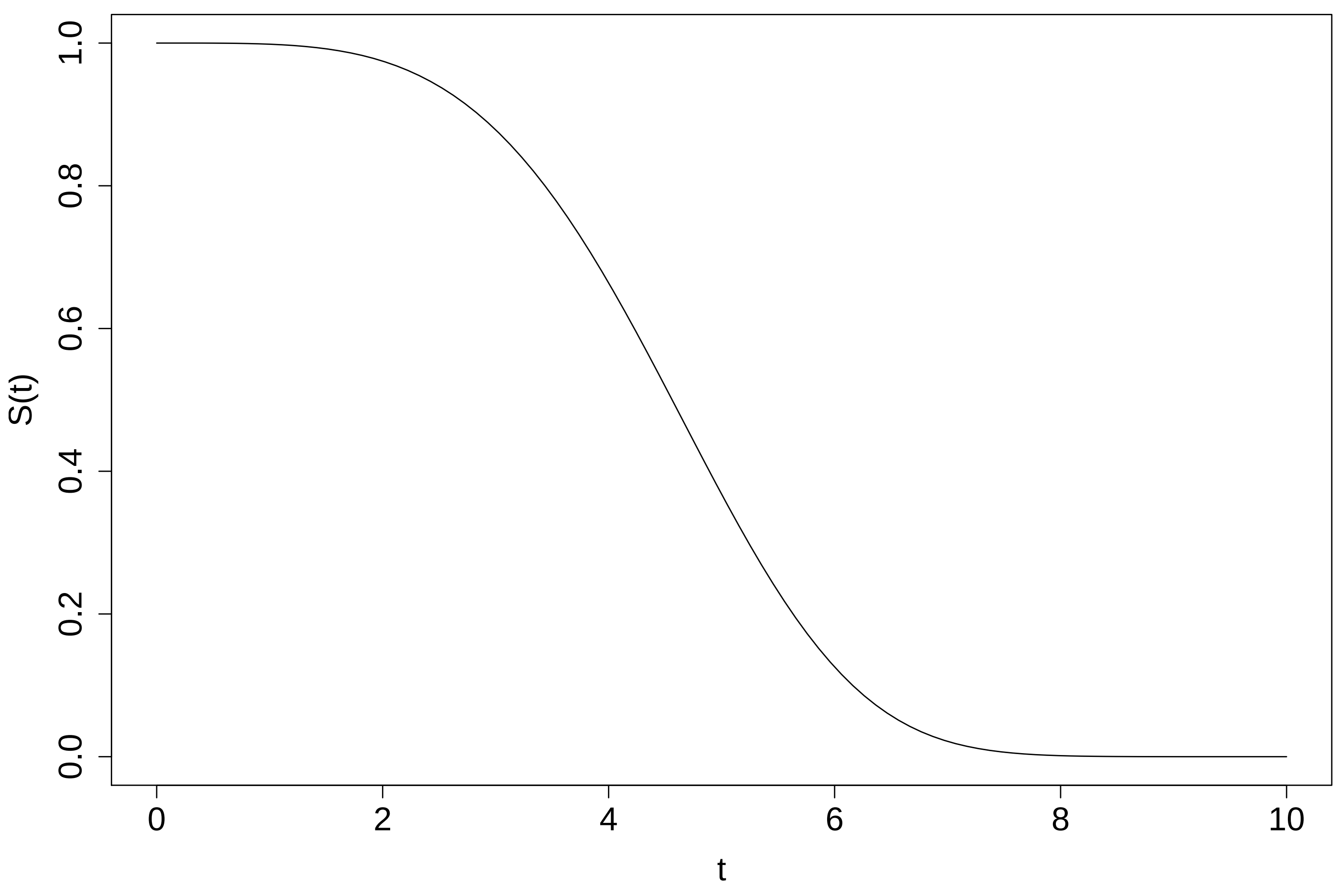
where $F(t)$ is the CDF of T^*

- Properties:

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

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

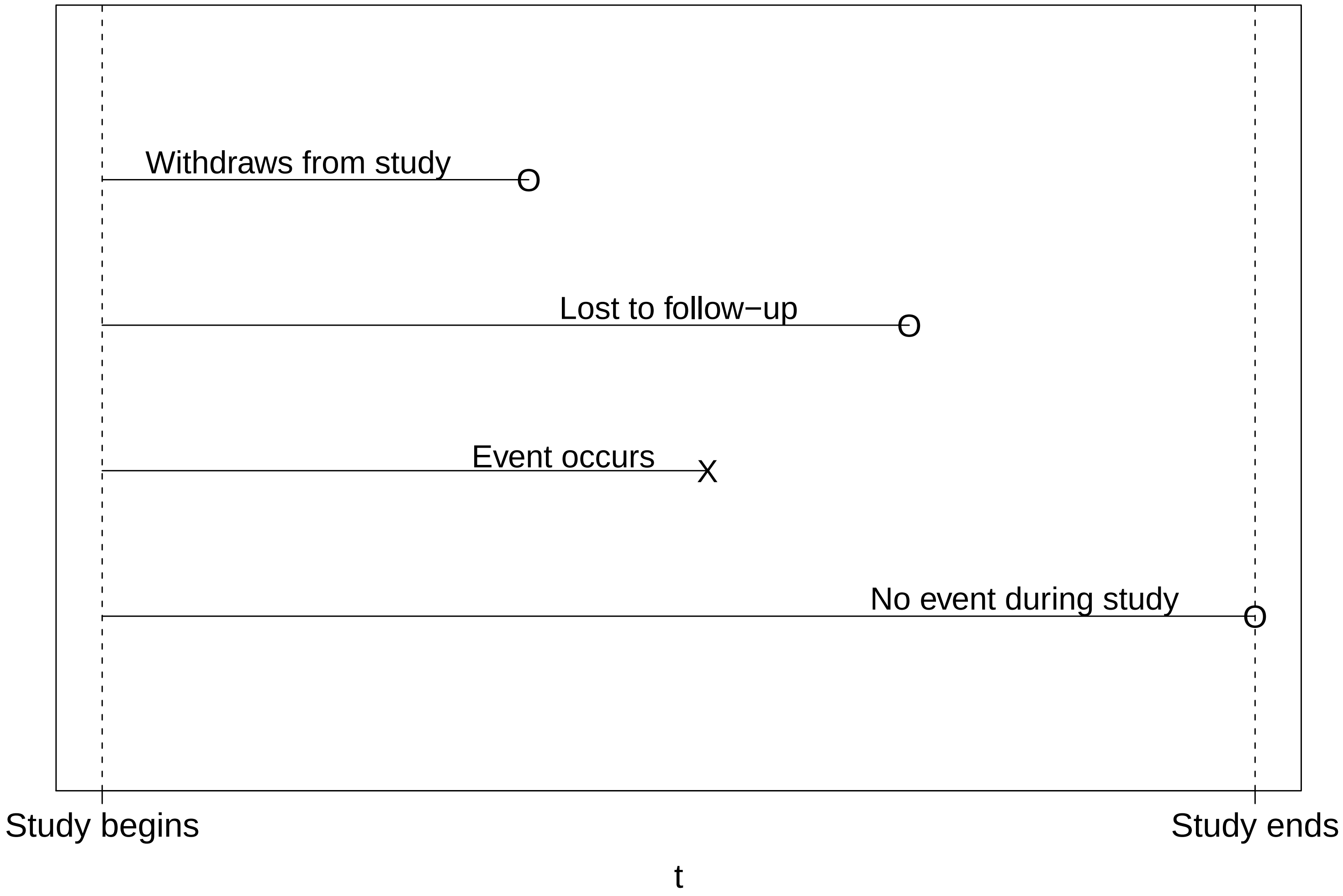
Example Survival Curve/Function



Censoring

- Often we do not know the exact time of failure of all subjects
- Reasons for **right** censoring:
  - subject does not experience the event of interest before the end of the study
  - subject is lost to follow-up during the study (e.g., withdraws from study, moves, dies from something other than the 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

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

- We 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 $S(t)$ with minimal assumptions?
- Answer 1: In the absence of censoring, use $1 - \mathrm{EDF}$
- Answer 2: Otherwise, use the Kaplan-Meier estimator

Tabular Summary of Data

- Let $t_{(1)}, t_{(2)}, \ldots, t_{(J)}$ be the distinct ordered failure times (censoring times are ignored)

Failure time $t_{(j)}$	Risk set $R(t_{(j)})$	No. of failures m_j	No. censored in $[t_{(j)}, t_{(j+1)})$ q_j
$t_{(0)} = 0$	$R(t_{(0)}) = N$	$m_0 = 0$	q_0
$t_{(1)}$	$R(t_{(1)})$	m_1	q_1
$t_{(2)}$	$R(t_{(2)})$	m_2	q_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(J)}$	$R(t_{(J)})$	m_J	q_J

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

Leukemia Example

$t_{(j)}$	$R(t_{(j)})$	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

Kaplan-Meier Estimator of $S(t)$

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

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

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

$$\begin{aligned}\hat{S}(t) &= \hat{S}(t_{(j-1)}) \cdot \widehat{\Pr}[T > t_{(j)} | T \geq t_{(j)}] \\ &= \hat{S}(t_{(j-1)}) \left(\frac{R(t_{(j)}) - m_j}{R(t_{(j)})} \right)\end{aligned}$$

- Assumes anyone censored at time $t_{(j)}$ has $T > t_{(j)}$

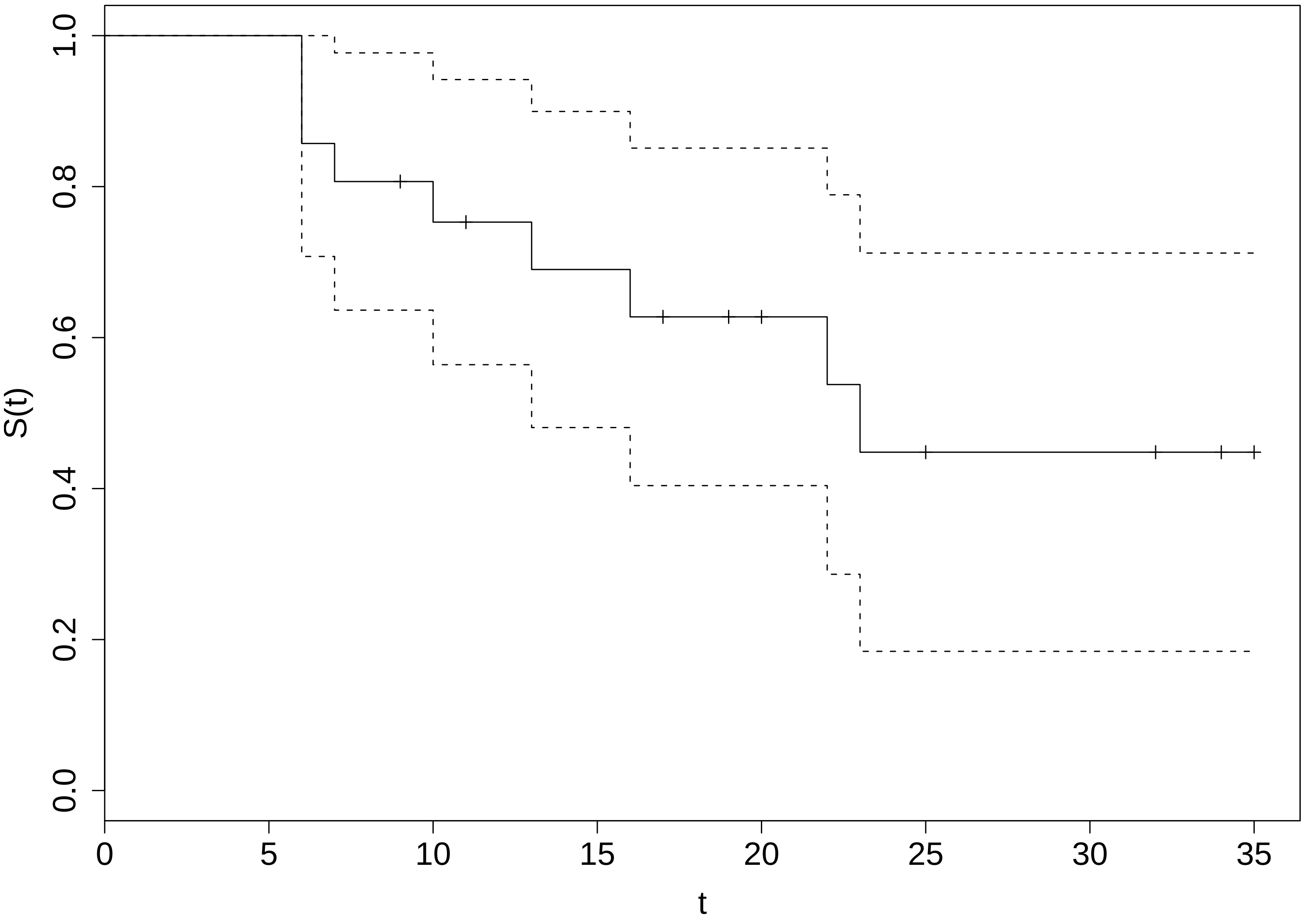
Kaplan-Meier Estimator

- KM is a nonparametric maximum likelihood estimator (NPMLE))
- Assumes independent censoring
- Also known as the *product limit estimator*
- If no censoring, KM equals 1 − EDF
- Alternative: *Life-table or actuarial method*

Leukemia Example

$t_{(j)}$	$R(t_{(j)})$	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	$18/21 = 0.857$
7	17	1	1	$0.857(16/17) = 0.807$
10	15	1	2	$0.807(14/15) = 0.753$
13	12	1	0	$0.753(11/12) = 0.690$
16	11	1	3	$0.690(10/11) = 0.627$
22	7	1	0	$0.627(6/7) = 0.538$
23	6	1	5	$0.538(5/6) = 0.448$

Kaplan-Meier Estimate for Leukemia Example



Kaplan-Meier Estimate: R

```
> t <- c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35)
> delta <- c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0)
> x <- rep(1,21)

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

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

> summary(fit)

Call: survfit(formula = Surv(t, delta) ~ x, conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.707        1.000
    7     17       1    0.807  0.0869        0.636        0.977
   10     15       1    0.753  0.0963        0.564        0.942
   13     12       1    0.690  0.1068        0.481        0.900
   16     11       1    0.627  0.1141        0.404        0.851
   22      7       1    0.538  0.1282        0.286        0.789
   23      6       1    0.448  0.1346        0.184        0.712
```

Kaplan-Meier Estimate: SAS

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

```
                         The LIFETEST Procedure
                     Product-Limit Survival Estimates

                                          Survival
                                          Standard     Number      Number
           t      Survival     Failure     Error       Failed       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
     13.0000        0.6902      0.3098      0.1068         6          11
     16.0000        0.6275      0.3725      0.1141         7          10
```

Kaplan-Meier Estimate: SAS cont.

```
17.0000*           .           .           .        7           9
19.0000*           .           .           .        7           8
20.0000*           .           .           .        7           7
22.0000       0.5378      0.4622      0.1282        8           6
23.0000       0.4482      0.5518      0.1346        9           5
25.0000*           .           .           .        9           4
32.0000*           .           .           .        9           3
32.0000*           .           .           .        9           2
34.0000*           .           .           .        9           1
35.0000*           .           .           .        9           0

     NOTE: The marked survival times are censored observations.


       Summary of the Number of Censored and Uncensored Values

                                               Percent
                 Total  Failed    Censored    Censored

                    21       9          12       57.14
```

Greenwood SE/CI of KM

- Let $n_j = R(t_{(j)})$
- Write the Kaplan-Meier estimator as

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

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 $t_{(j)}$

Greenwood SE/CI for KM

- Take logs

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

so that

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

- Binomial argument

$$\widehat{\mathrm{Var}}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$$

Greenwood SE/CI for KM

- Taylor series approximation

$$\widehat{\text{Var}}\big(g(X)\big) \approx \big(g'(\mu)\big)^2 \, \widehat{\text{Var}}(X)$$

implies

$$\widehat{\text{Var}}(\log \hat{p}_j) \approx \left(\frac{1}{\hat{p}_j}\right)^2 \left(\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}\right) = \frac{1-\hat{p}_j}{n_j \, \hat{p}_j}$$

$$= \frac{m_j}{n_j(n_j - m_j)}$$

- Thus

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

Greenwood SE/CI for KM

- Additional application of Taylor series approximation

$$\widehat{\mathrm{Var}}\big(\log \hat{S}(t)\big) \approx \big(\hat{S}(t)\big)^{-2}\widehat{\mathrm{Var}}\big(\hat{S}(t)\big)$$

  implying

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

- Thus

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

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

Greenwood SE/CI for KM

- For the leukemia example,

$$\widehat{\mathrm{SE}}(\hat{S}(6)) = 0.8571\sqrt{\frac{3}{21 \cdot (21-3)}} = 0.0764$$

$$\widehat{\mathrm{SE}}(\hat{S}(7)) = 0.8067\sqrt{\frac{3}{21 \cdot 18} + \frac{1}{17 \cdot 16}} = 0.0869$$

- An approximate $100(1-\alpha)\%$ CI is

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

Greenwood SE/CI for KM

- Greenwood based CIs are symmetric
- This is problematic when the survival function is near 0 or 1 because it is possible for part of the CI to lie outside the interval [0,1]
- Pragmatic solution: Set relevant end of interval to 0 or 1 in this case
- Many other methods exist to estimate the standard error and obtain confidence intervals
- All have pointwise interpretation; different methods exist to obtain *confidence bands*

Testing

- How do we test under minimal assumptions whether two survival functions are different?
- For example: Suppose 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 test)
- In the presence of right censoring, use the log-rank test

Log-Rank Test

- Suppose we have data from two samples

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

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

- We want to test

$$H_0 : S_1(t) = S_2(t) \text{ for all } t$$

where

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

Log-Rank Test

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

Group	At risk	Events	Survive
1	$R_1(t_{(k)})$	m_{1k}	$R_1(t_{(k)}) - m_{1k}$
2	$R_2(t_{(k)})$	m_{2k}	$R_2(t_{(k)}) - m_{2k}$
	$R(t_{(k)})$	m_k	$R(t_{(k)}) - m_k$

Log-Rank Test

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

$$E_{1k} = R_1(t_{(k)})\,\frac{m_k}{R(t_{(k)})}$$

- The hypergeometric variance is

$$V_{1k} = \frac{R_1(t_{(k)})R_2(t_{(k)})m_k\big(R(t_{(k)}) - m_k\big)}{R(t_{(k)})^2\,\big(R(t_{(k)}) - 1\big)}$$

Log-Rank Test

- The log-rank (Mantel-Haenszel) statistic uses

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

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

$$X = \frac{(O_1 - E_1)^2}{V_1} \sim \chi_1^2$$

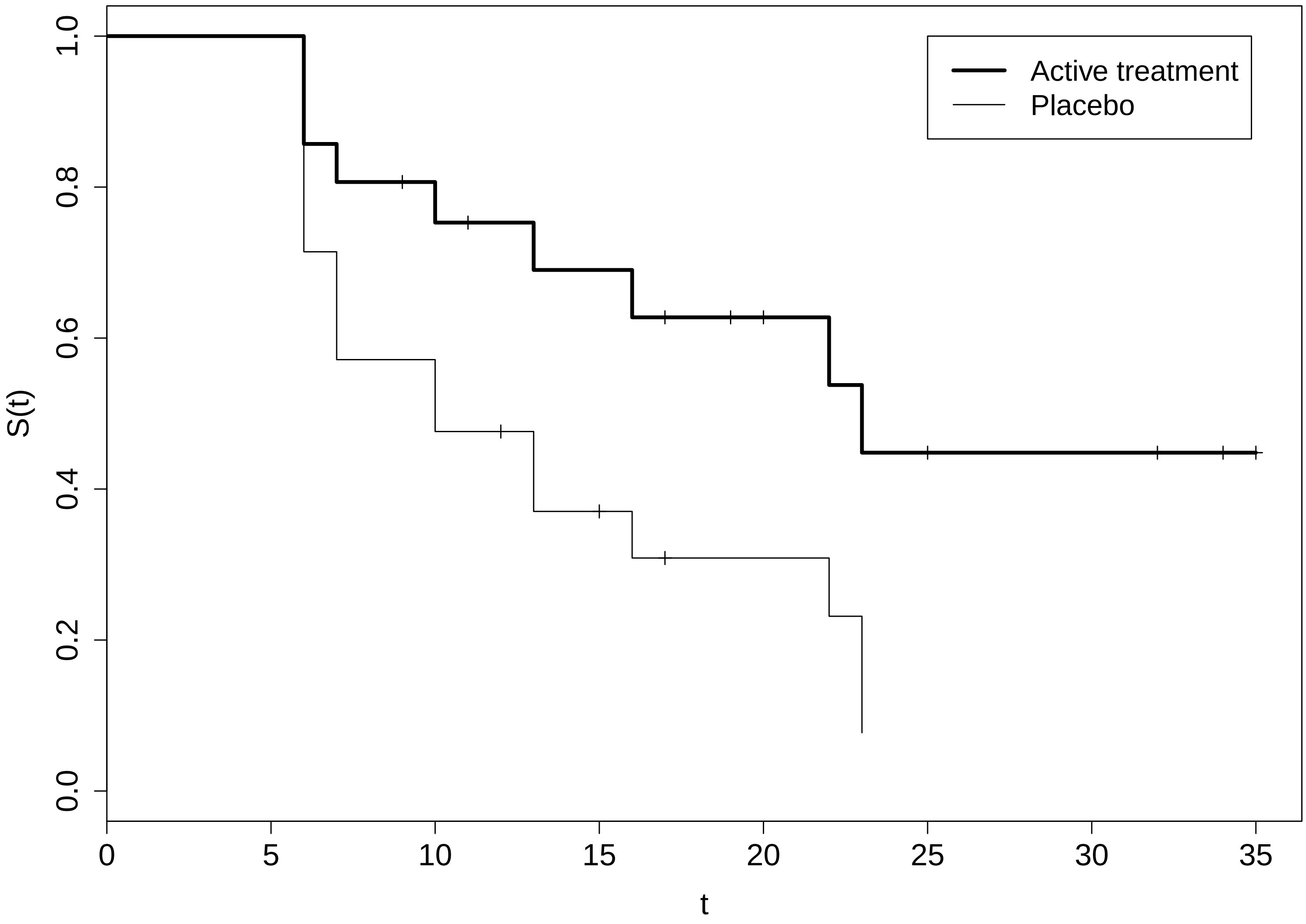
Log-Rank Test

- 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+

where + indicates that the person was censored at that time

Log-Rank 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("Active treatment","Placebo"),lwd=c(3,1))
dev.off()
```

- SAS

```
proc lifetest plots=(s) graphics;
  time t*delta(0);
  strata trt;
```

Log-Rank Test “By Hand”: Leukemia Example

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
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

Log-Rank Test: Leukemia Example

- Therefore

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

$$\Pr[\chi_1^2 > 5.75] = 0.0165$$

- R code:

```
> survdiff(Surv(t, delta)~rx)

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
```

Log-Rank Test: Leukemia Example cont.

- SAS code

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

```
                   Test of Equality over Strata

                                                Pr >
            Test        Chi-Square      DF    Chi-Square

            Log-Rank        5.7507       1      0.0165
            Wilcoxon        4.3357       1      0.0373
            -2Log(LR)       6.0441       1      0.0140
```

Log-Rank Test: SAS

- We can also use proc freq and the Mantel-Haenszel statistic, setting up a 2×2 table at each time point at which there is at least one event. All those in the risk set at such a time contribute to the table at that time

```
data;
  input time group remission 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
.
.
.
```

Log-Rank Test: SAS cont.

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

                      The FREQ Procedure


           Summary Statistics for group by remission
                      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
```

Cox / Proportional Hazards Model

- The *hazard function* $\lambda(t)$ is the instantaneous event rate at any time t

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$$

- The proportional hazards model is a linear model for the log of the hazard or, equivalently, a multiplicative model for the hazard

$$\log \lambda(t) = \log \lambda_0(t) + \beta X$$

or

$$\lambda(t) = \lambda_0(t) \exp(\beta X)$$

- $\lambda_0(t)$ is called the *baseline hazard*

Cox / Proportional Hazards Model

- Consider two values of X, x_1 and x_2; then

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t)\exp(\beta x_1)}{\lambda_0(t)\exp(\beta x_2)} = \frac{e^{\beta x_1}}{e^{\beta x_2}}$$

  independent of t

- This independence of t is an *assumption* and needs to be checked

- Let X be an indicator of being in one of two exposure or treatment groups, then if $x_1 = 1$ and $x_2 = 0$,

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{\beta \cdot 1}}{e^{\beta \cdot 0}} = e^{\beta}$$

- e^{β} is the *hazard ratio* comparing group 1 to group 2

Leukemia Treatment Example: R

- There are a substantial number of tied observations; R and SAS have different defaults for handling ties
- Using R's default method of handling ties (Efron)

```
> summary(coxph(Surv(t, delta)˜rx))
Call:
coxph(formula = Surv(t, delta) ˜ rx)

  n= 42

       coef exp(coef) se(coef)      z Pr(>|z|)
rxt -0.9684    0.3797   0.4164 -2.325   0.0200 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    exp(coef) exp(-coef) lower .95 upper .95
rxt    0.3797      2.634    0.1679    0.8588
```

Leukemia Treatment Example: SAS

- Using the "exact" method for ties rather than the SAS default (Breslow)

```
proc phreg;
   model t*delta(0) = active  / ties=exact;
```

```
                   Summary of the Number of Event and Censored Values

                                                            Percent
                         Total        Event     Censored    Censored

                            42           26           16       38.10

                        Analysis of Maximum Likelihood Estimates

                         Parameter    Standard                                Hazard
    Parameter    DF      Estimate        Error    Chi-Square  Pr > ChiSq       Ratio

    active        1      -0.97790      0.41896        5.4482      0.0196       0.376
```