Topics in Biostatistics
Analyzing Time-to-Event Data
Univariate Methods

B. Rosner

6/14/2018

Outline

- 1. Introduction to Survival Analysis
- (a) the Survival function
- (b) the Hazard function
- (c) Censored data

- 2. Kaplan-Meier Survival Estimates
- using the product-limit method

- 3. Log rank test for comparing 2 survival curves

Rationale

- Logistic regression can be used for either case-control studies, or cross-sectional studies to control for confounding where there is a binary outcome variable

- It can also be used for prospective studies, but usually in the setting where the follow-up time is the same for each subject.

Rationale (cont.)

- For example, suppose a group of asthmatics are assessed at baseline and at a 1 year follow-up visit.
- Endpoint = FEV % predicted < 80% at 1 year follow-up
- Study population = asthmatics with FEV % predicted > 80% at baseline

Rationale (cont.)

- However, in some prospective studies subjects are ascertained at multiple follow-up visits and the period of follow-up is variable for different subjects.

- Example: Time to MI in a clinical trial among high risk patients with unstable angina who are ascertained every 6 months for a 3-year period.

Rationale (cont.)

- The period of follow-up varies for different subjects due to
- (a) loss to follow-up,
- (b) mortality due to a non-CVD cause (e.g., lung cancer)

Rationale (cont.)

- Logistic regression could be used with variable follow-up times by including follow-up time as one of the covariates.

- However, we usually would have to assume a linear relationship between the logit of the probability of a positive outcome and follow-up time, which may not hold.

- Survival analysis is essential in this setting because it takes into account the time to an event, not just whether a subject has had an event.

Survival analysis- goals

- Estimate distribution of survival time for a population
- Test for the equality of survival distributions (e.g., active-treated vs. control group in a clinical trial)
- Estimate and control for the effects of other covariates

The Survival Function

- The random variable T = time to an event of interest (e.g., time to MI in a clinical trial, among subjects who have never had an MI)
- By definition, T must be > 0.
- The survival function S(t) is defined by:
- $S(t) = Pr(T > t)$

  = proportion of individuals who are event-free at time t.

The Survival Function (cont.)

- In the above example, S(t) = proportion of subjects free from MI at time t.

Survival function - properties

- By definition,

$(a)\ \ 0 \le S(t) \le 1$, for each $t \ge 0$

(b) $S(0) = 1$

(c) If $t_2 \ge t_1$, then $S(t_2) \le S(t_1)$,

(i.e., survival functions are non-increasing over time)

Example – Clinical trial to evaluate the efficacy of maintenance chemotherapy for leukemia

- The following study was reported by Embury, et al , Western Journal of Medicine (1977) and Miller, R.G., Survival Analysis (Wiley, 1998).

- The study was conducted among leukemia patients who were in remission.

Leukemia patients were randomized to 2 groups:

- (a) a maintenance chemotherapy group
- (b) a control group

Example – Clinical trial to evaluate the efficacy of maintenance chemotherapy for leukemia (cont.)

- The goal of the study was to compare the survival experience of the two treatment groups, where survival time is defined as time to relapse.

- Note that the endpoint is not just whether remission was maintained, but for how long it was maintained.

Clinical trial- preliminary data

- There were 11 patients randomized to maintenance chemotherapy, and 12 patients randomized to control.

- The times to relapse in each group are given as follows:

Clinical trial- preliminary data (cont.)

- Length of complete remission (wks)

- Maintenance chemotherapy group (n = 11)
- 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)
- 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Clinical trial- preliminary data (cont.)

- We notice, as is true in this example, that the distribution of survival times are frequently skewed.

- Thus, mean survival time is usually not a good summary measure.

- Instead, estimation of percentiles of the survival distribution is preferable.

Hazard function

- The hazard function is defined as the instantaneous rate of failure per unit time at time t, given that one has survived to time t.

- Note that h(t)>= 0, and has no bound.

- It is a rate, not a probability.

Relationship between the hazard function and the survival function

- The relationship between S(t) and h(t) is given by:

$$S(t) = \exp\left[-\int_{u=0}^{t} h(u)du \right].$$

Relationship between the hazard function and the survival function (cont.)

- In general, as h(t) increases, S(t) decreases,
- and conversely.

- Thus, if we compare two groups, where the
- hazard function for group 1 is always higher
- than for group 2, then the corresponding
- survival function for group 1 will be lower
- than for group 2.

Censored data

- Since most studies occur over a finite time period, the event of interest may not have occurred for some subjects during the study (e.g., time to MI)

- All that is known is that the time to an event (T) is greater than the period of follow-up C, where C is called the censoring time.

Censored data (cont.)

- Thus, for subjects who have had an event during the study, we have an actual event time T.

- For other subjects, who have not had an event, all that is known is that $T > C$, where C = censoring time.

- Thus, in general, we can only observe the follow-up time = $Y = \min(T,C)$ and the censoring indicator $\delta = 1$ if $T <= C$, $= 0$ if $T > C$.

Example of Censoring

- We note in the clinical trial data on slide 15 that the 3rd maintenance chemotherapy patient has a remission time of 13+ which means that the patient remained in remission for 13 weeks, did not have an event (i.e., relapse) and was censored at that point.

- The second maintenance chemotherapy patient had a remission time of 13 weeks, which indicates that they relapsed during the study at 13 weeks.

Causes of censoring

- 1. No event by the end of the study
- 2. Withdrawn from the study prior to the end of the study (i.e., loss to follow-up)

Causes of censoring (cont.)

- Loss to follow-up can be caused by either:

(a) dropout by the participant due to lack of interest, or

- (b) by an adverse event which presents a medical contraindication for continuing with the study

Causes of censoring (cont.)

- 3. Death from another (unrelated?) cause (e.g., death due to cancer in a study where the endpoint is MI).

- 4. Competing risks
- In a study where breast cancer is the outcome, subjects will usually be censored if they develop another type of cancer (e.g., lung cancer) before breast cancer. In this case, C = time to lung cancer.

Effects of censoring

- Standard statistical methods could be used if all event times T were observed, although distributions of event times would still typically be skewed.

- However, the presence of censoring makes this impossible (or very difficult), since the actual event time is unknown.

Types of censoring

- (a) Right censoring
- (b) Left censoring
- (c) Interval censoring

Right censoring

1. We know that the event time T is greater than the censoring time C. This is the most common form of censoring.

2. For example, if a subject is followed thru the end of a clinical trial and has not had an event.

Left censoring

- 1. We know that the event time T is less than the censoring time C, but we don't know the exact value of T.

- 2. For example, we know that a subject reports that they have diabetes, but the date of diagnosis of diabetes is unknown.

Interval censoring

- 1. We know that the event time is between C_1 and C_2, i.e. $C_1 <= T <= C_2$.

- 2. For example, if a disease occurs between two follow-up cycles (say 2 years apart), but the exact date of onset is unknown.

Effects of censoring on inference

- To make valid comparisons of survival time distributions between groups, we need to assume that the censoring time (C) is independent of the survival time (T).
- This assumption implies that :

$$\Pr(C|T = t) = \Pr(C), \text{ for all t.}$$

Effects of censoring on inference (cont.)

- An example when this is valid is administrative censoring, when a study is terminated at a fixed date and
- C = termination date minus enrollment date.

Effects of censoring on inference (cont.)

- An example when this is not valid is if subjects who are at higher risk for an endpoint (i.e., T low) tend to also be at higher risk for an adverse event and subsequent loss-to-follow-up (i.e. C low), than those at low risk for an endpoint (i.e., T high).

Estimation of Survival Probabilities

- The estimation of survival probabilities is complex because of the presence of censored data.

The survival times for the patients in the leukemia study are given in the next slide for easy reference.

Leukemia clinical trial data

- Length of complete remission (wks)

- Maintenance chemotherapy group (n = 11)
- 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)
- 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Estimation of Survival Probabilities (cont.)

- For example, in the leukemia example, there are 7 out of 11 patients (64%) in the maintained group who survived for at least 20 weeks.
- Does this mean that the estimated survival probability at 20 weeks for the maintained group should be:

$$\hat{S}(20) = 0.64?$$

Estimation of Survival Probabilities (cont.)

- The answer is probably not since it will be biased downwards because it doesn't take censoring into account.

Estimation of Survival Probabilities (cont.)

- Specifically, the 3rd subject in the maintained group survived for 13 weeks and withdrew from the study.

- The subject might have survived for 20+ weeks if his/her follow-up time was longer.

Estimation of Survival Probabilities (cont.)

- A nonparametric method for estimating survival probabilities that takes censoring into account is called the Kaplan-Meier or product-limit estimator.

Risk Sets

- We define a patient to be at risk for an event at time t if they have not experienced an event before time t and are not yet censored just before time t.

- Thus, for the maintenance chemotherapy group, the risk set consists of 11 patients at 9 weeks, 10 patients at 13 weeks, 8 patients at 18 weeks, etc.

Risk Sets (cont.)

- A subject who is censored at time t is assumed to have had no event up to time t and then was censored just after time t.

- Thus, the 3rd patient in the maintained group (coded as 13+) is in the risk set at 13 weeks, but is not followed beyond 13 weeks.

Product Limit Method

- Let t_i = distinct observed failure times (uncensored) in increasing order so that
- $t_1 < t_2 < \ldots < t_k$.
- Let k = number of distinct failure times
- Let n_i = number of subjects in the risk set at time t_i.
- Let d_i = number of failures (events) at time t_i.

Product Limit Method (cont.)

- Let

  $\hat{p}_i = d_i / n_i$

  = probability of failure at time t_i given that a subject is in the risk set at time t_i.

  Let $\hat{q}_i = 1 - \hat{p}_i$

  = probability of surviving beyond time t_i given that a subject is in the risk set at time t_i.

Kaplan-Meier Estimate of the Survival Distribution

- Thus,

1. $\hat{S}(0) = 1.$
2. It drops after each of the failure times.
3. Let t_k be the kth observed failure time.

In order for subject i to have $T_i > t_k$, subject i needs to
(1) be at risk at time t_1 and have $T_i > t_1$,
(2) be at risk at time t_2 and have $T_i > t_2$,
…
(k) be at risk at time t_k and have $T_i > t_k$.

Kaplan-Meier Estimate of the Survival Distribution (cont.)

- Thus,

$$\Pr(T_i > t_k) = \Pr(T_i > t_1 \,|\, \text{in risk set at time } t_1)$$

$$\text{x } \Pr(T_i > t_2 \,|\, \text{in risk set at time } t_2)$$

$$\text{x } \ldots$$

$$\text{x } \Pr(T_i > t_k \,|\, \text{in risk set at time } t_k).$$

Kaplan-Meier Estimate of the Survival Distribution (cont.)

- Thus,

$$\hat{S}(t) = 1 \text{ for } 0 \le t < t_1,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, \text{in risk set at time } t_1)$$

$$= 1 - d_1 / n_1 = \hat{q}_1 \quad \textit{for } t_1 \le t < t_2,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, \text{in risk set at time } t_1) \text{ x } \hat{\Pr}(T > t_2 \,|\, \text{in risk set at time } t_2)$$

$$= (1 - d_1 / n_1)(1 - d_2 / n_2) = \prod_{j=1}^{2} (1 - d_j / n_j) \quad \textit{for } t_2 \le t < t_3.$$

Kaplan-Meier Estimate of the Survival Distribution (cont.)

- In general,

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - d_j / n_j) = \prod_{j=1}^{k} \hat{q}_j \quad for\ t_k \le t < t_{k+1}.$$

Leukemia clinical trial data

- Length of complete remission (wks)

- Maintenance chemotherapy group (n = 11)
- 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)
- 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Estimation of survival time for the maintained group in the leukemia treatment study

- For the maintained group, $t_1 = 9$, $t_2 = 13$, $t_3 = 18$, etc.

- Note that the censored subject at 13 weeks is assumed to survive longer than the uncensored subject.

Estimation of survival time for the maintained group in the leukemia treatment study

n_j	j	t_j	$S(t_j)$
11	0	0	1
11	1	9	1-1/11 = 10/11 = 0.909
10	2	13	(10/11)(1-1/10) = 9/11 = 0.818
8	3	18	(9/11)(1-1/8) = (9/11)(7/8) = 0.716,
7	5	23	0.716 x (1-1/7) = 0.614
5	6	31	0.614 x (1-1/5) = 0.491
4	7	34	0.491 x (1-1/4) = 0.368
2	8	48	0.368 x (1-1/2) = 0.184

Use of Stata to estimate survival curves

- We can use Stata to obtain the Kaplan-Meier estimate of the survival curve.
- We first read in the data from a spreadsheet with separate columns for
- (a) the survival time (timevar),
- (b) the failure indicator(failvar),
- which = 1 if a subject fails and 0 if a subject is censored, and
- (c) a group indicator variable.

Use of Stata to estimate survival curves (cont.)

- After reading in the data Stata needs to be told that the data is survival time data, which we accomplish by using the stset command as follows:

- **.stset timevar, failure(failvar)**

Use of Stata to estimate survival curves (cont.)

- We can then use the sts list command to obtain a printed listing of the survival function for each group as follows:
- **.sts list , by (group)**

Use of Stata to estimate survival curves (cont.)

- We can then use the sts graph command to obtain plots of the survival function by group as follows:

- **.sts graph, by(group)**

- The results are given on the next few slides.

Stata output from leukemia example

- `. insheet using leukemia.names.csv, names comma clear`
- `(4 vars, 23 obs)`
-
- `. list`

Stata output from leukemia example (cont.)

	id	group	time	relapse
1.	1	maintained	9	1
2.	2	maintained	13	1
3.	3	maintained	13	0
4.	4	maintained	18	1
5.	5	maintained	23	1
6.	6	maintained	28	0
7.	7	maintained	31	1
8.	8	maintained	34	1
9.	9	maintained	45	0
10.	10	maintained	48	1
11.	11	maintained	161	0

Stata output from leukemia example (cont.)

```
    12. | 12      control      5          1 |
    13. | 13      control      5          1 |
    14. | 14      control      8          1 |
    15. | 15      control      8          1 |
        |-----------------------------------|
    16. | 16      control     12          1 |
    17. | 17      control     16          0 |
    18. | 18      control     23          1 |
    19. | 19      control     27          1 |
    20. | 20      control     30          1 |
        |-----------------------------------|
    21. | 21      control     33          1 |
    22. | 22      control     43          1 |
    23. | 23      control     45          1 |
        +-----------------------------------+
```

Stata output from leukemia example (cont.)

```
. stset time, failure (relapse)

     failure event:  relapse != 0 & relapse < .
obs. time interval:  (0, time]
 exit on or before:  failure

--------------------------------------------------------------------------
        23  total obs.
         0  exclusions
--------------------------------------------------------------------------
        23  obs. remaining, representing
        18  failures in single record/single failure data
       678  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =       161
```

Stata output from leukemia example (cont.)

```
. sts list, by(group)

         failure _d:  relapse
   analysis time _t:  time

           Beg.          Net            Survivor      Std.
  Time     Total    Fail   Lost           Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
control
     5        12       2      0             0.8333    0.1076     0.4817    0.9555
     8        10       2      0             0.6667    0.1361     0.3370    0.8597
    12         8       1      0             0.5833    0.1423     0.2701    0.8009
    16         7       0      1             0.5833    0.1423     0.2701    0.8009
    23         6       1      0             0.4861    0.1481     0.1919    0.7297
    27         5       1      0             0.3889    0.1470     0.1263    0.6498
    30         4       1      0             0.2917    0.1387     0.0724    0.5609
    33         3       1      0             0.1944    0.1219     0.0312    0.4614
    43         2       1      0             0.0972    0.0919     0.0057    0.3489
    45         1       1      0             0.0000         .          .         .
```

Stata output from leukemia example (cont.)

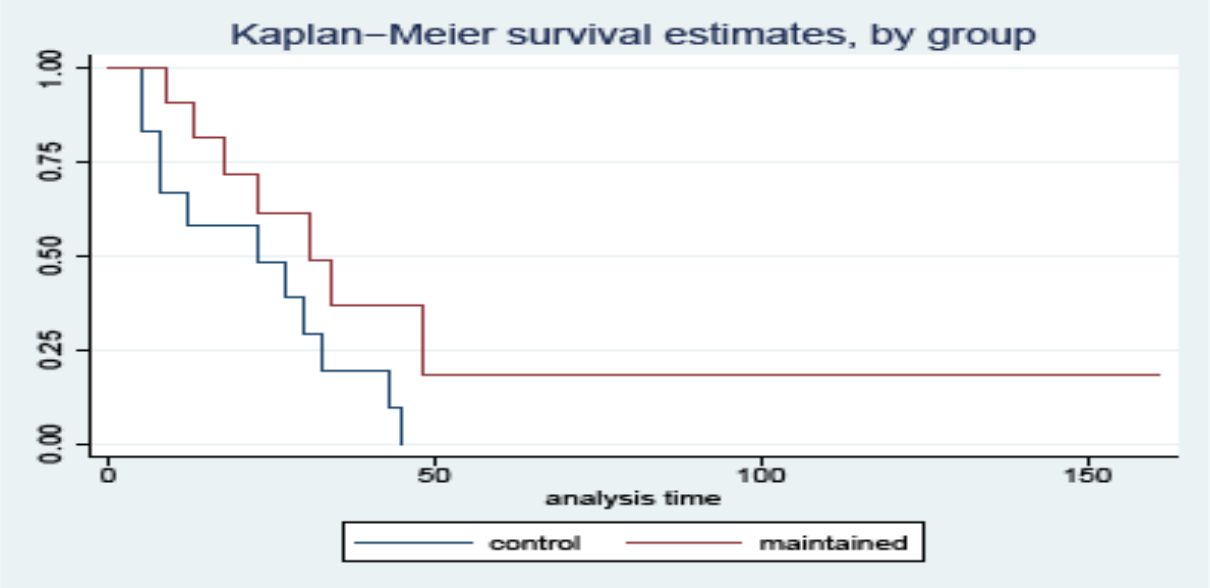
```
maintained
     9       11       1       0          0.9091    0.0867    0.5081    0.9867
    13       10       1       1          0.8182    0.1163    0.4474    0.9512
    18        8       1       0          0.7159    0.1397    0.3502    0.8990
    23        7       1       0          0.6136    0.1526    0.2658    0.8353
    28        6       0       1          0.6136    0.1526    0.2658    0.8353
    31        5       1       0          0.4909    0.1642    0.1673    0.7534
    34        4       1       0          0.3682    0.1627    0.0928    0.6570
    45        3       0       1          0.3682    0.1627    0.0928    0.6570
    48        2       1       0          0.1841    0.1535    0.0117    0.5250
   161        1       0       1          0.1841    0.1535    0.0117    0.5250
-------------------------------------------------------------------------------
```

This agrees with the calculations for the maintained group on slide 50.

Stata output from leukemia example (cont.)

```
. sts graph , by(group)

        failure _d:  relapse
  analysis time _t:  time
```


Kaplan–Meier survival estimates, by group
1.00
0.75
0.50
0.25
0.00
0
50
100
150
analysis time
control
maintained

Comparison of survival functions by group in leukemia example

- We see that at each time t, the survival function for the maintained group is consistently higher than the survival function for the control group.

Estimation of Survival Probabilities in SAS

- The SAS procedure

- PROC LIFETEST
- can estimate survival probabilities using the Kaplan-Meier approach
- can provide confidence limits for the survival probabilities
- can compare survival curves among 2 or more groups

Log rank test

- To compare 2 survival curves, we use a generalization of the Mantel-Haenszel test known as the log rank test.
- The log rank test corresponds to an application of the Mantel-Haenszel test to censored survival data.
- The test was first introduced by Mantel (Cancer Chemotherapy Reports, 1966).

Log rank test - hypotheses

- The log rank test is used to test the hypothesis

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

$$H_1 : S_1(t) \neq S_2(t) \text{ for some t.}$$

Log rank test – hypotheses (cont.)

- An equivalent statement of the hypotheses in terms of hazards is given by:

$$H_0 : h_1(t) = h_2(t) \text{ for all t, vs:}$$

$$\mathrm{H}_1 : h_1(t) \neq h_2(t) \text{ for some t.}$$

Log rank test – hypotheses (cont.)

If $h_2(t) / h_1(t)$ is defined as the hazard ratio at time t = HR(t),

then the log rank test has the most power when under H_1 :

$HR(t)$ is a constant = λ, in which case

$$S_2(t) = \left[S_1(t)\right]^{\lambda}.$$

If $\lambda > 0$, then $S_2(t) < S_1(t)$ for all t,

i.e., group 2 has a lower survival probability than group 1 for all t.

If $\lambda < 0$, then $S_2(t) > S_1(t)$ for all t,

i.e., group 2 has a higher survival probability than group 1 for all t.

Log rank test - procedure

- 1. We construct a 2 x 2 table at each observed failure time of treatment group x failure.

- 2. At each failure time, we calculate:
- (a) the observed number of failures in group 1,
- (b) the expected number of failures in group 1 under H_0,
- (c) the variance of the number of failures in group 1 under H_0.

Log rank test - procedure

- 3. We then sum the observed, expected and variance of the number of events over all failure times and construct an overall test statistic.

Log rank test –construction of 2 x 2 tables

- 1. Suppose there are K distinct failure times that are ordered as follows:
- $t_1 < t_2 < \ldots < t_K$.

- 2. At the i^{th} failure time (t_i), we have the following table:

Log rank test –construction of 2 x 2 table at time t_i (cont.)

Group	**Failure**			
	Yes	**No**	**Total**	
Maintained	d_{1i}	n_{1i}-d_{1i}	n_{1i}	
Control	d_{2i}	n_{2i}-d_{2i}	n_{2i}	
Total	d_i	n_i-d_i	n_i	

Log rank test –construction of 2 x 2 table at time t_i (cont.)

- n_{1i}= number of maintenance group subjects in the risk set at time t_i
- n_{2i} = number of control group subjects in the risk set at time t_i
- n_i = total number of subjects (both groups combined) in the risk set at time t_i

- d_{1i} = number of failures in the maintenance group at time t_i
- d_{2i} = number of failures in the control group at time t_i

  d_i = number of failures in both groups combined at time t_i

Log rank test –test statistic

- We add the observed number of failures, expected number of failures and the variance of the number of failures in the maintenance group (group 1) over all unique failure times.

- These are given by:

Log rank test –test statistic (cont.)

$$O = \sum_{i=1}^{K} O_i = \sum_{i=1}^{K} d_{1i} = \text{total observed failures in group 1,}$$

$$\text{Under } H_0,$$

$$E = \sum_{i=1}^{K} E_i = \sum_{i=1}^{K} n_{1i} d_i / n_i = \text{total expected failures in group 1,}$$

$$V = \sum_{i=1}^{K} V_i = \sum_{i=1}^{K} \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)} = \text{ variance of O.}$$

All margins are considered fixed at each failure time.

Thus, d_{1i} = the number of failures at time t_i in group 1 is considered random conditional on the total number of failures and the total number of subjects in the risk set in each group at time t_i

Log rank test –test statistic (cont.)

- Thus, we have the test statistic:

$$X^2_{LR} = \frac{\left(|O - E| - 0.5\right)^2}{V} \sim \chi^2_1 \text{ under } \mathrm{H}_0,$$

$$\text{p-value} = \Pr(\chi^2_1 > X^2_{LR}).$$

We should only use this test if $\mathrm{V} \geq 5$.

Comparison of survival curves in leukemia example

- The raw data are as follows:
- Length of complete remission (wks)

- Maintenance chemotherapy group (n = 11)
- 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)
- 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Comparison of survival curves in leukemia example (cont.)

- The 1st failure (relapse) occurred at 5 weeks.
- Maintained group:
- 11 people in the risk set and 0 failures at 5 weeks

- Control group:
- 12 people in the risk set and 2 failures at 5 weeks.
- Hence, we have the following 2 x 2 table:

Leukemia example, 2 x 2 table at 5 weeks

	Failure			
Group	**yes**	**no**	**total**	
Maintained	0	11	11	
Control	2	10	12	
total	2	21	23	

Leukemia example, 2 x 2 table at 5 weeks (cont.)

- Hence,

$$O_1 = 0,$$

$$E_1 = \frac{11(2)}{23} = 0.96,$$

$$V_1 = \frac{11(12)(2)(21)}{23^2(22)} = 0.476.$$

Leukemia example, 2 x 2 table at 8 weeks

- The next failure (relapse) was at 8 weeks.
- Maintained group:
- 11 people in the risk set and 0 failures at 8 weeks
- Control group:
- 10 people in the risk set and 2 failures at 8 weeks.
- Hence, we have the following 2 x 2 table:

Leukemia example, 2 x 2 table at 8 weeks (cont.)

	Failure			
Group	**yes**	**no**	**total**	
Maintained	0	11	11	
Control	2	8	10	
total	2	19	21	

Leukemia example, 2 x 2 table at 8 weeks (cont.)

- Hence,

$$O_2 = 0,$$

$$E_2 = \frac{11(2)}{21} = 1.05,$$

$$V_2 = \frac{11(10)(2)(19)}{21^2(20)} = 0.474.$$

Leukemia example, full data analysis

- The calculations over the 15 unique failure times are given in the spreadsheet on the next slide.

Comparison of Survival Curves in the Leukemia Example

t_j	d_1j	n_1j	d_2j	n_2j	d_j	n_j	E_j	V_j
5	0	11	2	12	2	23	0.96	0.476
8	0	11	2	10	2	21	1.05	0.474
9	1	11	0	8	1	19	0.58	0.244
12	0	10	1	8	1	18	0.56	0.247
13	1	10	0	7	1	17	0.59	0.242
18	1	8	0	6	1	14	0.57	0.245
23	1	7	1	6	2	13	1.08	0.456
27	0	6	1	5	1	11	0.55	0.248
30	0	5	1	4	1	9	0.56	0.247
31	1	5	0	3	1	8	0.63	0.234
33	0	4	1	3	1	7	0.57	0.245
34	1	4	0	2	1	6	0.67	0.222
43	0	3	1	2	1	5	0.60	0.240
45	0	3	1	1	1	4	0.75	0.188
48	1	2	0	0	1	2	1.00	0.000
Tot	7						10.69	4.008

Comparison of survival curves in leukemia example

- We have that :

$O = 7,\ \mathrm{E} = 10.69,\ \mathrm{V} = 4.008.$

We have ignored the condition that V should be ≥ 5, because V is only slightly less than 5 in this example.

Thus,

$$X^2_{LR} = \frac{\left(|7 - 10.69| - 0.5\right)^2}{4.008} = 2.54 \sim \chi^2_1 \text{ under } \mathrm{H}_0,$$

$$\text{p-value} = \Pr(\chi^2_1 > 2.54) = 0.111.$$

Using Stata to perform the log rank test

- We can also use Stata to perform the log rank test.
- For this purpose, we use:

- `.stset time, failure(relapse)`
- `.sts test group`

- The results are given on the next slide.

Using Stata to perform the log rank test (cont.)

```
. sts test group

           failure _d:  relapse
     analysis time _t:  time


Log-rank test for equality of survivor functions
----------------------------------------------

             |   Events          Events
group        |  observed        expected
-------------+-------------------------
control      |        11            7.31
maintained   |         7           10.69
-------------+-------------------------
Total        |        18           18.00

                  chi2(1) =        3.40
                  Pr>chi2 =      0.0653
```

Using Stata to perform the log rank test (cont.)

- Note that Stata does not use a continuity correction and hence has a slightly lower p-value.

Summary

- In this lecture, we discussed
- 1. Basic principles of analyzing time-to-event data including:

- (a) survival time (i.e.,S(t))
- (b) hazard rate (i.e., h(t))
- (c) censored data

Summary (cont.)

- 2. The Kaplan-Meier product limit method to estimate a survival curve.
- 3. The log rank test for comparing two survival curves.
- Multivariate methods for analyzing time-to-event data will be discussed in the next lecture.