Class 5 Outline

1. Time to event (survival) data, censoring survival times and the survivor function

2. The survivor function and the hazard rate
   - Uncensored data, Censored data

3. Kaplan-Meier estimates of the survivor function, *S(t),* for ungrouped survival data

4. Example using AML Data

5. Summary

1

0. Learning Objectives

- Describe ungrouped survival data in which the exact time to event or censoring is known

- Describe the survival function and hazard function

- Describe how to estimate the survivor function using the Kaplan-Meier survival curve and confidence intervals for ungrouped survival data

Key words - censoring, survival function, hazard function, Kaplan-Meier estimates

2

1. Time to Event Data

- The outcome (event) of interest is dichotomous
- The study design may not be able to assure that the outcome is know for all individuals at the endpoint of the study.

- Uncensored data: The event has occurred
- Censored data: The event has yet to occur
  - Event-free at the current follow-up time
  - A competing event that is not an endpoint stops follow-up
  - Death (if not part of the endpoint)
  - Clinical event that requires treatment, etc.

3

1.1 Survival Times

- Distribution of times to event – often called "survival times," even when the "event" is not "death"
- Survival times follow a continuous distribution with times ranging from zero to infinity
- The probability distribution of the survival times can be described by:
  - cumulative distribution function
  - density function
  - survivor function = 1 - cumulative distribution function
  - hazard function

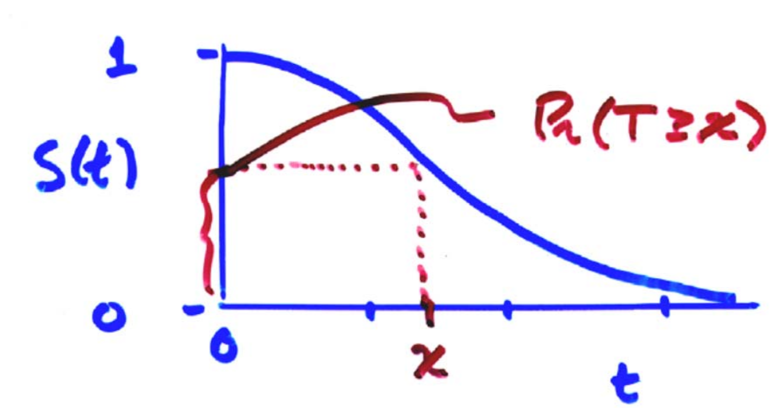
4

2. Survivor Function, *S(t)*

- The survival function, denoted S(t), is a useful way to represent the probability distribution of the survival time T, when some of the observed times are censored – only know that T>t, rather than T=t
- $S(t) = Pr(T > t) \qquad 0 < t < \infty$
- *S(t)* is the probability of surviving beyond *t*

5

2.1 Schematic of the Survivor Function



6

2.2a Handling Uncensored Survival Data

- Use the “usual” methods
  - — t-tests
  - — regression
  - — ANOVA
  - — transforms
- The survival curve as an important and complete summary of a population

$$S(t) = \frac{(\text{"alive"}^{*} \text{ at followup time t})}{(\text{alive at time 0})}$$

$$0 = \begin{Bmatrix} \text{Date of randomization} \\ \text{Birth} \end{Bmatrix}$$

*** Not had the event**

7

2.2b Handling Uncensored Survival Data

- The survival curve starts at 1.0 and decreases over time
- Estimating these curves and comparing them among groups constitutes a “survival analysis”
- Need to decide on what summary measure is important
  - – Mean survival time
  - – Median survival time
  - – Value at a specific time: S(12)
  - – Difference of curves: S1(12) - S2(12)
  - – Maximal difference

8

2.3 Example with Uncensored Data

- Data: 2, 3, 3, 5, 6, 9, 9, 10, 13, 16

Time	$\hat{S}(t)$
0	1.00
1	1.00
2	0.90
3	0.70
4	0.70
5	0.60
6	0.50
7	0.50
8	0.50
9	0.30
10	0.20
11	0.20
12	0.20
13	0.10
14	0.10
15	0.10
16	0.00

2.4 Ways of Handling Censoring

1. Ignore the incomplete cases; drop them (<u>never!</u>)
   - Produces bias in the estimated curve
   - Unbalanced censoring produces biased comparisons

2. Impute a missing event time
   - Depends on a detailed probability model
3. Calculate the overall event rate

4. Use the available information on each participant
   - Important issue: If no events are reported in the interval from last follow-up to "now", we need to choose between:
     - No news is good news?
     - No news is no news?

2.5 Overall Event Rate

- Overall event rate:

$$\text{Event rate} = \frac{\text{\# events}}{\text{total observation time}}$$

- Example: 5 events in 600 person months

= 1 event per 120 months
= 1 event per 10 years
= 0.1 events per year
= 10 events per 100 person-years

- Gives an average event rate over the follow-up period; actual event rate may vary over time

- For a finer time resolution, use small intervals

11

3.0 Kaplan-Meier Estimates of the Survivor Function, S(t)

Biostatistics Trivia: Professor Paul Meier was an assistant professor in the JHU Department of Biostatistics from 1952 to 1957. He teamed with E.L. Kaplan to write their seminal paper "Non-parametric Estimation from Incomplete Observations," which appeared in the *Journal of the American Statistical Association* in 1958. This paper was to lay the groundwork for modern survival analysis.

12

3.1a The Hazard Function

- Basic idea: Estimate the hazard of death at each event time t using available data and then use them to produce the survival curve by multiplying (1 - hazard) terms
- The hazard =
  Pr(event “now” | no event yet))/unit time
  where “now” means in the current unit interval
- Thus, (1-hazard) =
  Pr(no event “now”| no event yet)

13

3.1b The Hazard Function

- The hazards for time intervals i=1, 2, 3, ... are h_1, h_2, h_3,…
- Example: $S(3) = (1 - h_1) \times (1 - h_2) \times (1 - h_3)$
- If the hazard is large, the survival curve decreases rapidly
- Estimate $\mathbf{h_i} = \dfrac{\text{\# events at t=i}}{\text{\# at risk at t=i}} = \dfrac{Y_i}{n_i}$
- Kaplan-Meier estimate: $\hat{S}(t) = \prod_{i \text{ for } t_i < t} (1-h_i)$

14

3.2 Relationship between the Survivor and Hazard Functions

Example:

S(3) = Pr(survive for 3 months)
= Pr(survive 1st month) ×
Pr(survive 2nd month | survive 1st month) ×
Pr(survive 3rd month | survive 2nd month)

- Thus, $$S(3) = S(1) \cdot \frac{S(2)}{S(1)} \cdot \frac{S(3)}{S(2)}$$

$$S(3) = S(1) \cdot \frac{S(2)}{S(1)} \cdot \frac{S(3)}{S(2)}$$

$$S(3) = (1\text{-}h_1) \cdot (1\text{-}h_2) \cdot (1\text{-}h_3)$$

15

3.3a Calculating the Hazard – Uncensored Data

- Data: 2, 3, 3, 5, 6, 9, 9, 10, 13, 16 (uncensored data)

$h_1 = 0$	(0/10)
$h_2 = 0.10$	(1/10)
$h_3 = 0.22$	(2/9)
$h_4 = 0$	(0/7)
$h_5 = 0.14$	(1/7)
$h_6 = 0.17$	(1/6)
$h_7 = h_8 = 0$	(0/5)
$h_9 = 0.40$	(2/5)
$h_{10} = 0.33$	(1/3)
$h_{11} = h_{12} = 0$	(0/2)
$h_{13} = 0.50$	(1/2)
$h_{14} = h_{15} = 0$	(0/1)
$h_{16} = 1.00$	(1/1)

16

3.3b Calculating the Survivor Function - Uncensored Data

- Uncensored (complete) data example

$S(3) = (1 - h_1)(1 - h_2)(1 - h_3)$

- Estimate by

$= (1 - y_1/n_1)(1 - y_2/n_2)(1 - y_3/n_3)$

$= (1 - 0/10)(1 - 1/10)(1 - 2/9)$

$= 1 \times 9/10 \times 7/9 = 0.70$

where y_i is the number of events at time i and n_i is the number at risk at time i

17

3.4 Hazard Function and the Survival Curve

- If we know the hazard function, we know the survival curve:

$$S(t_i) = \prod_{i \le t} (1 - h_i) = (1 - h_i) S(t_{i-1})$$

- If we know the survival curve, we know the hazard function:

cumulative hazard = -log[S(t)]

hazard = increments of cumulative hazard

18

3.5a Calculating the Hazard – Censored Data

- Data: 2, 3, 3*, 5, 6*, 9, 9*, 10, 13, 16 (* = censored at end of the interval)

h1 = 0	(0/10)
h2 = 0.10	(1/10)
h3 = 0.11	(1/9)
h4 = 0	(0/7)
h5 = 0.14	(1/7)
h6 = 0	(0/6)
h7 = h8 = 0	(0/5)
h9 = 0.20	(1/5)
h10 = 0.33	(1/3)
h11 = h12 = 0	(0/2)
h13 = 0.50	(1/2)
h14 = h15 = 0	(0/1)
h16 = 1.00	(1/1)

19

3.5b Calculating the Survivor Function Censored Data

- Censored data example

$$S(3) = (1 - h_1)(1 - h_2)(1 - h_3)$$

- Estimate by

$$= (1-0)(1-0.10)(1-0.11)$$
$$= 1 \times .90 \times 8/9$$
$$= .80$$

20

4. Example- AML Data

- Example: Acute Myelogenous Leukemia (AML)

  Y - time from start of treatment to cancer relapse

  X - Indicates chemotherapy group
  (=0 for not maintained on chemotherapy group and
  =1 for maintained on chemotherapy)

21

4.1 Survival Data - AML Example

- Consider a clinical trial in patients with acute myelogenous leukemia (AML) comparing two groups of patients: no maintenance treatment with chemotherapy ($X=0$) -vs- maintenance chemotherapy treatment ($X=1$)

Group	Weeks in remission -- ie, time to relapse
Maintenance chemo ($X=1$)	*9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+*
No maintenance chemo ($X=0$)	*5, 5, 8, 8, 12, 16+, 23, 27, 30+, 33, 43, 45*

- \+ indicates a censored time to relapse; e.g., 13+ = more than 13 weeks to relapse

22

4.2 Schematic of AML Survival Data

4.3a Stata Commands for Survival Data

- There are many **Stata** commands for input, management, and analysis of survival data, most of which are found in the manual in the *st* section – all survival data commands start with *st*
- *st* can be used to analyze individual level data (Kaplan-Meier, Cox regression, etc) or to group the individual level data for grouped analysis (SMRs, output for Poisson regression, etc.)
- **Stata** 15 Reference manual

24

4.3b Stata Commands for Survival Data

- With ungrouped survival data on individuals:

    1. Use the ordinary **Stata** input commands to input and/or generate the following variables:
        - X variables
        - Person-time (denominator) variable (if applicable)
        - Time variable containing follow-up time
        - Censoring variable indicating status at the end of follow-up either “failed” or “censored”

    2. Then, use the *st* commands, as illustrated, to process and analyze the data

25

4.3c Stata Commands for Survival Data

- Define survival data:
  stset command

  Used to define the time variable, the status variable with the codes for “failures,” and an “id” variable the uniquely identifies each individual observation

  stset t , failure(failed==1) id(id)

- Descriptive statistics for survival data:

  stdes, stsum command

26

4.4 Listing of AML Data

```
. list id Chemo time failed
     +---------------------------+
     | id   Chemo   time   failed |
     |---------------------------|
  1. |  1       1      9        1 |
  2. |  2       1     13        1 |
  3. |  3       1     13        0 |
  4. |  4       1     18        1 |
  .
  .
  .
| .
  .
 19. | 19       0     27        1 |
 20. | 20       0     30        0 |
     |---------------------------|
 21. | 21       0     33        1 |
 22. | 22       0     43        1 |
 23. | 23       0     45        1 |
     +---------------------------+
```

27

4.5 Defining Survival Data

```
.stset time, failure(failed==1) id(id)

                id:  id
     failure event:  failed == 1
obs. time interval:  (time[_n-1], time]
 exit on or before:  failure

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

28

4.6 Description of Survival Data

```
. stdes if Chemo==0

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
------------------------------------------------------------------------------
no. of subjects               12
no. of records                12           1           1          1          1

(first) entry time                         0           0          0          0
(final) exit time                      21.25           5       19.5         45

subjects with gap              0
time on gap if gap             0           .           .          .          .
time at risk                 255       21.25           5       19.5         45

failures                      10    .8333333           0          1          1
------------------------------------------------------------------------------
```

29

4.7b Summary of Survival Data

```
.stsum

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

         |               incidence       no. of    |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
   total |          678   .0250737            23         12        27        43
```

30

4.7b Description of Survival Data

```
. stdes if Chemo==1

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

                                   |-------------- per subject --------------|
Category                   total        mean         min      median        max
------------------------------------------------------------------------------
no. of subjects               11
no. of records                11           1           1           1          1

(first) entry time                         0           0           0          0
(final) exit time                   38.45455           9          28        161

subjects with gap              0
time on gap if gap             0           .           .           .          .
time at risk                 423    38.45455           9          28        161

failures                       7    .6363636           0           1          1
------------------------------------------------------------------------------
```

31

4.8 Overall Incidence Rates by Group

```
. stir Chemo
  failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

note:  Exposed <-> Chemo==1 and Unexposed <-> Chemo==0

                 | Chemo                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
         Failure |         7          10  |         17
            Time |       423         255  |        678
-----------------+------------------------+------------
                 |                        |
  Incidence Rate |  .0165485    .0392157  |  .0250737
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |        -.0226672       |  -.0498895     .004555
 Inc. rate ratio |         .4219858       |   .1363296    1.228119 (exact)
 Prev. frac. ex. |         .5780142       |  -.2281186    .8636704 (exact)
 Prev. frac. pop |         .3606195       |
                 +-------------------------------------------------
                    (midp)   Pr(k<=7) =                     0.0418 (exact)
                    (midp) 2*Pr(k<=7) =                     0.0836 (exact)
```

32

4.9a Kaplan-Meier Estimates of the Survivor Function – Not Maintained Group

```
. sts list  if Chemo==0          "Not Maintained" Group

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

           Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     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      0      1             0.3889    0.1470     0.1263    0.6498
    33        3      1      0             0.2593    0.1442     0.0484    0.5478
    43        2      1      0             0.1296    0.1166     0.0079    0.4224
    45        1      1      0             0.0000         .          .         .
-------------------------------------------------------------------------------
```

33

4.9b Kaplan-Meier Estimates of the Survivor Function – Not Maintained Group

Time	Number at Risk	Events	$S(t_i) = (1 - h_i)S(t_{i-1})$
0	12	0	1.0
5	12	2	1.0(1- 2/12) = 0.833
8	10	2	0.833(1- 2/10)=0.666
12	8	1	0.666(1-1/8) = 0.583
23	6	1	0.583(1- 1/6) =0.486
27	5	1	0.486(1- 1/5) =0.389
33	3	1	0.389(1- 1/3) = 0.259
43	2	1	0.259(1 -1/2) = 0.130
45	1	1	0.130(1 – 1/1) = 0

34

4.10a Kaplan-Meier Estimates of the Survivor Function –Maintained Group

```
. sts list  if Chemo==1            "Not Maintained" Group

         failure _d:  failed == 1
   analysis time _t:  t
                id:  id

            Beg.              Net              Survivor      Std.
  Time     Total    Fail     Lost              Function     Error      [95% Conf. Int.]
-------------------------------------------------------------------------------
   -----------------------------------------------------------------
     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
```

35

4.10b Kaplan-Meier Estimates of the Survivor Function –Maintained Group

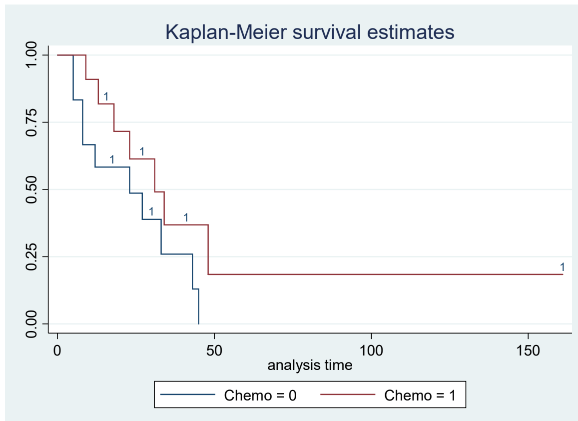
Time	Number at Risk	Events	$S(t_i) = (1 - h_i)S(t_{i-1})$
0	11	0	1.0
9	11	1	1.0(1- 1/11) = 0.909
13	10	1	0.909(1- 1/10) = 0.818
18	8	1	0.818(1- 1/8) = 0.716
23	7	1	0.716(1- 1/7) = 0.614
31	5	1	0.614(1- 1/5) = 0.491
34	4	1	0.491(1- 1/4) = 0.368
48	2	1	0.368(1- 1/2) = 0.184

36

4.11 Graph of Kaplan Meier Survival Curves

- Estimates of the survival function S(t) versus *time* -- separate curves for each group

`. sts graph, by(Chemo)lost` *or* `.sts graph, by(Chemo)risktable`



37

5.a Summary

- There are statistical techniques for describing and making inferences for time to event data (survival times) in the presence of censoring:
  - Overall incidence rate, survivor function

- The survivor function S(t) represents the probability distribution of the survival times; S(t) is the probability of surviving beyond t

- The hazard function, h_i = Pr(event "now"| no event yet)/unit time where "now" means in the current unit interval i

38

5.b Summary

- There is a relationship between the survivor function and hazard function in discrete time

$$S(t_i) = \prod_{i \leq t} (1 - h_i) = (1 - h_i) S(t_{i-1})$$

- The survivor function is a product of (1-hazard) terms

- Kaplan-Meier estimates of the survival curve for ungrouped data are calculated only at times that events occur

39