Session 6: Survival Analysis I

Levi Waldron

CUNY SPH Biostatistics 2

Learning objectives and outline

Learning objectives

1. Distinguish censored data from binary or continuous data
2. Define survival function, hazard functions, cumulative event function
3. Perform a Kaplan-Meier estimate
4. Perform, interpret, and identify assumptions of the logrank test
5. Define **potential follow-up time**
6. Calculate median survival time and potential follow-up time

Outline

1. Introduction to censored data
   - Outcome variable: time-to-event
   - Censored data
2. Survival function and Kaplan-Meier estimator
3. Comparing groups: Log-rank test

- Vittinghoff sections 3.1-3.5

Introduction to censored data

Outcome variable: time to event

- Generally time to the occurrence of a particular event, e.g.
  - death
  - disease recurrence
  - or other experience of interest
- Time: The time from the beginning of an observation period t0 (e.g. surgery) to:
  - an event, or
  - end of the study, or
  - loss of contact or withdrawal from the study

Typical research questions

- What is the median survival time (in years) of patients diagnosed with a certain disease?
- What is the probability of those patients surviving for at least 5 years?
- Are certain personal, behavioral, or clinical characteristics correlated with participant's chance of survival?
- Is there a survival difference between groups?
  - e.g. treatment vs. control
  - e.g. exposed vs. unexposed

Special considerations in survival analysis

- Survival data requires special techniques:
  - Survival data is generally not normally distributed
  - **Censoring** - observe individuals for differing lengths of time that may or may not result in an "event"
- Censoring is a key challenge in survival analysis. Consider a clinical study where:
  - patient 1 dies 1 month after diagnosis
  - patient 2 dies 12 years after diagnosis
  - patient 3 is lost to follow-up after 1 month
  - patient 4 is still alive after 12 years of follow-up

Question #1: which patients are "censored?"

Question #2: how would you rank these patients in order of disease severity?

Definitions

Definition: A survival time is said to be *right-censored* at time t if it is only known to be greater than t.

Definition: The *survival function* at time t, denoted $S(t)$, is the probability of being event-free at t. Equivalently, it is the probability that the survival time is greater than t.

leukemia Example: see leuk.csv

- Study of 6-mercaptopurine (6-MP) maintenance therapy for children in remission from acute lymphoblastic leukemia (ALL)
- 42 patients achieved remission from induction therapy and were then randomized in equal numbers to 6-MP or placebo.
- Survival time studied was from randomization until relapse.

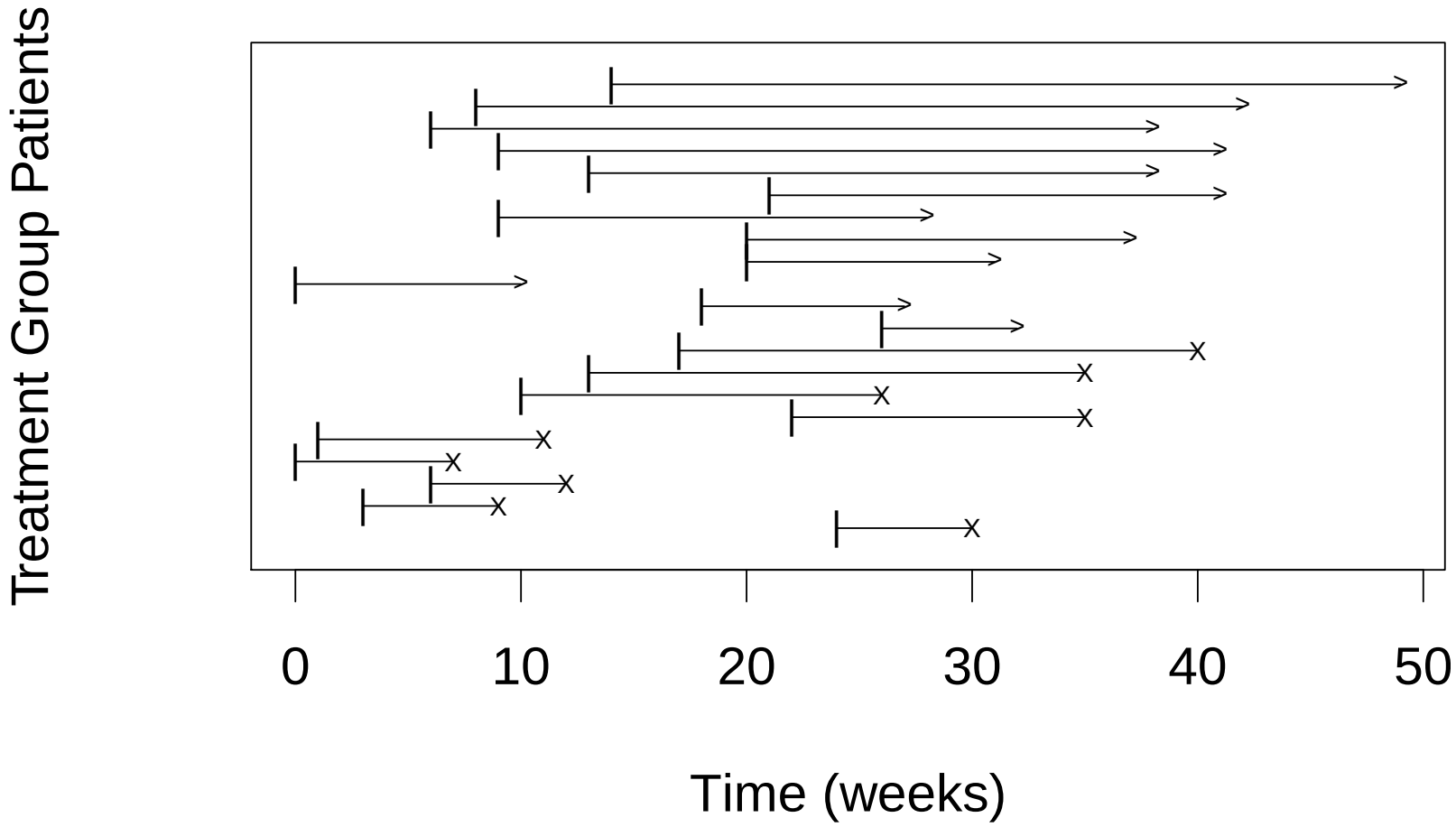
Survival times in weeks for Placebo group:

```
##  [1]  1  1  2  2  3  4  4  5  5  8  8  8  8 11 11
```

Survival times in weeks for Treatment group:

```
##  [1]  6   6   6   7  10  13  16  22  23   6+  9+ 1
## [20] 34+ 35+
```

A graphical look at the treatment group



(Initiation times (t0) are simulated between 0 and 26 weeks)

leukemia study follow-up table

Table 3.13 Follow-up table for placebo patients in the leukemia study

Week of follow-up	No. followed	No. relapsed	No. censored	Conditional prob. of remission	Survival function
1	21	2	0	$19/21 = 0.91$	0.91
2	19	2	0	$17/19 = 0.90$	$0.90 \times 0.91 = 0.81$
3	17	1	0	$16/17 = 0.94$	$0.94 \times 0.81 = 0.76$
4	16	2	0	$14/16 = 0.88$	$0.88 \times 0.76 = 0.67$
5	14	2	0	$12/14 = 0.86$	$0.86 \times 0.67 = 0.57$
6	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
7	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
8	12	4	0	$8/12 = 0.67$	$0.67 \times 0.57 = 0.38$
9	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$
10	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$

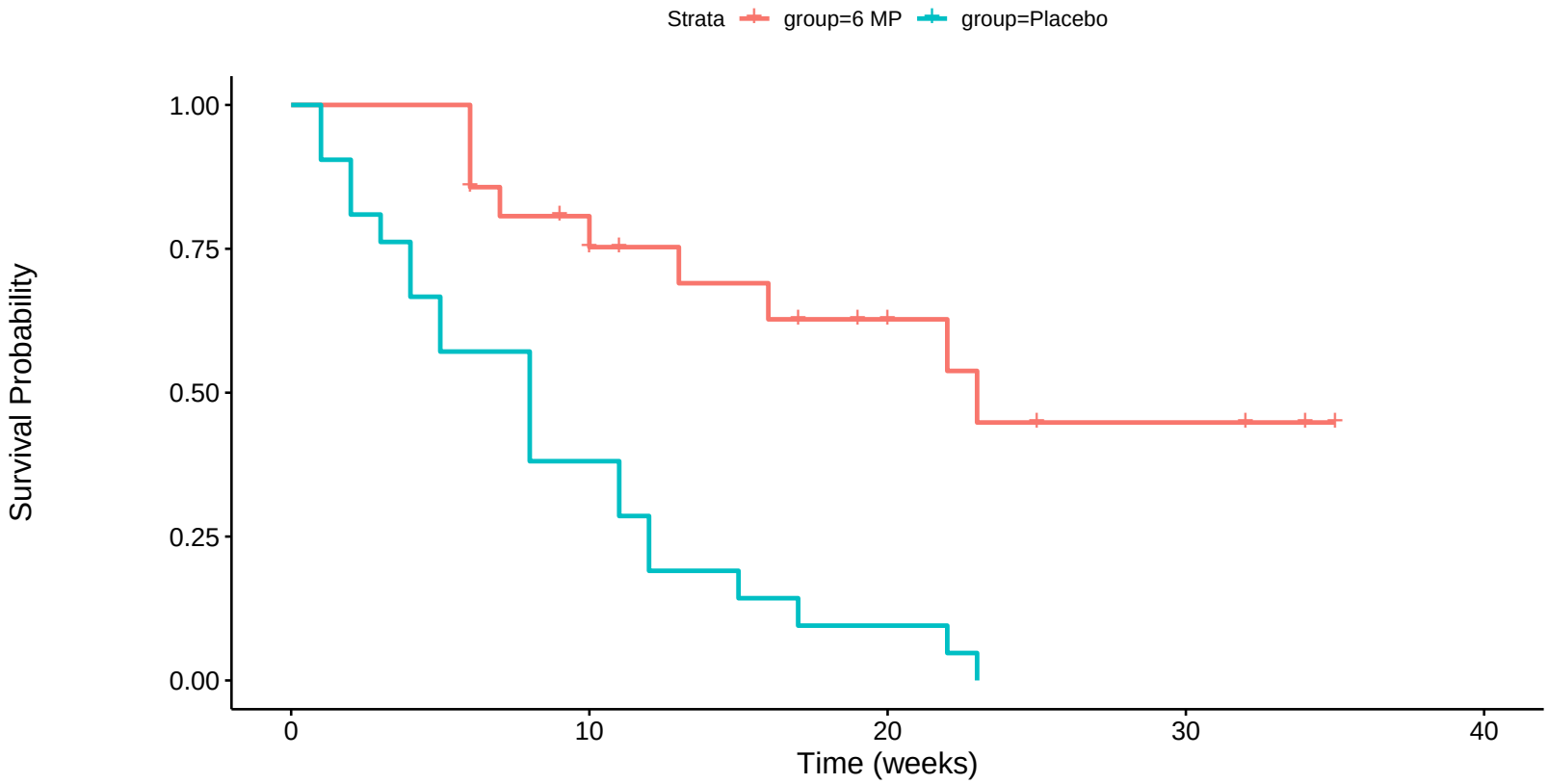
Figure 1: leukemia Follow-up Table

This is the **Kaplan-Meier Estimate** $\hat{S}(t)$ of the Survival function $S(t)$.

Survival function and Kaplan-Meier estimator

Kaplan-Meier Estimate vs. time

```
## Warning: Vectorized input to ‘element_text()‘ is n
## Results may be unexpected or may change in future
```



Number at risk

Strata					
group=6 MP	21	15	8	4	0
group=Placebo	21	8	2	0	0

Median Survival Time

Definition: *Median Survival Time* is the time at which half of a group (sample, population) is expected to experience an event (in this example, death)

- Without censoring, median survival time can be calculated the obvious way
- With censoring, we need to use the Kaplan-Meier estimate of the survival function $\hat{S}(t)$

```
survfit(Surv(time, cens)~group, data=leuk)
```

```
## Call: survfit(formula = Surv(time, cens) ~ group, data = leuk)
##
##                 n events median 0.95LCL 0.95UCL
## group=6 MP     21      9     23      16      NA
## group=Placebo 21     21      8       4      12
```

Median Potential Follow-Up Time

Definition: *Median Potential Follow-Up Time* is the time for which half of a sample would have been expected to be followe, *in the absence of events*.

- Without any events, median follow-up time can be calculated the obvious way
- With events, a simple median will *under-estimate* the potential follow-up time. Use a reverse Kaplan-Meier estimate instead:

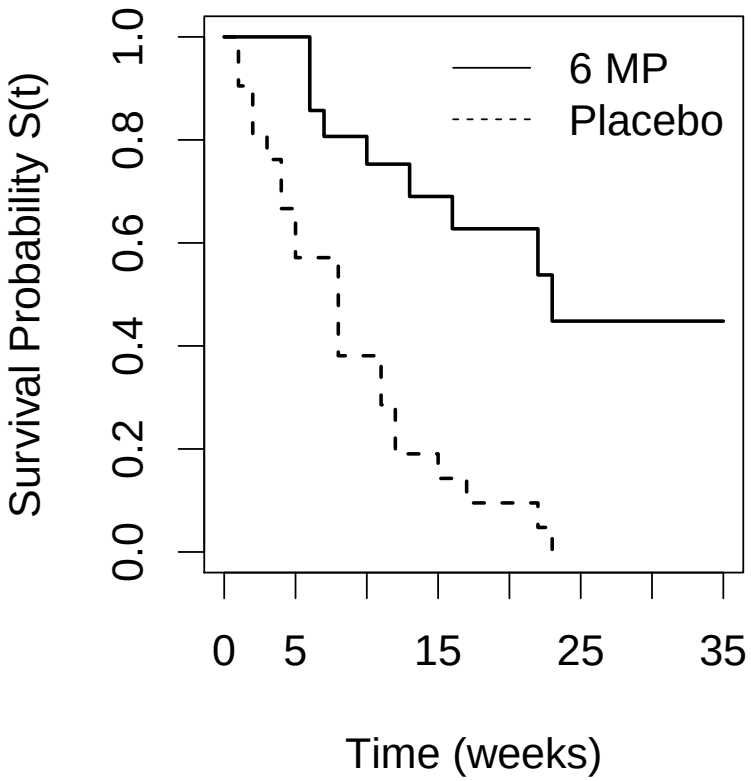
```
survfit(Surv(time, 1-cens)~group, data=leuk)
```

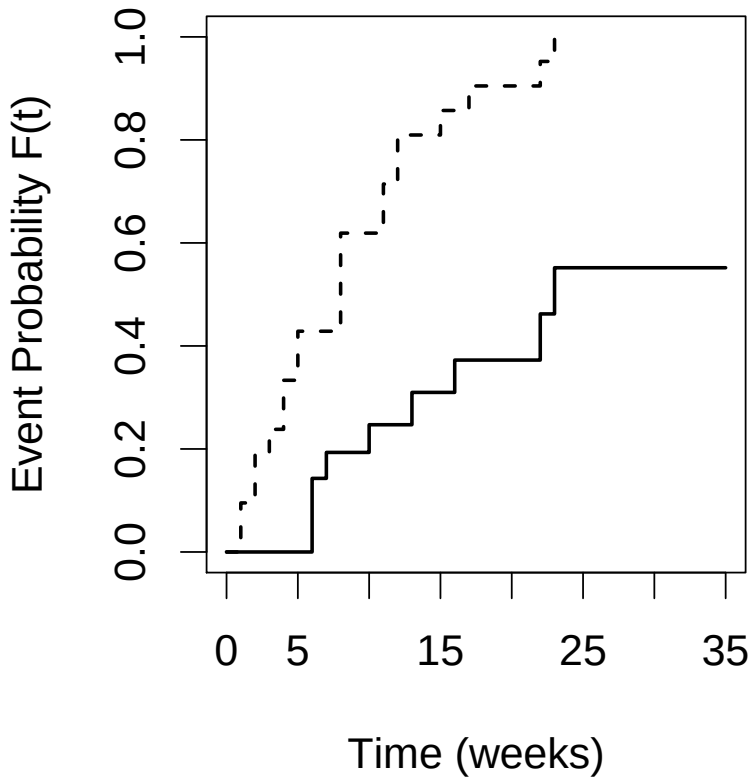
```
## Call: survfit(formula = Surv(time, 1 - cens) ~ group, data = l
##
##                 n events median 0.95LCL 0.95UCL
## group=6 MP     21     12     25      17      NA
## group=Placebo 21      0     NA      NA      NA
```

Note: *Actual* median follow-up time is half as long for the placebo group,

Cumulative Event Function

Definition: The *cumulative event function* at time t, denoted $F(t)$, is the probability that the event has occurred by time t, or equivalently, the probability that the survival time is less than or equal to t. Note $F(t) = 1 - S(t)$.





Hazard and Cumulative Hazard functions

- $h(t)$: hazard function, risk of event at a point in time
  - only calculated by software
- $H(t) = -log[S(t)]$: cumulative hazard function
  - not easily interpretable
  - cumulative force of mortality, or the number of events that would be expected for each individual by time t if the event were a repeatable process.
- Will be important next class for Cox Proportional Hazards

Comparing Groups Using the Logrank Test

Comparing Groups Using the Logrank Test

- *logrank test* is used to compare survival between two or more groups
  - H_0 is that the population survival functions are equal at all follow-up times
  - H_1 is that the population survival functions differ at at least one follow-up time
- logrank test is really just a *chi-square test* comparing expected vs. observed number of events in each group.
  - Observed is just what we see.
  - How to calculate expected?

Comparing groups: Log-rank test

Setting for the Logrank Test

H_0: The risk in two or more groups is the same. Observed differences in numbers of events only occur by chance.

```
survdiff(Surv(time, cens)~group, data=leuk)
```

```
## Call:
## survdiff(formula = Surv(time, cens) ~ group, data = leuk)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## group=6 MP     21        9     19.3      5.46      16.8
## group=Placebo 21       21     10.7      9.77      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

- Many alternatives are available, but log-rank should be the default unless you have good reason.
  - E.g. Wilcoxon (Breslow), Tarone-Ware, Peto tests

Notes about the Logrank Test

- Non-parametric: no assumptions on the form of $S(t)$
- Log-rank test and K-M curves don't work with continuous predictors
- Assumes *non-informative censoring*:
  - censoring is unrelated to the likelihood of developing the event of interest
  - for each subject, his/her censoring time is statistically independent from their failure time

Summary

Summary

- Censoring requires special methods to make full use of the data
- Kaplan-Meier estimate provides non-parametric estimate of the survival function
  - non-parametric meaning that no form of the survival function is assumed; instead it is empirically estimated
- Logrank test provides a non-parametric hypothesis test
  - H0: identical survival functions of multiple strata