# Lecture 14: Nonparametric Survival Analysis Methods

Lin Chen

Department of Public Health Sciences
The University of Chicago

# Non-parametric Estimators in Survival Analysis

ex) doesn't follow normal distribution...

- In this lecture, we will introduce non-parametric (distribution-free) methods to estimate the survivor function and hazard function
- We will assume non-informative right censoring is in effect

  not yet failed before study ends
- To motivate the derivation of these estimators, we will first consider a set of survival times where there is no censoring.

# If there is no censoring

- Consider a single sample of survival times, where each is no censoring completely observed to failure, so none of the observations are censored: $T_1, ..., T_n \ iid \sim F(t)$

- The survivor function $S(t) = P(T \geq t)$ is the probability that an individual survives to a time greater than or equal to cut-off value $t$. This could be estimated by the *empirical survivor function*, the proportion of individuals with survival times greater than or equal to $t$, given by

$$\hat{S}(t) = \frac{\textit{Number of individuals with } t_i \geq t}{\textit{Number of individuals in the (initial) sample}} \quad \text{Total} \tag{1}$$

# Example 1: *Pulmonary Metastasis*

One complication in the management of patients with a malignant bone tumor, or osteosarcoma, is that the tumor often metastasizes to the lungs. The following data give the survival times, in months, of eleven male patients in a study of treatment for pulmonary metastasis arising from osteosarcoma. Total of 11 individuals

$$\underset{\text{Months}}{11} \quad 13 \quad 13 \quad 13 \quad 13 \quad 13 \quad 14 \quad 14 \quad 15 \quad 15 \quad 17$$

This is a case where all the survival times are fully observed (no censoring, which does not often occur in medical studies).

# *Pulmonary metastasis*: estimate $S(t)$

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 13 | 13 | 13 | 13 | 13 | 14 | 14 | 15 | 15 | 17 |

1 2 3 4 5 6

- One approach: estimate $S(t)$ by computing the survival proportions following:

Numerator: # who survived

$$\hat{S}(t) = \frac{\textit{Number of individuals with } t_i \geq t}{\textit{Number of individuals in the data set}}$$

Target those with...

1 - $0 < t \leq 11$: $\hat{S}(t) = \hat{P}(T \geq 11) = \frac{11}{11} = 1$
2 - $11 < t \leq 13$: $\hat{S}(t) = \hat{P}(T \geq 13) = \frac{10}{11} = 0.909$ → exclude #1
3 - $13 < t \leq 14$: $\hat{S}(t) = \hat{P}(T \geq 14) = \frac{5}{11} = 0.455$ → exclude #1-6
4 - $14 < t \leq 15$: $\hat{S}(t) = \hat{P}(T \geq 15) = \frac{3}{11} = 0.273$ ⋮
5 - $15 < t \leq 17$: $\hat{S}(t) = \hat{P}(T \geq 17) = \frac{1}{11} = 0.091$
6 - $17 < t$: $\hat{S}(t) = \hat{P}(T \geq 17^{+}) = \frac{0}{11} = 0$

no censoring

# *Pulmonary metastasis*: estimate $S(t)$

```
. use pulmonary_metastasis.dta
. stset time
```
survival test

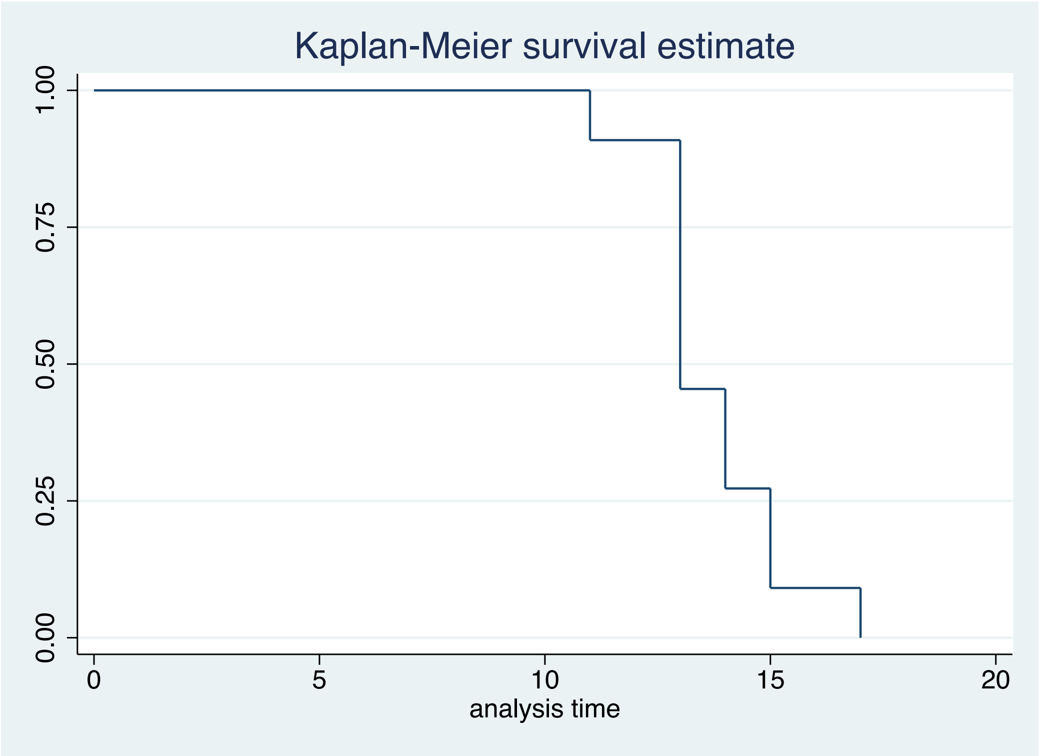
```
     failure event:  (assumed to fail at time=time)
obs. time interval:  (0, time]
 exit on or before:  failure

------------------------------------------------------------------------------
         11  total observations
          0  exclusions
```
no censoring, in this example, everyone died before the study ended.

```
------------------------------------------------------------------------------
         11  observations remaining, representing
         11  failures in single-record/single-failure data
        151  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =        17
```

The *stset* command tells Stata that this is a survival time variable - must have certain properties (non-negative, may have censoring var associated with it)

# *Pulmonary metastasis*: estimate $S(t)$

```
. sts graph

        failure _d: 1 (meaning all fail)
  analysis time _t: time
```




- $\hat{S}(t)$ is 1 from the time origin until the time of first death (11 months). Everyone started as alive
- $\hat{S}(t)$ is 0 after the last observed survival time (17 months). Everyone died before study ended
- $\hat{S}(t)$ is non-increasing in $t$. People only keep dying.

The *sts* command relates to a set of survival summaries that are available - *graph* is one of these.

# *Pulmonary metastasis*: estimate $S(t)$

```
. sts list

        failure _d:  1 (meaning all fail)
  analysis time _t:  time

                At                          Survivor      Std.
  Time        Risk    Fail    Lost         Function     Error       [95% Conf. Int.]
------------------------------------------------------------------------------------
    11          11       1       0           0.9091    0.0867      0.5081    0.9867
    13          10       5       0           0.4545    0.1501      0.1666    0.7069
    14           5       2       0           0.2727    0.1343      0.0652    0.5389
    15           3       2       0           0.0909    0.0867      0.0054    0.3329
    17           1       1       0           0.0000         .           .         .
------------------------------------------------------------------------------------
```

Passed Away

ends until everyone died

- Survival to the first failure time is 100% (S(t) =1.0). Stata does not show this (other pgms do by convention).
- First value change of the estimated survivor function occurs at time 11 months, $\hat{S}(11^+) = 0.9091$:
- Second value change of the estimated survivor function occurs at time 13 months, $\hat{S}(13^+) = 0.4545$:
- ...

# *Pulmonary Metastasis*: estimate $S(t)$

11 13 13 13 13 13 14 14 15 15 17

- The other approach: we can estimate $S(t)$ using conditional probabilities:
  - 1st $0 < t \leq 11 : \hat{S}(t) = \hat{P}(T \geq 11) = \frac{11}{11} = 1$
  - 2nd $11 < t \leq 13$:

$$\hat{S}(t) = \hat{P}(T \geq t) = \hat{P}(\overset{\text{2nd}}{T \geq 13}) = \hat{P}(T \geq 13, T \geq 11)$$

$$= \hat{P}(\overset{\text{2nd phase}}{T \geq 13} \mid \overset{\text{1st phase}}{T \geq 11}) \cdot \hat{P}(\overset{\text{1st phase}}{T \geq 11}) = \frac{10}{11} \times \frac{11}{11} = 0.909$$

=1

conditioning on previous phase

1st entering

Need to survived previous in order to survive until now!!

  - 3rd $13 < t \leq 14$:

$$\hat{S}(t) = \hat{P}(T \geq 14) = \hat{P}(T \geq 14, T \geq 13, T \geq 11)$$

$$= \hat{P}(\overset{\text{3rd}}{T \geq 14} \mid \overset{\text{2nd}}{T \geq 13}) \cdot \hat{P}(\overset{\text{2nd}}{T \geq 13} \mid \overset{\text{1st}}{T \geq 11}) \cdot \hat{P}(\overset{\text{1st}}{T \geq 11})$$

$$= \frac{5}{10} \times \frac{10}{11} \times \frac{11}{11} = 0.455$$

numerator of 2

numerator of 1

  - …
- This conditional probability idea allows for extension to the case where we have right censoring. Provide flexibility to censoring

# What if there is censoring?

- The method of estimating the survivor function using the *empirical survivor function* in Equation (1) on Slide 3 cannot be used when there are censored observations.
- The reason for this is that the method does not allow information provided by an individual whose survival time is censored before time $t$ to be used in computing the estimated survivor function at $t$.
- The best known non-parametric method that accounts for censoring is the Kaplan-Meier estimator.
  - Introduced in 1958 (JASA) - Paul Meier, Department of Statistics, University of Chicago 1950's-1990's.
  - Kaplan-Meier estimator is widely used today (their original paper has been cited over 65,000 times (Google Scholar) since its publication).

# Kaplan-Meier estimator

The New York Times

HEALTH

## Paul Meier, Statistician Who Revolutionized Medical Trials, Dies at 87

By DENNIS HEVESI AUG. 12, 2011

Paul Meier, a leading medical statistician who had a major influence on how the federal government assesses and makes decisions about new treatments that can affect the lives of millions, died on Sunday at his home in Manhattan. He was 87.

The cause was complications of a stroke, his daughter Diane Meier said.

As early as the mid-1950s, Dr. Meier was one of the first and most vocal proponents of what is called "randomization."

Under the protocol, researchers randomly assign one group of patients to receive an experimental treatment and another to receive the standard treatment. In that way, the researchers try to avoid unintentionally skewing the results by choosing, for example, the healthier or younger patients to receive the new treatment.

If the number of subjects is large enough, the two groups will be the same in every respect except the treatment they receive. Such randomized controlled trials are considered the most rigorous way to conduct a study and the best way to gather

# Kaplan-Meier estimate of $S(t)$: notation

- Consider $n$ individuals with observed survival times $t_1, t_2, ..., t_n$. Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time.
- Suppose that there are $r$ (distinct) survival times (event occurred, not censored) among those $n$ individuals where $r \leq n$. Let's arrange these $r$ survival times in ascending order, the $j^{th}$ is denoted $t_{(j)}$, for $j = 1, 2, ..., r$, and so the $r$ ordered survival times are $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$.
- For $t_{(j)}$, let
  - $n_j$ denote the total number at risk at time $t_{(j)}$ (the number of individuals who are known to be alive just before time $t_{(j)}$, including those who are about to fail at $t_{(j)}$)
  - $d_j$ denote the total number of deaths occurring at time $t_{(j)}$

# Kaplan-Meier estimate of $S(t)$: notation

- To compute the estimator, at each of the ordered survival times $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ :
  - Record $n_j$, by counting all those whose failure (event or censored) time is equal or greater than $t_{(j)}$
  - Record the number of failures $d_j$ at each $t_{(j)}$
- From this information alone, Kaplan-Meier estimate of $S(t)$ can be computed

World Health Organization (WHO) data from clinical trials involving a number of different types of contraceptive (WHO, 1987): The data in Table 1 are the number of weeks from the commencement of use of a particular type of intrauterine device (IUD), known as the Multiload 250, until discontinuation because of menstrual bleeding problems. Discontinuation times that are censored are labeled with an asterisk.

Table 1: Time in weeks to discontinuation of the use of an IUD

ex): event at week 10

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | 13* | 18* | 19 | 23* | 30 | 36 | 38* | 54* |
| 56* | 59 | 75 | 93 | 97 | 104* | 107 | 107* | 107* |

ex): * last observed at 13 weeks, but no event observed due to loss of follow-up

# *Time to discontinuation of the use of an IUD* (continued)

For analytic purposes, survival data (time-to-event data) are recorded using two variables:

① failure time ② Censored failure time

- A variable for the *observed* survival time (also called failure time, last status time or event time). For observations that are not censored, it is the failure time; for observations that are right censored, it is the censored failure time (the actual failure time is unknown and exceeds the censored time).
- A second variable is the event indicator, 1 if event is observed, 0 of not observed (censored).
- Each individual has data pair (`time,status`) as the response variable

# Time to discontinuation of the use of an IUD (continued)

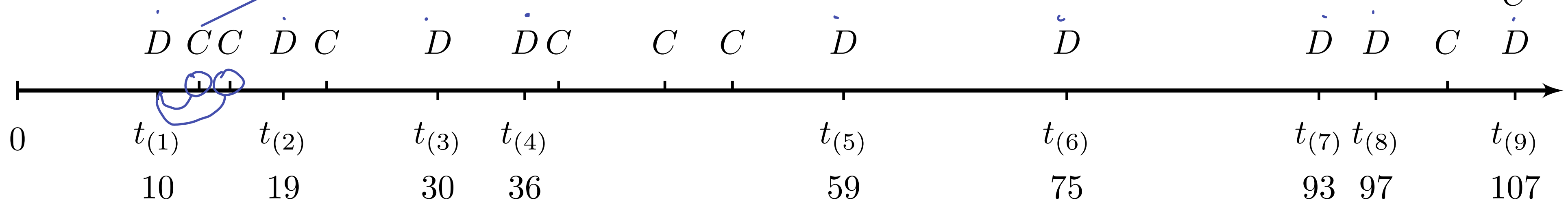
```
. use discontinuation_IUD.dta
. list

     +---------------+
     | time   status |
     |---------------|
  1. |   10        1 |
  2. |   13        0 |
  3. |   18        0 |
  4. |   19        1 |
  5. |   23        0 |
     |---------------|
  6. |   30        1 |
  7. |   36        1 |
  8. |   38        0 |
  9. |   54        0 |
 10. |   56        0 |
     |---------------|
 11. |   59        1 |
 12. |   75        1 |
 13. |   93        1 |
 14. |   97        1 |
 15. |  104        0 |
     |---------------|
 16. |  107        1 |
 17. |  107        0 |
 18. |  107        0 |
     +---------------+
```

censored, did not observe happening of event, therefore 0.

D: event (9)

*censoring denominator accounts for previous t
ex) $C_1$ for $t_{(1)}$

contribute to previous study, but not in the future.




- By convention, when censored survival times occur at the same time as one or more failures, the censored survival time is taken to occur immediately after the failure time.

| $t_{(j)}$ | $n_j$ | $d_j$ (Status) |
|---|---|---|
| 10 | 18 | 1 |
| 19 | 15 | 1 |
| 30 | 13 | 1 |
| 36 | 12 | 1 |
| 59 | 8 | 1 |
| 75 | 7 | 1 |
| 93 | 6 | 1 |
| 97 | 5 | 1 |
| 107 | 3 | 1 |

→ censored
−1 event

# Kaplan-Meier estimate of $S(t)$

Let's apply the conditional probability idea.

For $0 < t \le t_{(1)}, \hat{S}(t) = 1$.

For $t_{(k-1)} < t \le t_{(k)}$,

$$\hat{S}(t) = \hat{P}(\underset{\text{current}}{T \ge t_{(k)}} \mid \underset{\text{previous}}{T \ge t_{(k-1)}}) \overbrace{\hat{P}(\underset{\text{previous}}{T \ge t_{(k-1)}})}^{\text{conditioning on:}}$$

and so on all the way back to $t_{(1)}$

The Kaplan-Meier estimate of the survivor function is given by

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - \frac{d_j}{n_j}) = \prod_{j=1}^{k} (\frac{n_j - d_j}{n_j}) \tag{2}$$

for $t_{(k)} < t \le t_{(k+1)}, k = 1, 2, ..., r$, with $\hat{S}(t) = 1$ for $t \le t_{(1)}$, and where $t_{(r+1)} = \infty$.

"product" of all the conditional probability

Kaplan-Meier method is also called a product limit method. It re-estimated the survival probability each time an event occurs.

# Kaplan-Meier estimate of $S(t)$

- If the largest observed survival time, $t_{(r)}$, is an uncensored observation, $n_r = d_r$, then $\hat{S}(t)$ is 0 for $t > t_{(r)}$.
- Strictly speaking, if the largest observation is a censored survival time, $t^*$, say, $\hat{S}(t)$ is undefined for $t > t^*$.
- A plot of the Kaplan-Meier estimate of the survivor function is a step-function, in which the estimated survival probabilities are constant between adjacent ordered survival times and decrease at each ordered survival time.

# *Time to discontinuation of the use of an IUD* (continued)

Accounted for censoring

ex) "Week 10 Period"

ex) "Survive for 10 weeks"

*Final Exam* (fill in the blank)

Table 2: Kaplan-Meier estimate of the survivor function for the IUD data.

| | Risk | Event | Prob. of Survival | Survival Function |
|---|---|---|---|---|
| Time interval | $n_j$ | $d_j$ | $(n_j - d_j)/n_j$ | $\hat{S}(t)$ |
| 0- | 18 | 0 | 1.0000 $= 1 - \text{failure}$ | 1.0000 |
| 10- | 18 | 1 | 0.9444 $= 1 - \frac{1}{18}$ | 0.9444 $= 0.9444 \times \underline{1.0000}$ |
| 19- | 15 | 1 | 0.9333 $= 1 - \frac{1}{15}$ | 0.8815 $= 0.9333 \times \underline{0.9444}$ |
| 30- | 13 | 1 | 0.9231 $= 1 - \frac{1}{13}$ | 0.8137 $= 0.9231 \times \underline{0.8815}$ |
| 36- | 12 | 1 | 0.9167 | 0.7459 |
| 59- | 8 | 1 | 0.8750 | 0.6526 |
| 75- | 7 | 1 | 0.8571 | 0.5594 |
| 93- | 6 | 1 | 0.8333 | 0.4662 |
| 97- | 5 | 1 | 0.8000 | 0.3729 |
| 107 | 3 | 1 | 0.6667 | 0.2486 |
| 108 | | undefined | | undefined |

2nd conditioned on 1st

18 at risk, but 1 died

15 at risk (−1 died from $t_{10}$, and −2 censoring)

- Note that since the largest discontinuation time of 107 days is censored, $\hat{S}(t)$ is not defined beyond $t = 107$.

Stop after the largest censored failure time

# Time to discontinuation of the use of an IUD (continued)

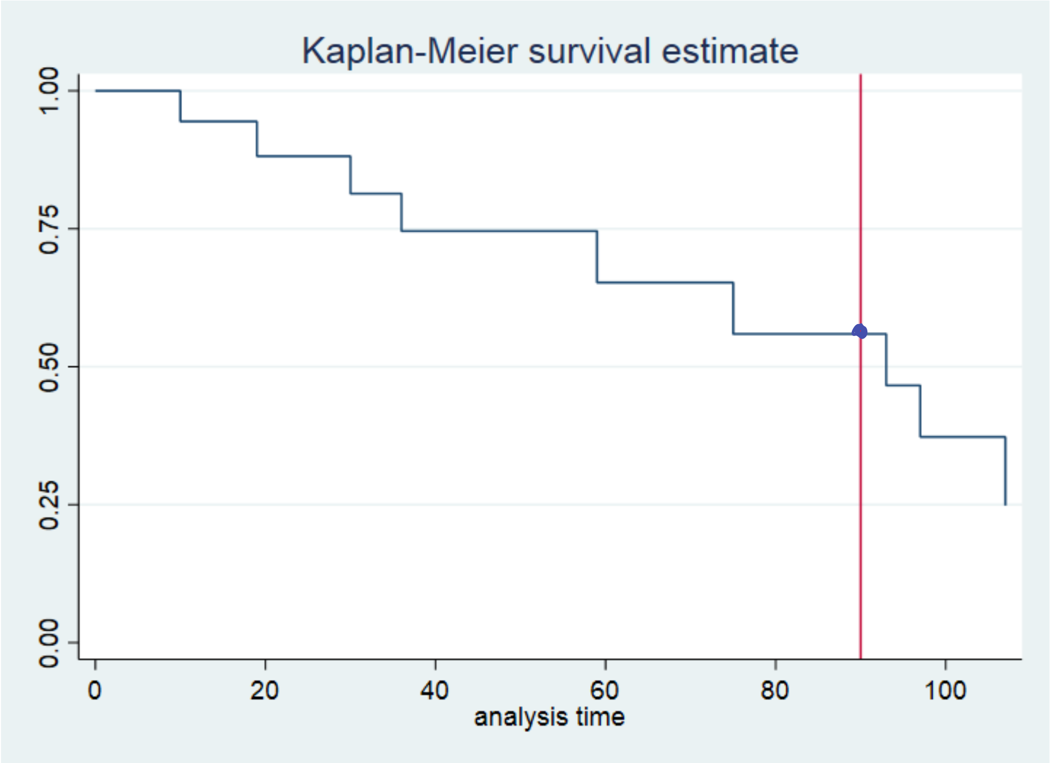
Add this to account for censoring (*)

```
. stset time, failure(status)

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

------------------------------------------------------------------------------
         18  total observations
          0  exclusions
------------------------------------------------------------------------------
         18  observations remaining, representing
          9  failures in single-record/single-failure data   d=1
       1046  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =       107
```

# Time to discontinuation of the use of an IUD (continued)

specify values (t) of interest

```
. sts graph, xline(90)

        failure _d:  status
  analysis time _t:  time
```




What is the probability of survival beyond 90 weeks?

By checking the table on the next page, $\hat{S}(90) = P(T > 90) = P(T \geq 75) = 0.5594$.

# Time to discontinuation of the use of an IUD (continued)

. sts list

failure _d: status
analysis time _t: time

STATA didn't include 1.000

| Time | At Risk | Fail | Lost Censoring | Survivor Function | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|
| | | | | 1.000 | | | |
| 10 | 18 | 1 | 0 | 0.9444 | 0.0540 | 0.6664 | 0.9920 |
| 13 | 17 | 0 | 1 | 0.9444 | 0.0540 | 0.6664 | 0.9920 |
| 18 | 16 | 0 | 1 | 0.9444 | 0.0540 | 0.6664 | 0.9920 |
| 19 | 15 | 1 | 0 | 0.8815 | 0.0790 | 0.6019 | 0.9691 |
| 23 | 14 | 0 | 1 | 0.8815 | 0.0790 | 0.6019 | 0.9691 |
| 30 | 13 | 1 | 0 | 0.8137 | 0.0978 | 0.5241 | 0.9363 |
| 36 F(0.25) closest below | 12 | 1 | 0 | 0.7459 | 0.1107 | 0.4536 | 0.8970 |
| 38 | 11 | 0 | 1 not affected | 0.7459 | 0.1107 | 0.4536 | 0.8970 |
| 54 | 10 | 0 | 1 by censoring | 0.7459 | 0.1107 | 0.4536 | 0.8970 |
| 56 | 9 | 0 | 1 hence prob. is the same | 0.7459 | 0.1107 | 0.4536 | 0.8970 |
| 59 | 8 | 1 | 0 | 0.6526 | 0.1303 | 0.3438 | 0.8432 |
| 75 | 7 | 1 | 0 | 0.5594 | 0.1412 | 0.2564 | 0.7804 |
| 93 F(0.5) closest below | 6 | 1 | 0 | 0.4662 | 0.1452 | 0.1830 | 0.7097 |
| 97 | 5 | 1 | 0 | 0.3729 | 0.1430 | 0.1209 | 0.6310 |
| 104 | 4 | 0 | 1 | 0.3729 | 0.1430 | 0.1209 | 0.6310 |
| 107 F(0.75) closest below | 3 | 1 | 2 | 0.2486 | 0.1392 | 0.0468 | 0.5313 |

# Standard error of the Kaplan-Meier estimate

The Kaplan-Meier estimate of the survivor function for any value of $t$ in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be written as

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

Variance estimate by Greenwood's Formula:

$$\mathrm{Var}\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \tag{3}$$

Thus, the standard error is given by

$$\mathrm{se}\{\hat{S}(t)\} \approx \hat{S}(t) \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}} \tag{4}$$

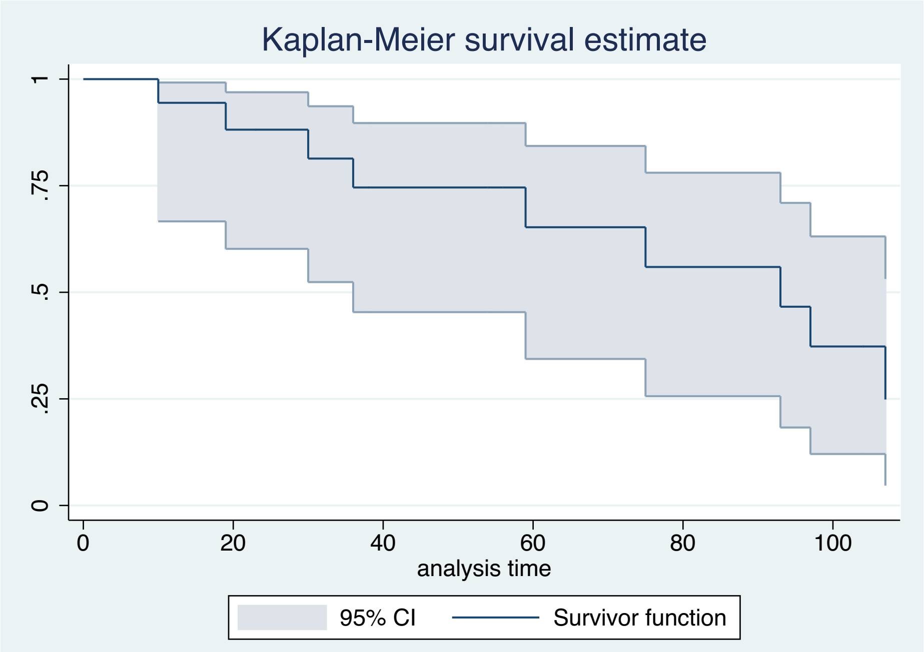
# Confidence interval for survivor function

1. $(\hat{S}(t) - z_{\frac{\alpha}{2}} \cdot \mathrm{se}\{\hat{S}(t)\},\ \hat{S}(t) + z_{\frac{\alpha}{2}} \cdot \mathrm{se}\{\hat{S}(t)\})$

   One difficulty with Greenwood's CI is that when the estimated survivor function is close to $0$ or $1$, this method can lead to confidence limits for the survivor function that lie outside the interval (0,1).

Kaplan−Meier graph        CI

```
. sts graph, gwood
        failure _d:  status
  analysis time _t:  time
```



# Confidence interval for survivor function

- An alternative procedure is to transform $\hat{S}(t)$ to a value in the range $(-\infty, \infty)$, and obtain a CI for the transformed value. The resulting confidence limits are then back-transformed to give a CI for $S(t)$ itself.
- For example, use complementary log-log transform of $S(t)$
- Using Taylor series approximation to the variance of a function of a random variable, the standard error for $\log[-\log(\hat{S}(t))]$ is approximately

$$\text{SE}\{\log[-\log S(t)]\} \approx \frac{1}{-\log \hat{S}(t)} \cdot \sqrt{\sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}}$$

- The confidence interval for $S(t)$ is

$$(\hat{S}(t)^{\exp[-z_{\alpha/2} \cdot \text{SE}\{\log[-\log \hat{S}(t)]\}]}, \ \hat{S}(t)^{\exp[z_{\alpha/2} \cdot \text{SE}\{\log[-\log \hat{S}(t)]\}]})$$

# Estimating the median and other percentiles of survival times

- The median and other percentiles are frequently used as summary measure of the distribution and survival experience
- *Median survival time* $t_{.50}$ is that time such that half of all survival times are larger than $t_{.50}$ and half are smaller, i.e.

$$F(t_{.50}) = S(t_{.50}) = 0.5 \tag{5}$$

- The *pth percentile* of survival times $t_{\frac{p}{100}}$ is that time such that a fraction $p/100$ of survival times are less than $t_{\frac{p}{100}}$ and the other remaining fraction $1 - p/100$ of times are larger than $t_{\frac{p}{100}}$, i.e.

$$F(t_{\frac{p}{100}}) = p/100 \Leftrightarrow S(t_{\frac{p}{100}}) = 1 - p/100 \tag{6}$$

0% percentile : Start
ex) HW: 75% percentile, 25% percentile

# Estimating the median and other percentiles of survival times

- To deal with the discreteness in $\hat{S}(t)$, define

$$\hat{t}_{\frac{p}{100}} = \inf\{t \mid \hat{S}(t) \leq 1 - p/100\} \tag{7}$$

  i.e., the earliest time $t$ where $\hat{S}(t)$ dips below $1 - p/100$.

- In the example of time to discontinuation of the use of an IUD, the smallest discontinuation time where the estimated probability of discontinuation dips below 0.5 is 93, $\hat{t}_{.50} = 93$;
  the smallest discontinuation time where $\hat{S}(t)$ dips below $\underbrace{1 - 0.25}_{0.75}$ is 36, $\hat{t}_{.25} = 36$;
  the smallest discontinuation time where $\hat{S}(t)$ dips below $\underbrace{1 - 0.75}_{0.25}$ is 107, $\hat{t}_{.75} = 107$

Refer to slide 23

# Estimating the median and other percentiles of survival times

The *stsum* command in Stata is useful to summarize survival data. The incidence rate is calculated as the number of events (9) divided by total time at risk = 9/1046.

```
. stsum

         failure _d:  status
   analysis time _t:  time

         |               incidence       no. of    |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
   total |         1046   .0086042            18         36        93       107
```

# Kaplan-Meier estimate of the cumulative hazard function

Instataneous Risk

- Since $H(t) = -\log(S(t))$, we can estimate $H(t)$ by

$$\hat{H}(t) = -\log(\hat{S}(t)) = -\sum_{j=1}^{k} \log(\frac{n_j - d_j}{n_j}) \tag{8}$$

for $t$ in the interval from $t_{(k)}$ to $t_{(k+1)}$.

- If the hazard function is assumed to be constant between successive death times, then the hazard function in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be estimated by

$$\hat{h}(t) = \frac{d_k}{n_k \tau_k} \tag{9}$$

where $\tau_k = t_{(k+1)} - t_{(k)}$. It is calculated as the observed death in the interval divided by the average time survived in the interval.

# Life-table Estimate of the Survivor Function

**Actuarial Method**

- Method used by actuaries, demographers, etc.
- The life-table method was developed before the Kaplan-Meier method. It was once popular and is still used by insurance companies for very large data.
- The life-table method competes with the Kaplan-Meier product-limit method as a technique for survival analysis.
- Motivated by the case where the survival data are grouped into intervals, in which case the estimation of the survivor function is complicated by the fact that we don't know exactly when during each time interval an event occurs.
- Could be applied to ungrouped survival data by first grouping survival data into intervals.

# Life-table estimate of the survivor function: notation

- The $j^{th}$ time interval is $[t_j, t_{j+1})$
- $c_j$: the number of censored survival times in the $j^{th}$ interval
- $d_j$: the number of deaths (events) in the $j^{th}$ interval
- $n_j$: the number of individuals who are alive, and therefore at risk of death, at the start of the $j^{th}$ interval.

# Life-table estimate of the survivor function: notation (continued)

Table 3: Time in weeks to discontinuation of the use of an IUD

1st 2nd 3rd 4th 5th 18 Total

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | 13* | 18* | 19 | 23* | 30 | 36 | 38* | 54* |
| 56* | 59 | 75 | 93 | 97 | 104* | 107 | 107* | 107* |

- Breaking down into time interval assumes censoring occur in uniform.

| Time interval | $j$ | $d_j$ death | $c_j$ censored | $n_j$ alive |
|---|---|---|---|---|
| $[0-10)$ | 1st | 0 | 0 | 18 up to start |
| $[10-20)$ | 2nd | 2 | 2 | 18 up to 1st |
| $[20-30)$ | 3rd | 0 | 1 | 14 up to 2nd |
| $[30-40)$ | 4th | 2 | 1 | 13 up to 3rd |
| $[40-50)$ | 5 ⋮ | 0 | 0 | 10 up to 4th |
| $[50-60)$ | 6 | 1 | 2 | 10 ⋮ |
| $[60-70)$ | 7 | 0 | 0 | 7 |
| $[70-80)$ | 8 | 1 | 0 | 7 |
| $[80-90)$ | 9 | 0 | 0 | 6 |
| $[90-100)$ | 10 | 2 | 0 | 6 |
| $[100-110)$ | 11 | 1 | 3 | 4 |

# Life-table estimate of the survivor function (continued)

- We could apply the Kaplan-Meier formula directly to the numbers in the table on the previous page, estimating $S(t)$ by

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - \frac{d_j}{n_j})$$

  for $t$ in the $k^{th}$ interval from $t_k$ to $t_{k+1}$

- However, this approach is unsatisfactory for grouped data: it treats the problem as if it were in discrete time, with events happening only at 10 weeks, 20 weeks, 30 week, etc. In fact, what we are trying to calculate here is the conditional probability of dying (event) <u>within</u> the interval, given survival to the beginning of it.

# Life-table estimate of the survivor function (continued)

- What should we do with the censored individuals? We should assume that
  - at the beginning of each interval: $n'_j = n_j - c_j$
  - at the end of each interval: $n'_j = n_j$
  - on average, number of subjects at risk within the interval: $n'_j = n_j - c_j/2$
- The last assumption yields the Life-table (Actuarial) estimator. It is appropriate if censorings occur uniformly throughout the interval, which is reasonable to assume in absence of evidence otherwise:

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - \frac{d_j}{n_j - c_j/2}) \tag{10}$$

for the $j^{th}$ interval.

# *Time to discontinuation of the use of an IUD*

Table 4: Life-table estimate of the survivor function for the data of *Time to discontinuation of the use of an IUD*

Conditional Prob. of current time

Conditional Prob. of current time + previous time

| Time interval | $j$ | $d_j$ | $c_j$ | $n_j$ | $1 - \frac{d_j}{n_j - c_j/2}$ | $\hat{S}(t)$ | |
|---|---|---|---|---|---|---|---|
| $[0-10)$ | 1 | 0 | 0 | 18 | 1 | 1 | |
| $[10-20)$ | 2 | 2 | 2 | 18 | 0.8824 | 0.8824 | $= 0.8824 \times 1$ |
| $[20-30)$ | 3 | 0 | 1 | 14 | 1 | 0.8824 | $= 1 \times 0.8824$ |
| $[30-40)$ | 4 | 2 | 1 | 13 | 0.8400 | 0.7412 | $= 0.8400 \times 0.8824$ |
| $[40-50)$ | 5 | 0 | 0 | 10 | 1 | 0.7412 | |
| $[50-60)$ | 6 | 1 | 2 | 10 | 0.8889 | 0.6588 | |
| $[60-70)$ | 7 | 0 | 0 | 7 | 1 | 0.6588 | |
| $[70-80)$ | 8 | 1 | 0 | 7 | 0.8571 | 0.5647 | |
| $[80-90)$ | 9 | 0 | 0 | 6 | 1 | 0.5647 | |
| $[90-100)$ | 10 | 2 | 0 | 6 | 0.6667 | 0.3765 | |
| $[100-110)$ | 11 | 1 | 3 | 4 | 0.6000 | 0.2259 | |

# Time to discontinuation of the use of an IUD

The intervals do not contain an event nor censoring have the same $\hat{S}(t)$ as the previous interval and are omitted.

```
. ltable time status, interval(10)
```

| Interval | | Beg. Total | Deaths | Lost | Survival | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|---|
| 10 | 20 | 18 | 2 | 2 | 0.8824 | 0.0781 | 0.6060 | 0.9692 |
| 20 | 30 | 14 | 0 | 1 | 0.8824 | 0.0781 | 0.6060 | 0.9692 |
| 30 | 40 | 13 | 2 | 1 | 0.7412 | 0.1126 | 0.4451 | 0.8951 |
| 50 | 60 | 10 | 1 | 2 | 0.6588 | 0.1267 | 0.3572 | 0.8444 |
| 70 | 80 | 7 | 1 | 0 | 0.5647 | 0.1392 | 0.2642 | 0.7824 |
| 90 | 100 | 6 | 2 | 0 | 0.3765 | 0.1429 | 0.1234 | 0.6337 |
| 100 | 110 | 4 | 1 | 3 | 0.2259 | 0.1448 | 0.0314 | 0.5276 |

# Life-table estimates of the cumulative hazard function

- Since $H(t) = -\log(S(t))$, we can estimate $H(t)$ by

$$\hat{H}(t) = -\log(\hat{S}(t)) = -\sum_{j=1}^{k} \log(1 - \frac{d_j}{n'_j}) \tag{11}$$

where $n'_j = n_j - c_j/2$, and $t$ is in the interval of $t_k$ to $t_{k+1}$.

- The life-table estimate of the hazard function in the $k^{th}$ time interval (from $t_k$ to $t_{k+1}$) is given by Instataneous incidence risk

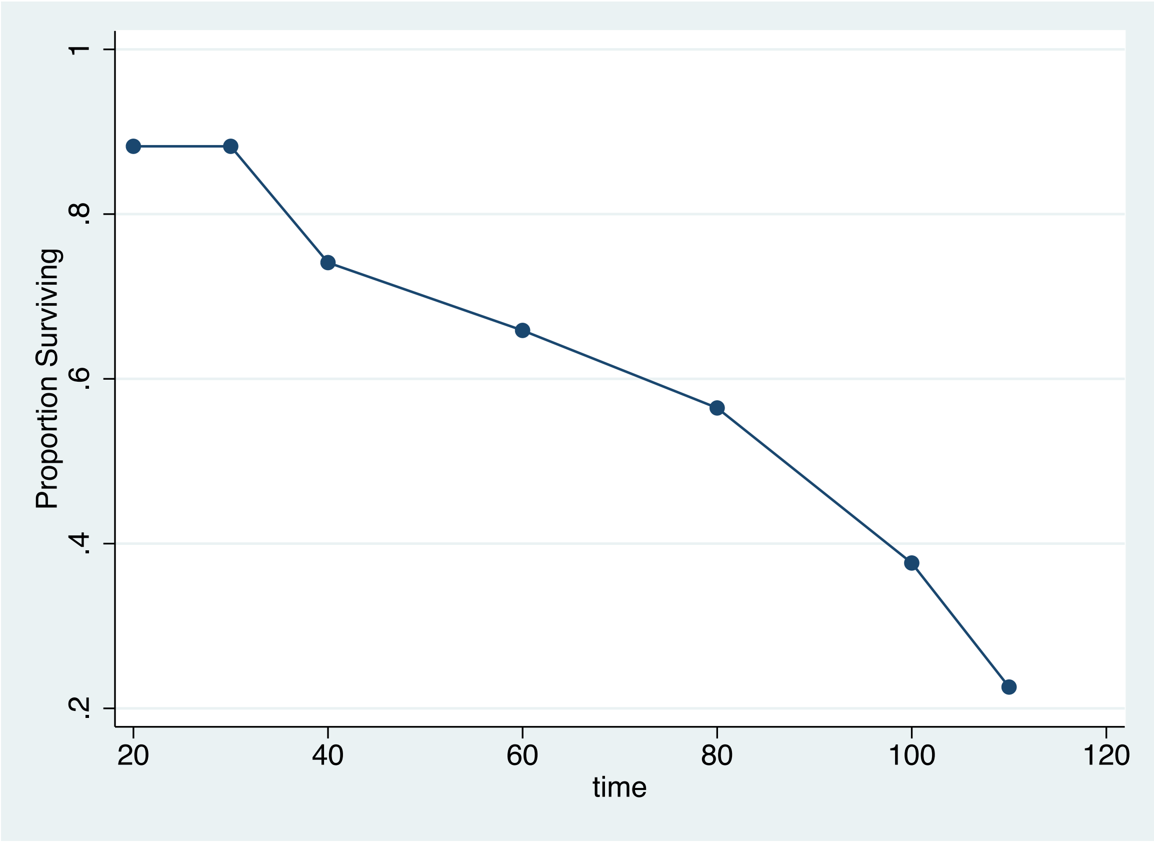
$$\hat{h}(t) = \frac{d_k}{(n'_k - d_k/2)\tau_k} \tag{12}$$

where $\tau_k = t_{k+1} - t_k$.

## Actuarial Estimate

Life Table

```
. ltable time status, interval(10) graph
```



# Time to discontinuation of the use of an IUD Actuarial Estimate

```
. ltable time status, interval(10) hazard
```

| Interval | | Beg. Total | Cum. Failure | Std. Error | Hazard | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|---|
| 10 | 20 | 18 | 0.1176 | 0.0781 | 0.0125 | 0.0088 | 0.0000 | 0.0298 |
| 20 | 30 | 14 | 0.1176 | 0.0781 | 0.0000 | . | . | . |
| 30 | 40 | 13 | 0.2588 | 0.1126 | 0.0174 | 0.0123 | 0.0000 | 0.0414 |
| 50 | 60 | 10 | 0.3412 | 0.1267 | 0.0118 | 0.0117 | 0.0000 | 0.0348 |
| 70 | 80 | 7 | 0.4353 | 0.1392 | 0.0154 | 0.0153 | 0.0000 | 0.0454 |
| 90 | 100 | 6 | 0.6235 | 0.1429 | 0.0400 | 0.0277 | 0.0000 | 0.0943 |
| 100 | 110 | 4 | 0.7741 | 0.1448 | 0.0500 | 0.0484 | 0.0000 | 0.1449 |

*Final Exam*

Kaplan-Meier : • Survival Prob. for given "t"
(focusing) • Median Survival time (50%)
(75%)
(25%)
⋮

## Estimating survivor function S(t)

- S(t) Non-parametric KM estimator is most commonly used and reported Probability of Surviving beyond certain "t".
- H(t) Actuarial estimator is still relevant, used in public health life tables

  https://www.cdc.gov/nchs/products/life_tables.htm

  Instataneous rate of experiencing event at particular "t", given survival up to that time.
- Next: Inference for survival data

• Censoring : Survival Analyzing : previous contribution
hazard Analyzing : middle contribution