Biost 517
Applied Biostatistics I

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 6:
(Right) Censored Data Descriptives

October 12, 2005

Lecture Outline

- Graphical Depiction of the Entire Distn
- Methods for Right Censored Data
  - Setting
  - Motivating example
  - Estimation of survivor functions
    - Life table methods
    - Kaplan-Meier estimates

Graphical Characterizations of an Entire Distribution

Probability Distribution Function

- For ordered variables, we define
  - Cumulative distribution function (cdf):
    - $F(x) = Pr(X \le x)$
  - Survivor function:
    - $S(x) = Pr(X > x) = 1 - F(x)$

Empirical Distribution Function

- Sample cumulative distribution function or survivor function can be used as an estimate
  - (Just treat the sample as if it were the population)
- These functions can sometimes be estimated for censored data (unlike histograms, densities, etc.)

5

Empirical CDF: No Censoring

- Definition:

For uncensored data $\{X_1, X_2, \ldots, X_n\}$

Empirical cumulative distributi on function

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \le x]} = \frac{\#\,\text{observatio ns} \le \text{x}}{\text{n}}$$

Empirical survivor function

$$\hat{S}(x) = 1 - \hat{F}(x)$$

6

Empirical CDF: Properties

- The empirical cdf assigns probability mass of 1/n at each observation
  - Step function:
    - jumps at each observation
    - level between observations
- The empirical cdf can be graphed for an ordered variable
    - Because we draw conclusions from the spacing of the x-axis, this makes most sense when the measurements are on an interval or ratio scale

7

Stata: Empirical CDF

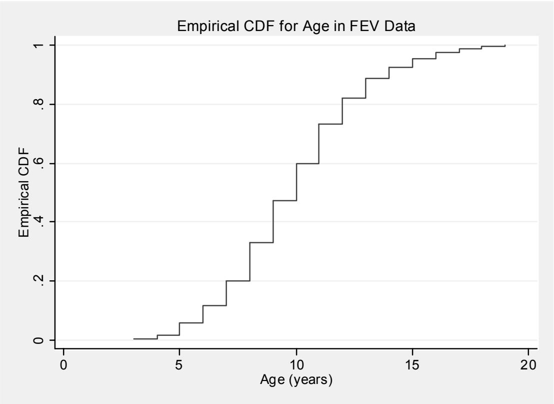
- "`cumul` *`var`*`, gen(`*`Fvar`*`) equal`"
  - Generates a new variable named *Fvar* with empirical CDF
  - (Note the need to use the "equal" option to handle ties)
- "`line` *`Fvar`* `var, sort connect(stairstep)`"
  - Produces empirical CDF (as a step function)
  - (Note the need to use the "sort" option)

8

Stata Ex: Age CDF (FEV data)

```
• cumul age, gen(Fage) equal
• line Fage age,
  connect(stairstp) sort
  xtitle("Age (years)")
  ytitle("Empirical CDF")
  t1("Empirical CDF for Age in
  FEV Data")
```


Stata Ex: Age CDF (FEV data)



Setting for Right Censored Data

Missing Data

- Ideal: "Just say no."
- Real life: "Missing data happens"
  - Ignorable
    - We can safely throw out the cases with missing data without biasing our results
  - Nonignorable
    - Omitting cases with missing data leads to erroneous conclusions

Sad Facts of Life

- "Bloodsuckers hide beneath my bed"
  - *Eyepennies",* Mark Linkous (Sparklehorse)
- Typically, nothing in your data can tell you whether missing data is ignorable or nonignorable
  - You just have to deal with what you worry about

Censored Data

- Special type of nonignorable missing data
  - The value is known to be in some interval, but the exact value is not always known
  - Commonly arises when measuring time to some event
  - Can also arise when measuring laboratory values due to nondetectable levels or saturation of the device

Types of Censored Data

- Right censoring:
  - For some observations it is only known that the true value exceeds some threshold
- Left censoring:
  - For some observations it is only known that the true value is below some threshold
- Interval censoring:
  - For some observations it is only known that the true value is between some thresholds

Example: Setting

- A clinical trial of aspirin in prevention of cardiovascular mortality
  - 10,000 subjects are randomized equally to receive either aspirin or placebo
  - Subjects are randomized over a three year period
  - Subjects are followed for fatal events for an additional three year period following accrual of the last subject

Example: Right Censoring

- Problem:
  - At the end of the clinical trial, some subjects have been observed to die
    - True time to death is known for these subjects
  - At the end of the clinical trial, most subjects are likely to be still alive
    - Death times of these subjects are only known to be longer than the observation time
    - "(Right) Censored observations"

Example: Wrong Approach

- Cannot ignore censored data
  - These are our treatment successes
  - If we throw these cases out of the dataset, we will underestimate the probability of longer survival

Example: Bad Solution #1

- Cannot just treat as binary (live/die) data
  - Potential time of follow-up (censoring time) differs across subjects
    - Administrative censoring (alive at time of analysis)
    - Loss to follow-up due to adverse events
  - Confounding vs loss of precision

Example: Bad Solution #2

- Should not just treat as binary (live/die) data at time of earliest censoring
  - May not answer the scientific question
    - Detecting short term versus long term effects
  - Statistically less efficient

Right Censored Data

- Notation:

Unobserved:

True times to event: $\{T_1^0, T_2^0, \ldots, T_n^0\}$

Censoring Times: $\{C_1, C_2, \ldots, C_n\}$

Observed data:

Observation Times: $T_i = \min(T_i^0, C_i)$

Event indicators: $D_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

21

Motivating Example

22

Motivating Example

- Hypothetical study of subject survival
  - Subjects accrued to study and followed until time of analysis
    - Study done at three centers, which started the studies in three successive years
    - Censoring time thus differs across centers

23

Data by Date (Real Time)

Staggered study entry by site

Year		Accrual Group A	B	C
1990	On study	100	--	--
	Died	43		
	Surviving	57		
1991	On study	57	100	--
	Died	27	53	
	Surviving	30	47	
1992	On study	30	47	100
	Died	13	22	55
	Surviving	17	25	45

24

Data by Study Time

Realign data according to time on study

Year		Accrual Group A	B	C
1	On study	100	100	100
	Died	43	53	55
	Surviving	57	47	45
2	On study	57	47	--
	Died	27	22	
	Surviving	30	25	
3	On study	30	--	--
	Died	13		
	Surviving	17		

Combined Data

Year		Accrual Group A	B	C	Combined
1	On study	100	100	100	300
	Died	43	53	55	151
	Surviving	57	47	45	149
2	On study	57	47	--	104
	Died	27	22		49
	Surviving	30	25		55
3	On study	30	--	--	30
	Died	13			13
	Surviving	17			17

Problem Posed by Missing Data

- Sampling scheme causes (informative) missing data
  - Potentially, we might want to estimate three year survival probabilities
  - Different centers contribute information for varying amounts of time
    - One year survival can be estimated at A, B, C
    - Two year survival can be estimated at A, B
    - Three year survival can be estimated at A

Possible Remedies

- WRONG: Ignore missing
  - E.g., 17 of 300 subjects alive at three years
- RIGHT BUT WRONG QUESTION: Use data only up to earliest censoring time
  - E.g., 149 of 300 subjects alive at one year
- RIGHT BUT INEFFICIENT: Use only center A
  - E.g., 17 of 100 subjects alive at three years

Best Approach

- RIGHT AND EFFICIENT
  - Use all available data to estimate that portion of survival for which it is informative
    - Use Centers A, B, and C to estimate one year survival
    - Use Centers A and B to estimate proportion of one-year survivors who survive to two years
    - Use Center A to estimate proportion of two-year survivors who survive to three years

Theoretical Basis for Approach

- Properties of probabilities
  - Probability of event A and B occurring is product of
    - Probability that A occurs when B has occurred
    - Probability that B has occurred

Application of Theory to Survival

- For times $T_1 < T_2$, probability of surviving beyond time T_2 is the product of
  - Probability of surviving beyond time T_2 given survival beyond time T_1, and
  - Probability of surviving beyond time T_1

$$\text{For } t_0 \le t_1 \le t_2 \le \cdots \le t_k$$

$$\Pr\left(T^0 \ge t_j\right) = \Pr\left(T^0 \ge t_j \cap T^0 \ge t_{j-1}\right)$$

$$= \Pr\left(T^0 \ge t_j \mid T^0 \ge t_{j-1}\right) \Pr\left(T^0 \ge t_{j-1}\right)$$

Estimate Conditional Survival

- Condition on surviving up until the start of the time interval
  - Denominator is number of subjects at start of interval
  - Numerator is deaths during the interval
- Requirement for validity
  - Subjects available at the start of each time interval are a random sample of the population surviving to that time
    - “Noninformative censoring”

Estimate Survival Probability

- Estimate probability of survival at the endpoint of each time interval
  - Multiply the conditional probabilities for all intervals prior to the time point of interest

Application to Example

- Within interval conditional probabilities
  - Use A, B, C to estimate $Pr\ (T^0 \geq 1)$
  - Use A, B to estimate $Pr\ (T^0 \geq 2 \mid T^0 \geq 1)$
  - Use A to estimate $Pr\ (T^0 \geq 3 \mid T^0 \geq 2)$
- Multiply to obtain unconditional cumulative survival
  - $Pr\ (T^0 \geq 1)$
  - $Pr\ (T^0 \geq 2) = Pr\ (T^0 \geq 2 \mid T^0 \geq 1)\ Pr\ (T^0 \geq 1)$
  - $Pr\ (T^0 \geq 3) = Pr\ (T^0 \geq 3 \mid T^0 \geq 2)\ Pr(T^0 \geq 2)$

Motivating Example Results

```
                         Survival Probabilities

Yr  Combined        Each Year                Cumulative

1  On study 300
       Died 151
  Surviving 149  149/300 = 49.67%                     49.67%

2  On study 104
       Died  49
  Surviving  55   55/104 = 52.88%     .4967*.5288 = 26.27%

3  On study  30
       Died  13
  Surviving  17   17/ 30 = 56.67%     .2627*.5667 = 14.88%
```


Estimation of Survivor Functions

Noninformative Censoring

- When estimating survivor functions using censored data:
  - Censoring must not be informative
    - Censored subjects neither more nor less likely to have an event in the immediate future
  - Censored individuals must be a random sample of those at risk at time of censoring
  - (Later: a random sample from all subjects at risk having similar modeled covariates)

Informative Censoring Examples

- Subjects in a clinical trial are withdrawn due to treatment failure (likely they would die sooner than those remaining)
- Subjects in a clinical trial in a fatal condition are lost to follow up when they go on vacation (likely they are healthier than those remaining)

Informative Censoring Examples

- Leukemia patients in a clinical trial of bone marrow transplantation are censored if they die of infections rather than dying of cancer (the subjects who died of infections might have had a more effective regimen to wipe out existing cancer)

Detecting Informative Censoring

- As a general rule it is impossible to use the data to detect informative censoring
  - The necessary data is almost certainly missing in the data set
  - In some cases, it is impossible to ever observe the missing data
    - Nonfelines can only die once
    - We cannot observe whether subjects dying of one cause are more or less likely to die of another if we cure them of the first cause

Life Table Methods

- In the actuarial (e.g., insurance) setting
  - The time intervals are often chosen by years, decades, etc.
  - The data are presented for each year as
    - N_j: Number of subjects at risk at start of interval
    - C_j: Number censored during interval (these will contribute half a person)
    - D_j: Number of events in interval

Life Table Methods: Notation

- Number at risk, censored, failed in each interval

$$\begin{aligned} &\text{Time interval}: && (t_{j-1}, t_j] \\ &\text{Number at risk}: && N_j \\ &\text{Number censored}: && C_j \\ &\text{Number of events}: && D_j \end{aligned}$$

Life Table Methods: Formula

- Computation of probability of survival

Conditional probability of survival in interval:

$$\Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right) = 1 - \frac{D_j}{N_j - 0.5 \times C_j}$$

Cumulative probability of survival:

$$\Pr\left(T^0 \geq t_j\right) = \Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right)\Pr\left(T^0 \geq t_{j-1}\right)$$

Kaplan-Meier Estimates

- Kaplan-Meier (Product Limit) Estimates
  - With more precisely measured individual data
    - The time intervals are defined by unique observation times
    - The data are presented for each year as
      - N_j: Number of subjects at risk at start of interval
      - D_j: Number of events at end of interval
      - (Note no censoring or events during interval by definition)
      - (Note also that for ties, censoring occurs after deaths)

Kaplan-Meier Notation

- Definition of intervals, number at risk, failures

Ordered distinct observation times:

$t_1 \le t_2 \le \cdots \le t_k$

Time interval: $\left(t_{j-1}, t_j\right]$

Number at risk at t_j: N_j

Number of events at t_j: D_j

Kaplan-Meier Hazard Estimates

- Computation of hazard and conditional probability of survival in interval

Hazard for event in interval: $\frac{D_j}{N_j}$

Conditional probability of survival in interval:

$$\Pr\left(T^0 \ge t_j \mid T^0 \ge t_{j-1}\right) = 1 - \frac{D_j}{N_j}$$

Kaplan-Meier Survival Estimate

- Estimating survival probability

S(t) = Pr (T⁰ > t)

Cumulative probability of survival:

$$\Pr\left(T^0 > t_j\right) = \Pr\left(T^0 > t_j \mid T^0 > t_{j-1}\right)\Pr\left(T^0 > t_{j-1}\right)$$

$$\hat{S}\left(t_j\right) = \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \cdots \times \left(1 - \frac{D_1}{N_1}\right)$$

$$= \prod_{i=1}^{j}\left(1 - \frac{D_i}{N_i}\right)$$

If Last Observation Censored

- Note that in the above definition, for an interval which ends in a censored observation with no observed events, the conditional probability of surviving within the interval is 1.
- Note also that if the largest observation time is censored, the KM (PLE) survivor function never goes to zero
  - We generally regard the KM (PLE) survivor function to be undefined for times beyond the largest observation time in this situation

Kaplan-Meier Properties

- The KM (PLE) survivor functions can be shown to be
  - Consistent: As sample sizes go to infinity, they estimate the true value
  - Nonparametric maximum likelihood estimates
    - But usual asymptotic (large sample) theory for regular, parametric MLE's does not apply
    - Asymptotic (large sample) normal distribution for estimates was established differently

Other Derivations of KM

- The KM (PLE) survivor functions can also be derived as the
  - Self-consistent estimator (see Miller, Survival Analysis)
  - "Redistribute to the right" estimator

Redistribute to the Right

- Basic idea
  - Recall the empirical cdf assigns probability 1/n to each observation
  - A censored observation should be equally likely to have event time like any of the remaining uncensored observations
    - Recursively redistribute the mass of each censored observation among the subjects remaining at risk

Ex: Redistribute to the Right

- Data: 1, 3, 4*, 5, 7*, 9, 10
  - (asterisk means censored)
- Initially: each point has mass 1/7
- Determine probability of events at earliest observed (uncensored) event times
  - Pr (T^0 = 1) = 1/7
  - Pr (T^0 = 3) = 1/7

Ex: Redistribute to the Right

- Censored observation at 4
  - Divide the mass at 4 equally among the remaining subjects at risk
    - Now mass of 1/7 + 1/28 = 5/28 for each of 5, 7, 9, 10
- Determine probability of events at next observed (uncensored) event times
  - Pr (T^0 = 5) = 5/28

Ex: Redistribute to the Right

- Censored observation at 7
  - Divide the mass at 7 equally among the remaining subjects at risk
    - Now mass of 5/28 + 5/56 = 15/56 for each of 9, 10
- Determine probability of events at next observed (uncensored) event times
  - Pr (T^0 = 9) = 15/56
  - Pr (T^0 = 10) = 15/56

Ex: Redistribute to the Right

Kaplan-Meier estimate of Survival

t	Pr (T^0 = t)	Pr (T^0 > t)
0		1.000
1	1/ 7 = 0.143	.857
3	1/ 7 = 0.143	.714
4	0.000	.714
5	5/28 = 0.179	.536
7	0.000	.536
9	15/56 = 0.268	.268
10	15/56 = 0.268	.000

Stata: Kaplan-Meier Commands

- First step is declaring data to be of censored survival type
  - Potentially three variables may be used
    - Start of interval
      - Assumed to be at time 0 if nothing supplied
    - End of interval
    - Status at end of interval
      - 0 = censored
      - Nonzero = event occurred at end of interval

Stata: Kaplan-Meier Commands

- Syntax for “setting survival data”
  - `“stset` *`endtime eventind`*`, t0(`*`entrytime`*`)”`
    - *endtime*: name of the variable measuring the time at the end of the interval
    - *eventind* : name of an indicator (0 or 1) variable indicating event status at the end of the interval
    - *entrytime*: name of the variable specifying the time at the start of the interval
      - (does not need to be supplied)
  - `“stset, clear”` resets the data set

Stata: Kaplan-Meier Commands

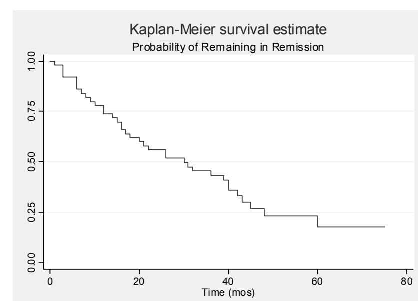
- Syntax for getting estimates, plots
  - Plotting survival curves
    - `“sts graph”`
    - `“sts graph, atrisk”`
    - `“sts graph, cens(s)”`
  - Listing survival estimates
    - `“sts list”`
  - Saving survival estimates
    - `“sts gen` *`newvar`* `= s”`

Example: PSA Data

- PSA data set
  - infile … obstime **str8 inrem** using psa.txt
  - g relapse = 0
  - replace relapse = 1 if inrem==“no”
  - stset obstime relapse
  - sts graph, xtitle(“Time from Treatment (mos)”)
  - sts list
  - sts gen estremt = s

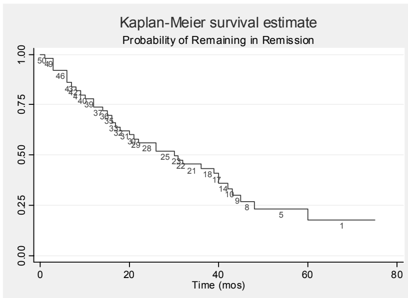
Example: KM Graph

- sts graph, xtitle(“Time (mos)”) t1(“Probability of Remaining in Remission”)



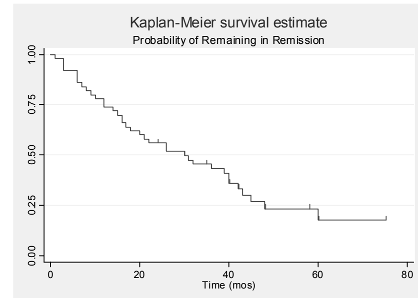
Example: KM Graph

- sts graph, atrisk xtitle("Time (mos)") t1("Probability of Remaining in Remission")



Example: KM Graph

- sts graph, cens(s) xtitle("Time (mos)") t1("Probability of Remaining in Remission")



Example: KM Listing

- sts list

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	50	1	0	0.9800	0.0198	0.8664	0.9972
3	49	3	0	0.9200	0.0384	0.8007	0.9692
6	46	3	0	0.8600	0.0491	0.7286	0.9307
7	43	1	0	0.8400	0.0518	0.7054	0.9166
8	42	1	0	0.8200	0.0543	0.6826	0.9020
9	41	1	0	0.8000	0.0566	0.6602	0.8870
10	40	1	0	0.7800	0.0586	0.6381	0.8716
12	39	2	0	0.7400	0.0620	0.5947	0.8399
14	37	1	0	0.7200	0.0635	0.5735	0.8236
15	36	1	0	0.7000	0.0648	0.5525	0.8070
16	35	2	0	0.6600	0.0670	0.5114	0.7730
17	33	1	0	0.6400	0.0679	0.4911	0.7557

--more--

Example: KM Listing

- sts list, at(24 27 30 33 36)

Time	Beg. Total	Fail	Survivor Function	Std. Error	[95% Conf.	Int.]
24	28	22	0.5600	0.0702	0.4124	0.6842
27	27	2	0.5185	0.0709	0.3725	0.6461
30	25	1	0.4978	0.0710	0.3529	0.6267
33	22	2	0.4545	0.0711	0.3124	0.5860
36	20	1	0.4318	0.0711	0.2913	0.5645