BIOST 514/517
Biostatistics I / Applied Biostatistics I

Kathleen Kerr, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 6:
Descriptive Statistics for
(Right) Censored data

October 11-14, 2013

2003

Lecture Outline

- Probability distribution function, cumulative distribution function, and survivor curves
- Setting for censored data
- Standard notation for censored data
- Motivating example
- Kaplan-Meier “math” explanation
- Kaplan-Meier “redistribute to the right” explanation
- RMST – Restricted Mean Survival Time

Probability Distribution Function

- For ordered variables, we define
  - Cumulative distribution function (cdf):
    - $F(x) \equiv F_X(x) \equiv P(X \leq x)$
  - Survivor function:
    - $S(x) \equiv S_X(x) \equiv P(X > x) = 1 - F_X(x)$

Empirical Distribution Function

- Sample cumulative distribution function or survivor function can be used as an estimate of the population cdf or survivor function
- These functions can sometimes be estimated for censored data (unlike histograms, densities, etc.)

Empirical CDF: No Censoring

- Definition:

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

Empirical cumulative distribution function

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

Empirical survivor function

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

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 quantitative (not just ordered categorical)

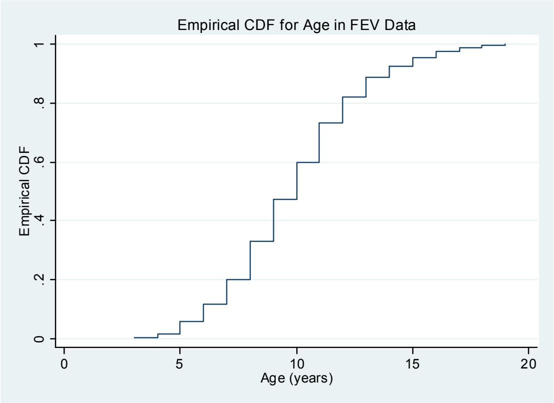
STATA: Empirical CDF

- “`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)

STATA Ex: Age CDF (FEV data)

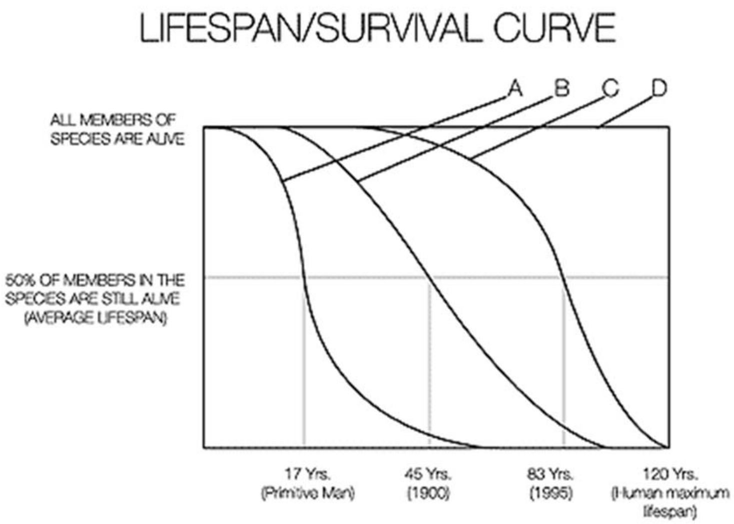
- `cumul age, gen(Fage) equal`
- `line Fage age, connect(stairstp) sort xtitle(“Age (years)”) ytitle(“Empirical CDF”) t1(“Empirical CDF for Age in FEV Data”)`

STATA Ex: Age CDF (FEV data)



Survival Curves

- Curves that estimate the probability of surviving for a Time > t
  - horizontal axis is time
  - vertical axis is P(Survival > t)

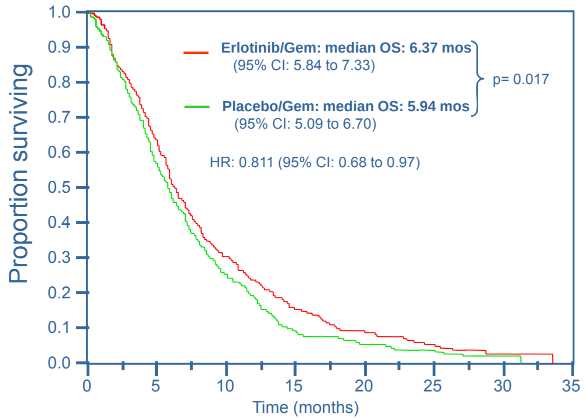


Survival Curves

- In biomedicine, we typically look at the “survivor” or “survival” curves for times to an event, rather than the CDF
- We use Kaplan-Meier methods to get a survival curve
- Note that we can “see” some common sample statistics from a survival curve
- Next slide example: survival in a clinical trial for advanced prostate cancer
  - 569 patients randomly assigned to new treatment or control 1:1
  - Journal of Clinical Oncology 2007

Kaplan Meier Survival Curve

Erlotinib/Gem vs Placebo/Gem (504 deaths)



* Stratified log-rank test

Setting for Right Censored Data

Missing Data Classifications

- Mechanistic classification
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
    - Missingness can depend on other observed data
  - Missing not at random (MNAR)
- Functional classification
  - Ignorable (MCAR and sometimes MAR)
    - Discarding cases with missing data does not bias results
  - Nonignorable (MNAR and most times MAR)
    - Omitting cases with missing data leads to erroneous conclusions

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
  - "Left censoring" can arise with measurement technologies that have a lower limit of detection
  - "Interval censoring" example: screening a cancer patient for recurrent tumors every year. If tumors are seen, the time of recurrence is known to be sometime between the detection visit and the previous visit
  - "right censoring" commonly arises when measuring time to some event

Censored Data

- Special type of nonignorable missing data
  - "right censoring" commonly arises when measuring time to some event
    - Time to death (survival time)
      - Some patients live
    - Time to relapse
      - Some patients don't relapse during the study
    - Among transplant patients, time to first ambulation
      - Some patients observed not to have walked for the study period

Right Censored Data

- If it weren't for censoring, we would almost certainly work with the "time to" variable as a continuous variable
  - Summaries using mean, SD, etc.
  - Visualize with scatterplot, histogram, etc.
- But, in the presence of censoring, we need special methods, and are somewhat limited to the kind of descriptives we can compute

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
  - Observation time differs across subjects
  - Potential for bias:
    - If pattern of censoring is different in the groups you will compare, then this approach can introduce bias
    - E.g. imagine if taking the aspirin made people sick; people in the aspirin group quickly dropped out the study. Can't just analyze them as "alive" – must account for observation time

Example: Bad Solution #2

- To avoid this bias, we could analyze the outcome as binary (live/die) data at time of earliest censoring
  - This would be valid, BUT
  - Probably does not answer the scientific question
    - Detecting short term versus long term effects
  - Statistically less efficient

Right Censored Data

- Notation:

Unobserved:

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

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

Observed data:

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

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

Motivating Example

Motivating Example

- Hypothetical study of subject survival
- Subjects accrued to study and followed until time of analysis
- Study done at three centers; each center started the studies in three successive years
- Censoring time thus differs across centers
  - Only administrative censoring in this example, no other drop-outs

Data by Date (Real Time)

Staggered study entry by site

Year		A	B	C
		Accrual Group		
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

Data by Study Time

Realign data according to time on study

Year		A	B	C
		Accrual Group		
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

$$\Pr(A \text{ and } B) = \Pr(A \mid B) \times \Pr(B)$$
$$\Pr(A \cap B) = \Pr(A \mid B) \times \Pr(B)$$

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

$$\text{For } t_0 \le t_1 \le t_2 \le \cdots \le t_k$$
$$\Pr(T^0 \ge t_j) = \Pr(T^0 \ge t_j \cap T^0 \ge t_{j-1}) = \Pr(T^0 \ge t_j \mid T^0 \ge t_{j-1}) \Pr(T^0 \ge t_{j-1})$$
$$\text{For } 1 \le 2 \le 3 \le \cdots \le 100$$
$$\Pr(T^0 \ge 2) = \Pr(T^0 \ge 2 \cap T^0 \ge 1) = \Pr(T^0 \ge 2 \mid T^0 \ge 1) \Pr(T^0 \ge 1)$$

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:

Informative Censoring Examples

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

Detecting Informative Censoring

- Guiding principle: it is impossible to use the data to detect informative censoring
  - The necessary data are almost certainly missing in the data set
- In some cases, it is impossible to ever observe the missing data: "Competing Risks"
  - Consider the aspirin example. Suppose the outcome is "death from MI."
  - Some people will die from other causes (e.g., cancer) and suppose these are treated as censored.
  - We cannot observe whether subjects dying of other causes are more or less likely to die of another if we cure them of the first cause

Kaplan-Meier Estimates

- Kaplan-Meier (Product Limit) Estimates
- Extends the idea from the motivating example to precisely measured individual data
  - The time intervals are defined by unique observation times
  - N_j: Number of subjects at risk at start of interval
  - D_j: Number of events at end of interval
    - (If a censoring time is exactly the same as a death time, the convention is to treat the censoring as having occurred momentarily after the death)

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: $\dfrac{D_j}{N_j}$

Conditional probability of survival in interval:

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

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

- For an interval that 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 does not go to zero
  - We generally regard the KM (PLE) survivor function to be undefined for times beyond the largest observation (death) 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
    - Censoring must be noninformative

Redistribute to the Right

- The KM (PLE) survivor functions can also be derived as the “redistribute to the right” estimator
- 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
  - (plus sign 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

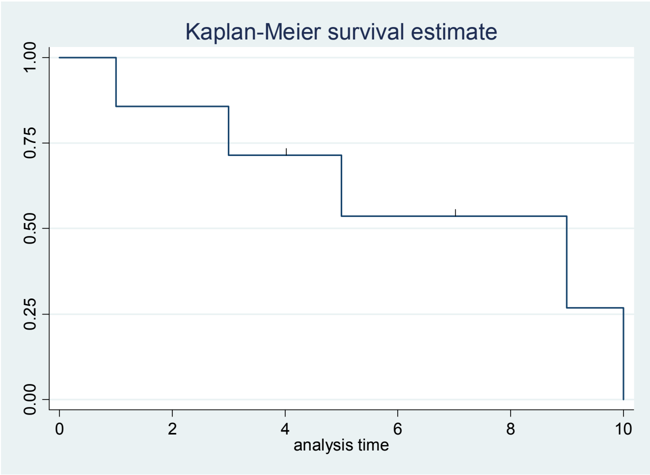
Ex: Redistribute to the Right

- 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

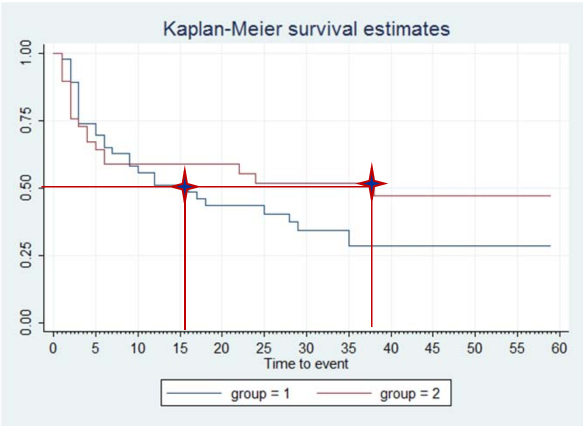
t	$Pr(T^0 = t)$	$Pr(T^0 > t)$
0		1.000
1	1/ 7 = 0.143	.857
3	1/ 7 = 0.143	.714
4	0.000	.714
5	5/28 = 0.179	.536
7	0.000	.536
9	15/56 = 0.268	.268
10	15/56 = 0.268	.000



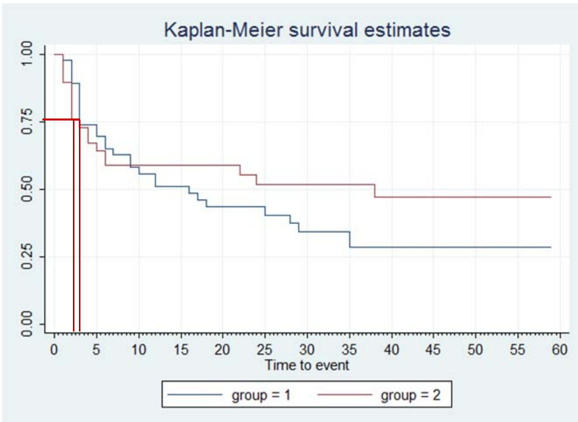
Comparing Survival Curves

- With censored data, we cannot use sample means, sample standard deviations, sample medians, etc.
- Using Kaplan-Meier methods, it is possible to compare population:
  - A. Median (horizontal difference)
  - B. 25th and 75th Percentiles (horizontal difference)
  - C. Prob of exceeding thresholds (vertical difference)
  - D. Restricted Mean (area under curve)
  - E. Hazard ratio (related to slopes)

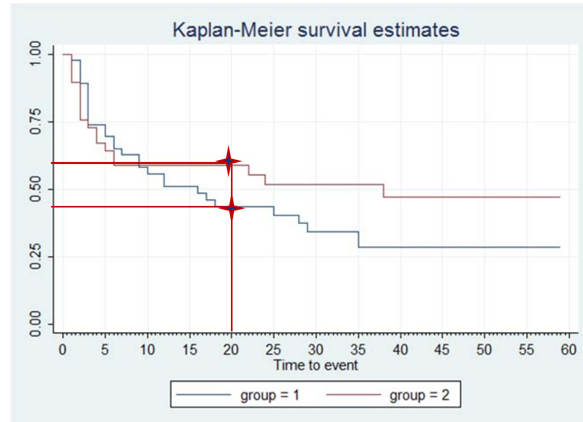
Median: Median survival is about 38 months in Group 2 and 16 months in Group 1



25th percentile: Almost identical in Group 1 and Group 2



Proportion surviving 20 months: Roughly 60% in Group 2 and 45% in Group 1



Restricted Mean

- Can be shown that the area under S(t) is the mean of a positive random variable
- Select a time, t* , up to which we wish to compute the restricted mean survival time
- Formally: restricted mean survival time:

  $$RMST = E[\min(T, t^*)] = \int_0^{t^*} S(t)\,dt$$

- Area under the survival curve up until time t*
- A patient might be told that "your life expectancy over the next 5 years with Z disease on this treatment regimen is 4 years"
- Area under the Kaplan-Meier curve up until time t* gives the RMST for the dataset

STATA, R: 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

R: Kaplan-Meier Commands

- Syntax for creating a “survival object”

  “*svarnm* `<- Surv(`*endtime, eventind*`)`

  “*svarnm* `<- Surv(`*entrytime, endtime, eventind*`)`
    - *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 if 0)
- Any command will specify which survival data you will want to use

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`”
  - Listing restricted mean (up to maximum observation time)
    - “`stci, rmean`”

R: Kaplan-Meier Commands

- Syntax for getting estimates, plots
  - Creating a survfit object
    - `sfitnm <- survfit(svarnm ~ 1)`
  - Plotting survival estimates
    - “`plot(sfitnm,…)`”
  - Listing survival estimates
    - “`summary(sfitnm)`”
  - Listing restricted mean
    - “`print(sfitnm,rmean=#)`”

Example: PSA Data

- PSA data set

  gen relapse = 0

  replace relapse = 1 if inrem==“no”

  stset obstime relapse

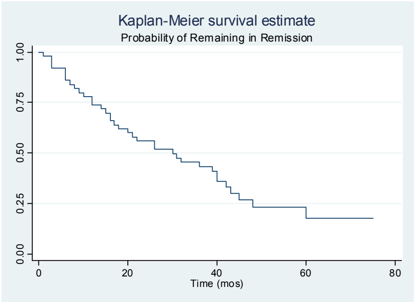
  sts graph, xtitle(“Time from Treatment (months)”)

  sts list

  sts gen estremt = s

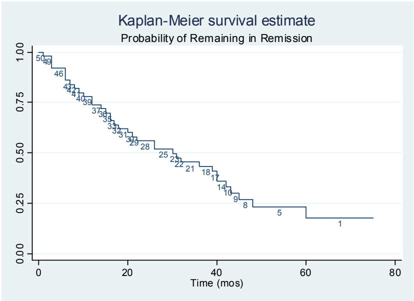
Example: KM Graph

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



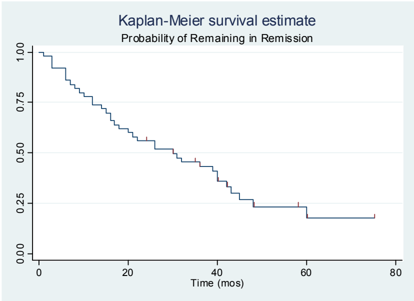
Example: KM Graph

- 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: Restricted Means

- **STATA will give an estimate of the restricted mean**
  - **Appears not to be possible to give RMST for shorter times**

```
. stci, rmean
        failure _d:  relapse
   analysis time _t:  obstime

       | no. of  restrctd
       |  subj    mean      Std. Err.  [95% Conf. Intrvl]
-------+---------------------------------------------
 total |   50    33.92(*)    3.65323   26.76    41.08

(*) largest observed analysis time is censored, mean is
   underestimated
```

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