Topic 3 - Survival Analysis –

1. Topics

- Introduce survival analysis with grouped data
- Estimation of the hazard rate and survivor function
- Kaplan-Meier curves to estimate the survival function, S(t)
- Standard errors and 95% CI for the survival function
- Cox proportional hazards model
- Key words: survival function, hazard, grouped data, Kaplan-Meier, log-rank test, hazard regression, relative hazard

2. Learning objectives

- Describe the survival time density function, survival function, and hazard function
- Describe how to estimate and use the Kaplan-Meier survival curve and confidence intervals
- Describe and use a log-rank test to compare two survival curves
- Describe and use the Cox proportional hazards model to compare survival experience

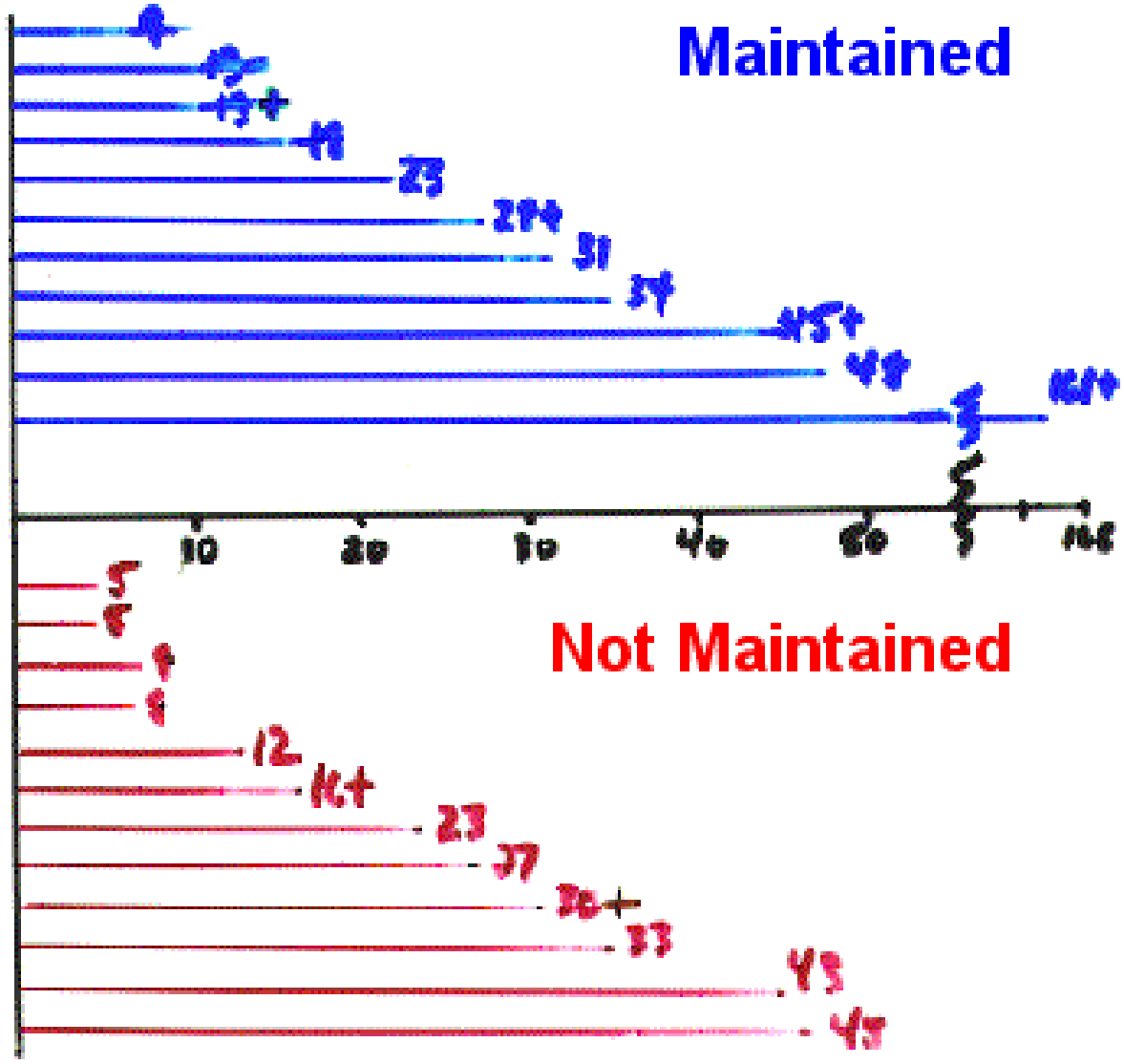
3. Grouped survival data - leukemia example

- 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

3.1 Cohort survival data schematic



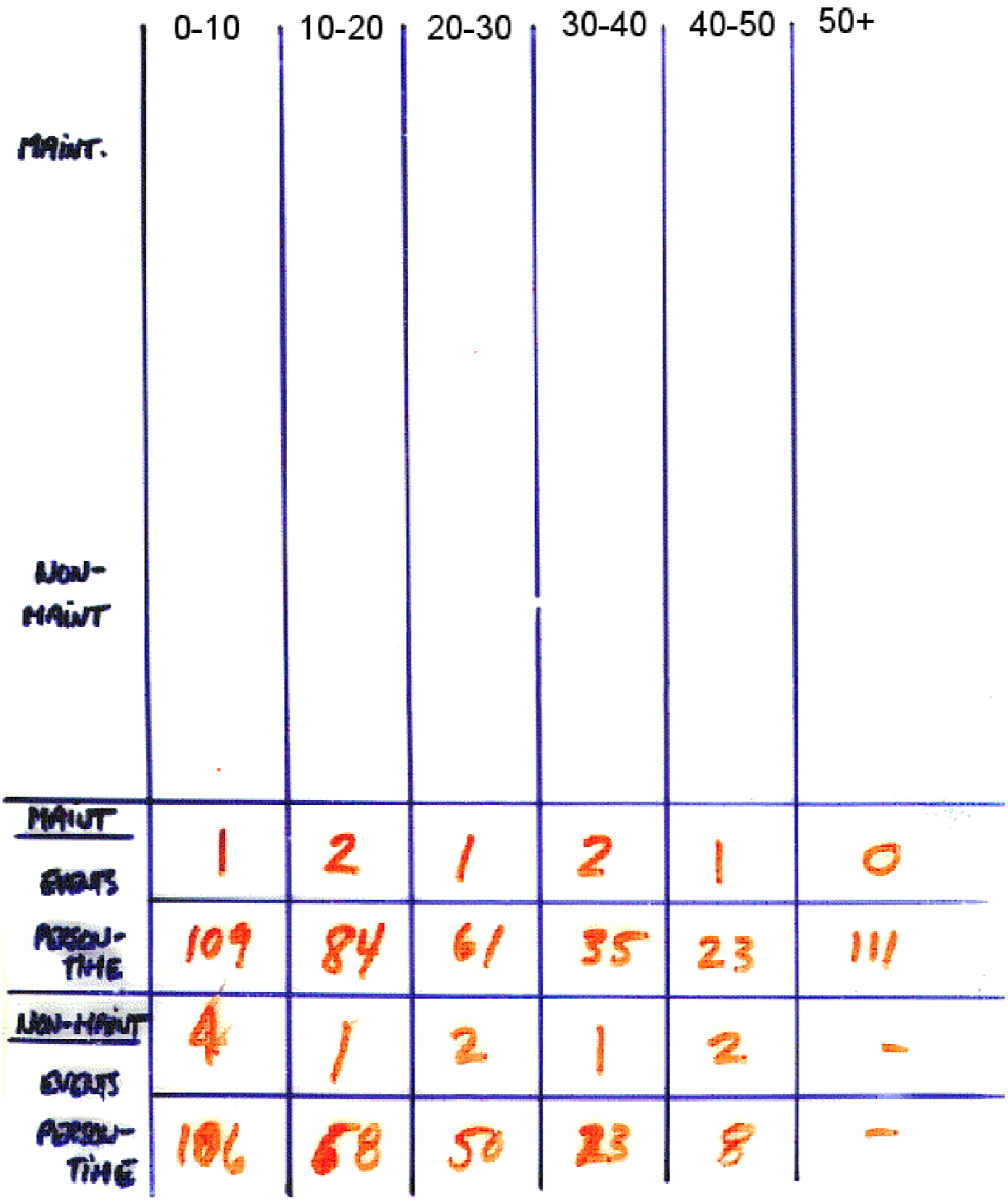
3.2 Tabulation of events and time at risk

- Divide the time period into intervals appropriate for the data
  - use more intervals in periods of changing incidence

- For each person, tally time spent at risk (person-years) in each interval
  - these are the denominators for rates

- Tally the events in each interval
  - these counts are the numerators for the rates and are the values of the response variable

3.2 Tabulation of events and time at risk



3.2 Tabulation of events and time at risk

- Divide follow-up time in the AML example into 15 intervals (defined below in the table) and hand-tally each patient's follow-up time in weeks to produce the following summary table of events and person-time

	Maintained on chemo		Not maintained on chemo	
Interval	Events	Person-time (weeks)	Events	Person-time (weeks)
0-2.5	*0*	*27.5*	*0*	*30*
2.5-5	*0*	*27.5*	*2*	*30*
5-7.5	*0*	*27.5*	*0*	*25*
7.5-10	*1*	*26.5*	*2*	*21*
10-12.5	*0*	*25*	*1*	*19.5*
12.5-15	*1*	*21*	*0*	*17*
15-17.5	*0*	*20*	*0*	*16*
17.5-20	*1*	*18*	*0*	*15*
20-25	*1*	*33*	*1*	*28*
25-30	*0*	*28*	*1*	*22*
30-35	*2*	*20*	*1*	*13*
35-40	*0*	*15*	*0*	*10*

3.2 Tabulation of events and time at risk

40-45	*0*	*15*	*2*	*8*
45-50	*1*	*8*	*-*	*-*
50+	*0*	*111*	*-*	*-*

3.4 Displaying incidence (hazard) rates

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)

- Table of contents for *st* command, **Stata** 7 Reference manual

3.4 Displaying incidence (hazard) rates

Title

st — Survival-time data

Description

The term st refers to survival-time data and the commands—all of which begin with the letters st—for analyzing this data. If you have data on individual subjects with observations recording that this subject came under observation at time t_0, and then, later, at t_1, a failure or censoring was observed, you have what we call survival-time data.

If you have subject-specific data, with observations recording not a span of time, but measurements taken on the subject at that point in time, then you have what we call a snapshot dataset, see [R] **snapspan**.

If you have data on populations, with observations recording the number of units under test at time t (subjects alive) and the number of subjects that failed or were lost due to censoring, you have what we call count-time data; see [R] **ct**.

The st commands are

`stset`	[R] **st stset**	Declare data to be survival-time data
`stdes`	[R] **st stdes**	Describe survival-time data
`stsum`	[R] **st stsum**	Summarize survival-time data
`stvary`	[R] **st stvary**	Report which variables vary over time
`stfill`	[R] **st stfill**	Fill in by carrying forward values of covariates
`stgen`	[R] **st stgen**	Generate variables reflecting entire histories
`sts`	[R] **st sts**	Generate, graph, list, and test the survivor and cumulative hazard functions
`stir`	[R] **st stir**	Report incidence-rate comparison
`strate`	[R] **st strate**	Tabulate failure rate
`stmh`	[R] **st strate**	Calculates rate ratios using Mantel–Haenszel method
`stmc`	[R] **st strate**	Calculates rate ratios using Mantel–Cox method
`stcox`	[R] **st stcox**	Estimate Cox proportional hazards model
`stphtest`	[R] **st stcox**	Test of Cox proportional hazards assumption
`stphplot`	[R] **st stphplot**	Graphical assessment of the Cox proportional hazards assumption
`stcoxkm`	[R] **st stphplot**	Graphical assessment of the Cox proportional hazards assumption
`streg`	[R] **st streg**	Estimate parametric survival models
`stcurv`	[R] **st streg**	Plot fitted survival functions
`stsplit`	[R] **st stsplit**	Split time-span records
`stjoin`	[R] **st stsplit**	Join time-span records
`stbase`	[R] **st stbase**	Form baseline dataset
`sttocc`	[R] **st sttocc**	Convert survival-time data to case–control data
`sttoct`	[R] **st sttoct**	Convert survival-time data to count-time data
`cttost`	[R] **ct cttost**	Convert count-time data to survival-time data
`snapspan`	[R] **snapspan**	Convert snapshot data to time-span data
`st_*`	[R] **st st_is**	Survival analysis subroutines for programmers

The st commands are used for analyzing time-to-absorbing-event (single failure) data and for analyzing time-to-repeated-event (multiple failure) data.

3.4 Displaying incidence (hazard) rates

- Outline for survival data input and analysis:

With data that are already grouped into appropriate time intervals:

1. Enter the data on counts, denominators, and *X*s into **Stata** (bypass the *st* commands)

With ungrouped survival data on individuals:

1. Use the ordinary **Stata** input commands to input and/or generate the following variables:

   X variables

   Denominator variable (if applicable)

   Time variable containing follow-up time

   Censoring variable indicating status at the end of follow-up either "failed" or "censored"

3.4 Displaying incidence (hazard) rates

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

3.4 Displaying incidence (hazard) rates

```
. stdes if x==0

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

. stdes if x==1
ETC

. stsum , by(x)

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

         |                incidence       no. of    |------ Survival time -----|
x        | time at risk      rate        subjects        25%       50%       75%
---------+--------------------------------------------------------------------
       1 |          423   .0165485            11         18        31        48
       0 |          255   .0392157            12          8        23        43
---------+--------------------------------------------------------------------
   total |          678   .0250737            23         12        27        43
```

3.4 Displaying incidence (hazard) rates

- Compare overall incidence by groups:

stir command

```
. stir x

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

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

                | x                       |
                |   Exposed   Unexposed   |     Total
----------------+-------------------------+----------
        Failure |        10            7  |        17
           Time |       255          423  |       678
----------------+-------------------------+----------
                |                         |
  Incidence Rate |  .0392157    .0165485   |  .0250737
                |                         |
                |      Point estimate     |  [95% Conf. Interval]
                |-------------------------+----------------------
 Inc. rate diff. |         .0226672        |  -.004555    .0498895
 Inc. rate ratio |         2.369748        |  .8141934    7.334788  (exact)
 Attr. frac. ex. |         .5780142        | -.2282095    .8636634  (exact)
 Attr. frac. pop |         .3400083        |
                +-------------------------------------------------
                  (midp)   Pr(k>=10) =                    0.0418  (exact)
                  (midp) 2*Pr(k>=10) =                    0.0836  (exact)
```

- Bin the time for grouped survival analysis:

stsplit command

```
* Specify ends of intervals, last interval extends to
               infinity

 stsplit tbin , at( 2.5(2.5)20,  25, 30, 35, 40, 45, 50, 161 )
```

- Tabulate rates by a categorical variable group(x) and bins (groups) of follow-up time:

strate command

```
* Output to new dataset: _D=events  _Y=time at risk
                                    _Rate=rate


 strate tbin x , output(binrates.dta,replace)
```

3.4 Displaying incidence (hazard) rates

- Incidence rates -- also called hazard rates

  simply estimated as the ratio of the number of events to the total time at risk in an interval:

$$\hat{\lambda} = \frac{\#\ of\ events}{person - time}$$

- To display the incidence rates:

  — Plot

  log incidence -vs- time

  stratified by groups of interest

  (plotting incidence -vs- time on a semi- log scale has the same effect and preserves the original units for the rates)

  — Plots are especially useful when the person-time denominators are large in each group; ie, when the estimates $\hat{\lambda}$ are not too noisy

3.6 Displaying survivor function, S(t)

- The "Survivor Function" is defined as

  S(t) = Pr (Survived beyond time t)

- For example, suppose t = end of follow-up time bin 3

S(t) = Pr (Survived > t)

= Pr (survived through bin 1 ***and***
survived through bin 2 ***and***
survived through bin 3)

= Pr(survived bin 1) x
Pr(survived bin 2 given survived bin 1) x
Pr(survived bin 3 given survived bin 1 and bin 2)

- Calculate probabilities of surviving through *bin j* of follow-up time by finding the complement of the probability of dying in *bin j*

3.6 Displaying survivor function, S(t)

Pr (Survived bin j) = 1 - Pr(died in bin j)

- *Pr ("Die" in bin j)* is approximated by

$$P_j = \frac{\#\ Events\ in\ Bin\ j}{\frac{Average\ number\ of\ people\ at\ risk\ in\ Bin\ j}{Length\ of\ Bin\ j}}$$

$$P_j = \frac{y_j}{N_j / L_j} = \frac{y_j\ L_j}{N_j}$$

where

y_j = # of events in bin j

N_j = time at risk (person-time) in bin j

L_j = length of bin j (must be small for the approximation to work well)

3.6 Displaying survivor function, S(t)

- Then, use P_j , the probabilities of dying in bin *j,* to estimate the survivor function, *S(t):*

$$\hat{S}(t) = \prod_{j=1}^{t} [\, 1 - Pr(\text{Die in } j) \,]$$

$$= \prod_{j=1}^{t} (\, 1 - P_j \,)$$

$$\approx \prod_{j=1}^{t} \left(1 - \frac{y_j \cdot L_j}{N_j} \right)$$

- The calculations needed for $\hat{S}(t)$, the estimated survivor function, are usually organized into a "life table, as follows:

S_j = *Pr (Survived beyond the end of bin j)*
$S_0 = 1$

3.6 Displaying survivor function, S(t)

		Maintained on chemo				Not maintained on chemo			
j	L_j	N_j	y_j	$1-P_j = 1- y_i\ L_j/N_j$	S_j	N_j	y_j	$1-P_j = 1-y_i\ L_j/N_j$	S_j
1	2.5	27.5	0	1	1	30	0	1	1
2	2.5	27.5	0	1	1	30	2	.833	.833
3	2.5	2.5	0	1	1	25	0	1	.833
4	2.5	26.5	1	.905	.905	21	2	.762	.635
5	2.5	25	0	1	.905	19.5	1	.872	.553
6	2.5	21	1	.881	.798	17.5	0	1	.553
7	2.5	20	0	1	.798	16	0	1	.553
8	2.5	18	1	.861	.686	15	0	1	.553
9	5	33	1	.848	.582	28	1	.821	.454
10	5	28	0	1	.582	22	1	.773	.351
11	5	20	2	.50	.291	13	1	.615	.216
12	5	15	0	1	.291	10	0	1	.216
13	5	15	0	1	.291	8	2	0*	0
14	5	8	1	.375	.109	-			0
15	111	111	0	1	.109	-			0

3.6 Displaying survivor function, S(t)

- Trouble with follow-up time bins that are too wide:

$$1\text{-}P_j = 1\text{-}y_i\, L_j / N_j = 1\text{-}(10/8) = -0.25$$

Work-around: set the probability, $1 - P_j$, to zero whenever the estimate is negative

---

3.6 Displaying survivor function, S(t)

- To display the estimated survivor,

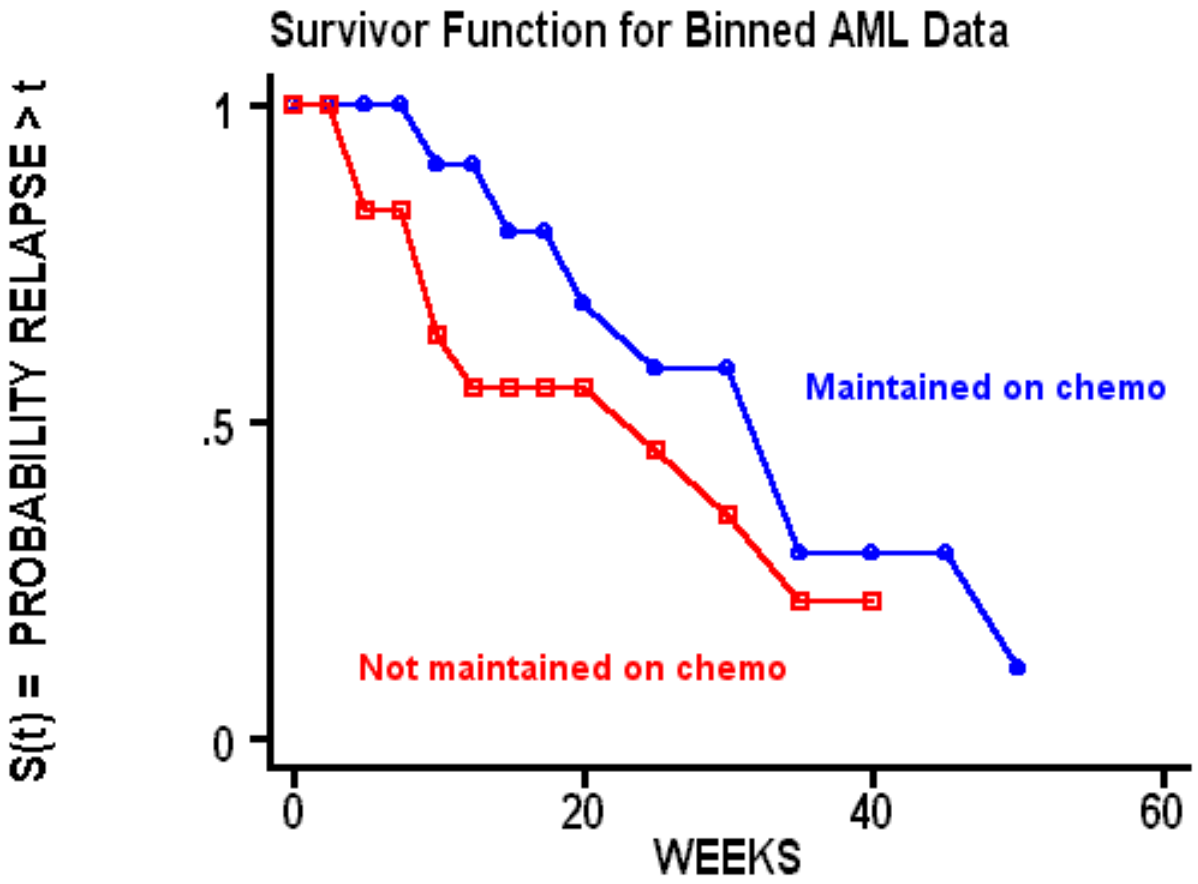
  plot $\hat{S}(t)$ *-vs-* t

  — For grouped data:

  Plot $\hat{S}(t)$ at the end of each time interval connecting the points with line segments (not steps like Kaplan-Meier)

  At *time=0*, plot $\hat{S}(t) \equiv 1$

3.6 Displaying survivor function, S(t)



4. Stata do-file scripts: cl10ex1.do, cl10ex1a.do, cl10ex2.do

---

- The **Stata** script for the AML data example, including commands for inputting survival data and grouping the survival data on the course web site:

  cl10ex1.do

  (The raw data are contained in the script)

- Another related script for the AML data shows how to input the grouped survival data directly into **Stata** as you might had you tabulated the grouped data by hand:

  cl10exa.do

4.1 AML example – cl10ex1.do

```
version 7.0

*  CL10EX1.DO  Grouped survival data

*  AML data:  weeks in remission -vs-  treatment group
*
*


*  Raw data:  AML data included below




* Assumes files are in folder  [path]\bio623

*    If files are in another folder, change cd  command below to
*      point Stata to the correct folder

* To run this program,  use the following Stata commands:

*          cd [path]\bio623    ... change directory to folder bio623

*          do  cl10ex1




* OUTLINE:


* Part a.  Input data, define as a survival dataset
* Part b.  Define survival variables:  stset
* Part c.  Descriptive summaries:  stdes, stsum
* Part d.  Bin the time for grouped survival analysis:  stsplit
* Part e.  Tabulate rates by categorical variable group(x) and bins:  strate
* Part f.  Calculate survivor function, S(t)  from grouped data
* Part g.  Plot Survivor function S(t) for grouped data
* Part h.  Fit different log-linear models for group(x) , get deviances and AIC
* Part i.  Plot estimated hazard functions for models A-E
* Part j.  Fit non-proportional hazard for group effect -- Model F
* Part k.  Use Model D to estimate and plot smoothed S(t)




* Housekeeping

* Clear workspace
```

```
clear


* Turn off -more- pause
set more off


* Save log file on disk, use .log so Notepad will open it

capture log close
log using cl10ex1.log, replace


* Make subfolder for graphs
shell md cl10ex1


* Extend linesize for log

set log linesize 100




* Part a.  Input data, define as a survival dataset


*   id,  x(0=no maint 1=maint), t = time to relapse,  failed=(1=relaspsed 0=censored)

input id x t failed
1   1  9   0
2   1  13  0
3   1  13  1
4   1  18  0
5   1  23  0
6   1  28  1
7   1  31  0
8   1  34  0
9   1  45  1
10  1  48  0
11  1  161 1
12  0  5   0
13  0  5   0
14  0  8   0
15  0  8   0
16  0  12  0
17  0  16  1
18  0  23  0
19  0  27  0
20  0  30  1
21  0  33  0
22  0  43  0
23  0  45  0
end
```

```
* Part b.  Define survival variables:  stset

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

* Save as Stata dataset

save cl10ex1.dta , replace



* Part c.  Descriptive summaries:  stdes, stsum

* Simple counts of persons, events, time at risk

stdes if x==1
stdes if x==0

* Summary stats:  time at risk, rates, subject, 25,50,75 %tiles (K-M estimates)

stsum , by(x)

* Compare overall incidence rates by group:  stir

stir x



* Part d.  Bin the time for grouped survival analysis:  stsplit

*   Note:  Expands dataset, 1 record for each person-time interval combination

* Specify ends of intervals, last interval extends to infinity

stsplit tbin , at( 2.5 (2.5) 20, 25,30,35,40,45,50,161 )

* Part e.  Tabulate rates by categorical variable group(x) and bins:  strate

*             Output to new dataset: _D=events  _Y=time at risk  _Rate=rate

* NOTE:  The strate command  REQUIRES STATA 6 or 7

strate tbin x , output(binrates.dta,replace)
```

```
* Part f.  Calculate survivor function, S(t)  from grouped data


* Access rates, time at risk dataset

use binrates.dta , clear

* UGH:  For some reason, _Y was created as a string!!  Convert to numeric

gen temp=real(_Y)
drop _Y
gen _Y=temp
drop temp


* First, calculate interval lengths, L, for grouped survival analysis

* Make sure in order by group(x) and time bin

sort x tbin


*   L = subtract lower limtits for interval _n+1 -vs- _n ; last(_N) interval is
undefined

by x:  gen L = cond( _n < _N , tbin[_n+1] - tbin , . )



* Calculate midpoints f intevals for log-linear models -- last intervals must be
*                          treated as special cases

gen midT = tbin + L/2


replace midT = 42.5 if  (x==1 & midT==.)

replace midT = 105.5 if (x==0 & midT==.)

* Calculate survival probs P for each interval: rate x length  , (correct if P <0)

gen P = min(1 - _Rate*L,1)



* Calculate S(t) = Prob (Surviving beyond t) = Product P1 P2 ... Pt

gen S = P

by x: replace S = cond( _n>1, P*S[_n-1] , S)


* Show results _Y=time at risk  _D=failures

list midT x _Y _D P S


* Part g.  Plot Survivor function S(t) for grouped data
```

```
*    Plot S(t) for grouped data at end of intervals; connect with lines

* To plot S(t) for each of two groups, need two variables

* Plot S(t) at end of interval = lower limit + length/2 , last interval not used
*   by convention, plot S(0)=1

gen T = tbin

by x: gen MAINT     = cond(_n>1, S[_n-1], 1)  if x==1

by x: gen NOMAINT   = cond(_n>1, S[_n-1], 1)  if x==0

* Check shifted plotting points

*list tbin L T S MAINT NOMAINT

set textsize 140

#delimit ;
graph MAINT NOMAINT T , symbol(OS) connect(ll) xlab ylab
l1(" ")  l2("S(t) =  PROBABILITY RELAPSE > t ")
b1(" ") b2("WEEKS")
t2("Survivor Function for Binned AML Data");
#delimit cr;

gphprint , saving(cl10ex1\figg1.wmf,replace)

drop T P S MAINT NOMAINT

* Close log file -- Only when all errors have been fixed

*log close
```

5.Kaplan-Meier estimate of survivor function, S(t)

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. He recently retired as chair of the Department of Statistics at Columbia University, where he made important contributions to the methods for and practice of clinical trials.

- In the “not maintained on chemotherapy” group:

time (t)	times of events/censoring						# at risk	# events	fraction of events	fraction no events	fraction surviving after t
	5	5	8	8	12	16+					
0	0	0	0	0	0	0	6	0	0	1	1
1	0	0	0	0	0	0	6	0	0	1	1
2	0	0	0	0	0	0	6	0	0	1	1
3	0	0	0	0	0	0	6	0	0	1	1
4	0	0	0	0	0	0	6	0	0	1	1
5	1	1	0	0	0	0	6	2	2/6	4/6	4/6
6			0	0	0	0	4	0	0	1	4/6
7			0	0	0	0	4	0	0	1	4/6
8			1	1	0	0	4	2	2/4	2/4	4/6 x 2/4 = 2/6
9					0	0	2	0	0	1	2/6
10					0	0	2	0	0	1	2/6
11					0	0	2	0	0	1	2/6
12					1	0	2	1	½	½	2/6 x ½ = 1/6
13						0	1	0	1	1	1/6
14						0	1	0	1	1	1/6
15						0	1	0	1	1	1/6
16						0	1	0	1	1	1/6
⋮						⋮					

5.1 Kaplan-Meier estimate of the survivor function, *S(t)*

- For grouped survival data,

$\hat{S}(t)$ = Estimated *Pr (Survive beyond t)*

$$(*) \qquad = \prod_{j:\ \text{bins 1 thru t}} \left(1 - \frac{y_j L_j}{N_j} \right)$$

- Let interval lengths L_j become very small - all of length $L=\Delta t$ and let $t_1, t_2,$ ... be times of events (survival times)

- 2 cases to consider in (*)

Case 1. No event in bin (interval)

$$\frac{y_j L_j}{N_j} = \frac{0 \times L_j}{N_j} = 0$$

$$\Rightarrow \quad 1 - \frac{y_j L_j}{N_j} = 1$$

$\hat{S}(t)$ does not change - which means that we can ignore bins with no events

5.1 Kaplan-Meier estimate of the survivor function, *S(t)*

Case 2. y_j events occur in a bin (interval)

Also: n_j persons enter the bin

assume any censored times that occur in the bin **occur at the end of the bin**

$$1 - \frac{y_j L_j}{N_j} = 1 - \frac{y_j \Delta t}{n_j \Delta t}$$

$$= \frac{n_j - y_j}{n_j}$$

5.1 Kaplan-Meier estimate of the survivor function, *S(t)*

- So, as $\Delta t \to 0$, we get the Kaplan- Meier estimate of the survivor function, *S(t)*:

(IMPORTANT)
$$\hat{S}(t) = \prod_{j:\, t_j \le t} \left(\frac{n_j - y_j}{n_j} \right)$$

$$\hat{S}(0) \equiv 1 \quad \text{(by convention)}$$

Also called the "product-limit estimate" of the survivor function, *S(t)*

5.2 Example: Kaplan-Meier survival curves for the AML data

- Calculation of Kaplan-Meier estimates:

In the "not maintained on chemotherapy" group:

Time	At risk	Events	$\hat{S}(t)$
t_j	n_j	y_j	$\hat{S}(t_j)=\hat{S}(t_{j-1})\times\frac{n_j - y_j}{n_j}$
0	12	0	1.0
5	*12*	*2*	*1.0 x ((12-2)/2)=0.833*
8	*10*	*2*	*0.833 x ((10-2)/10)=0.666*
12	*8*	*1*	*0.666 x ((8-1)/8)=0.583*
23	*6*	*1*	*0.583 x ((6-1)/6)=0.486*
27	*5*	*1*	*0.486 x ((5-1)/5)=0.389*
33	*3*	*1*	*0.389 x ((3-1)/3)=0.259*
43	*2*	*1*	*0.259 x ((2-1)/2)=0.130*
45	*1*	*1*	*0.130 x ((1-1)/1)=0*

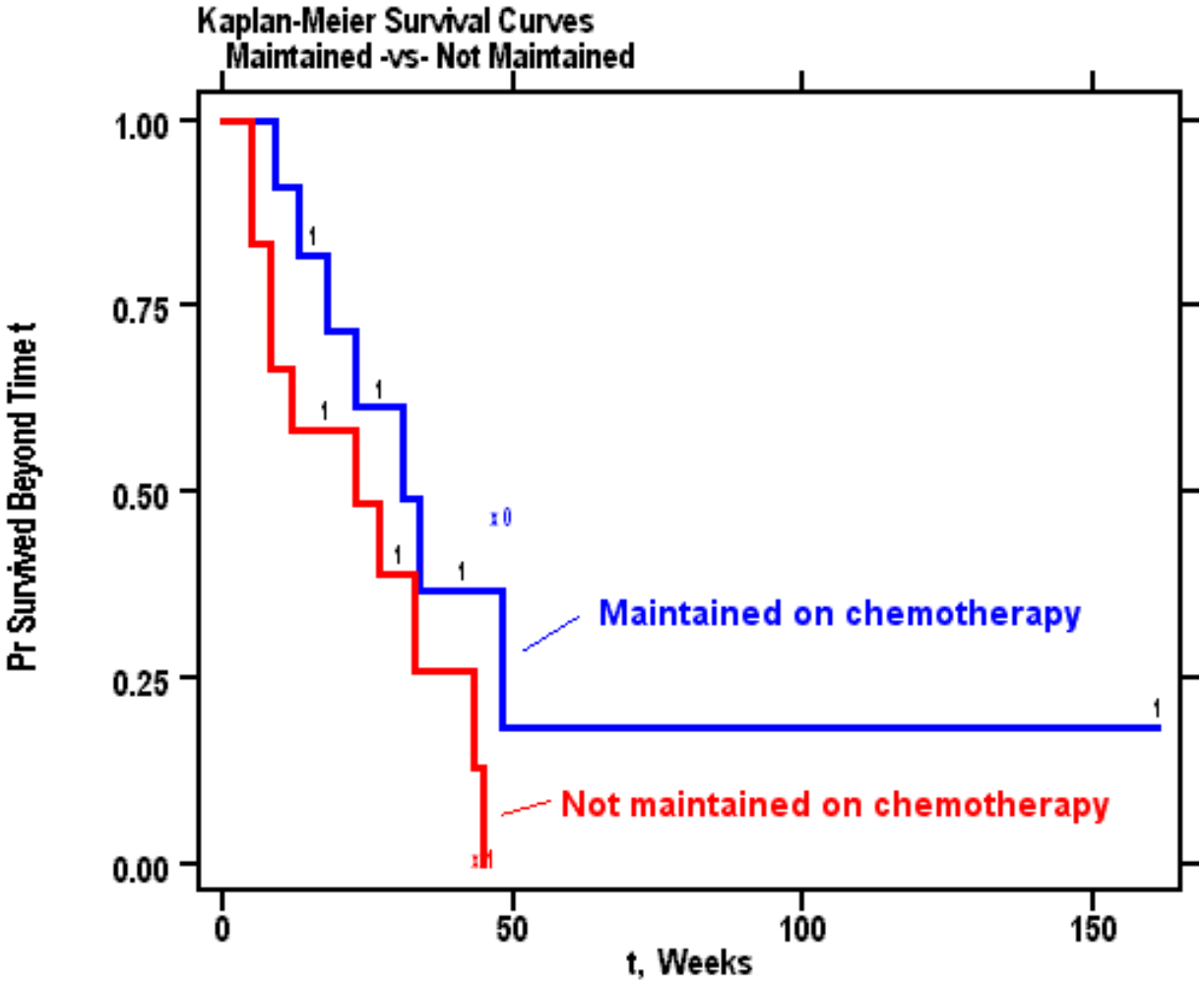
5.2 Example: Kaplan-Meier survival curves for the AML data

In the "maintained on chemotherapy" group:

Time	At risk	Events	$\hat{S}(t)$
t_j	n_j	y_j	$\hat{S}(t_j) = \hat{S}(t_{j-1}) \times \frac{n_j - y_j}{n_j}$
0	11	0	1.0
9	*11*	*1*	$.909 = 1.0 \times \frac{11-1}{11}$
13	*10*	*1*	$.818 = .909 \times \frac{10-1}{10}$
18	*8*	*1*	*.716*
23	*7*	*1*	*.614*
31	*5*	*1*	*.491*
34	*4*	*1*	*.368*
48	*2*	*1*	*.184*

5.2 Example: Kaplan-Meier survival curves for the AML data

- The "Kaplan-Meier curve" plots the estimated survival function $\hat{S}(t)$ -vs- *time* -- separate curves for each group



5.2 Example: Kaplan-Meier survival curves for the AML data

- Notes

  — Can count the total number of events by counting the number of steps (times)

  — If feasible, picture the censoring times on the graph as shown above

- **Stata** code for Kaplan-Meier estimates and plots

  — Input data and define as a survival dataset

5.2 Example: Kaplan-Meier survival curves for the AML data

```
*  Raw data:  id,  x(0=no maint 1=maint),
       t = time to relapse,
       failed=(1=relapsed 0=censored)

input id x t failed
1   1  9   1
2   1  13  1
3   1  13  0
...
end

*  Define survival variables:  stset

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

— Calculate and print Kaplan-Meier estimates for each group

```
sts list  if x==1

sts list  if x==0
```

Stata log:

```
. sts list  if x==1

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

5.2 Example: Kaplan-Meier survival curves for the AML data

```
                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
-------------------------------------------------------------------------------
```

— Plot Kaplan-Meier curves; list counts of censored on plots

** Plot Kaplan-Meier estimates*

sts graph , by(x) lost

(Graph shown above)

5.3 Confidence interval for *S(t)* -- Greenwood's formula

- Greenwood's formula for the variance of $\hat{S}(t)$:

$$\widehat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j: t_j \le t} \frac{y_j}{n_j(n_j - y_j)}$$

$$SE_{GW}(t) = \sqrt{\widehat{Var}[\hat{S}(t)]}$$

- Using Greenwood's formula, an approximate 95% CI for *S(t)* is

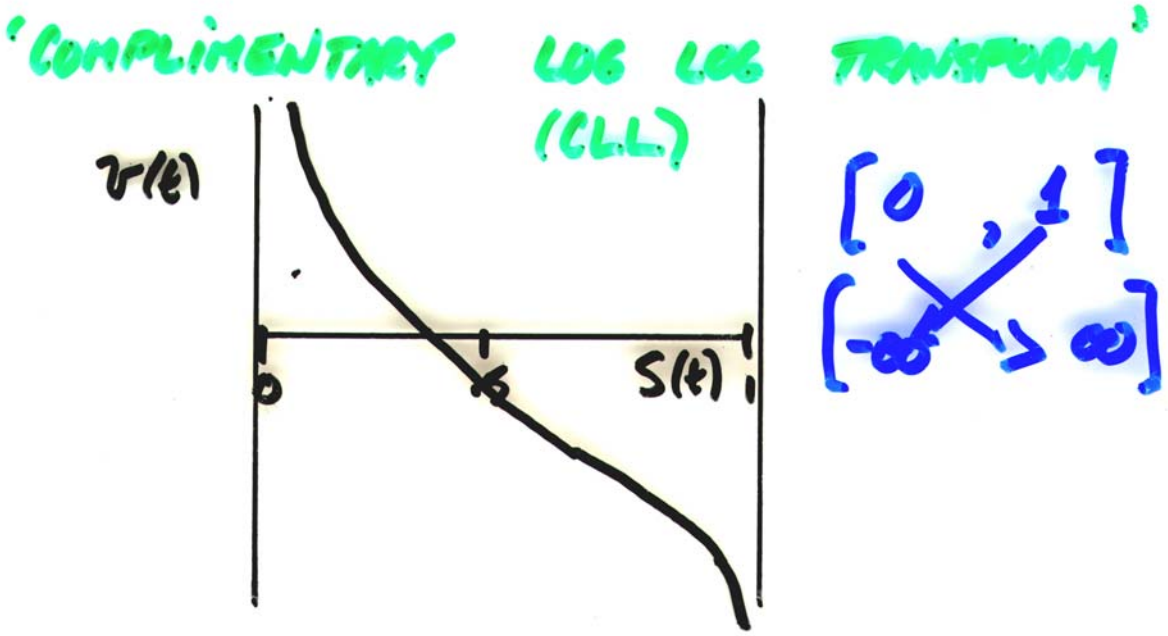
$$\hat{S}(t) \pm 2\, SE_{GW}(t)$$

- There is a "problem": the 95% CI is not constrained to lie within the interval (0,1)

5.4 Better CI for S(t) -- complementary log-log transform

- Consider the "Complementary *log log* transform" (*CLL*):

$$\hat{v}(t) = \log [- \log \hat{S}(t)]$$



5.4 Better CI for S(t) -- complementary log-log transform

- Variance of *CLL*:

$$\hat{v}(t) = \log[-\log \hat{S}(t)]$$

$$\widehat{Var}(\hat{v}(t)) = \frac{\sum_{j: t_j \le t} \frac{y_j}{n_j(n_j - y_j)}}{\left[\sum_{j: t_j \le t} \log\left(\frac{n_j - y_j}{n_j}\right)\right]^2}$$

$$SE_{CLL}(t) = \sqrt{\widehat{Var}(\hat{v}(t))}$$

5.4 Better CI for S(t) -- complementary log-log transform

- Use *CLL* to obtain 95% CI on *S(t)*

1. Get 95% CI for *v(t)*:

$$\hat{v}(t) \pm 2\, SE_{CLL}(t)$$

2. Transform back to get 95% CI for *S(t)*:

Use the inverse transformation

$$S(t) = e^{(-e^{v(t)})}$$

to get the 95% CI for *S(t)*:

$$\left[e^{(-e^{\hat{v}(t)+2SE_{CLL}(t)})}, e^{(-e^{\hat{v}(t)-2SE_{CLL}(t)})} \right]$$

$$= [\hat{S}(t)]^{e^{(\pm 2 SE_{CLL}(t))}}$$

(NOTE: **Stata** uses the *CLL* transformation for 95% CI on *S(t)* -- see log above)

- Example: Back to the AML data

5.4 Better CI for S(t) -- complementary log-log transform

Time	At risk	Events	
t_j	n_j	y_j	$\hat{S}(t)$
9	11	1	.909
13	10	1	.818
18	8	1	.716
23	7	1	.614
31	5	1	.491
34	4	1	.368
48	2	1	.184

- $\hat{Var}_{Greenwood}[\hat{S}(13)] = .818^2 \left(\frac{1}{11\times10} + \frac{1}{10\times9}\right)$

  $= (.116)^2$

- 95% $CI_{Greenwood} = .818 \pm 2\,(.116)$

  $= (.586, \underline{1.05})$

  1.05 is out of range

5.4 Better CI for S(t) -- complementary log-log transform

- Better 95% CI using the *CLL* transformation:

$$\hat{v}(t) = \log(-\log \hat{S}(t)) = -1.605$$

$$\widehat{Var}(\hat{v}(13)) = \frac{\left(\frac{1}{110} + \frac{1}{90}\right)}{\left(\log\frac{10}{11} + \log\frac{9}{10}\right)^2}$$

$$= \frac{.0202}{.04027} = .502$$

$$SE_{CLL}(13) = .708$$

- 95% CI for S(13) $= [.818]^{e^{\pm 2(.708)}}$

$= (.437, .952)$

5.4 Better CI for S(t) -- complementary log-log transform

- 95% CI for *S(t)* in the maintained on chemotherapy group

Time	At risk	Events	$\hat{S}(t)$	$\hat{se}_{GW}$	95% CI*
t_j	n_j	y_j			
9	*11*	*1*	*.91*	*.087*	*.51,.99*
13	*10*	*1*	*.82*	*.116*	*.45,.95*
18	*8*	*1*	*.72*	*.140*	*.35,.90*
23	*7*	*1*	*.61*	*.152*	*.27,.84*
31	*5*	*1*	*.49*	*.164*	*.17,.75*
34	*4*	*1*	*.37*	*.163*	*.09,.65*
48	*2*	*1*	*.18*	*.154*	*.01,.53*

*Based on complementary log-log transform

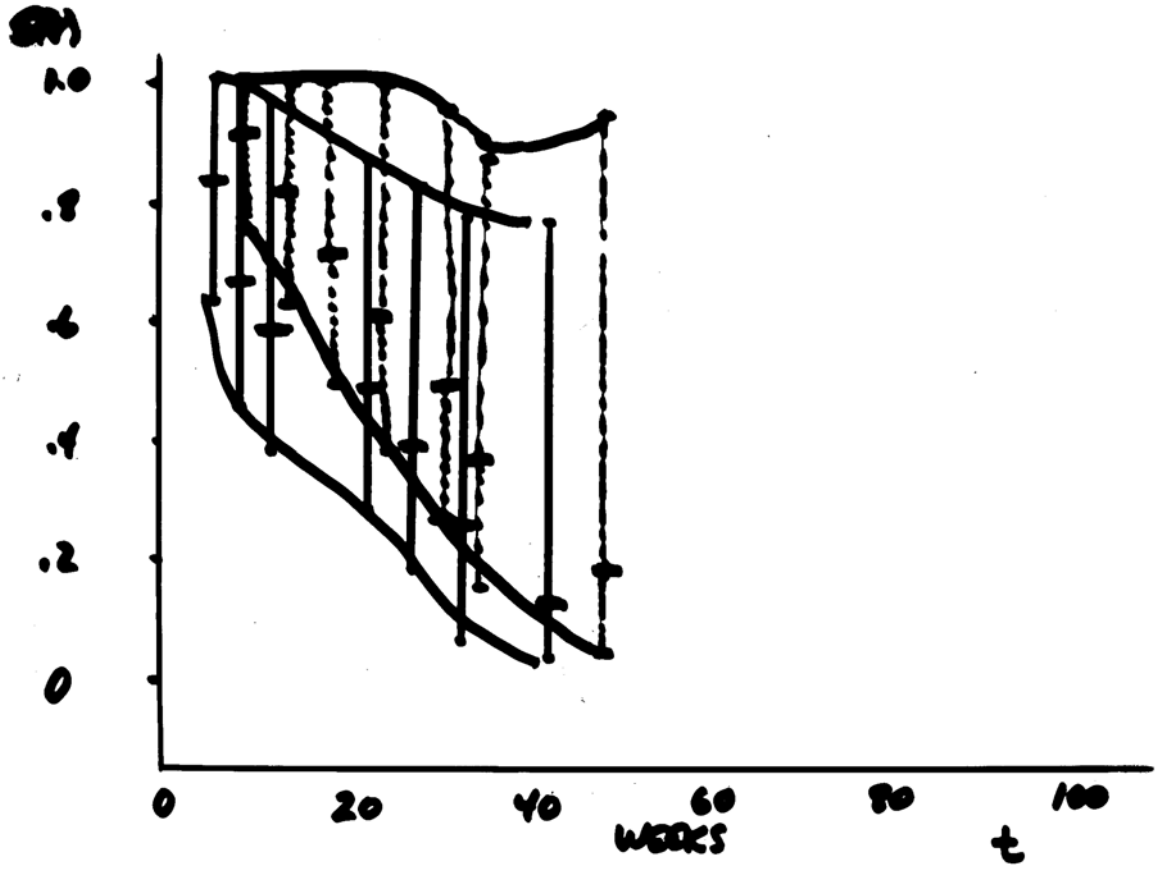
5.4 Better CI for S(t) -- complementary log-log transform

- 95% CI for *S(t)* in the not maintained on chemotherapy group

Time	At risk	Events	$\hat{S}(t)$	$\hat{se}_{GW}$	95% CI*
t_j	n_j	y_j			
5	*12*	*2*	*.83*	*.11*	*.48,.96*
8	*10*	*2*	*.67*	*.14*	*.34,.86*
12	*8*	*1*	*.58*	*.14*	*.27,.80*
23	*6*	*1*	*.49*	*.15*	*.19,.73*
27	*5*	*1*	*.39*	*.15*	*.13,.65*
33	*3*	*1*	*.26*	*.14*	*.05,.55*
43	*2*	*1*	*.13*	*.12*	*.01,.42*
45	*1*	*1*	*.00*	-	-

*Based on complementary log-log transform

5.4 Better CI for S(t) -- complementary log-log transform



6. Log-rank test for comparing survivor curves

- Are two survivor curves the same?

- Use the times of events: $t_1, t_2, \ldots$

  (do not include censoring times)

- Treat each event and its "set of persons still at risk" (i.e., risk set) at each time t_j as an independent table

- Make a 2×2 table ***at each*** t_j

	Event	No Event	Total
Group A	a_j	n_{jA}- a_j	n_{jA}
Group B	c_j	n_{jB}-c_j	n_{jB}
Total	d_j	n_j-d_j	n_j

6. Log-rank test for comparing survivor curves

- At each event time t_j, under assumption of equal survival (IE, $S_A(t) = S_B(t)$), the expected number of events in Group A out of the total events ($d_j = a_j + c_j$) is in proportion to the numbers at risk in group A to the total at risk at time t_j:

$$E\, a_j = d_j \cdot \frac{n_{jA}}{n_j}$$

Differences between a_j and Ea_j represent evidence against the null hypothesis of equal survival in the two groups

Use the Cochran Mantel-Haenszel idea of pooling over events j to get the log-rank χ^2 with one degree of freedom

$$\chi^2_{LR} = \frac{\left[\sum_j (a_j - Ea_j)\right]^2}{\sum_j \hat{Var}\, a_j} \sim \chi^2_1$$

6. Log-rank test for comparing survivor curves

where

$$E\, a_j = d_i \cdot n_{iA} / n_j$$

$$\hat{Var}\,(a_j) = \frac{d_j(n_j - d_j)n_{jA}n_{jB}}{n_j^2(n_j - 1)}$$

- **Stata** log

```
. * Part f.  Log-rank test for comparing survival experience across groups
.
. sts test x

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


Log-rank test for equality of survivor functions
------------------------------------------------

      |  Events
x     |  observed       expected
------+-------------------------
1     |         7          10.13
0     |        10           6.87
------+-------------------------
Total |        17          17.00

            chi2(1) =       2.61
            Pr>chi2 =     0.1061
```

7. Stata do-file script

- Below, is the **Stata** do-file script for the AML data examples in this lecture, including commands to input individual survival data and calculate and plot Kaplan-Meier estimates. The script is available on the course website:

  cl12ex1.do

  (The raw AML data are contained in the script)

```
version 7.0

*  Cl12EX1.DO  Survival analysis

*  Kaplan-Meier curves, log-rank test, Cox PH regression model

*  Data:  AML Weeks in remission -vs-  Treatment group
*
*


*  Raw data:  AML data included below



*  Contents:

* Part a.  Input data, define as a survival dataset
* Part b.  Define survival variables:  stset
* Part c.  Descriptive summaries:  stdes, stsum
* Part d.  Calculate and print Kaplan-Meier estimates for each group
* Part e.  Plot Kaplan-Meier estimates for each group
* Part f.  Log-rank test for comparing survival experience across groups
* Part g.  Fit Cox proportional hazards model

* Assumes files are in folder  [path]\bio623

*    Use Stata command  cd "[path]\bio623"  to point to the correct folder
```

```
* To run this program,  use the following Stata command:

*          do  cl12ex1




* Housekeeping

* Clear workspace
clear


* Turn off -more- pause
set more off


* Save log file on disk, use .txt so Notepad will open it

capture log close
log using cl12ex1.txt, replace


* Make subfolder for graphs
shell md cl12ex1


* Extend linesize for log

set log linesize 100




* Part a.  Input data, define as a survival dataset


*   id,  x(0=no maint 1=maint), t = time to relapse,  failed=(1=relapsed 0=censored)

input id x t failed
1   1  9   1
2   1  13  1
3   1  13  0
4   1  18  1
5   1  23  1
6   1  28  0
7   1  31  1
8   1  34  1
9   1  45  0
10  1  48  1
11  1  161 0
12  0  5   1
13  0  5   1
14  0  8   1
```

```
15  0  8   1
16  0  12  1
17  0  16  0
18  0  23  1
19  0  27  1
20  0  30  0
21  0  33  1
22  0  43  1
23  0  45  1
end
```

```
* Part b.  Define survival variables:  stset

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

* Save as Stata dataset

save cl12ex1.dta , replace
```

```
* Part c.  Descriptive summaries:  stdes, stsum

* Summary stats:  time at risk, rates, subject, 25,50,75 %tiles (K-M estimates)

stsum , by(x)
```

```
* Part d.  Calculate and print Kaplan-Meier estimates for each group

sts list  if x==1

sts list  if x==0

sts list , by(x) compare
```

```
* Part e.  Plot Kaplan-Meier curves for each group
```

```
sts graph , by(x) lost  t1("Kaplan-Meier Survival Curves") t2("    Maintained -vs- Not
Maintained") b2("t,  Weeks") l2("Pr Survived Beyond Time t")

gphprint , saving(cl12ex1\fige2.wmf,replace)



* Part f.  Log-rank test for comparing survival experience across groups

sts test x



* Part g.  Fit Cox proportional hazards model

stcox x

* Show coefficients

stcox , nohr



* Close log file -- Only when all errors have been fixed

*log close
```

8.1 Cox proportional hazards regression model

The regression model for the hazard function (instantaneous incidence rate) as a function of *p* explanatory (*X*) variables is specified as follows:

log hazard:

$$\log \lambda(t; \mathbf{X}) = \log \lambda_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

hazard:

$$\lambda(t; \mathbf{X}) = \lambda_0(t)\,(e^{\beta_1 X_1})\,(e^{\beta_2 X_2}) \ \ldots \ (e^{\beta_p X_p})$$

$$= \lambda_0(t)\, e^{X\beta} \qquad (\text{ vector of } X\text{s})$$

Interpretation of $\lambda_0(t)$:

Hazard (incidence) rate as a function of time when all *X*'s are zero – often must center *X*s to make $\lambda_0(t)$ interpretable

8.1 Cox proportional hazards regression model

Interpretation of e^{β_1}:

e^{β_1} is the relative hazard associated with a *1* unit change in X_1 (*IE,* X_1*+1 -vs-* X_1), holding other *X*s constant, independent of time

or, in relative risk terms,

e^{β_1} is the relative risk for X_1*+1 -vs-* X_1, holding other *X*s constant, independent of time

Other βs have similar interpretations

Note:

$e^{X\beta}$ "multiplies" the baseline hazard $\lambda_0(t)$ by the same amount regardless of the time t. This is therefore a " proportional hazards" model – the effect of any (fixed) *X* is the same at any time during follow-up

8.1 Cox proportional hazards regression model

- Applying the formula relating *S(t)* to the cumulative hazard to the proportional hazards model,

$$S(t) = e^{-\int_0^t \lambda(u)\,du},$$

gives,

$$S(t;X) = e^{-\int_0^t \lambda_0(u) e^{X\beta}\,du}$$

$$= e^{\left(-\int_0^t \lambda_0(u)\,du\right) e^{X\beta}}$$

$$= [S_0(t)]^{e^{X\beta}}$$

- β is the focus whereas $\lambda_0(t)$ is a nuisance variable

- David Cox (1972) showed how to estimate β without having to assume a model for $\lambda_0(t)$

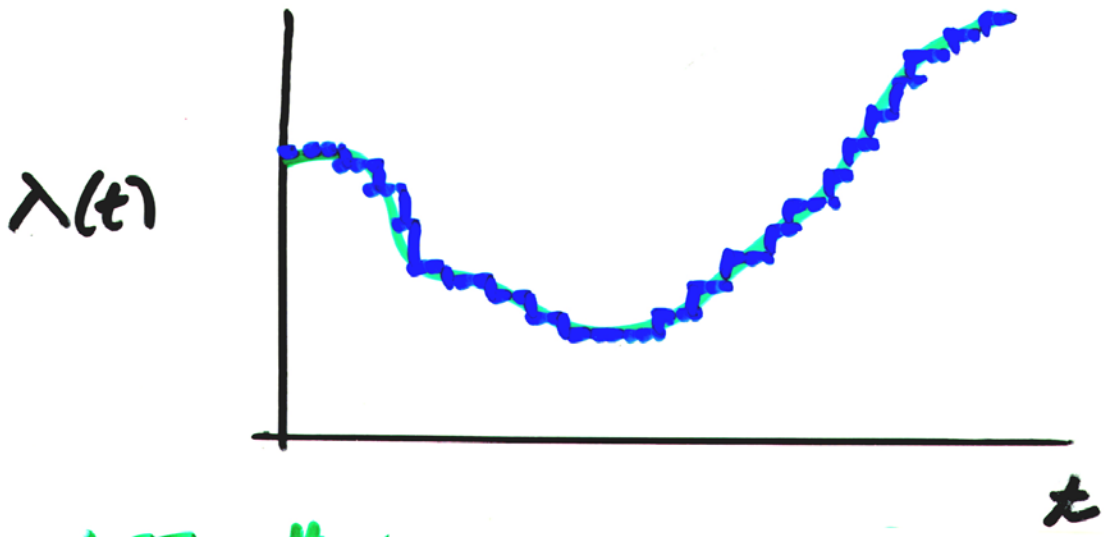
8.1 Cox proportional hazards regression model

- "Semi-parametric"

  — $\lambda_0(t)$ is the baseline hazard - "non-parametric" part of the model

  — $X\beta$ are the regression coefficients - "parametric" part of the model

- Think of estimating $\lambda_0(t)$ with a step function



- Let # steps get large ⇒ "partial likelihood" for β depends on β, not $\lambda_0(t)$

8.2 Partial likelihood

- Let the survival times (times to failure) be:

  $t_1 < t_2 < \ldots < t_k$

- And let the "risk sets" corresponding to these times be:

  $R_1, R_2, \ldots, R_k$

  R_j = list of persons at risk just before t_j

- Then, the "partial likelihood" for β is

$$L(\beta) = \prod_{i=1}^{k} \left(\frac{e^{X_i\beta}}{\sum_{j \in R_i} e^{X_j\beta}} \right)$$

  (Assumes no ties in event times)

- To estimate β, find the values of βs that minimize $L(\beta)$ above!

8.2 Partial likelihood

- Why does the partial likelihood make sense?

$$\frac{e^{x_i\beta}}{\sum_{j \in R_i} e^{x_j\beta}} = \frac{\lambda_0(t_i)\, e^{x_j\beta}}{\sum_{j \in R_i} \lambda_0(t_i)\, e^{x_j\beta}}$$

$$= \frac{\textit{hazard of failed person}}{\textit{hazards of ones who could have failed at } t_i}$$

- Choose β so that the one who failed at each time was most likely - relative to others who might have failed!

8.3 Example: Cox PH model for AML data

- Semi-parametric model for the hazard (incidence) rate for the AML data

$$\lambda_i(t) = \lambda_0(t)\, e^{X_i\beta}$$

where $\lambda_i(t)$ is the hazard for person I at week t, $\lambda_0(t)$ is the hazard if $X_i = 0$ (not maintained group), and $e^{X_i\beta}$ is the multiplicative effect of $X_i=1$ (maintained group)

Variable	β	SE_β	z
Group	-0.812	.521	1.56

$e^{\hat{\beta}} = 0.44$ -- relative rate of AML relapse maintained vs not maintained

1/0.44 = 2.25 – relative rate of AML relapse not-maintained vs maintained

95% CI: $[e^{.812 - 2(.521)}, e^{.812 + 2(.521)}]$

(.81, 6.26)

8.3 Example: Cox PH model for AML data

- Stata log

```
.
. * Part g.  Fit Cox proportional hazards model
. * X coded 1 for not maintained; 0 for maintained
. stcox x

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

Iteration 0:   log likelihood = -40.700899
Iteration 1:   log likelihood = -39.438723
Iteration 2:   log likelihood = -39.438713
Refining estimates:
Iteration 0:   log likelihood = -39.438713

Cox regression -- Breslow method for ties

No. of subjects =           23                     Number of obs   =        23
No. of failures =           17
Time at risk    =          678
                                                   LR chi2(1)      =      2.52
Log likelihood  =   -39.438713                     Prob > chi2     =    0.1121

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   2.251808   1.174376      1.556   0.120       .8102293    6.258279
------------------------------------------------------------------------------

.
.
```

8.3 Example: Cox PH model for AML data

```
. * Show coefficients
.
. stcox , nohr

Cox regression -- Breslow method for ties

No. of subjects =           23                     Number of obs   =        23
No. of failures =           17
Time at risk    =          678
                                                   LR chi2(1)      =      2.52
Log likelihood  =   -39.438713                     Prob > chi2     =    0.1121

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .8117336   .5215257      1.556   0.120       -.210438    1.833905
------------------------------------------------------------------------------
```

8.4 Example: Cox PH model for CABG surgery

- Fisher and Van Belle use a Cox model to compare two treatments, controlling for several predictors

  — Compare surgical (CABG) with medical treatment for left main coronary heart disease

  — Use mortality (time to death) as the response variable

  — Control for 7 risk factors (age at baseline and 6 coronary status measures) in making the comparison

  — Time variable is time from treatment initiation to death or censoring due to the end of the study or lost to follow-up

8.4 Example: Cox PH model for CABG surgery

- Variables

X_1 = *THRPY*	1=med 2=surgical
X_2 = *CHFSCR*	Congestive heart failure score: 0-4
X_3 = *LMCA*	% lowering of diameter of left main coronary artery
X_4 = *LVSCR*	Left ventricular function score: 5-30
X_5= *DOM*	Dominant side of heart: 0=right/balanced 1=left
X_6 = *AGE*	Patient's age in years (at baseline)
X_7 = *HYPTEN*	History of hypertension (1=yes 0=no)
X_8 = *RCA*	Right coronary artery stenosis: 1= $\geq$ 70% stenosis 0 = otherwise

8.4 Example: Cox PH model for CABG surgery

- Model for the log hazard rate (incidence of death):

$$\log \lambda(t;X) = \log \lambda_0(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_8 X_8$$

- Model for the hazard rate

$$\lambda(t;X) = \lambda_0(t)\, e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_8 X_8}$$

8.4 Example: Cox PH model for CABG surgery

- Cox model results:

Variable	β	*SE*	*z*
THRPY	*-1.0777*	*.1668*	*-6.46*
CHFSCR	*.2985*	*.0667*	*4.48*
LMCA	*.0178*	*.0049*	*3.63*
LVSCR	*.1126*	*.0182*	*6.19*
DOM	*1.2331*	*.3564*	*3.46*
AGE	*.0423*	*.0098*	*4.32*
HYPTEN	*-.5428*	*.1547*	*-3.51*
RCA	*.5285*	*.2923*	*1.81*

8.4 Example: Cox PH model for CABG surgery

- What is the relative risk of death for the CABG group compared to the medical group, adjusting for age and other risk factors?

  $e^{-1.0777} = .34$ – 66% reduction in the risk of death for otherwise comparable patients treated with CABG compared with patients treated medically

  – note the coding 2=CABG, 1=Medical gives the same results as 1=CABG and 0=Medical

- What is the interpretation of each coefficient?

  CHFSCR: Controlling for type of treatment and other risk factors, the risk of death, as estimated from a Cox model, is $e^{.2985} = 1.35$ times higher per unit difference in CHF score

8.4 Example: Cox PH model for CABG surgery

— *AGE*: Controlling for type of treatment and other risk factors, the risk of death, as estimated from a Cox model, is $e^{.0423} = 1.04$ times higher per year of age

— *HYPTEN*: Controlling for type of treatment and other risk factors, the risk of death, as estimated from a Cox model, is $e^{-.5428} = 0.58$ times lower for patients who have a history of hypertension compared with those who do not -- anyone know why a history of hypertension should lower risk following treatment?

— ETC -- You do!

8.4 Example: Cox PH model for CABG surgery

- What is the relative risk death for

(A) a medically treated 45-year old

-vs-

(B) a surgically treated 75 year old

who otherwise have comparable risk factors?

log hazard for (A) =

const + 1·(-1.0777) + 45·(.0423) =

const + .8258

log hazard for (B) =

const + 2·(-1.0777) + 75·(.0423) =

const + 1.017

Difference in log hazards, (B) -vs- (A):

(const+1.017) - (const+.8258)

= .1913

8.4 Example: Cox PH model for CABG surgery

Relative Risk, (B) -vs- (A):

$e^{.1913} = 1.21$ – higher risk for older, surgically treated patient than for younger, medically treated patient

— Is the assumption of "otherwise comparable risk factors" reasonable?

8.4 Example: Cox PH model for CABG surgery

- How much higher is the risk of a 70 year old patient compared with a 60 year old patient, assuming treatment and other coronary risk factors are the same?

  The estimated difference in log hazards for two patients whose ages differ by 10 years, holding other predictors fixed is

  $10 \times \hat{\beta}_{AGE} = 10 \times .0423 = .423$

  $RR = e^{.423} = 1.53$

  – A ten year difference in the age at initiation of treatment increases the risk of subsequent mortality by 50%

- How would you determine whether the mortality advantage of CABG over medical treatment was greater for younger patients than for older patients?

8.5 Stata do-file

- Examples for this lecture included in:

  cl12ex1.do