V. INTRODUCTION TO SURVIVAL ANALYSIS

- Survival data: time to event
  - Right censored data
- Kaplan-Meier survival curves
- Kaplan-Meier cumulative mortality curves
  - Greenwood confidence bands for survival and mortality curves
  - Displaying censoring times and numbers of patients at risk
- Estimating survival probabilities
- Censoring and biased Kaplan-Meier survival curves
- Log rank test for comparing survival curves
- Hazard functions and cumulative mortality
- Simple proportional hazards regression model
  - Hazard rate ratios and relative risk
  - Estimating relative risks from proportional hazards models
- Tied failure times and biased relative risk estimates

1. Survival and Cumulative Mortality Functions

Suppose we have a cohort of n people.

Let

t_i be the age that the i^{th} person dies,

$m[t]$ be the number of patients for whom $t < t_i$, and

$d[t]$ be the number of patients for whom $t_i \le t$.

Then the **survival function** is

$S[t] = \Pr\left[t_i > t\right]$ = the probability of surviving until at least age t.

The **cumulative mortality function** is

$D[t] = \Pr[t_i \le t]$ = the probability of dying before age t.

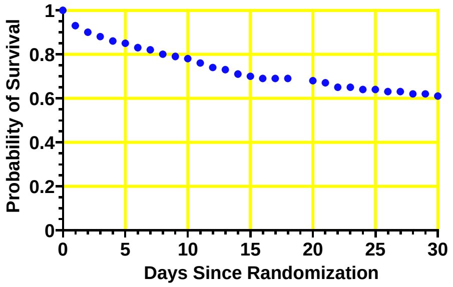
If t_i is known for all members of the cohort we can estimate $S(t)$ and $D(t)$ by

$\hat{S}[t] = m[t]/n$ the proportion of subjects who are alive at age t, and

$\hat{D}[t] = d[t]/n$ the proportion who have died by age t.

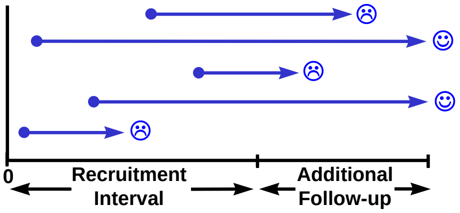
a) Example: Survival among sepsis patients

Days Since Entry	Number of Patients Alive	Number of Deaths	Proportion Alive
0	n = m(0) = 455	0	m(0) / n = 1.00
1	m(1) = 423	32	m(1) / n = 0.93
2	m(2) = 410	45	m(2) / n = 0.90
3	m(3) = 400	55	m(3) / n = 0.88
4	m(4) = 392	63	m(4) / n = 0.86
5	m(5) = 386	69	m(5) / n = 0.85
6	m(6) = 378	77	m(6) / n = 0.83
7	m(7) = 371	84	m(7) / n = 0.82
8	m(8) = 366	89	m(8) / n = 0.80
9	m(9) = 360	95	m(9) / n = 0.79
10	m(10) = 353	102	m(10) / n = 0.78
.	.	.	.
.	.	.	.
.	.	.	.
21	m(21) = 305	150	m(21) / n = 0.67
22	m(22) = 296	159	m(22) / n = 0.65
23	m(23) = 295	160	m(23) / n = 0.65
24	m(24) = 292	163	m(24) / n = 0.64
25	m(25) = 290	165	m(25) / n = 0.64
26	m(26) = 288	167	m(26) / n = 0.63
27	m(27) = 286	169	m(27) / n = 0.63
28	m(28) = 283	172	m(28) / n = 0.62
29	m(29) = 280	175	m(29) / n = 0.62
30	m(30) = 279	176	m(30) / n = 0.61



2. Right Censored Data

In clinical studies, patients are typically recruited over a recruitment interval and then followed for an additional period of time.



Let

t_i = the time from entry to exit for the i^{th} patient

and

$$f_i = \begin{cases} 1: & i^{th} \text{ patient dies at exit} \\ 0: & i^{th} \text{ patient alive at exit} \end{cases}$$

Patients who are alive at exit are said to be **right censored**. This means that we know that they survived until at least time t_i but do not know how much longer they lived thereafter.

With censored data, the proportion of patients who are known to have died by time t underestimates the true cumulative mortality since some patients will die after their censoring times.

3. Kaplan-Meier (Product Limit) Survival Curves

Suppose that we have censored survival data on a cohort of patients. We divide the follow-up time into intervals that are small enough that few patients die in any one interval.

Suppose this interval is days.

Let

n_i be the number of patients known to be at risk at the beginning of day *i*.

d_i be the number of patients who die on day *i*

Then for patients alive at the beginning of the i^{th} day, the estimated probability of surviving the day is

$$p_i = \frac{n_i - d_i}{n_i}$$

The probability that a patient survives the first *t* days is the joint probability of surviving days 1, 2, …,*t* which is estimated by

$$\hat{S}[t] = p_1 p_2 p_3 \ldots p_t$$

Note that $p_i = 1$ on all days that no deaths are observed. Hence, if t_k denotes the k^{th} day on which deaths are observed then

$$\hat{S}[t] = \prod_{\{k:t_k<t\}} p_k \qquad \{7.1\}$$

This estimate is the **Kaplan-Meier survival curve**.

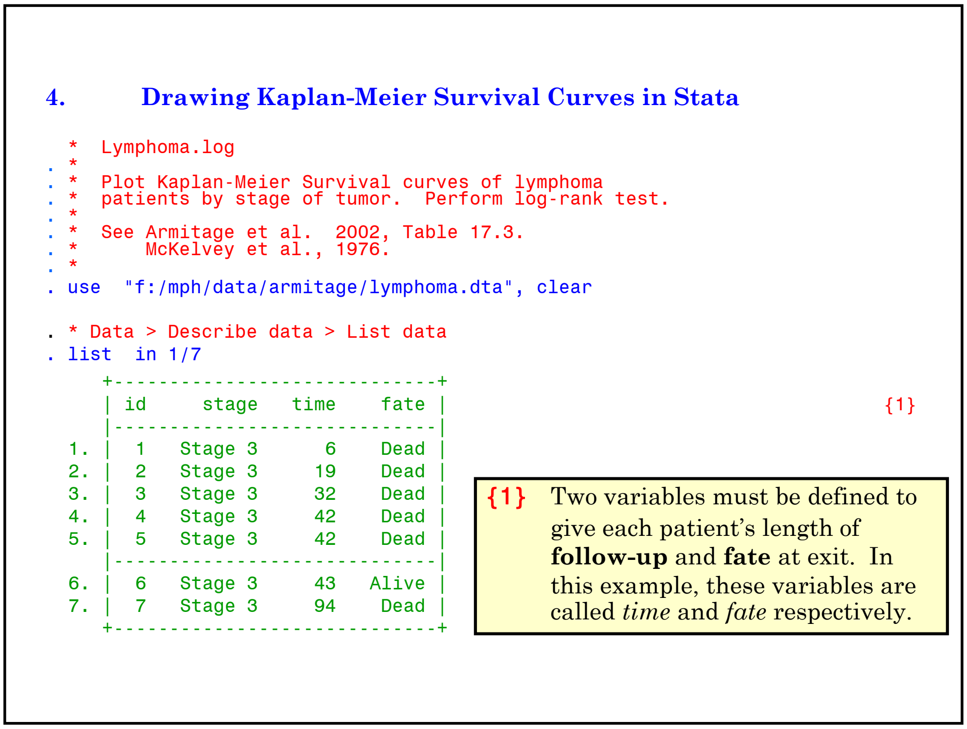
The **Kaplan-Meier cumulative mortality curve** is

$$\hat{D}[t] = 1 - \hat{S}[t]$$

a) Example: Survival in lymphoma patients

Armitage et al. (2002: p. 579) discuss the following data on patient survival after recruitment into a clinical of patients with diffuse histiocytic lymphoma (KcKelvey et al. *Cancer* 1976; **38**: 1484 – 93).

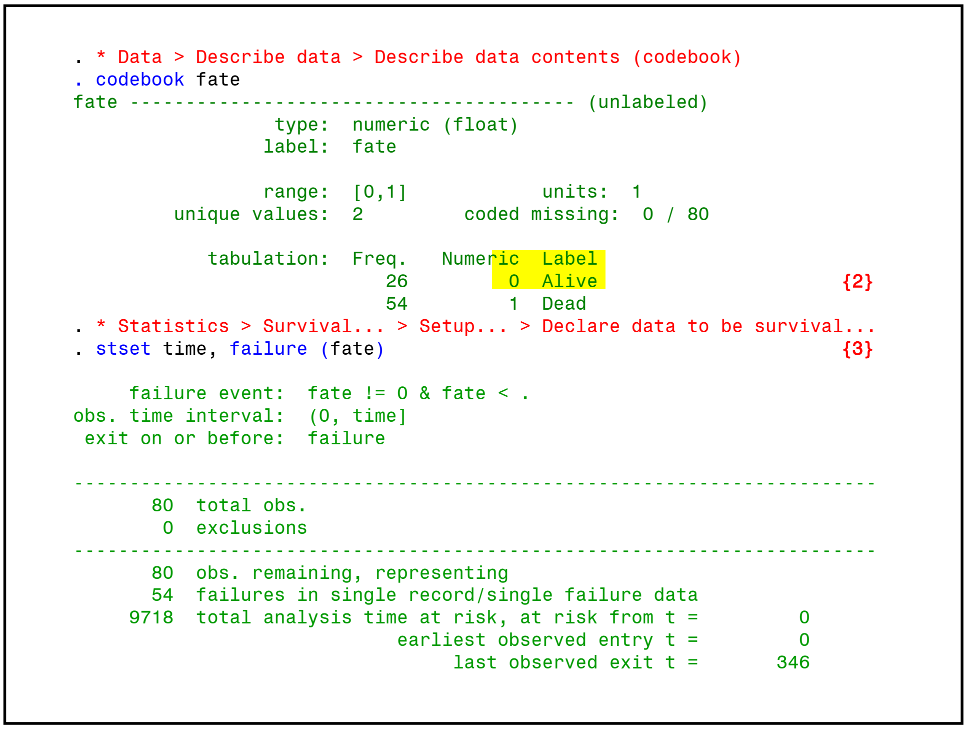
	Follow-up (days)							
	Dead at end of follow-up				Alive at end of follow-up			
Stage 3								
	6	19	32	42	43	126	169	211
	42	94	207	253	227	255	270	310
					316	335	346	
Stage 4								
	4	6	10	11	41	43	61	61
	11	11	13	17	160	235	247	260
	20	20	21	22	284	290	291	302
	24	24	29	30	304	341	345	
	30	31	33	34				
	35	39	40	45				
	46	50	56	63				
	68	82	85	88				
	89	90	93	104				
	110	134	137	169				
	171	173	175	184				
	201	222						



4. Drawing Kaplan-Meier Survival Curves in Stata

```
  *  Lymphoma.log
. *
. *  Plot Kaplan-Meier Survival curves of lymphoma
. *  patients by stage of tumor.  Perform log-rank test.
. *
. *  See Armitage et al.  2002, Table 17.3.
. *      McKelvey et al., 1976.
. *
. use  "f:/mph/data/armitage/lymphoma.dta", clear

. * Data > Describe data > List data
. list  in 1/7

     +----------------------------+
     | id     stage   time    fate |                                {1}
     |----------------------------|
  1. |  1   Stage 3      6    Dead |
  2. |  2   Stage 3     19    Dead |
  3. |  3   Stage 3     32    Dead |
  4. |  4   Stage 3     42    Dead |
  5. |  5   Stage 3     42    Dead |
     |----------------------------|
  6. |  6   Stage 3     43   Alive |
  7. |  7   Stage 3     94    Dead |
     +----------------------------+
```

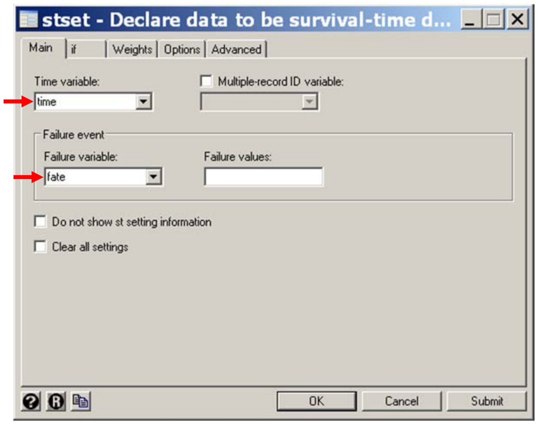
{1} Two variables must be defined to give each patient’s length of **follow-up** and **fate** at exit. In this example, these variables are called *time* and *fate* respectively.



```
. * Data > Describe data > Describe data contents (codebook)
. codebook fate
fate ---------------------------------------- (unlabeled)
                  type:  numeric (float)
                 label:  fate

                 range:  [0,1]            units:  1
         unique values:  2        coded missing:  0 / 80

            tabulation:  Freq.   Numeric  Label
                            26         0  Alive                    {2}
                            54         1  Dead
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset time, failure (fate)                                        {3}

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

------------------------------------------------------------------------
       80  total obs.
        0  exclusions
------------------------------------------------------------------------
       80  obs. remaining, representing
       54  failures in single record/single failure data
     9718  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       346
```

{2} The ***fate*** variable is coded as 0 = alive and 1 = dead at exit

{3} ***stset*** specifies that the data set contains survival data, with each patient's **exit time** denoted by ***time*** and **status at exit** denoted by ***fate***. Stata interprets ***fate*** = 0 to mean that the patient is **censored** at exit and **fate** ≠ 0 to mean that she suffered the **event** of interest at exit.



stset - Declare data to be survival-time d...

Time variable: time

□ Multiple-record ID variable:

Failure event

Failure variable: fate

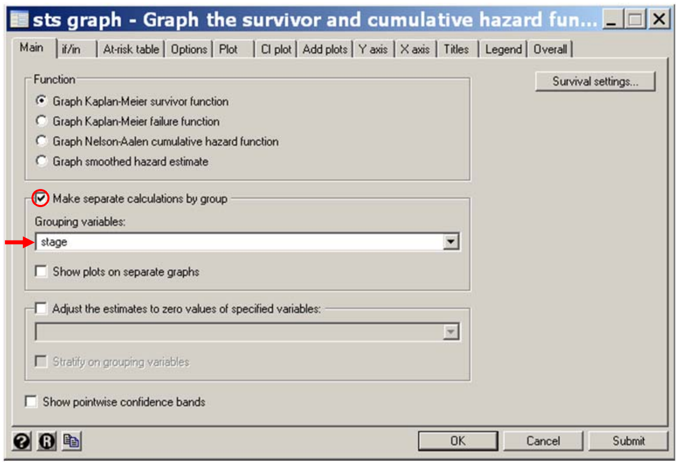
Failure values:

□ Do not show st setting information

□ Clear all settings

```
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(stage) ytitle(Probability of Survival)             {4}

    failure time:  time
  failure/censor:  fate
```

{4} ***sts graph*** plots Kaplan-Meier survival curves. ***by(stage)*** specifies that **separate** plots will be generated for each value of *stage*. The *y*-axis title is *Probability of Survival*.



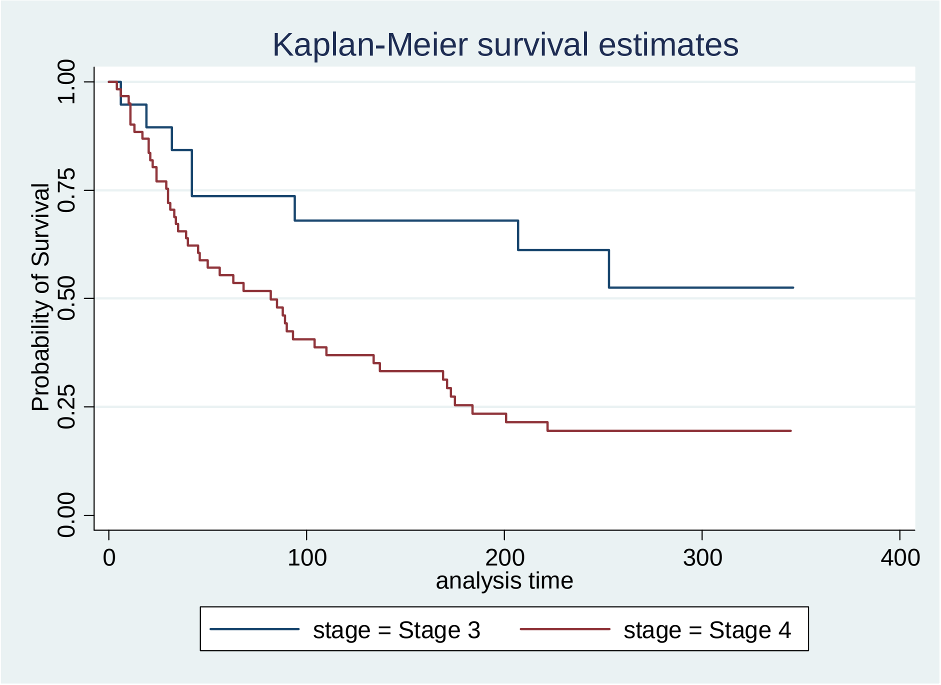
sts graph - Graph the survivor and cumulative hazard fun...

Function
● Graph Kaplan-Meier survivor function
○ Graph Kaplan-Meier failure function
○ Graph Nelson-Aalen cumulative hazard function
○ Graph smoothed hazard estimate

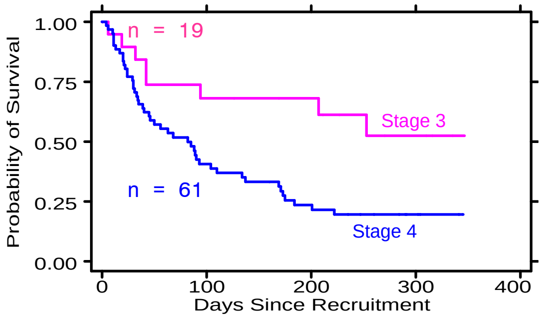
☑ Make separate calculations by group
Grouping variables: stage
□ Show plots on separate graphs

□ Adjust the estimates to zero values of specified variables:
□ Stratify on grouping variables

□ Show pointwise confidence bands

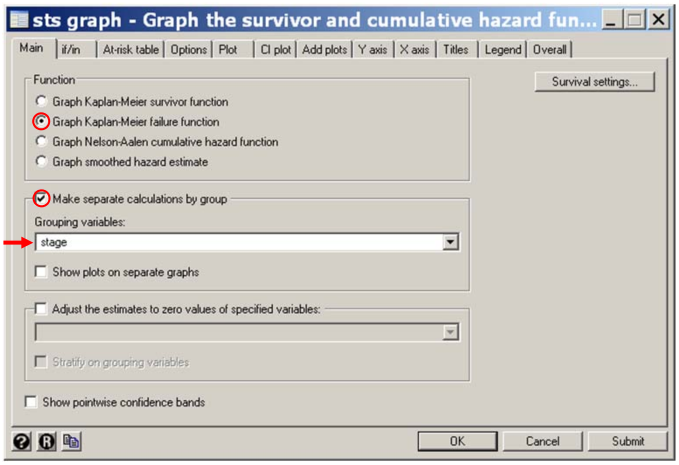


- In the preceding graph, $\hat{S}(t)$ is **constant** over days when **no deaths** are observed and **drops** abruptly on days when **deaths occur.**
- If the time interval is short enough that there is rarely more than one death per interval, then the **height** of the drop at each death day indicates the **size** of the cohort remaining on that day.
- The **accuracy** of the survival curve gets **less** as we move towards the right, as it is based on **fewer** and fewer **patients.**



We can also plot the cumulative mortality curve using the **failure** option as follows

```
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(stage) ytitle(Cumulative Mortality) failure
```



Function

○ Graph Kaplan-Meier survivor function
● Graph Kaplan-Meier failure function
○ Graph Nelson-Aalen cumulative hazard function
○ Graph smoothed hazard estimate

☑ Make separate calculations by group

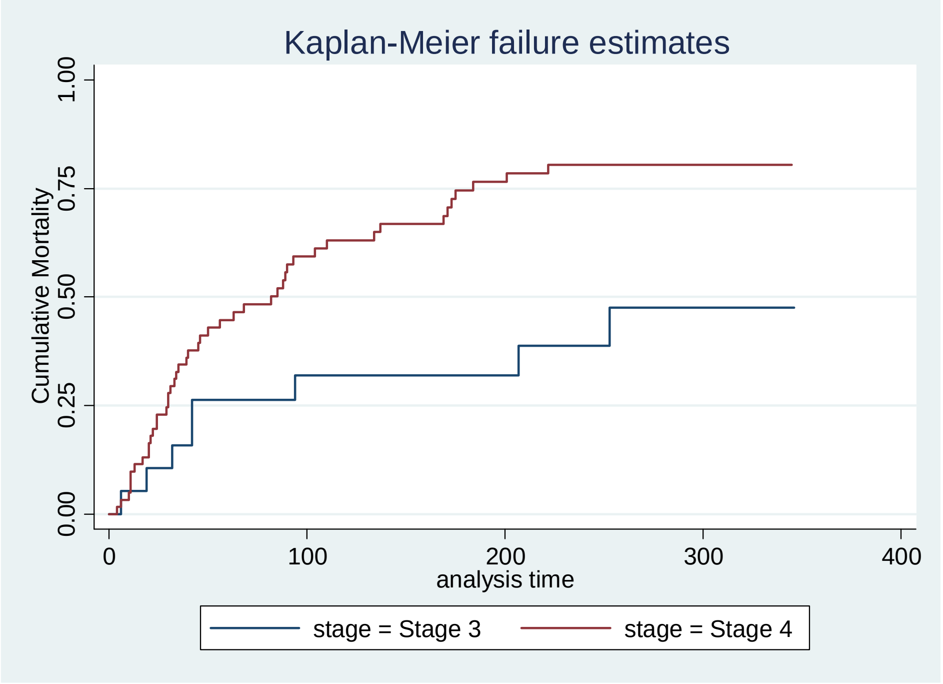
Grouping variables: stage

□ Show plots on separate graphs

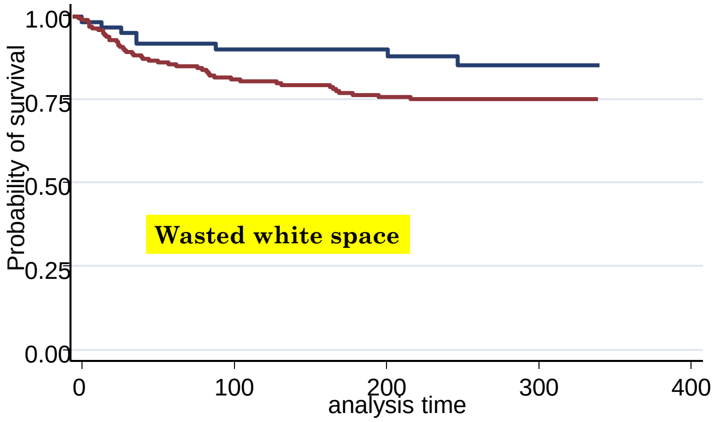
□ Adjust the estimates to zero values of specified variables:

□ Stratify on grouping variables

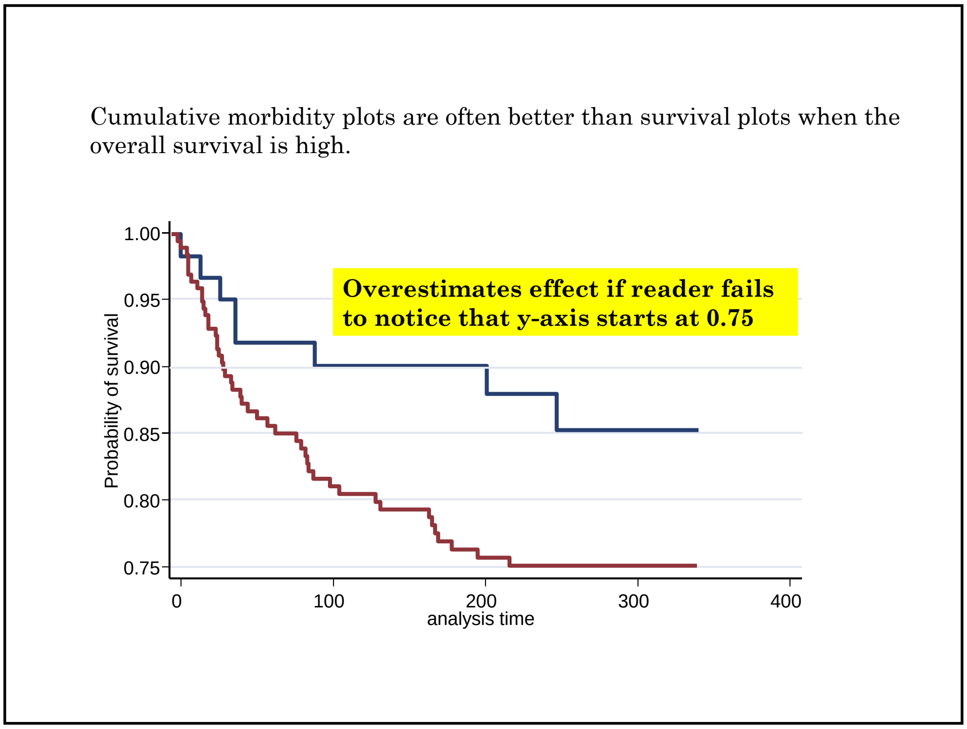
□ Show pointwise confidence bands



Cumulative morbidity plots are often better than survival plots when the overall survival is high.

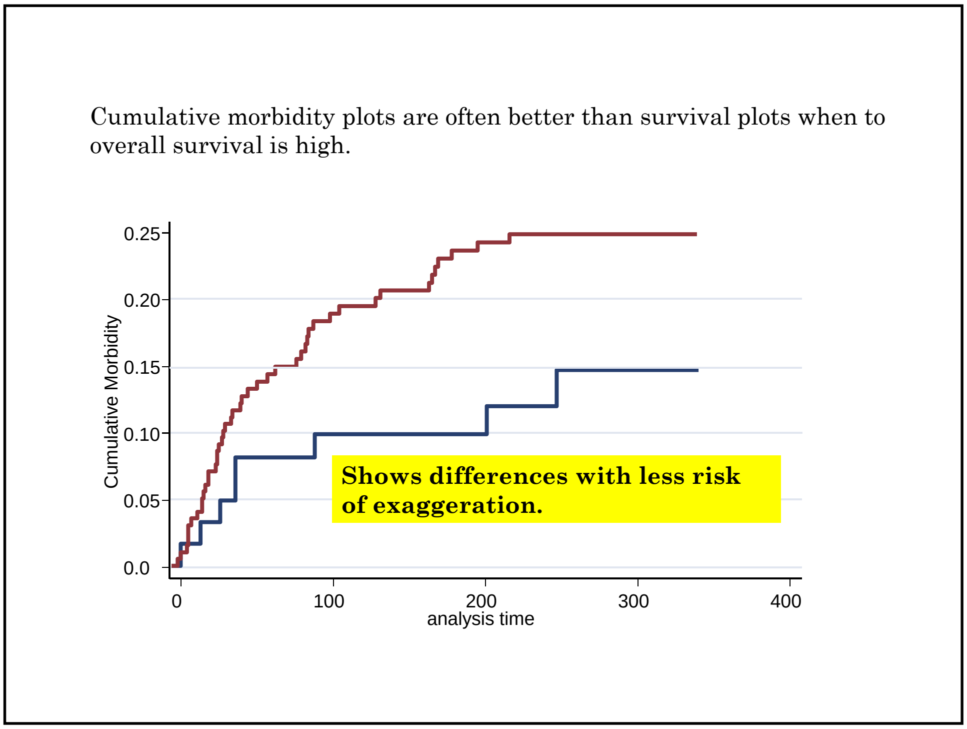


Wasted white space



Cumulative morbidity plots are often better than survival plots when the overall survival is high.

Overestimates effect if reader fails to notice that y-axis starts at 0.75



Cumulative morbidity plots are often better than survival plots when to overall survival is high.

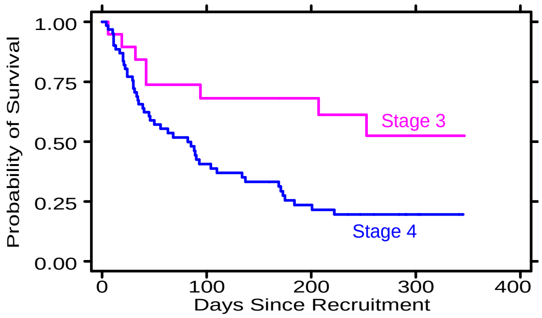
Shows differences with less risk of exaggeration.

- If there is no censoring and there a q death days before time t then

$$\hat{S}(t) = \left(\frac{n_1 - d_1}{n_1}\right)\left(\frac{n_2 - d_2}{n_1 - d_1}\right)....\left(\frac{n_q - d_q}{n_{q1} - d_{q1}}\right)$$

$$= \frac{n_q - d_q}{n_1} = \frac{m(t)}{n}$$

Hence the **Kaplan-Meier** survival curve reduces to the **proportion** of patients alive at time t if there is no **censoring**.



a) Life Tables

A life table is a table that gives estimates of $S(t)$ for different values of t. The term is slightly old fashioned but is still used.

5. 95% Confidence Intervals for Survival Functions

The variance of $\hat{S}(t)$ is estimated by Greenwood's formula

$$s^2_{\hat{S}(t)} = \hat{S}(t)^2 \sum_{\{k:t_k<t\}} \frac{d_k}{n_k(n_k - d_k)} \qquad \{7.2\}$$

A 95% confidence interval for $S(t)$ could be estimated by

$$\hat{S}(t) \pm 1.96 s_{\hat{S}(t)}$$

However, this interval does not optimal when $\hat{S}(t)$ is near 0 or 1 since this statistic will have a skewed distribution near these extreme values (the true survival curve is never less than 0 or greater than 1).

The variance of $\log\left[-\log\left[\hat{S}(t)\right]\right]$ has variance

$$\hat{\sigma}^2(t) = \frac{\sum_{\{k:t_k<t\}} \frac{d_k}{n_k(n_k-d_k)}}{\left[\sum_{\{k:t_k<t\}} \log\left[\frac{(n_k-d_k)}{d_k}\right]\right]^2} \qquad \{7.3\}$$

and a 95% confidence interval $\log\left[-\log\left[\hat{S}(t)\right]\right] \pm 1.96\hat{\sigma}(t)$.

Exponentiating twice gives a 95% confidence interval for $\hat{S}(t)$ of

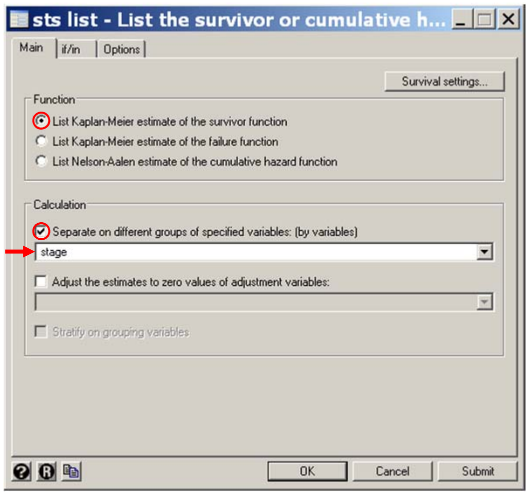
$$\hat{S}(t)^{\exp(\mp 1.96\hat{\sigma}(t))} \qquad \{7.4\}$$

which behaves better for extreme values of $\hat{S}(t)$. We can either list or plot these values with Stata. *Lymphoma.log* continues as follows:

```
. *
. *  List survival statistics
. *
. * Statistics > Survival... > Summary statistics... > List survivor...
. sts list, by(stage)                                                         {1}
    failure time:  time
  failure/censor:  fate
           Beg.          Net         Survivor      Std.
  Time    Total   Fail   Lost        Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
stage=3
     6       19      1      0          0.9474    0.0512     0.6812    0.9924
    19       18      1      0          0.8947    0.0704     0.6408    0.9726
    32       17      1      0          0.8421    0.0837     0.5865    0.9462
    42       16      2      0          0.7368    0.1010     0.4789    0.8810
    43       14      0      1          0.7368    0.1010     0.4789    0.8810
    94       13      1      0          0.6802    0.1080     0.4214    0.8421   {2}
         .
         .
         .
   335        2      0      1          0.5247    0.1287     0.2570    0.7363
   346        1      0      1          0.5247    0.1287     0.2570    0.7363
```

{1} ***sts list*** provides the same data that is plotted by *sts graph*.

{2} For example, of the original **19** stage three patients there are **13** still alive at the beginning of the 94 days of follow-up. There were **5 deaths** in this group before day **94** and one death on day **94**. The survivor Function $\hat{S}(94) = 0.68$, with standard error $s_{\hat{S}(t)}$ = 0.11. The 95 % confidence interval for $\hat{S}(94)$ is (0.42, 0.84)



sts list - List the survivor or cumulative h...

Function

● List Kaplan-Meier estimate of the survivor function

○ List Kaplan-Meier estimate of the failure function

○ List Nelson-Aalen estimate of the cumulative hazard function

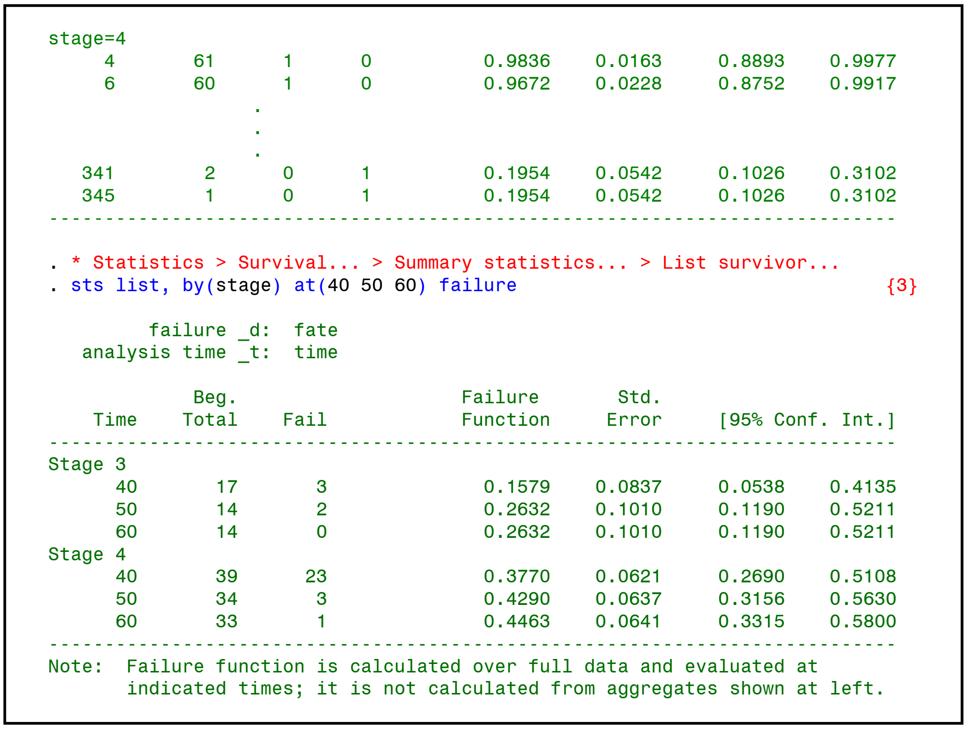
Calculation

☑ Separate on different groups of specified variables: (by variables)

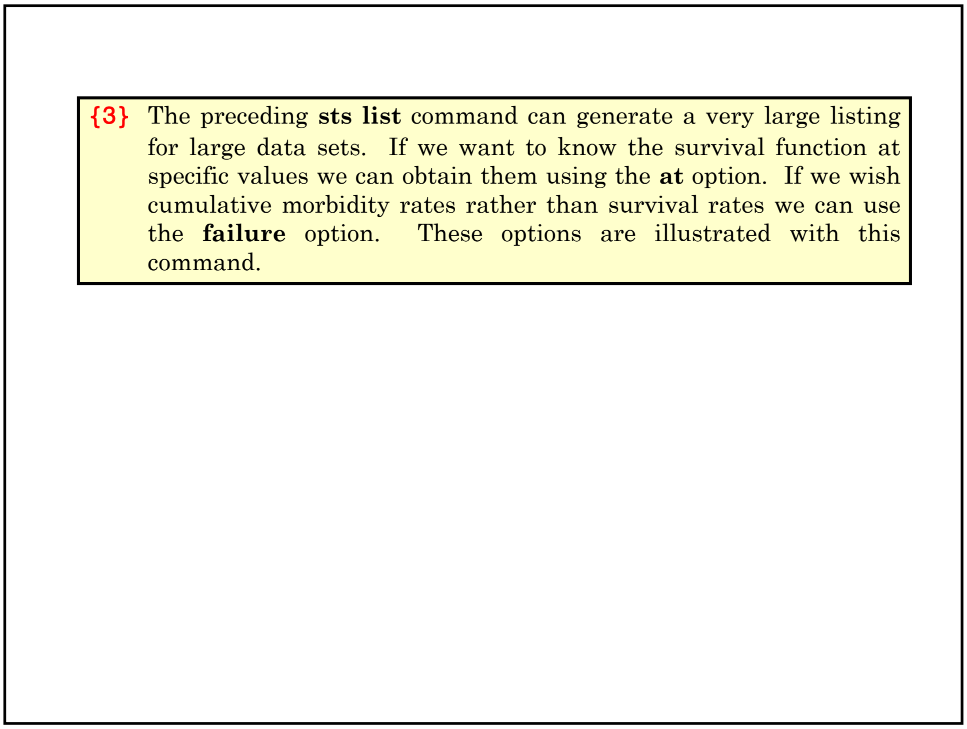
stage

□ Adjust the estimates to zero values of adjustment variables:

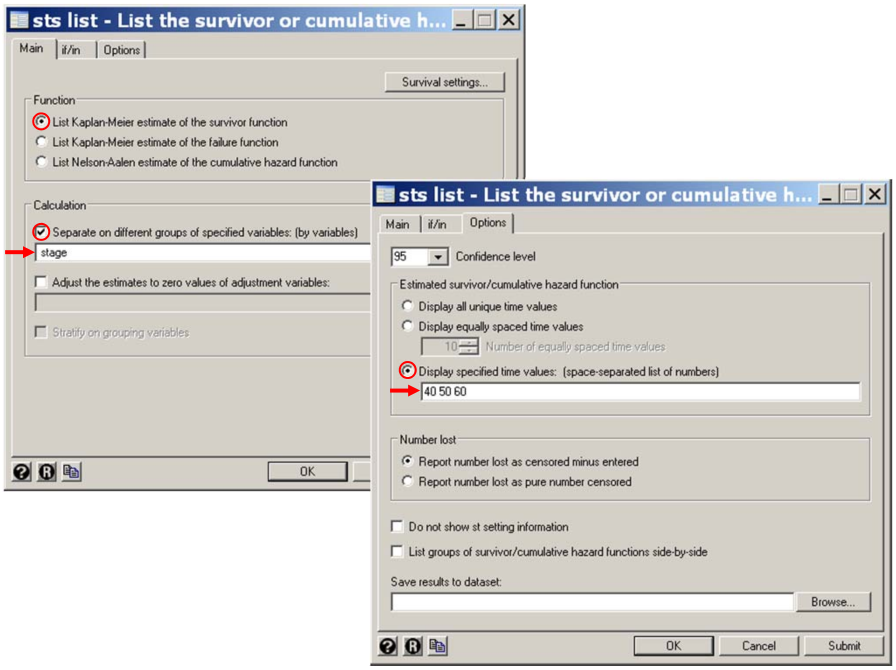
□ Stratify on grouping variables



```
stage=4
     4       61      1      0          0.9836    0.0163     0.8893    0.9977
     6       60      1      0          0.9672    0.0228     0.8752    0.9917
                  .
                  .
                  .
   341        2      0      1          0.1954    0.0542     0.1026    0.3102
   345        1      0      1          0.1954    0.0542     0.1026    0.3102
----------------------------------------------------------------------------

. * Statistics > Survival... > Summary statistics... > List survivor...
. sts list, by(stage) at(40 50 60) failure                                 {3}

         failure _d:  fate
   analysis time _t:  time

            Beg.                   Failure       Std.
    Time    Total    Fail          Function      Error     [95% Conf. Int.]
----------------------------------------------------------------------------
Stage 3
      40       17       3            0.1579    0.0837     0.0538    0.4135
      50       14       2            0.2632    0.1010     0.1190    0.5211
      60       14       0            0.2632    0.1010     0.1190    0.5211
Stage 4
      40       39      23            0.3770    0.0621     0.2690    0.5108
      50       34       3            0.4290    0.0637     0.3156    0.5630
      60       33       1            0.4463    0.0641     0.3315    0.5800
----------------------------------------------------------------------------
Note:  Failure function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```



{3} The preceding **sts list** command can generate a very large listing for large data sets. If we want to know the survival function at specific values we can obtain them using the **at** option. If we wish cumulative morbidity rates rather than survival rates we can use the **failure** option. These options are illustrated with this command.



sts list - List the survivor or cumulative h...

Function
● List Kaplan-Meier estimate of the survivor function
○ List Kaplan-Meier estimate of the failure function
○ List Nelson-Aalen estimate of the cumulative hazard function

Calculation
☑ Separate on different groups of specified variables: (by variables)
stage
□ Adjust the estimates to zero values of adjustment variables:
□ Stratify on grouping variables

sts list - List the survivor or cumulative h...

95 Confidence level

Estimated survivor/cumulative hazard function
○ Display all unique time values
○ Display equally spaced time values
10 Number of equally spaced time values
● Display specified time values: (space-separated list of numbers)
40 50 60

Number lost
● Report number lost as censored minus entered
○ Report number lost as pure number censored

□ Do not show st setting information
□ List groups of survivor/cumulative hazard functions side-by-side

Save results to dataset:

```
. *
. *  Kaplan-Meier survival curves by stage with 95% CIs
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(stage)   ci censored(single) separate     ///       {4}
>     xlabel(0 (50) 350) xmtick(0 (25) 350)               ///
>     byopts(title(, size(0)) legend(off))                ///       {5}
>     ytitle(Probability of Survival)                     ///
>     ylabel(0 (.1) 1, angle(0)) ciopts(color(yellow))    ///       {6}
>     xtitle(Days Since Recruitment)  ymtick(0 (.05) 1)
```

{4} Stata also permits users to graph confidence bounds for $\hat{S}(t)$ and to indicate when subjects lost to follow-up with tick marks. This is done with the ***ci*** and ***censored(single)*** options, respectively. The ***separate*** option causes the survival curves to be drawn in separate panels.

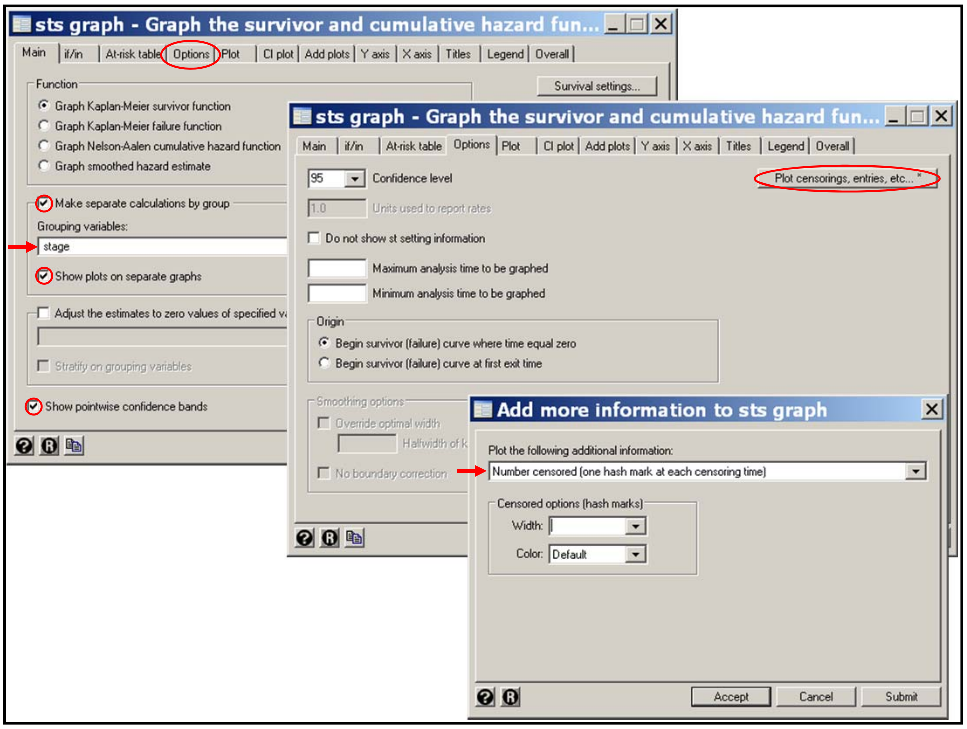
{5} The ***byopts*** option controls attributes related to having multiple curves on the same graph; ***title(" ", size(0))*** suppresses the graph's default title; ***legend(off)*** suppresses the legend. When the ***separate*** option is given ***title*** and ***legend*** must be suboptions of ***byopts*** rather than separate options.

{6} The **ciopts** option allows control of the confidence bands. Here we choose yellow bands.

{4} Stata also permits users to graph confidence bounds for $\hat{S}(t)$ and to indicate when subjects lost to follow-up with tick marks. This is done with the ***ci*** and ***censored(single)*** options, respectively. The ***separate*** option causes the survival curves to be drawn in separate panels.

{5} The ***byopts*** option controls attributes related to having multiple curves on the same graph; ***title(" ", size(0))*** suppresses the graph's default title; ***legend(off)*** suppresses the legend. When the ***separate*** option is given ***title*** and ***legend*** must be suboptions of ***byopts*** rather than separate options.

{6} The **ciopts** option allows control of the confidence bands. Here we choose yellow bands.



sts graph - Graph the survivor and cumulative hazard fun...

Function

● Graph Kaplan-Meier survivor function

○ Graph Kaplan-Meier failure function

○ Graph Nelson-Aalen cumulative hazard function

○ Graph smoothed hazard estimate

☑ Make separate calculations by group

Grouping variables:

stage

☑ Show plots on separate graphs

□ Adjust the estimates to zero values of specified v

□ Stratify on grouping variables

☑ Show pointwise confidence bands

sts graph - Graph the survivor and cumulative hazard fun...

95 Confidence level

1.0 Units used to report rates

□ Do not show st setting information

Maximum analysis time to be graphed

Minimum analysis time to be graphed

Origin

● Begin survivor (failure) curve where time equal zero

○ Begin survivor (failure) curve at first exit time

Smoothing options

□ Override optimal width

Halfwidth of k

□ No boundary correction

Add more information to sts graph

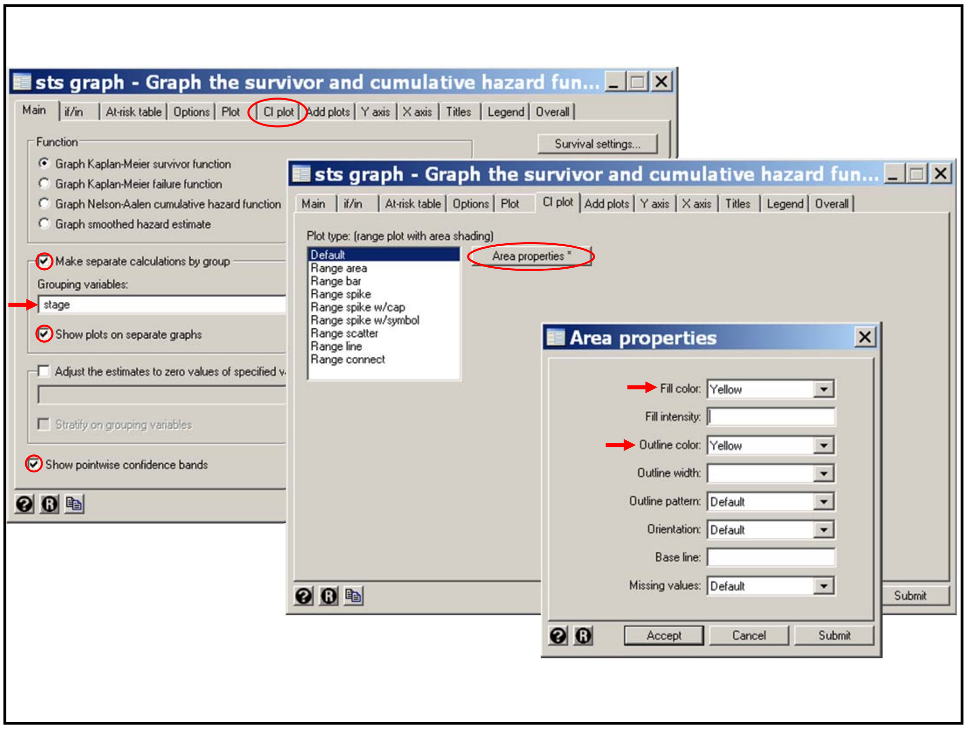
Plot the following additional information:

Number censored (one hash mark at each censoring time)

Censored options (hash marks)

Width:

Color: Default



sts graph - Graph the survivor and cumulative hazard fun...

Function

● Graph Kaplan-Meier survivor function

○ Graph Kaplan-Meier failure function

○ Graph Nelson-Aalen cumulative hazard function

○ Graph smoothed hazard estimate

☑ Make separate calculations by group

Grouping variables: stage

☑ Show plots on separate graphs

□ Adjust the estimates to zero values of specified v

□ Stratify on grouping variables

☑ Show pointwise confidence bands

sts graph - Graph the survivor and cumulative hazard fun...

Plot type: (range plot with area shading)

- Default
- Range area
- Range bar
- Range spike
- Range spike w/cap
- Range spike w/symbol
- Range scatter
- Range line
- Range connect

Area properties *

Area properties

Fill color: Yellow

Fill intensity:

Outline color: Yellow

Outline width:

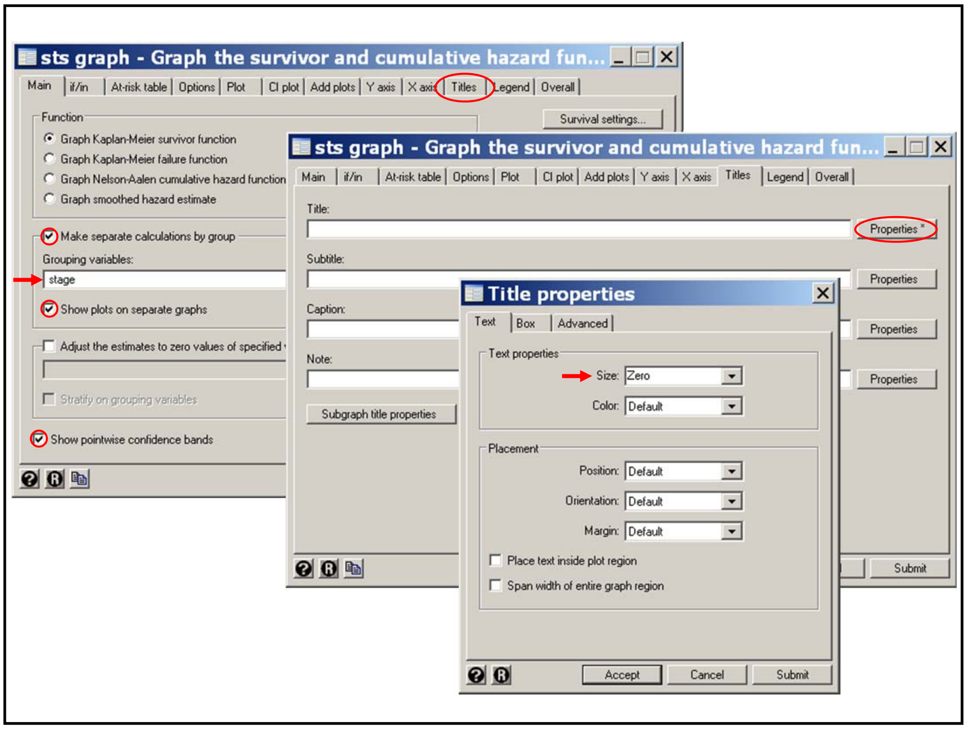
Outline pattern: Default

Orientation: Default

Base line:

Missing values: Default

Accept Cancel Submit



sts graph - Graph the survivor and cumulative hazard fun...

Function

● Graph Kaplan-Meier survivor function

○ Graph Kaplan-Meier failure function

○ Graph Nelson-Aalen cumulative hazard function

○ Graph smoothed hazard estimate

☑ Make separate calculations by group

Grouping variables: stage

☑ Show plots on separate graphs

□ Adjust the estimates to zero values of specified v

□ Stratify on grouping variables

☑ Show pointwise confidence bands

sts graph - Graph the survivor and cumulative hazard fun...

Title: Properties *

Subtitle: Properties

Caption: Properties

Note: Properties

Subgraph title properties

Title properties

Text | Box | Advanced

Text properties

Size: Zero

Color: Default

Placement

Position: Default

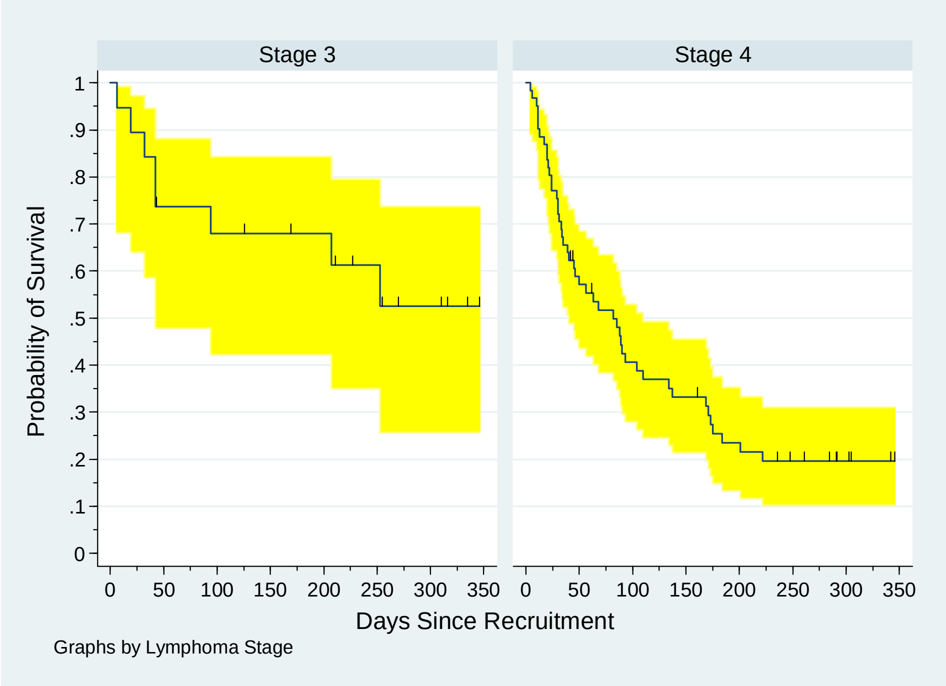
Orientation: Default

Margin: Default

□ Place text inside plot region

□ Span width of entire graph region

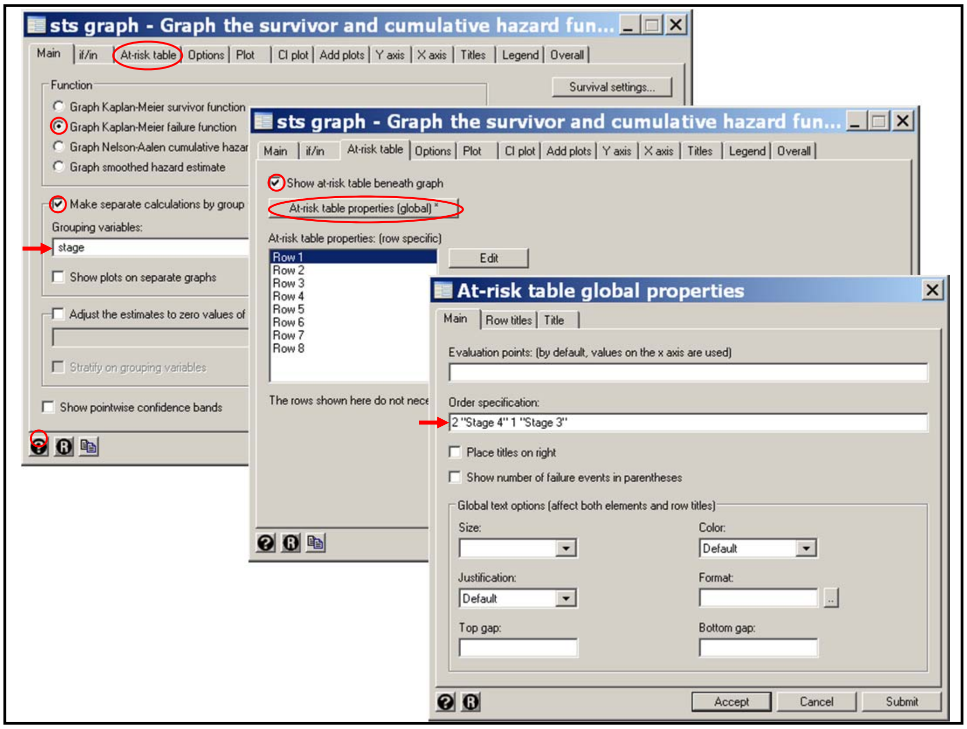
Accept Cancel Submit



Some journals require a table showing the number of subjects at risk at different survival times given below the survival curve. In Stata this can be done as follows.

```
. *
. *  Kaplan-Meier morbidity curves by stage with risk table
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(stage)  failure                           ///
>     risktable(,order(2 "Stage 4" 1 "Stage 3"))          /// {7}
>     ytitle(Cumulative Mortality)                        ///
>     xlabel(0 (50) 350) xmtick(0 (25) 350)               ///
>     ylabel(0 (.1) .8, angle(0))                        ///
>     xtitle(Days Since Recruitment)  ymtick(0 (.05) .8)  ///
>     title(" ",size(0)) legend(ring(0) cols(1)           ///
>        position(11) order(2 "Stage 4" 1 "Stage 3"))
```

{7} The ***risktable*** option creates a risk table below the graph with one row for each curve that is drawn. The ***order*** suboption orders and labels these rows. Its syntax is identical to that of the ***order*** suboption of the ***legend*** option.


sts graph - Graph the survivor and cumulative hazard fun...
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Survival settings...
Edit
At-risk table global properties
Main | Row titles | Title
Accept
Cancel
Submit

Function
○ Graph Kaplan-Meier survivor function
● Graph Kaplan-Meier failure function
○ Graph Nelson-Aalen cumulative hazar
○ Graph smoothed hazard estimate

☑ Make separate calculations by group
Grouping variables:
stage

□ Show plots on separate graphs

□ Adjust the estimates to zero values of

□ Stratify on grouping variables

□ Show pointwise confidence bands

☑ Show at-risk table beneath graph
At-risk table properties (global) *
At-risk table properties: (row specific)
Row 1
Row 2
Row 3
Row 4
Row 5
Row 6
Row 7
Row 8

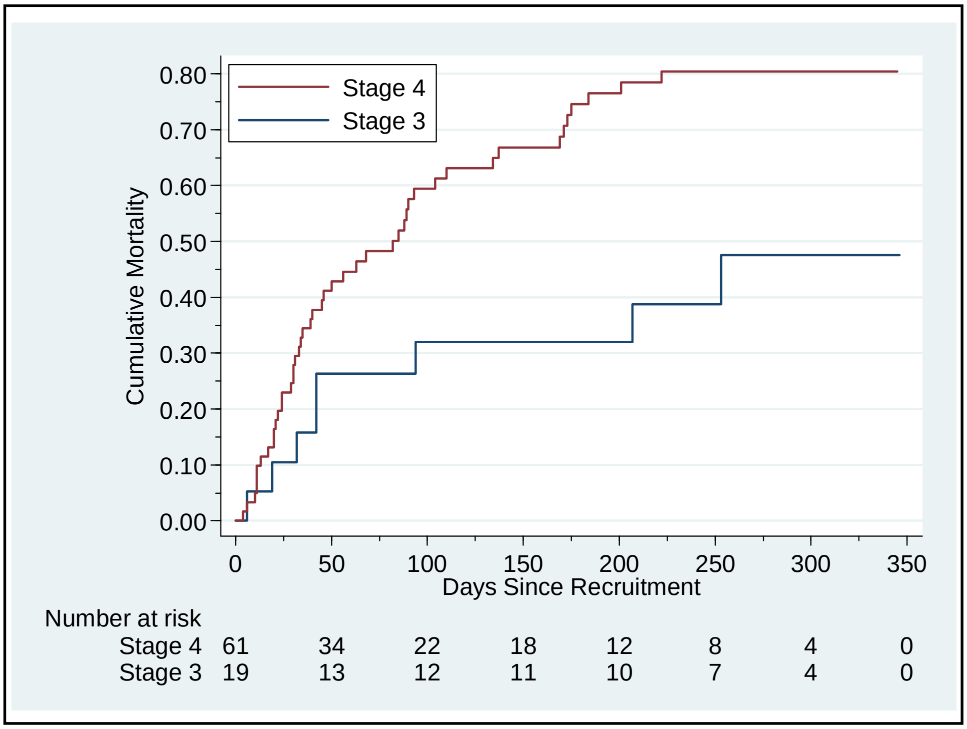
The rows shown here do not nece

Evaluation points: (by default, values on the x axis are used)

Order specification:
2 "Stage 4" 1 "Stage 3"

□ Place titles on right
□ Show number of failure events in parentheses

Global text options (affect both elements and row titles)
Size:
Color: Default
Justification: Default
Format:
Top gap:
Bottom gap:



6. Censoring and Bias

Kaplan-Meier survival curves will be unbiased estimates of the true survival curve as long as

1. The patients are representative of the underlying population and
2. Patients who are censored have the same risk of suffering the event of interest as are patients who are not.

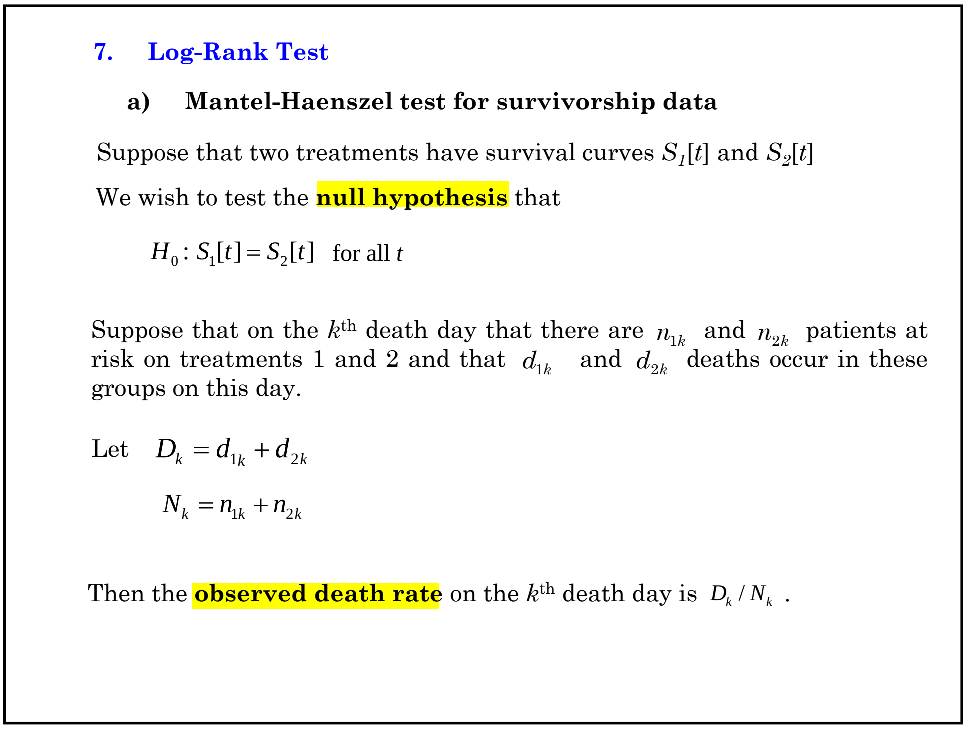
If censored patients are more likely to die than uncensored patients with equal follow-up then our survival estimates will be biased.

Such bias can occur for many reasons, not the least of which is that dead patients do not return for follow-up visits.

Survival curves are often derived for some endpoint other than death. In this case, some deaths may be treated as censoring events.

For example, if the event of interest is developing of breast cancer, then we may treat death due to heart disease as a censoring event. This is reasonable as long as there is no relationship between heart disease and breast cancer. That is, when we censor a woman who died of heart disease, we are assuming that she would have had the same subsequent risk of breast cancer as other women if she had lived.

If we were studying lung cancer, then treating death from heart disease as a censoring event would bias our results since smoking increases the risk of both lung cancer morbidity and cardiovascular mortality and patients who die of heart disease are more likely to have smoked and hence would have been more likely to develop lung cancer if they had not died of heart disease first.



7. Log-Rank Test

a) Mantel-Haenszel test for survivorship data

Suppose that two treatments have survival curves $S_1[t]$ and $S_2[t]$

We wish to test the **null hypothesis** that

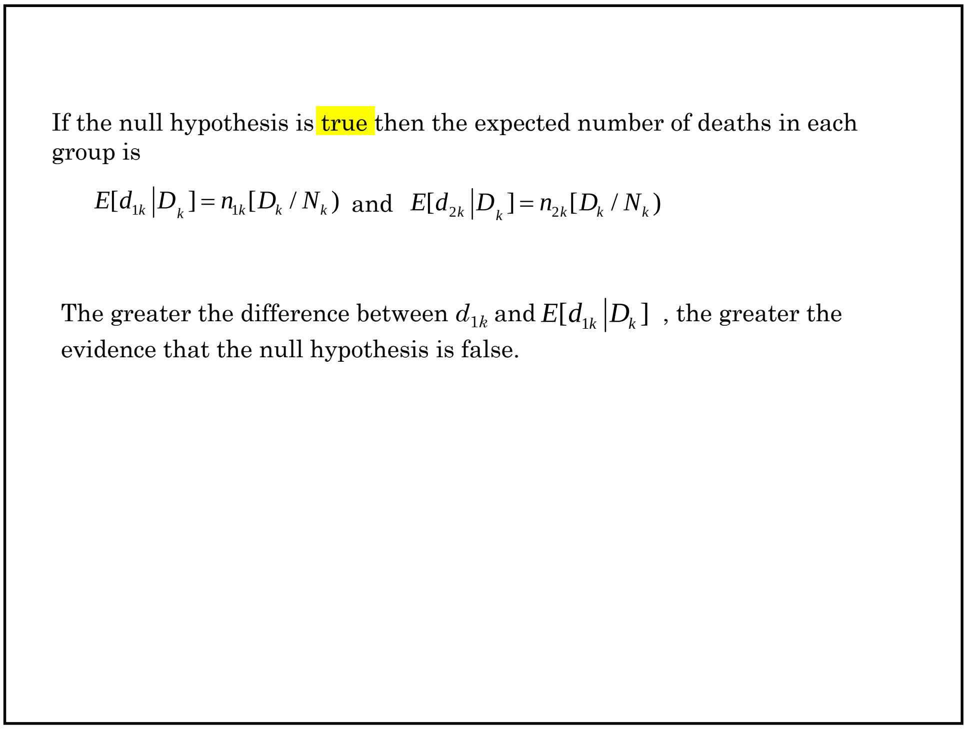
$$H_0: S_1[t] = S_2[t] \quad \text{for all } t$$

Suppose that on the k^{th} death day that there are n_{1k} and n_{2k} patients at risk on treatments 1 and 2 and that d_{1k} and d_{2k} deaths occur in these groups on this day.

Let $D_k = d_{1k} + d_{2k}$

$$N_k = n_{1k} + n_{2k}$$

Then the **observed death rate** on the k^{th} death day is D_k / N_k .



If the null hypothesis is true then the expected number of deaths in each group is

$$E[d_{1k} \left| D_k \right.] = n_{1k}[D_k / N_k) \text{ and } E[d_{2k} \left| D_k \right.] = n_{2k}[D_k / N_k)$$

The greater the difference between d_{1k} and $E[d_{1k} \left| D_k \right.]$, the greater the evidence that the null hypothesis is false.

Mantel proposed forming the 2x2 contingency tables

k^{th} death day	Treatment 1	Treatment 2	Total
Died	d_{1k}	d_{2k}	D_k
Survived	$n_{1k} - d_{1k}$	$n_{2k} - d_{2k}$	$N_k - D_k$
Total	n_{1k}	n_{2k}	N_k

on each death day and performing a Mantel-Haenszel χ^2 test.

This test was renamed the **log-rank test** by Peto who studied its mathematical properties.

If the time interval is short enough that $d_k \le 1$ for each interval, then the test of H_0 depends only on the order in which the deaths occur and not on their time of occurrence.

It is in this sense that the test is a rank test.

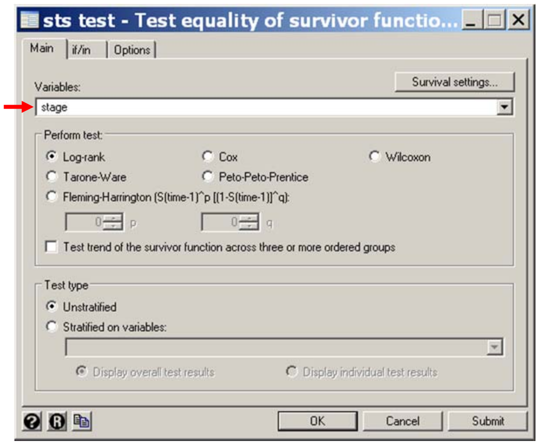
b) Example: Tumor stage in lymphoma patients

Lymphoma.log continues as follows:

```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test stage
                                                                   {1}
     failure _d:  fate
   analysis time _t:  time

Log-rank test for equality of survivor functions

      |  Events         Events
stage |  observed       expected
------+-------------------------
3     |         8          16.69
4     |        46          37.31
------+-------------------------
Total |        54          54.00

            chi2(1) =       6.71
            Pr>chi2 =     0.0096                                   {2}
```

{1} Perform a **log-rank** test for equality of survivor functions in patient groups defined by different values of *stage*. In this example, stage 3 patients are compared to stage 4 patients.

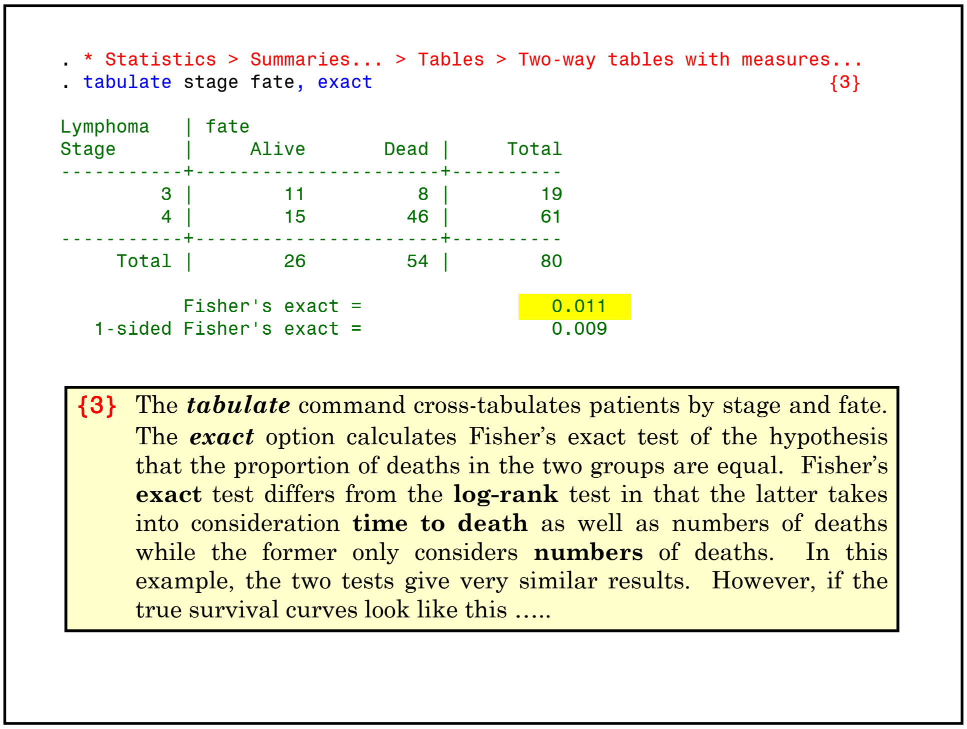
{2} In this example, the log-rank P value = **0.0096**, indicating that the marked **difference** in survivorship between stage **3** and stage **4** lymphoma patients is not likely to be due to chance.



Variables: stage

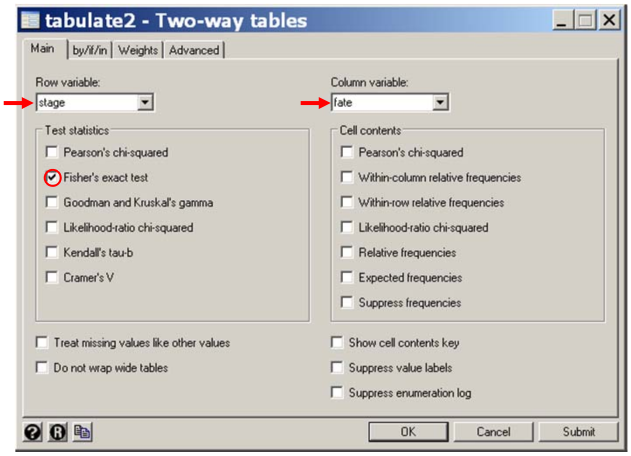
Perform test:
● Log-rank ○ Cox ○ Wilcoxon
○ Tarone-Ware ○ Peto-Peto-Prentice
○ Fleming-Harrington (S(time-1)^p [(1-S(time-1)]^q):
0 p 0 q
□ Test trend of the survivor function across three or more ordered groups

Test type
● Unstratified
○ Stratified on variables:
● Display overall test results ○ Display individual test results



```
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate stage fate, exact                                          {3}

Lymphoma   | fate
Stage      |     Alive       Dead |     Total
-----------+----------------------+----------
         3 |        11          8 |        19
         4 |        15         46 |        61
-----------+----------------------+----------
     Total |        26         54 |        80

           Fisher's exact =                 0.011
   1-sided Fisher's exact =                 0.009
```

{3} The ***tabulate*** command cross-tabulates patients by stage and fate. The ***exact*** option calculates Fisher's exact test of the hypothesis that the proportion of deaths in the two groups are equal. Fisher's **exact** test differs from the **log-rank** test in that the latter takes into consideration **time to death** as well as numbers of deaths while the former only considers **numbers** of deaths. In this example, the two tests give very similar results. However, if the true survival curves look like this .....



tabulate2 - Two-way tables

Row variable: stage

Column variable: fate

Test statistics

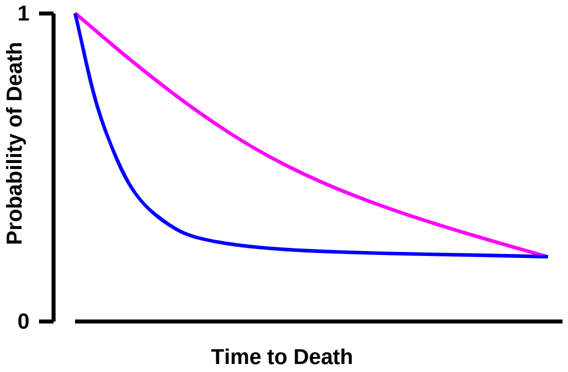
□ Pearson's chi-squared
☑ Fisher's exact test
□ Goodman and Kruskal's gamma
□ Likelihood-ratio chi-squared
□ Kendall's tau-b
□ Cramer's V

Cell contents

□ Pearson's chi-squared
□ Within-column relative frequencies
□ Within-row relative frequencies
□ Likelihood-ratio chi-squared
□ Relative frequencies
□ Expected frequencies
□ Suppress frequencies

□ Treat missing values like other values
□ Do not wrap wide tables

□ Show cell contents key
□ Suppress value labels
□ Suppress enumeration log



…the log-rank test may be highly significant even though the observed death rates in each group are equal. Fisher's exact test, however, will not be significant if the death rates are the same.

c) Log-rank test for multiple patient groups

The log-rank test generalizes to allow the comparison of survival in several groups.

These groups are defined by the number of distinct levels taken by the variable specified in the *sts test* command. E.g. in the preceding example if there were four different lymphoma stages define by *stage* then *sts test stage* would compare the four survival curves for these groups of patients. The test statistic has an asymptotic χ^2 distribution with one degree of freedom less than the number of patient groups being compared.

8. Hazard Functions

Suppose that a patient is alive at time t and that her probability of dying in the short time interval $(t, t+\Delta t)$ is

$$\lambda[t]\Delta t$$

Then $\lambda[t]$ is said to be the hazard function for the patient at time t.

More precisely

$$\lambda[t] = \frac{\Pr\left[\begin{array}{l|l}\text{Patient dies by} & \text{Patient alive} \\ \text{time } t+\Delta t & \text{at time } t\end{array}\right]}{\Delta t} \qquad \{7.5\}$$

For a very large population

$$\lambda[t]\Delta t \cong \frac{\text{The number of deaths in the interval } (t, t+\Delta t)}{\text{Number of people alive at time } t}$$

$\lambda[t]$ is the instantaneous rate per unit time at which people are dying at time t.

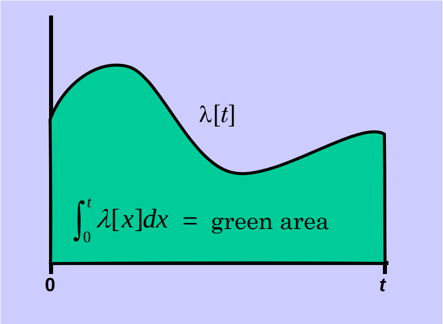
$\lambda[t] = 0$ implies that there is no risk of death at time t and $S[t]$ is flat at time t.

Large values of $\lambda[t]$ imply a rapid rate of decline in $S[t]$.

The hazard function is related to the survival function through the equation

$$S[t] = \exp\left[-\int_0^t \lambda[x]dx\right]$$

where $\int_0^t \lambda[x]dx$ is the area under the curve $\lambda[x]$ between 0 and t.



a) Proportional hazards

Suppose that $\lambda_0[t]$ and $\lambda_1[t]$ are the hazard functions for control and experimental for treatments, respectively.

Then these treatments have proportional hazards if

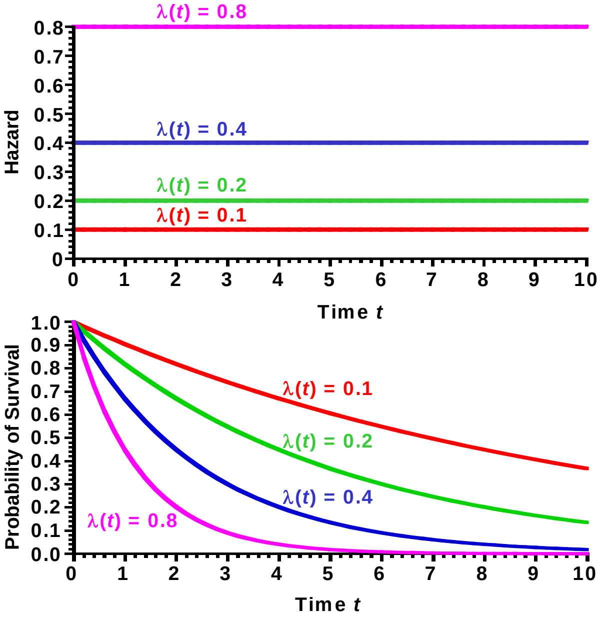
$$\lambda_1[t] = R\,\lambda_0[t]$$

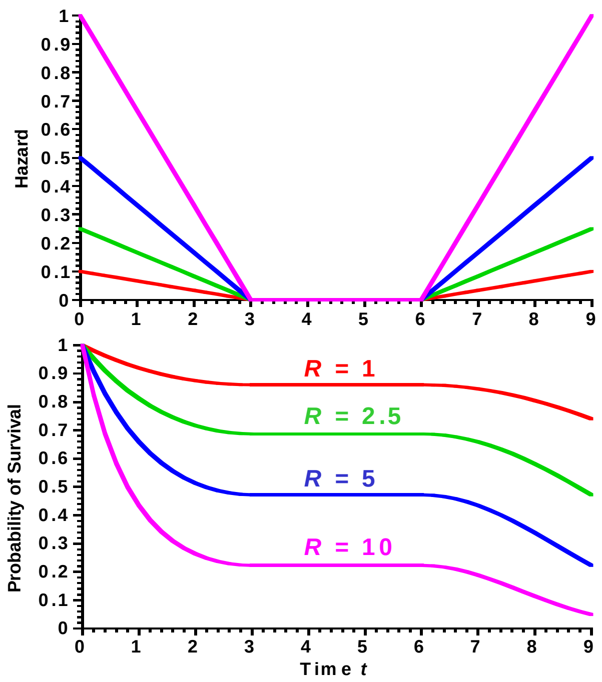
for some constant *R*.

The proportional hazards assumption places no restrictions on the shape of $\lambda_0(t)$ but requires that

$$\lambda_1[t]/\lambda_0[t] = R$$

Examples:





b) Relative risks and hazard ratios

Suppose that the risks of death by time $t + \Delta t$ for patients on control and experimental treatments who are alive at time t are $\lambda_0[t]\Delta t$ and, $\lambda_1[t]\Delta t$ respectively.

Then the risk of experimental subjects at time t relative to control is

$$\frac{\lambda_1[t]\Delta t}{\lambda_0[t]\Delta t} = \frac{\lambda_1[t]}{\lambda_0[t]}$$

If $\lambda_1[t] = R\lambda_0[t]$ at all times, then this relative risk is

$$\frac{\lambda_1[t]}{\lambda_0[t]} = \frac{R\lambda_0[t]}{\lambda_0[t]} = R$$

Thus the ratio of two hazard functions can be thought of as an instantaneous relative risk, or as a relative risk if this ratio is constant.

9. Proportional Hazards Regression Analysis

a) The model

Suppose that $\lambda_0[t]$ and $\lambda_1[t]$ are the hazard functions for the control and experimental therapies and β is an unknown parameter. The **proportional hazards** model assumes that

$$\lambda_1[t] = \lambda_0[t]\exp[\beta]$$

This model is said to be semi-nonparametric in that it makes no assumptions about the shape of the control hazard function.

If $\hat{\beta}$ is an estimate of β then $\exp[\hat{\beta}]$ estimates the relative risk of the experimental therapy relative to controls since

$$R = \frac{\lambda_1[t]}{\lambda_0[t]} = \frac{\exp[\beta]\lambda_0[t]}{\lambda_0[t]} = \exp[\beta]$$

b) Example: Risk of stage 3 vs. stage 4 lymphoma

In Stata proportional hazards regression analysis is performed by the *stcox* command. The *Lymphoma.log* file continues as follows.

```
. *
. *  Preform proportional hazards regression analysis of
. *  lymphoma patients by stage of tumor.
. *
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox stage                                                                   {1}

         failure _d:  fate
   analysis time _t:  time

Iteration 0:   Log Likelihood = -207.5548
Iteration 1:   Log Likelihood =-203.86666
Iteration 2:   Log Likelihood =-203.73805
Iteration 3:   Log Likelihood =-203.73761
Refining estimates:
Iteration 0:   Log Likelihood =-203.73761

Cox regression -- Breslow method for ties

No. of subjects =           80                     Number of obs   =        80
No. of failures =           54
Time at risk    =         9718
                                                   LR chi2(1)      =      7.63
Log likelihood   =   -203.73761                    Prob > chi2     =    0.0057

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   2.614362   1.008191     2.49   0.013     1.227756    5.566976 {2}
------------------------------------------------------------------------------
```

{1} This command fits the **proportional hazards** regression model.

$$\lambda(t, stage) = \lambda_0(t)\exp(\beta \times stage)$$

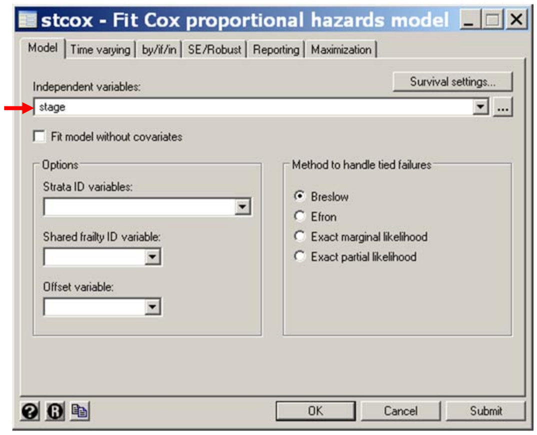
A **stset** command must precede the **stcox** command to define the fate and follow-up variables.

This model can be written $\lambda(t,3) = \lambda_0(t)e^{3\beta}$and $\lambda(t,4) = \lambda_0(t)e^{4\beta}$for stage 3 and 4 patients, respectively. Hence the hazard ratio for stage **4** patients **relative** to stage **3** patients is

$$\frac{\lambda(t,4)}{\lambda(t,3)} = \frac{\lambda_0(t)e^{4\beta}}{\lambda_0(t)e^{3\beta}} = e^{4\beta-3\beta} = e^{\beta}$$

which we interpret as the **relative risk** of death for stage **4** patients compared to stage **3** patients. Note that we could have redefined stage to be an indicator variable that equals 1 for stage 4 patients and 0 for stage 3 patients. Had we done that, the hazard for stage 3 and 4 patients would have been $\lambda_0(t)$ and $\lambda_0(t)e^{\beta}$ respectively. The **hazard ratio,** however, would still be e^{β}

{2} This **hazard ratio** or **relative risk** equals 2.61 and is significantly different from zero (P=0.013)



stcox - Fit Cox proportional hazards model

Independent variables: stage

☐ Fit model without covariates

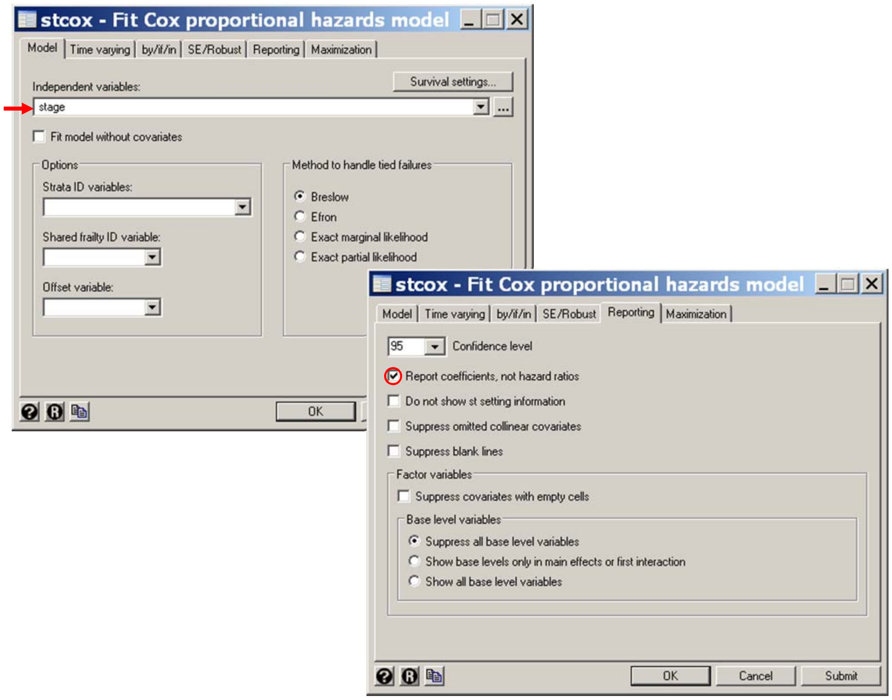
Method to handle tied failures

● Breslow
○ Efron
○ Exact marginal likelihood
○ Exact partial likelihood

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox stage,nohr                                                          {3}

         failure _d:  fate
   analysis time _t:  time

  Iteration 0:  Log Likelihood = -207.5548
  Iteration 1:  Log Likelihood =-203.86666
  Iteration 2:  Log Likelihood =-203.73805
  Iteration 3:  Log Likelihood =-203.73761
  Refining estimates:
  Iteration 0:  Log Likelihood =-203.73761

Cox regression -- Breslow method for ties

No. of subjects =          80                     Number of obs   =        80
No. of failures =          54
Time at risk    =        9718
                                                  LR chi2(1)      =      7.63
Log likelihood  =   -203.73761                    Prob > chi2     =    0.0057

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   .9610202   .3856356     2.49   0.013     .2051884    1.716852 {4}
------------------------------------------------------------------------------
```

{3} It is often useful to obtain direct estimates of the parameters of a hazard regression model. We do this with the ***nohr*** option, which stands for *no hazards ratios.*

{4} The estimate of β is 0.961. Note that exp(0.961) = 2.61, the hazard ratio obtained previously.



stcox - Fit Cox proportional hazards model

Independent variables: stage

Method to handle tied failures: ● Breslow

stcox - Fit Cox proportional hazards model

95 Confidence level

☑ Report coefficients, not hazard ratios

Model
Time varying
by/if/in
SE/Robust
Reporting
Maximization
Survival settings...
□ Fit model without covariates
Options
Strata ID variables:
Shared frailty ID variable:
Offset variable:
○ Efron
○ Exact marginal likelihood
○ Exact partial likelihood
OK
□ Do not show st setting information
□ Suppress omitted collinear covariates
□ Suppress blank lines
Factor variables
□ Suppress covariates with empty cells
Base level variables
● Suppress all base level variables
○ Show base levels only in main effects or first interaction
○ Show all base level variables
OK
Cancel
Submit

c) Estimating relative risks together with their 95% confidence intervals

The mortal risk of stage 4 lymphoma patients relative to stage 3 patients is exp(0.9610) = 2.61.

The **95% confidence interval** for this risk is

(2.61exp(-1.96*0.3856), 2.61exp(1.96*0.3856))

= (1.2, 5.6).

Note that Stata gave us this confidence interval when we did not specify the *nohr* option.

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   .9610202   .3856356     2.492   0.013      .2051884    1.716852
------------------------------------------------------------------------------


------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   2.614362   1.008191     2.492   0.013      1.227756    5.566976
------------------------------------------------------------------------------
```

d) Tied failure times

The most straight forward computational approach to the proportional hazards model can produce biased parameter estimates if a large proportion of the failure times are identical**.** For this reason it is best to record failure times as precisely as possible to avoid ties in this variable.

If there are extensive ties in the data, the *exactm, exactp,* or *efron* options of the *stcox* commands may be used to reduce this bias.

exactm and *exactp* are the most accurate, but can be computationally intensive.

An alternate approach is to use Poisson regression, which will be discussed in Chapters 7 and 8.

10. What we have covered

- ❖Survival data: time to event
  - ➢Right censored data
- ❖Kaplan-Meier survival curves: the ***sts graph*** command
- ❖Kaplan-Meier cumulative mortality curves: the ***failure*** option
  - ➢Greenwood confidence bands for survival and mortality curves
    the ***ci*** option
  - ➢Displaying censoring times
    the ***censored(single)*** option
  - ➢Displaying numbers of patients at risk
    the ***risktable*** option
- ❖Estimating survival probabilities: the ***sts list*** command
- ❖Censoring and biased Kaplan-Meier survival curves
- ❖Log rank test for comparing survival curves: the ***sts test*** command
- ❖Hazard functions and cumulative mortality
  - ➢Hazard rate ratios and relative risk
  - ➢Estimating relative risks from proportional hazards models
- ❖Simple proportional hazards regression model: the ***stcox*** command
- ❖Tied failure times and biased relative risk estimates

Cited References

Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research.* Malden MA: Blackwell Science, Inc. 2002.

McKelvey EM, Gottlieb JA, Wilson HE, Haut A, Talley RW, Stephens R, Lane M, Gamble JF, Jones SE, Grozea PN, Gutterman J, Coltman C, Moon TE. Hydroxyldaunomycin (Adriamycin) combination chemotherapy in malignant lymphoma. *Cancer* 1976;38:1484-93.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.