SOC 7717 Event History Analysis and Sequence Analysis

Week 4: Prepare Survival Data for Analysis in Stata
Wen Fan

Spring 2019

I. Overview

In order to analyze survival data it is necessary to specify at least (1) a variable representing survival time (or analysis time) and (2) a variable specifying whether or not the event of interest was observed (called the failure or the censoring variable). Depending on variable availability, sometimes, instead of specifying a variable representing survival time, we can specify the entry and exit dates. This is necessary if subjects enter the study at different times.

In many statistical software programs (such as SAS), these variables must be specified every time a new analysis is performed. In Stata, by contrast, these variables are specified once using the `stset` command and then used for all subsequent survival analysis (`st`) commands (until the next `stset` command).

Stata's `st` (survival time) suite of commands provide sophisticated tools for survival analysis. In short, with continuous survival time data, once you have `stset` them—declared the variables summarizing the spell length and failure/censoring status—you can go straight ahead and summarize and analyze the data without referring to those key variables again. It should be noted, however, that the Stata `st` suite is designed with an emphasis on analysis of continuous survival time data. Although discrete (grouped duration) data may be usefully summarized using `st` tools, estimation of discrete time hazard models is typically done outside this framework.

II. Assumptions about the original data structure

For now we assume the data are structured in such a way that there is a single record per "subject". And there are no complications arising from left censoring, gaps, left truncation, or multiple events, etc. (These complications can also be handled using Stata's `st` suite though. See the textbook for details.) There are no missing values for simplicity. And the data do not need to be weighted.

We shall also assume that variables related to survival time and censoring already exist. Depending on your particular research question, you may have to generate your own survival time and censoring variables in practice.

Finally, for now we assume that there are no time-varying covariates. In this case, all the explanatory variables in our regressions have a fixed value for each subject.

III. Using `stset`: An example

We use the `hrs` data as an example. This data set contains a sample of respondents from the Health and Retirement Study, with 33,918 observations and 17 variables.

```
. use "hrs.dta", clear
. describe
Contains data from hrs.dta
  obs:        33,918
 vars:            17                          4 Feb 2019 18:47
 size:     1,187,130
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
hhidpn          long    %12.0g                hhidpn: hhold id + person number
                                                /num
deathw12        float   %9.0g                 R dead by wave 12
deathyr         float   %9.0g                 R year of death
firstinyr       float   %9.0g                 first interview year
bpw2            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 2
bpw3            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 3
bpw4            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 4
bpw5            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 5
bpw6            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 6
bpw7            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 7
bpw8            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 8
bpw9            byte    %9.0g                 had high blood pressure since
                                                last wave, wave 9
bpw10           byte    %9.0g                 had high blood pressure since
                                                last wave, wave 10
bpw11           byte    %9.0g                 had high blood pressure since
                                                last wave, wave 11
bpw12           byte    %9.0g                 had high blood pressure since
                                                last wave, wave 12
```

```
byear           float   %9.0g                 year of birth
female          float   %9.0g                 female
-----------------------------------------------------------------------------
Sorted by: hhidpn
```

Let's first take a look at a few observations in the data, and the distribution of the failure variable.

```
. list in 10/20, linesize(75) compress

     +------------------------------------------------------------------+
 10. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10013040 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1947 |     1 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 11. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10038010 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1936 |     0 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 12. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10038040 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1943 |     1 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 13. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10050010 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1941 |     1 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 14. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10059020 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1935 |     1 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 15. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10059030 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1928 |     0 |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
 16. |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
     | 10063010 |     0 |  2015 |  1992 |    0 |    0 |    . |    . |
     |----------+-------+-------+-------+------+------+------+------|
     | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
     |    . |    0 |    0 |    . |     0 |     0 |     0 |  1938 |     1 |
     +------------------------------------------------------------------+
```

```
17. +-----------------------------------------------------------------------+
    |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
    | 10075020 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
    |-----------------------------------------------------------------------|
    | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
    |    0 |    0 |    0 |    0 |     0 |     0 |     0 |  1937 |     1 |
    +-----------------------------------------------------------------------+

18. +-----------------------------------------------------------------------+
    |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
    | 10075030 |     1 |  2007 |  1996 |    . |    . |    0 |    0 |
    |-----------------------------------------------------------------------|
    | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
    |    0 |    0 |    0 |    . |     . |     . |     . |  1927 |     0 |
    +-----------------------------------------------------------------------+

19. +-----------------------------------------------------------------------+
    |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
    | 10083020 |     0 |  2015 |  1992 |    0 |    0 |    0 |    0 |
    |-----------------------------------------------------------------------|
    | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
    |    0 |    0 |    0 |    0 |     0 |     0 |     . |  1941 |     1 |
    +-----------------------------------------------------------------------+

20. +-----------------------------------------------------------------------+
    |   hhidpn | de~12 | dea~r | fir~r | bpw2 | bpw3 | bpw4 | bpw5 |
    | 10090010 |     1 |  1994 |  1992 |    0 |    . |    . |    . |
    |-----------------------------------------------------------------------|
    | bpw6 | bpw7 | bpw8 | bpw9 | bpw10 | bpw11 | bpw12 | byear | fem~e |
    |    . |    . |    . |    . |     . |     . |     . |  1934 |     1 |
    +-----------------------------------------------------------------------+

. tab deathw12, m

  R dead by |
    wave 12 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     23,202       68.41       68.41
          1 |     10,716       31.59      100.00
------------+-----------------------------------
      Total |     33,918      100.00
```

We see that out of the 33,918 respondents in the data, about 68% are censored whereas the other 32% had experienced the event (i.e., death) by the twelfth wave.

Treating year of death as a continuous variable, we `stset` the data:

```
. stset deathyr, failure(deathw12) id(hhidpn) origin(time byear) enter(time fir
> stinyr)

                id:  hhidpn
     failure event:  deathw12 != 0 & deathw12 < .
obs. time interval:  (deathyr[_n-1], deathyr]
 enter on or after:  time firstinyr
 exit on or before:  failure
    t for analysis:  (time-origin)
            origin:  time byear

------------------------------------------------------------------------------
      33918  total observations
          0  exclusions
------------------------------------------------------------------------------
      33918  observations remaining, representing
      33918  subjects
```

```
     10716  failures in single-failure-per-subject data
    441292  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =        18
                                          last observed exit t =       115
```

As will become clear soon, this setup says that each individual enters the study (becomes "at risk") at the date specified by `firstinyr`. Here we use attained age as the clock, so the origin is year of birth (`byear`). To use calendar time as the clock, we can specify a fixed date as the time origin (e.g., `origin(1980)`).

1. Specifying analysis time (`origin()` and `scale()`)

Analysis time can be obtained in two ways:

- Construct the analysis time yourself and then `stset` the data;
- Specify `stset`'s `origin()` and/or `scale()` options and then `stset` the data.

If you go with the second approach, analysis time is calculated by Stata as $t = \frac{time-origin}{scale}$. In our case, the analysis time is in fact age, rather than calendar year—it does not make much sense to think that people living in the same calendar year have the same risk of dying, but it is reasonable to assume that people with the same age have roughly similar risk of dying. Therefore, we use `origin()` to define year of birth as the origin. See below for more discussion on the choice of origin. The scale is just the units in which time is measured. In our example, we want to talk about survival in years, so we keep it as it is. But suppose that we want to evaluate survival risks in decades, we could specify `scale(10)`.

Some other ways to specify `origin()` include: `origin(time 20)`, `origin(time td(15feb2018)`, `origin(time min(diagdate1, diagdate2)`, `origin(event == 3 4)`. If the time-related variables in your data are recorded using Stata's date format, please read section 6.1 of the textbook carefully.

2. Specifying failure (`failure()`)

In the `stset` command, `failure()` specifies the failure event. Note that there is a failure whenever the variable in the parenthesis is not equal to zero and not missing. Also note that if the failure variable contains missing values, it is treated as if it contains 0. Alternatively use the syntax `failure(failvar == numlist)` in which case the failure cases are those with `failvar` equal to `numlist`. In other words, we could have specified `stset deathyr, failure(deathw12 == 1) id(hhidpn) origin(time byear) enter(time firstinyr)` to get the same results. If `failure(.)` is not specified, every record is assumed to end in a failure.

3. Specifying the subject-ID variables (`id()`)

If there are multiple records per subject in your data ("long format"), you must use the `id()` option to specify a subject-ID variable. Even in single-record data ("wide format"), as in our HRS example, it is still recommended to specify an ID variable.

4. Specifying when subjects enter the analysis (`enter()`)

Ideally we want subjects to enter our analysis at the onset of risk, but sometimes our data may contain records reporting values before the subject was really under observation. In the HRS example, respondents entered the survey in different calendar years. To incorporate that information, we add `enter(time firstinyr)`. There are other ways to specify `enter()` (see the 6.4.5 section of the textbook for more details).

5. Specifying when subjects exit from the analysis (`exit()`)

If you don't specify `exit()`, Stata assumes a subject exits either when the subject's data run out or upon first failure. There are situations in which you may want subjects to exit earlier or later. Read the 6.3.4 section in our textbook carefully if that's the case.

IV. After `stset`

After `stset`, we see that a set of new variables are created in the data, all prefaced with "_". These always have the same names: that's how Stata's survival time estimation commands are able to work, because it knows that, if the data have been `stset`, then the key variables (duration, failure indicator, etc.) are available. Here are the variables:

```
. sum _*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_st	33,918	1	0	1	1
_d	33,918	.3159384	.4648954	0	1
_origin	33,918	1938.15	15.20826	1890	1995
_t	33,918	72.98458	11.98638	20	115
_t0	33,918	59.97403	11.21397	18	103

The `_st` variable is a 0/1 variable, equal to 1 for observations whose data have been `stset` (it would be zero if one had excluded some cases with an `if` qualifier, for instance). The `_d` variable is the failure indicator, another 0/1 variable, and corresponds to the variable `deathw12` in this case. Finally, `_t0` and `_t` are variables recording the time span for each case. Each record starts at `_t0` and concludes at `_t`. This dataset has delayed entry, so the entry times are different. If, however, all cases entered at the same time, then `_t0 = 0`. In our example, we also have a `_origin` variable to denote the time of origin, but note that unlike the other four, this variable is not generated by Stata after every `stset`—it is created when and only when we specify `origin()` when `stset` out data.

Typing `st` by itself shows how the data are currently set.

That there are 10,716 failures, as the output from the `stset` command says, can be easily verified using `tab deathw12`. Note that, however, this equality does not have to be the case, especially when there are tied events. The `441292` in the `stset` output refers to the total number of time periods for which this sample was observed at risk of failure since time $t = 0$: the sum of study time across all persons. You can get almost all this information more directly using `stsum`:

```
. stsum
        failure _d:  deathw12
  analysis time _t:  (deathyr-origin)
            origin:  time byear
 enter on or after:  time firstinyr
                id:  hhidpn

         |               incidence       no. of     |------ Survival time -----
         | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
   total |       441292   .0242832         33918        75        84        91
```

The incidence rate, 0.024 = 10716/441292. See also `stdes`, which comes into its own for description of more complicated survival data structures.

```
. stdescribe
        failure _d:  deathw12
  analysis time _t:  (deathyr-origin)
            origin:  time byear
 enter on or after:  time firstinyr
                id:  hhidpn

                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
------------------------------------------------------------------------------

no. of subjects            33918
no. of records             33918           1           1          1          1

(first) entry time                  59.97403          18         56        103
(final) exit time                   72.98458          20         74        115

subjects with gap              0
time on gap if gap             0           .           .          .          .
time at risk              441292    13.01055           1         11         23

failures                   10716    .3159384           0          0          1
------------------------------------------------------------------------------
```

The `stptime` command tabulates the number of events and person time-at risk and calculates event rates:

```
. stptime, by(female) per(1000)
```

```
       failure _d:  deathw12
 analysis time _t:  (deathyr-origin)
           origin:  time byear
enter on or after:  time firstinyr
               id:  hhidpn
```

female	person-time	failures	rate	[95% Conf.	Interval]
0	180972	5079	28.065115	27.30379	28.84766
1	260320	5637	21.654118	21.09615	22.22684
total	441292	10716	24.283241	23.8278	24.74739

Note that you need a subject-ID variable to use the `stptime` command. ALso note that person-time is in years but the rates are per 1000 person-years.

The `strate` command performs similar calculations.

```
. strate female, per(1000)
       failure _d:  deathw12
 analysis time _t:  (deathyr-origin)
           origin:  time byear
enter on or after:  time firstinyr
               id:  hhidpn
Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(33918 records included in the analysis)
```

female	D	Y	Rate	Lower	Upper
0	5079	180.9720	28.065	27.304	28.848
1	5637	260.3200	21.654	21.096	22.227

The incidence rate ratio (IRR) for men versus women is 28.065/21.654 = 1.30. That is, without controlling for any possible confounding factors, we estimate that men's risk of dying is approximately 30% higher compared with women's risk of dying. This is sometimes called a "crude estimate"; it is not adjusted for potential confounders.

Some of the Stata survival analysis (`st`) commands are given below. Further details can be found in the manuals or online help.

Command	**Function**
`stset`	Declare data to be survival-time data
`stdes`	Describe survival-time data
`stsum`	Summarize survival-time data
`stsplit`	Split time-span records
`sts`	Generate, graph, list, and test the survivor and cumulative hazard functions
`strate`	Tabulate failure rate
`stptime`	Calculate person-time at risk and failure rates
`stcox`	Estimate Cox proportional hazards model
`stphtest`	Test of Cox proportional hazards assumption

Command	Function
`stphplot`	Graphical assessment of the Cox proportional hazards assumption
`stcoxkm`	Graphical assessment of the Cox proportional hazards assumption
`streg`	Estimate parametric survival models

Once the data have been `stset` we can use any of these commands without having to specify the survival time or failure time variables.

V. The origin of time

There are several time dimensions along which rates might vary. These differ from one another only in the choice of time origin, the point at which time is zero.

This naturally leads to the question: In which units should we specify time? Could different units have been used? Determining the origin of time is important, because:

- It does make a difference, often substantial, in coefficient estimates and fit of the models.
- The preferred origin is sometimes unavailable, in which case you must use some proxy.
- Many situations occur in which two or more possible time origins are available, but there is no unambiguous criterion for deciding among them.

Some commonly used time scales are listed below:

Origin	Time Clock
Birth	Age
Any fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Disease onset	Time since onset
Diagnosis	Time since diagnosis
Start of treatment	Time on treatment

Some suggestions offered by Allison (2010) in choosing the principal time origin:

- Choose a time origin that marks the onset of continuous exposure to risk of the event.
- In experimental studies, choose the time of randomization to treatment as

the time origins.

- Choose the time origin that has the strongest effect on the hazard (e.g., age vs. time since diagnosis).

VI. Nonparametric estimates

1. The life table method

To obtain life table estimation, we use the `ltable` command, which is not an `st` command. Thus, the data do not have to be `stset` to use this command but we need to specify the survival time and failure variable every time we use the command.

```
. gen deathage = deathyr - byear
. ltable deathage deathw12
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
20	21	33918	0	1	1.0000	0.0000	.	.
23	24	33917	0	2	1.0000	0.0000	.	.
27	28	33915	0	1	1.0000	0.0000	.	.
30	31	33914	0	2	1.0000	0.0000	.	.
31	32	33912	0	3	1.0000	0.0000	.	.
32	33	33909	0	3	1.0000	0.0000	.	.
33	34	33906	0	4	1.0000	0.0000	.	.
34	35	33902	0	4	1.0000	0.0000	.	.
35	36	33898	0	5	1.0000	0.0000	.	.
36	37	33893	0	8	1.0000	0.0000	.	.
37	38	33885	0	6	1.0000	0.0000	.	.
38	39	33879	1	9	1.0000	0.0000	0.9998	1.0000
39	40	33869	0	8	1.0000	0.0000	0.9998	1.0000
40	41	33861	0	9	1.0000	0.0000	0.9998	1.0000
41	42	33852	1	15	0.9999	0.0000	0.9998	1.0000
42	43	33836	1	23	0.9999	0.0001	0.9997	1.0000
43	44	33812	0	29	0.9999	0.0001	0.9997	1.0000
44	45	33783	1	21	0.9999	0.0001	0.9997	1.0000
45	46	33761	2	37	0.9998	0.0001	0.9996	0.9999
46	47	33722	1	34	0.9998	0.0001	0.9996	0.9999
47	48	33687	3	47	0.9997	0.0001	0.9994	0.9998
48	49	33637	3	63	0.9996	0.0001	0.9993	0.9998
49	50	33571	5	73	0.9995	0.0001	0.9992	0.9997
50	51	33493	4	85	0.9993	0.0001	0.9990	0.9996
51	52	33404	7	117	0.9991	0.0002	0.9988	0.9994
52	53	33280	16	147	0.9987	0.0002	0.9982	0.9990
53	54	33117	25	163	0.9979	0.0003	0.9973	0.9983
54	55	32929	43	211	0.9966	0.0003	0.9959	0.9972
55	56	32675	58	285	0.9948	0.0004	0.9940	0.9955
56	57	32332	58	820	0.9930	0.0005	0.9920	0.9939
57	58	31454	71	866	0.9907	0.0005	0.9896	0.9917
58	59	30517	96	824	0.9876	0.0006	0.9863	0.9887
59	60	29597	97	829	0.9843	0.0007	0.9829	0.9856
60	61	28671	104	810	0.9807	0.0008	0.9791	0.9822
61	62	27757	131	790	0.9760	0.0009	0.9742	0.9776
62	63	26836	121	752	0.9715	0.0010	0.9696	0.9733

63	64	25963	143	813	0.9661	0.0011	0.9639	0.9681
64	65	25007	141	709	0.9606	0.0012	0.9582	0.9628
65	66	24157	184	748	0.9531	0.0013	0.9506	0.9555
66	67	23225	145	737	0.9471	0.0014	0.9443	0.9497
67	68	22343	206	679	0.9382	0.0015	0.9352	0.9410
68	69	21458	197	557	0.9295	0.0016	0.9263	0.9325
69	70	20704	188	544	0.9209	0.0017	0.9175	0.9242
70	71	19972	181	441	0.9125	0.0018	0.9089	0.9159
71	72	19350	229	464	0.9016	0.0019	0.8978	0.9052
72	73	18657	258	526	0.8889	0.0020	0.8849	0.8928
73	74	17873	285	574	0.8745	0.0022	0.8702	0.8787
74	75	17014	285	793	0.8595	0.0023	0.8549	0.8640
75	76	15936	301	760	0.8429	0.0025	0.8380	0.8476
76	77	14875	291	720	0.8260	0.0026	0.8208	0.8310
77	78	13864	343	704	0.8050	0.0028	0.7995	0.8104
78	79	12817	337	690	0.7833	0.0029	0.7774	0.7890
79	80	11790	352	614	0.7593	0.0031	0.7531	0.7653
80	81	10824	372	592	0.7324	0.0033	0.7259	0.7388
81	82	9860	349	558	0.7057	0.0035	0.6989	0.7125
82	83	8953	414	454	0.6723	0.0037	0.6650	0.6794
83	84	8085	367	466	0.6408	0.0039	0.6332	0.6483
84	85	7252	395	451	0.6048	0.0040	0.5968	0.6127
85	86	6406	396	354	0.5664	0.0042	0.5580	0.5746
86	87	5656	372	284	0.5282	0.0044	0.5195	0.5367
87	88	5000	389	292	0.4858	0.0045	0.4769	0.4947
88	89	4319	376	257	0.4422	0.0046	0.4331	0.4513
89	90	3686	355	240	0.3982	0.0047	0.3889	0.4075
90	91	3091	351	179	0.3516	0.0048	0.3423	0.3610
91	92	2561	298	173	0.3093	0.0048	0.2999	0.3187
92	93	2090	257	141	0.2699	0.0048	0.2606	0.2793
93	94	1692	266	132	0.2258	0.0047	0.2166	0.2350
94	95	1294	160	107	0.1967	0.0046	0.1877	0.2058
95	96	1027	162	96	0.1641	0.0045	0.1554	0.1731
96	97	769	145	77	0.1315	0.0044	0.1232	0.1402
97	98	547	96	45	0.1075	0.0042	0.0994	0.1158
98	99	406	86	38	0.0836	0.0040	0.0760	0.0916
99	100	282	61	32	0.0644	0.0037	0.0573	0.0720
100	101	189	44	19	0.0486	0.0035	0.0421	0.0558
101	102	126	29	10	0.0370	0.0033	0.0310	0.0438
102	103	87	26	10	0.0252	0.0029	0.0200	0.0315
103	104	51	16	7	0.0167	0.0026	0.0122	0.0225
104	105	28	7	6	0.0121	0.0024	0.0080	0.0175
105	106	15	6	0	0.0072	0.0021	0.0040	0.0124
106	107	9	1	1	0.0064	0.0020	0.0033	0.0114
107	108	7	2	0	0.0046	0.0018	0.0020	0.0094
108	109	5	1	0	0.0036	0.0017	0.0014	0.0083
109	110	4	1	0	0.0027	0.0015	0.0009	0.0072
112	113	3	1	0	0.0018	0.0012	0.0004	0.0060
115	116	2	0	2	0.0018	0.0012	0.0004	0.0060

Many of the differences between `sts` and `ltable` derive from the underlying assumptions about the nature of the survival time data. With `sts`, survival times are treated as observations on a continuous variable. In the `ltable` case, the technique is based on survival data that have been grouped into intervals (or implicitly assumed to be).

2. The Kaplan-Meier method

Estimation of the Kaplan-Meier empirical hazard and survival functions is done very easily in Stata by using either the `sts` collection of commands or by using the `ltable` (life table) command omitting the `intervals` option and including the `noadjust` option.

For example, after `stset` the data,

```
. sts list

         failure _d:  deathw12
   analysis time _t:  (deathyr-origin)
             origin:  time byear
  enter on or after:  time firstinyr
                 id:  hhidpn

           Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
    18        0      0     -1             1.0000         .          .         .
    19        1      0     -1             1.0000         .          .         .
    20        2      0      1             1.0000         .          .         .
    22        1      0     -2             1.0000         .          .         .
    24        3      0     -3             1.0000         .          .         .
    25        6      0     -7             1.0000         .          .         .
    26       13      0     -4             1.0000         .          .         .
    27       17      0     -2             1.0000         .          .         .
    28       19      0     -7             1.0000         .          .         .
    29       26      0     -9             1.0000         .          .         .
    30       35      0     -9             1.0000         .          .         .
    31       44      0    -14             1.0000         .          .         .
    32       58      0    -17             1.0000         .          .         .
    33       75      0    -16             1.0000         .          .         .
    34       91      0    -23             1.0000         .          .         .
    35      114      0    -25             1.0000         .          .         .
    36      139      0    -40             1.0000         .          .         .
    37      179      0    -38             1.0000         .          .         .
    38      217      1    -66             0.9954    0.0046     0.9677    0.9993
    39      282      0    -70             0.9954    0.0046     0.9677    0.9993
    40      352      0    -81             0.9954    0.0046     0.9677    0.9993
    41      433      1    -77             0.9931    0.0051     0.9706    0.9984
    42      509      1   -112             0.9911    0.0055     0.9704    0.9974
    43      620      0   -144             0.9911    0.0055     0.9704    0.9974
    44      764      1   -196             0.9898    0.0056     0.9701    0.9966
    45      959      2   -227             0.9878    0.0058     0.9692    0.9952
    46     1184      1   -296             0.9869    0.0059     0.9687    0.9946
    47     1479      3   -343             0.9849    0.0060     0.9674    0.9931
    48     1819      3   -415             0.9833    0.0060     0.9663    0.9918
    49     2231      5   -556             0.9811    0.0061     0.9646    0.9900
    50     2782      4   -712             0.9797    0.0061     0.9634    0.9888
    51     3490      7  -2406             0.9777    0.0061     0.9618    0.9871
    52     5889     16  -2437             0.9751    0.0062     0.9596    0.9847
    53     8310     25  -2196             0.9722    0.0062     0.9571    0.9820
    54    10481     43  -2096             0.9682    0.0062     0.9535    0.9783
    55    12534     58  -1943             0.9637    0.0062     0.9494    0.9740
    56    14419     58  -1183             0.9598    0.0062     0.9457    0.9703
    57    15544     71   -190             0.9554    0.0062     0.9416    0.9660
    58    15663     96   -253             0.9496    0.0062     0.9360    0.9603
    59    15820     97    -75             0.9437    0.0062     0.9303    0.9546
```

60	15798	104	-159	0.9375	0.0061	0.9243	0.9485
61	15853	131	-93	0.9298	0.0061	0.9167	0.9409
62	15815	121	315	0.9227	0.0061	0.9098	0.9338
63	15379	143	517	0.9141	0.0061	0.9013	0.9253
64	14719	141	447	0.9053	0.0061	0.8927	0.9166
65	14131	184	481	0.8935	0.0061	0.8810	0.9048
66	13466	145	533	0.8839	0.0061	0.8715	0.8952
67	12788	206	448	0.8697	0.0060	0.8573	0.8810
68	12134	197	45	0.8556	0.0060	0.8433	0.8669
69	11892	188	45	0.8420	0.0060	0.8299	0.8534
70	11659	181	-354	0.8290	0.0060	0.8169	0.8404
71	11832	229	-420	0.8129	0.0060	0.8009	0.8243
72	12023	258	-307	0.7955	0.0059	0.7835	0.8068
73	12072	285	-176	0.7767	0.0059	0.7649	0.7880
74	11963	285	62	0.7582	0.0059	0.7465	0.7695
75	11616	301	301	0.7385	0.0058	0.7269	0.7498
76	11014	291	318	0.7190	0.0058	0.7075	0.7302
77	10405	343	325	0.6953	0.0057	0.6839	0.7064
78	9737	337	351	0.6713	0.0057	0.6600	0.6823
79	9049	352	279	0.6452	0.0056	0.6340	0.6561
80	8418	372	234	0.6166	0.0056	0.6056	0.6275
81	7812	349	223	0.5891	0.0055	0.5782	0.5998
82	7240	414	177	0.5554	0.0054	0.5447	0.5660
83	6649	367	226	0.5248	0.0054	0.5142	0.5352
84	6056	395	213	0.4905	0.0053	0.4801	0.5008
85	5448	396	155	0.4549	0.0052	0.4447	0.4650
86	4897	372	123	0.4203	0.0051	0.4103	0.4303
87	4402	389	164	0.3832	0.0050	0.3734	0.3929
88	3849	376	149	0.3457	0.0049	0.3362	0.3553
89	3324	355	153	0.3088	0.0047	0.2996	0.3181
90	2816	351	101	0.2703	0.0046	0.2614	0.2793
91	2364	298	135	0.2362	0.0044	0.2277	0.2449
92	1931	257	106	0.2048	0.0042	0.1966	0.2131
93	1568	266	97	0.1701	0.0040	0.1623	0.1780
94	1205	160	86	0.1475	0.0039	0.1400	0.1551
95	959	162	79	0.1226	0.0037	0.1155	0.1299
96	718	145	58	0.0978	0.0035	0.0912	0.1047
97	515	96	38	0.0796	0.0033	0.0733	0.0862
98	381	86	26	0.0616	0.0031	0.0558	0.0678
99	269	61	28	0.0476	0.0028	0.0423	0.0534
100	180	44	15	0.0360	0.0026	0.0311	0.0414
101	121	29	7	0.0274	0.0024	0.0229	0.0325
102	85	26	10	0.0190	0.0022	0.0151	0.0236
103	49	16	5	0.0128	0.0019	0.0094	0.0171
104	28	7	6	0.0096	0.0018	0.0066	0.0136
105	15	6	0	0.0058	0.0016	0.0032	0.0097
106	9	1	1	0.0051	0.0016	0.0027	0.0090
107	7	2	0	0.0037	0.0014	0.0016	0.0074
108	5	1	0	0.0029	0.0013	0.0011	0.0066
109	4	1	0	0.0022	0.0012	0.0007	0.0057
112	3	1	0	0.0015	0.0010	0.0003	0.0048
115	2	0	2	0.0015	0.0010	0.0003	0.0048

By specifying the `failure` option, you can list the estimate of the cumulative distribution function $F(t)$. Below we also specify `at()` to produce less detailed output (here, 5 equally spaced survival times).

```
. sts list, failure at(5)
         failure _d:  deathw12
```

```
 analysis time _t:  (deathyr-origin)
           origin:  time byear
enter on or after:  time firstinyr
               id:  hhidpn

              Beg.                         Failure      Std.
    Time     Total     Fail               Function     Error     [95% Conf. Int.]
----------------------------------------------------------------------------------
      18         0        0                 0.0000         .          .         .
      50      2782       22                 0.0203    0.0061     0.0112    0.0366
      82      7240     6028                 0.4446    0.0054     0.4340    0.4553
     114         3     4666                 0.9985    0.0010     0.9952    0.9997
     146         2        0                      .         .          .         .
----------------------------------------------------------------------------------
Note: Failure function is calculated over full data and evaluated at
      indicated times; it is not calculated from aggregates shown at left.
```

The Kaplan-Meier estimates of `S(t)` can be plotted by `sts graph`. Here we add the Greenwood standard errors

```
. sts graph, gw
          failure _d:  deathw12
    analysis time _t:  (deathyr-origin)
              origin:  time byear
   enter on or after:  time firstinyr
                  id:  hhidpn

. graph export km1.png, width(500) replace
(file km1.png written in PNG format)
```

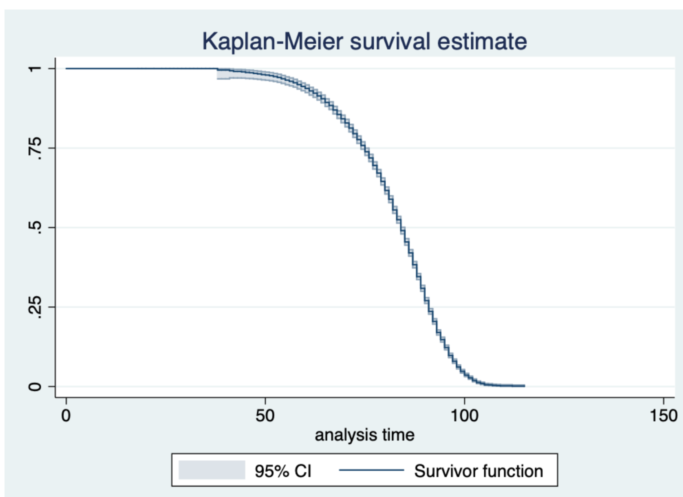

Figure 1: Kaplan-Meier Estimates of S(t)

If you want the graph to be plotted separately by, for instance, gender, add the `by()` option. We also add the `censored` option to display the number of censored observations.

```
. sts graph, by(female) censored(number)
```

```
        failure _d:  deathw12
  analysis time _t:  (deathyr-origin)
            origin:  time byear
 enter on or after:  time firstinyr
                id:  hhidpn
```

```
. graph export km2.png, width(500) replace
(file km2.png written in PNG format)
```

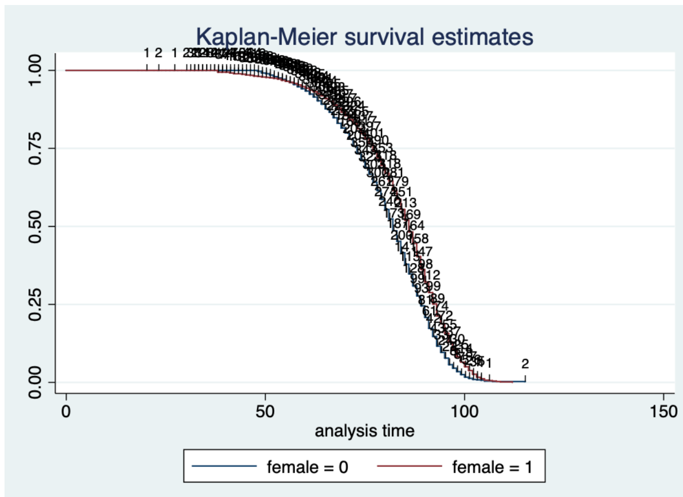

Figure 2: Kaplan-Meier Estimates of S(t), by Gender

3. The Nelson-Aalen estimator

```
. sts list, cumhaz

        failure _d:  deathw12
  analysis time _t:  (deathyr-origin)
            origin:  time byear
 enter on or after:  time firstinyr
                id:  hhidpn
```

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf. Int.]	
18	0	0	-1	0.0000	0.0000	.	.
19	1	0	-1	0.0000	0.0000	.	.
20	2	0	1	0.0000	0.0000	.	.
22	1	0	-2	0.0000	0.0000	.	.
24	3	0	-3	0.0000	0.0000	.	.
25	6	0	-7	0.0000	0.0000	.	.
26	13	0	-4	0.0000	0.0000	.	.
27	17	0	-2	0.0000	0.0000	.	.
28	19	0	-7	0.0000	0.0000	.	.
29	26	0	-9	0.0000	0.0000	.	.
30	35	0	-9	0.0000	0.0000	.	.
31	44	0	-14	0.0000	0.0000	.	.
32	58	0	-17	0.0000	0.0000	.	.

33	75	0	−16	0.0000	0.0000	.	.
34	91	0	−23	0.0000	0.0000	.	.
35	114	0	−25	0.0000	0.0000	.	.
36	139	0	−40	0.0000	0.0000	.	.
37	179	0	−38	0.0000	0.0000	.	.
38	217	1	−66	0.0046	0.0046	0.0006	0.0327
39	282	0	−70	0.0046	0.0046	0.0006	0.0327
40	352	0	−81	0.0046	0.0046	0.0006	0.0327
41	433	1	−77	0.0069	0.0052	0.0016	0.0298
42	509	1	−112	0.0089	0.0055	0.0026	0.0300
43	620	0	−144	0.0089	0.0055	0.0026	0.0300
44	764	1	−196	0.0102	0.0057	0.0034	0.0303
45	959	2	−227	0.0123	0.0059	0.0048	0.0313
46	1184	1	−296	0.0131	0.0059	0.0054	0.0318
47	1479	3	−343	0.0151	0.0060	0.0069	0.0331
48	1819	3	−415	0.0168	0.0061	0.0082	0.0343
49	2231	5	−556	0.0190	0.0062	0.0101	0.0360
50	2782	4	−712	0.0205	0.0062	0.0113	0.0372
51	3490	7	−2406	0.0225	0.0063	0.0130	0.0389
52	5889	16	−2437	0.0252	0.0063	0.0154	0.0412
53	8310	25	−2196	0.0282	0.0063	0.0182	0.0438
54	10481	43	−2096	0.0323	0.0064	0.0220	0.0476
55	12534	58	−1943	0.0369	0.0064	0.0263	0.0519
56	14419	58	−1183	0.0410	0.0064	0.0301	0.0557
57	15544	71	−190	0.0455	0.0064	0.0345	0.0601
58	15663	96	−253	0.0517	0.0065	0.0404	0.0660
59	15820	97	−75	0.0578	0.0065	0.0463	0.0721
60	15798	104	−159	0.0644	0.0065	0.0528	0.0786
61	15853	131	−93	0.0726	0.0066	0.0608	0.0867
62	15815	121	315	0.0803	0.0066	0.0683	0.0944
63	15379	143	517	0.0896	0.0067	0.0774	0.1036
64	14719	141	447	0.0992	0.0067	0.0868	0.1132
65	14131	184	481	0.1122	0.0068	0.0997	0.1263
66	13466	145	533	0.1230	0.0068	0.1103	0.1371
67	12788	206	448	0.1391	0.0069	0.1261	0.1533
68	12134	197	45	0.1553	0.0070	0.1421	0.1697
69	11892	188	45	0.1711	0.0071	0.1577	0.1856
70	11659	181	−354	0.1866	0.0072	0.1730	0.2013
71	11832	229	−420	0.2060	0.0073	0.1921	0.2209
72	12023	258	−307	0.2274	0.0074	0.2133	0.2425
73	12072	285	−176	0.2511	0.0076	0.2366	0.2663
74	11963	285	62	0.2749	0.0077	0.2602	0.2904
75	11616	301	301	0.3008	0.0078	0.2858	0.3166
76	11014	291	318	0.3272	0.0080	0.3119	0.3433
77	10405	343	325	0.3602	0.0082	0.3445	0.3766
78	9737	337	351	0.3948	0.0084	0.3786	0.4116
79	9049	352	279	0.4337	0.0087	0.4170	0.4510
80	8418	372	234	0.4779	0.0090	0.4606	0.4958
81	7812	349	223	0.5226	0.0093	0.5047	0.5410
82	7240	414	177	0.5797	0.0097	0.5610	0.5990
83	6649	367	226	0.6349	0.0101	0.6154	0.6551
84	6056	395	213	0.7002	0.0106	0.6796	0.7213
85	5448	396	155	0.7728	0.0112	0.7511	0.7952
86	4897	372	123	0.8488	0.0119	0.8258	0.8725
87	4402	389	164	0.9372	0.0127	0.9126	0.9624
88	3849	376	149	1.0349	0.0137	1.0084	1.0620
89	3324	355	153	1.1417	0.0148	1.1130	1.1711
90	2816	351	101	1.2663	0.0162	1.2349	1.2985
91	2364	298	135	1.3924	0.0178	1.3579	1.4277
92	1931	257	106	1.5255	0.0196	1.4874	1.5644
93	1568	266	97	1.6951	0.0222	1.6521	1.7392

94	1205	160	86	1.8279	0.0246	1.7803	1.8767
95	959	162	79	1.9968	0.0279	1.9428	2.0523
96	718	145	58	2.1988	0.0326	2.1358	2.2636
97	515	96	38	2.3852	0.0377	2.3123	2.4603
98	381	86	26	2.6109	0.0449	2.5243	2.7004
99	269	61	28	2.8376	0.0535	2.7348	2.9444
100	180	44	15	3.0821	0.0649	2.9574	3.2120
101	121	29	7	3.3218	0.0787	3.1710	3.4797
102	85	26	10	3.6276	0.0990	3.4387	3.8269
103	49	16	5	3.9542	0.1283	3.7105	4.2138
104	28	7	6	4.2042	0.1593	3.9032	4.5284
105	15	6	0	4.6042	0.2282	4.1780	5.0738
106	9	1	1	4.7153	0.2538	4.2432	5.2399
107	7	2	0	5.0010	0.3244	4.4040	5.6789
108	5	1	0	5.2010	0.3811	4.5053	6.0042
109	4	1	0	5.4510	0.4558	4.6271	6.4216
112	3	1	0	5.7843	0.5646	4.7771	7.0040
115	2	0	2	5.7843	0.5646	4.7771	7.0040

Similarly, you can add the `cumhaz` option in `sts graph` to plot the Nelson-Aalen cumulative hazard.

4. The hazard function

You can get a estimated hazard function by:

```
. sts graph, hazard

         failure _d:  deathw12
   analysis time _t:  (deathyr-origin)
             origin:  time byear
  enter on or after:  time firstinyr
                 id:  hhidpn

. graph export hazard.png, width(500) replace
(file hazard.png written in PNG format)
```

This graph uses the "optimal bandwidth", which is the distance on either side of each time point used in estimating the hazard. You can change the bandwidth with `width` option:

```
. sts graph, hazard width(3)

         failure _d:  deathw12
   analysis time _t:  (deathyr-origin)
             origin:  time byear
  enter on or after:  time firstinyr
                 id:  hhidpn

. graph export hazard2.png, width(500) replace
(file hazard2.png written in PNG format)
```

5. Comparison of survival between groups

To run a log rank test in Stata, use `sts test`. For example:

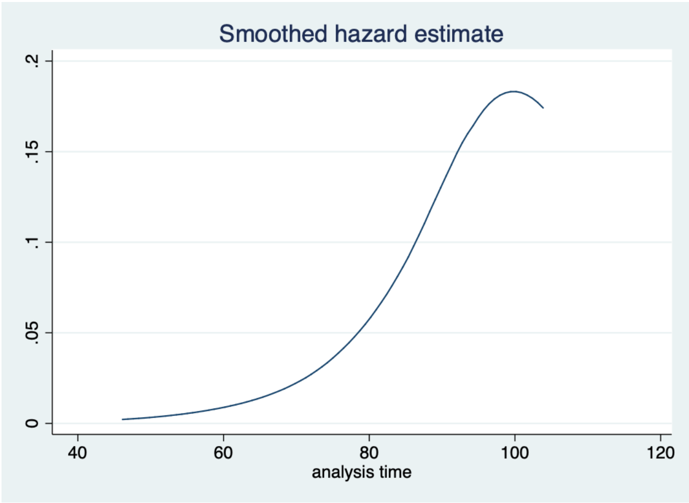

Figure 3: Hazard Plot h(t), with Optimal Bandwidth

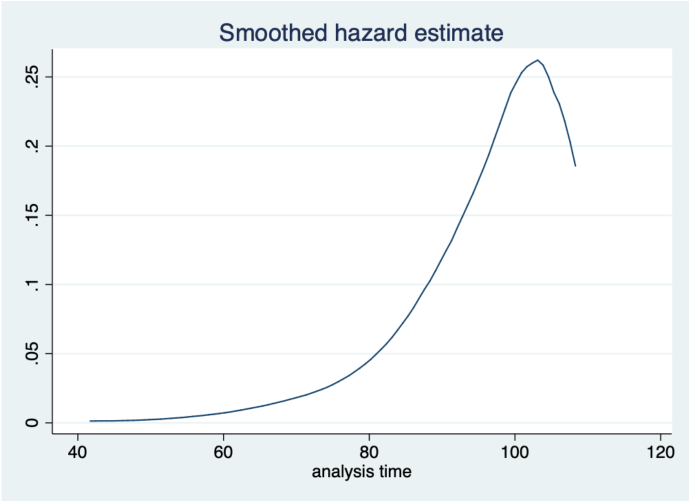

Figure 4: Hazard Plot h(t), with Alternative Bandwidth

```
. sts test female, logrank

         failure _d:  deathw12
   analysis time _t:  (deathyr-origin)
             origin:  time byear
  enter on or after:  time firstinyr
                 id:  hhidpn

Log-rank test for equality of survivor functions

       |   Events         Events
female |  observed       expected
-------+-------------------------
0      |      5079        4150.30
1      |      5637        6565.70
-------+-------------------------
Total  |     10716       10716.00

             chi2(1) =     366.67
             Pr>chi2 =     0.0000
```

To get the Wilcoxon test instead, use:

```
. sts test female, wilcoxon

         failure _d:  deathw12
   analysis time _t:  (deathyr-origin)
             origin:  time byear
  enter on or after:  time firstinyr
                 id:  hhidpn

Wilcoxon (Breslow) test for equality of survivor functions

       |   Events         Events          Sum of
female |  observed       expected          ranks
-------+----------------------------------------
0      |      5079        4150.30        8514584
1      |      5637        6565.70       -8514584
-------+----------------------------------------
Total  |     10716       10716.00              0

             chi2(1) =     340.31
             Pr>chi2 =     0.0000
```