Max Planck Odense Center on the Biodemography of Aging
University of Southern Denmark

Non-parametric Survival Analysis

Adam Lenart

MaxO - SDU

The survival function $S(t)$

The survival function at time t is the probability of surviving longer than time t or, alternatively, the probability of failing after time t.

$$S(t) = P(T > t)$$

where T is the time to event of interest (for example, time until death after cancer diagnonis).

The Kaplan-Meier estimator $\hat{S}(t)$

The Kaplan-Meier estimator is a non-parametric estimate of the survival function from the observed data, using information about censoring.

Example:
A cohort of 12 cancer patients followed up from diagnosis until death: 7 of them die, 5 drop out of the study (so their observations are right censored).

Ordered times: 5, 6*, 7, 8, 9*, 12*, 14, 15, 16, 20*, 22*, 23

The Kaplan-Meier estimator $\hat{S}(t)$

id	time	status
1	5	1
2	6	0
3	7	1
4	8	1
5	9	0
6	12	0
7	14	1
8	15	1
9	16	1
10	20	0
11	22	0
12	23	1

The Kaplan-Meier estimator $\hat{S}(t)$

S(12)=P(survive in [0,12])

=P(survive in [0,5])·

P(survive in [5,6]|survive[0,5])·

P(survive in [6,7]|survive[5,6])·

P(survive in [7,8]|survive[6,7])·

P(survive in [8,9]|survive[7,8])·

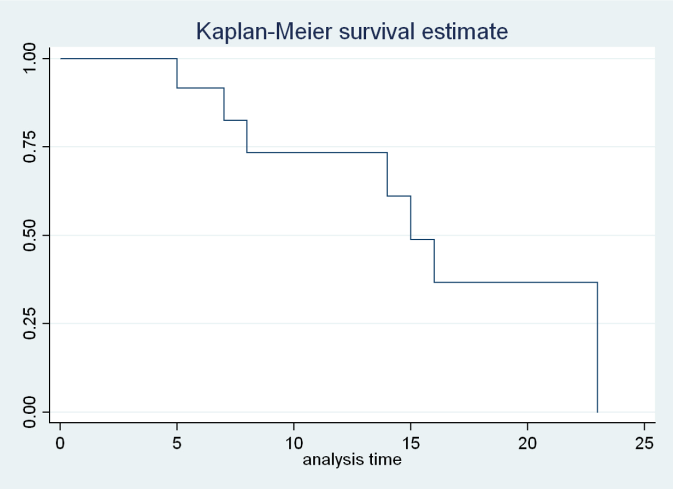
P(survive in [9,12]|survive[8,9])

Probabilities of surviving in an interval conditional on having survived until that interval.

The Kaplan-Meier estimator $\hat{S}(t)$

$(t-1,t)$	at risk	failed	censored	$p(t-1,t)$	$\hat{S}(t)$
(0,5]	12	1	0	$1-\frac{1}{12}$	$\frac{11}{12}=0.92$
(5,6]	11	0	1	1	$\frac{11}{12}\cdot 1=0.92$
(6,7]	10	1	0	$1-\frac{1}{10}$	$\frac{11}{12}\cdot 1\cdot\frac{9}{10}=0.82$
(7,8]	9	1	0	$1-\frac{1}{9}$	$\frac{11}{12}\cdot 1\cdot\frac{9}{10}\cdot\frac{8}{9}=0.73$
(8,9]	8	0	1	1	$\frac{11}{12}\cdot 1\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot 1=0.73$
(9,12]	7	0	1	1	$\frac{11}{12}\cdot 1\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot 1\cdot 1=0.73$
(12,14]	6	1	0	...	...
(14,15]	5	1	0	...	...
(15,16]	4	1	0	...	...
(16,20]	3	0	1		
(20,22]	2	0	1		
(22,23]	1	1	0		

The Kaplan-Meier estimator $\hat{S}(t)$



The Kaplan-Meier estimator $\hat{S}(t)$

Kaplan-Meier formula:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{1}{Y(t_i)}\right)$$

with death times $t_1, ..., t_d$ and $Y(t_i) =$ number at risk just before t_i.

For multiple deaths at the same time (also called **ties**):

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{m_i}{Y(t_i)}\right)$$

with $m_i =$ number of deaths at time t_i.

Kaplan-Meier is also known as the **product limit** estimate.

Kaplan-Meier in Stata

The first step for any type of survival analysis in Stata (so not only for Kaplan-Meier) is to **stset** the data, to define them as **survival time data.**

Basic syntax of the command: `stset` time var, failure var

```
. list

     +--------------------+
     | id   status   time |
     |--------------------|
  1. |  1        1      5 |
  2. |  2        0      6 |
  3. |  3        1      7 |
   .    .        .      .
 11. | 11        0     22 |
 12. | 12        1     23 |
     +--------------------+

. stset time, failure(status=1)
```

The `failure` option allows us to specify if an observation is right censored.

Kaplan-Meier in Stata

Once you have prepared the data for survival analysis with `stset`, the syntax for the Kaplan-Meier command in Stata is very easy:

for one curve for the whole population:
`sts list` estimates the survival function
`sts graph` plots the survival function

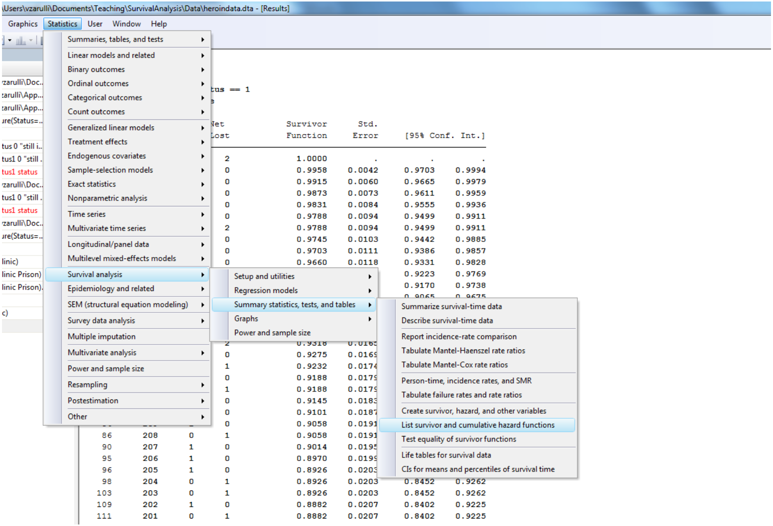
for curves by subgroups:

```
sts list, by(varname)
sts graph, by(varname)
```

There is no need to specify the data, because it always refers to the last survival-time dataset available in the memory.

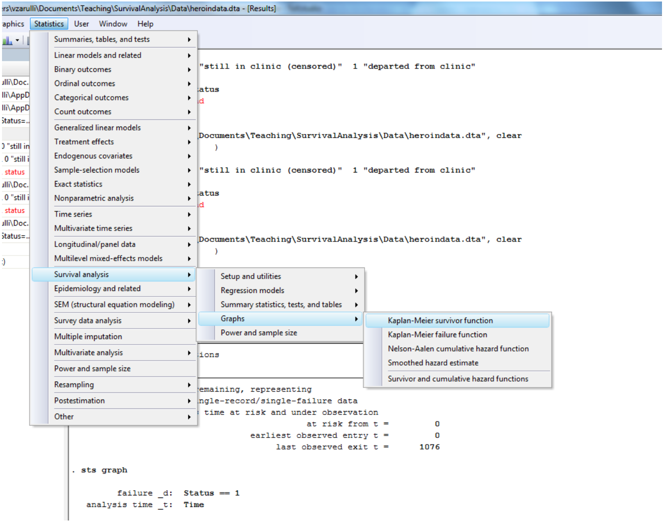
Kaplan-Meier in Stata

You can also use the drop-down menu:



Kaplan-Meier in Stata

You can also use the drop-down menu:



Exercise 1

Exercise 1:

Heroindata.dta contain information about times, in days, that heroin addicts spend in a clinic. The variables are:

- ID (individual identification number)
- Clinic (1 or 2)
- Status (0 = still in clinic at end of study (censored) or 1 = departed from clinic)
- Time (days spent in clinic)
- Prison (1 = prison record or 0 = no record)
- Dose (methadone dosage (mg/day))

Exercise 1

Tasks:

1. Open the data "heroindata.dta".
2. Prepare the data for the analysis by declaring them survival-time data.
3. Estimate and plot the Kaplan-Meier curve for the whole population of drug addicts.
4. Estimate and plot the clinic specific Kaplan-Meier curve.
5. Add censoring time to the previous plot with the `censored(number)` option.

Logrank test

To compare the survival curves of groups by testing their equality (H_0: Equality of the survival functions).

Comparison of the overall survival curve by comparing, at each failure time, the expected versus observed number of failures for each group and then combining the comparisons over all observed failure times.

Logrank test

For every time failure t_i

time t_i	failed	survived	at risk
group 1	f_{i1}	$r_{i1} - f_{i1}$	r_{i1}
group 2	f_{i2}	$r_{i2} - f_{i2}$	r_{i2}
total	f_i	$r_i - f_i$	r_i

under the validity of H_0: $E_{i1} = f_i \frac{r_{i1}}{r_i}$ and $V_i = \frac{r_{i1} r_{r2} f_i (r_i - f_i)}{r_i^2 (r_i - 1)}$

Test statistic:

$$LR = \frac{(O_1 - E_1)^2}{V} \sim \chi_1^2$$

where

$E_1 = \sum_i E_{i1}$ expected number of failures in group 1

$O_1 = \sum_i f_{i1}$ observed failures in group 1

$V = \sum_i V_i$ variance

Logrank test in Stata

Logrank test:

`sts test` *`varname`*

Stratified logrank test:

`sts test` *`varname`*, `strata(`*`varname`*`) detail`

When, while testing the equality of the survival curves of Group 1 and Group 2, we think that an additional variable, for example Sex, is worth to be taken into account (because males and females survival experiences differ...)

Logrank test in Stata

The interpretation of the test follows the usual hypothesis testing: the p-value tells us whether we can or can not reject the null hypothesis.

Log-rank test for equality of survivor functions

Clinic	Events observed	Events expected
1	122	90.91
2	28	59.09
Total	150	150.00

chi2(1) = 27.89
Pr>chi2 = 0.0000

Exercise 2

Tasks:

1. Test the equality of the survival curves of drug addicts by Clinic.
2. It may be that those who have been in prison and those who have not been have different survival patterns. This may be a confounding factor in the analysis of the performance of the clinic. Plot Kaplan-Meier curves by Clinic and Prison. What is your impression?
3. Perform a logrank test between survival curves of the two clinic, stratified by the variable Prison, to confirm the Kaplan-Meier result.

The hazard and cumulative hazard functions

The hazard function, $h(t)$ indicates the risk of death at the instant t, given survival until t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The cumulative hazard function, $H(t)$ is:

$$H(t) = \int_0^t h(x)dx$$

The survival function and the hazard functions are related by the following relation:

$$S(t) = e^{-\int_0^t h(x)dx} = e^{-H(t)}$$

The Nelson-Aalen estimator

In virtue of their relation, we could derive the cumulative hazard and the hazard from the survival function in the following way:

$$H(t) = -\ln S(t) \text{ and } h(t) = -\frac{d}{dt}\ln S(t)$$

However, when $S(t)$ is a step function and the sample size is rather small, a more appropriate method is the Nelson-Aalen estimator of the cumulative hazard $\hat{H}(t)$.
With death times $t_1, ..., t_d$:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{m_i}{Y(t_i)}$$

where $Y(t_i) =$ number at risk just before t_i and $m_i =$ number of deaths at time t_i.

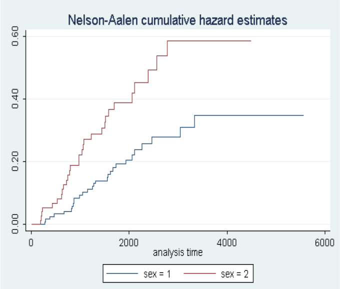
The Nelson-Aalen estimator in Stata

Cumulative hazard estimate:

```
sts list, cumhaz
sts list, by(varname) cumhaz
```

Cumulative hazard plot:

```
sts graph, cumhaz
sta graph, by(varname) cumhaz
```



The Nelson-Aalen estimator in Sata

And for the hazard function?

When the survival is a step function it can not be directly differentiated to obtain the hazard.

A non-straightforward way to estimate it is by taking the steps of the Nelson-Aalen cumulative hazard and smoothing them with some smoothing algorithm.

In Stata you need to specify the option `hazard` in the `sts graph` and which kernel function (epanechnikov, gaussian...) and bandwidth to use for the smoothing.

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

Exercise 3

Tasks:

1. Estimate the cumulative hazard of leaving the clinic for the drug addicts by clinic.
2. Plot the cumulative hazard by clinic, the survival function by clinic and save the plots in the memory.
   (just add `saving(nameplot)` to the `sts graph` command)
3. Combine the plots in a single graph.
   (use `gr combine "nameplot1" "nameplot2"`).

Exercise 3

Tasks:

4. Estimate the median survival time (time at which 50% of the individuals have failed and 50% is still surviving, also known as the 50th percentile) of the drug addicts in the two clinics (`stci, by(varname)`). Why the median survival time is not available for clinic 2?
5. Estimate the time at which $\frac{1}{4}$ of the patients have left the clinics (use `stci, by(varname) p(#)`, where # is the desired percentile).

Exercise 3

Tasks:

6. List the comparison of the survival function by clinic at 2 weeks, 1 month, 3 months, 6 months, 1 year, 2 years (`sts list, by(varname), at(#)`, where # is the desired survival time).