

Royal Veterinary and Agricultural University

<http://www.matfys.kvl.dk/~torbenm/DINA/survival>

Introduction to statistics for time to event data

Torben Martinussen

torbenm@dina.kvl.dk

Outline

- Introduction
- Kaplan-Meier estimator
- Nelson-Aalen Estimator
- Analysis using software package R

Statistics

Statistics is concerned with making conclusions based on **data** about population of interest.

The recipe when doing statistical analysis :

- **Scientific hypotheses** are formulated.
- Data is collected.
- **Statistical model** is formulated and validated.
 - **Systematic variation**. Parameters about which the hypotheses are formulated.
 - **Random variation** is described by model.
- **Inference** about parameters may be drawn.

Survival Data

Time to **death** or **event**.

Choose time-scale:

- Time from start of randomized clinical trial to death.
- Time from starting to get pregnant to pregnancy is obtained.
- Time from start to stop of optimal training of horse.
- Time to hatching for cocoons (earthworm).

What is special about survival data?

- Right-skewed. No problem.
- **Censoring!** For some we will only know a lower bound of event time.

Hatching of cocoons

Tina Svendsen, KVL.

894 cocoons (earthworms) were followed until hatching or end of study. Study period \approx 2 years.

- 791 did hatch in the study period
- 103 had not hatched but were considered alive.

Purpose: Study effect on time to hatching of treatment, age (of parents). Treatments were antiparasitic agents: Ivermectin and Fenbendazole.

Hatching of cocoons

treat	pair	age	time	status	kvart
K	2	6.0	10.99527	1	1
F	33	16.0	11.00598	1	3
K	2	6.0	10.99926	1	1
F	43	13.5	12.00651	1	1
F	33	5.0	13.00394	1	1
K	2	6.0	12.99521	1	1
F	34	6.0	12.99002	1	1
K	2	6.0	12.99007	1	1
K	3	6.0	13.00182	1	1
I	19	11.0	13.00739	1	2
F	34	6.0	12.99228	1	1
F	34	6.0	13.00159	1	1
K	13	9.5	13.99667	1	3
F	34	6.0	14.00489	1	1
F	33	6.0	13.99564	1	1
F	33	6.0	13.99577	1	1
I	20	16.0	13.99997	1	3
K	2	6.0	14.00329	1	1
I	17	9.0	14.00112	1	2
F	33	6.0	14.00028	1	1
		.			
		.			
		.			

Malignant melanoma

In the period 1962-77 205 patients had their tumour removed and were followed until 1977.

At the end of 1977:

- 57 died of mgl. mel.
- 14 died of non-related mgl. mel.
- 134 were still alive.

Purpose: Study effect on survival of sex, age, thickness of tumour, ulceration, ..

Malignant melanoma

N	time	status	sex	age	year	thickness	ulcer
1	10	3	1	76	1972	6.76	1
2	30	3	1	56	1968	0.65	0
3	35	2	1	41	1977	1.34	0
4	99	3	0	71	1968	2.90	0
5	185	1	1	52	1965	12.08	1
6	204	1	1	28	1971	4.84	1
7	210	1	1	77	1972	5.16	1
8	232	3	0	60	1974	3.22	1
9	232	1	1	49	1968	12.88	1
.
.
.
203	4688	2	0	42	1965	0.48	0
204	4926	2	0	50	1964	2.26	0
205	5565	2	0	41	1962	2.90	0

Survival data

For survival data employ techniques that take censoring into account. One can **not**

1. Calculate mean, standard deviation. Why?
2. Base inference on usual procedures : T-tests, Analysis of Variance, etc ...

because of **censored values!**

- **Right-censoring**. Ex: Patient still alive at end of study. T is above lower limit.
- **Left-censoring**. Ex: Time to recurrence of tumour after removal. At first inspection recurrence has already taken place. T is below upper limit.

Survival analysis

1. Kaplan-Meier estimates of chance of survival.
2. Log-Rank tests for comparisons.
3. Cox-Regression for regression analysis.

The **survival-function**:

Instead of usual procedures for quantitative data, consider the survival-function

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

= probability of survival to time t

where T is the lifetime of interest.

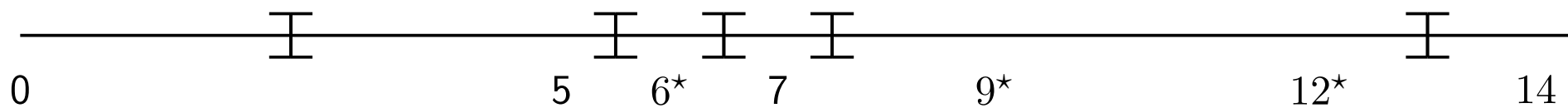
The survival-function is estimated by the non-parametric **Kaplan-Meier estimator**.

Kaplan-Meier estimator

Kaplan-Meier estimator $\hat{S}(t)$ of survival function $S(t)$. Observations :

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

The Kaplan-Meier estimator $\hat{S}(t)$ may be derived as follows : Partition the time-axis into fine intervals such that each interval contains at most one observation :



Consider the i th interval I_i . The chance of surviving I_i given that subject is alive at the start of the interval is

$$p_i = \begin{cases} 1 & \text{if nobody dies in } I_i \\ \frac{Y_{i-1}}{Y_i} & \text{if 1 subject dies in } I_i \end{cases}$$

where $Y_i = \#$ alive at start of I_i

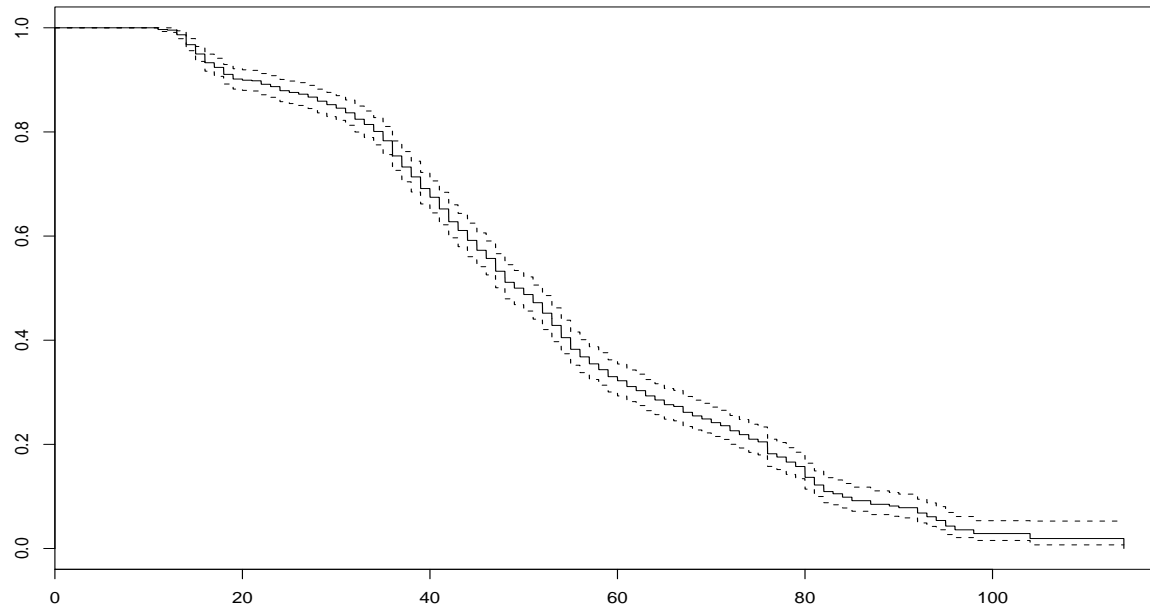
Kaplan-Meier estimator

Death times t_1, \dots, t_d . $Y(t_i) = \#$ alive just before t_i . Kaplan-Meier estimator:

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

Kaplan-Meier computation in R for worm-data:

```
> surv.all<-survfit(Surv(time,status==1))  
> par(mfrow=c(1,1))  
> plot(surv.all,mark.time=F)
```



Estimate median from KM plot.

The Nelson-Aalen Estimator

A simple estimator of the cumulative intensity is available for right censored data. Call the observed and ordered death times t_i and let $Y(t)$ denote the number of people at risk just before time t . Then

$$\Lambda(t) = \int_0^t \lambda(s) ds \approx \lambda(t_0)t_1 + \lambda(t_1)(t_2 - t_1) + \dots + \lambda(t_{k-1})(t_k - t_{k-1})$$

Estimate

$$\lambda(t_{j-1})(t_j - t_{j-1}) \approx P(\text{dying in }]t_{j-1}, t_j] \text{ given alive at } t_{j-1})$$

by

$$\frac{m_{j-1}}{Y(t_{j-1})}$$

where m_{j-1} is the number of people dying in $]t_{j-1}, t_j]$. Therefore

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

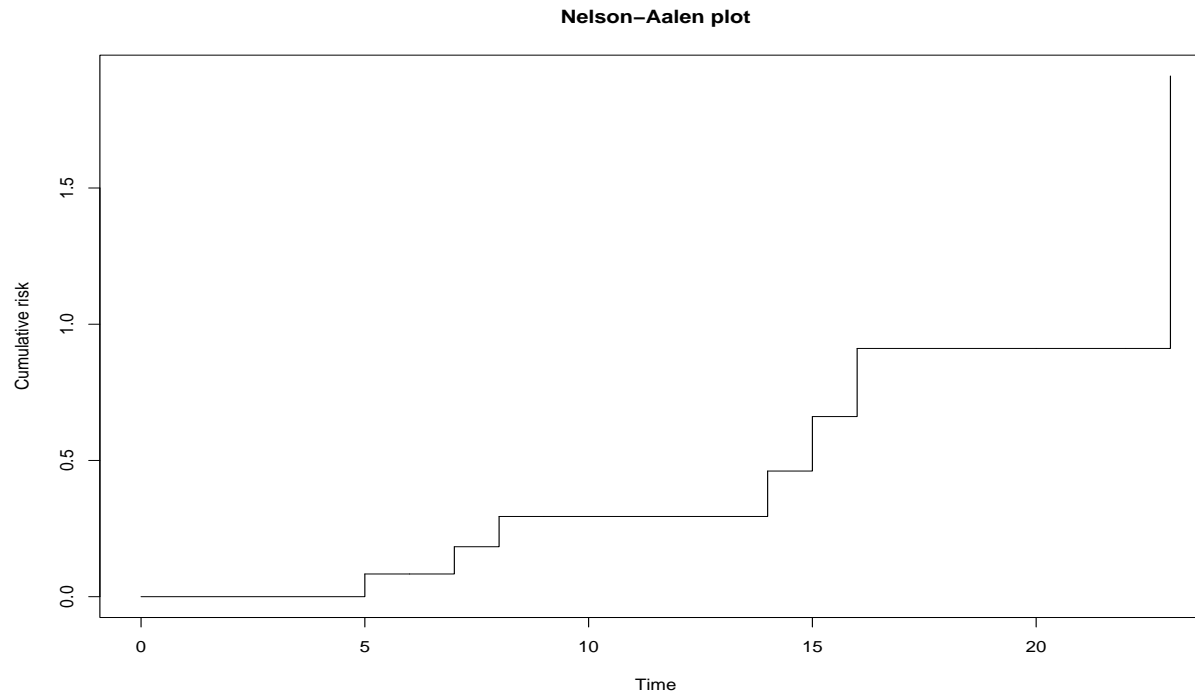
The Nelson-Aalen estimator may also be provided with standard errors and hence confidence intervals may be constructed.

The Nelson-Aalen Estimator

For the data Ordered survival times (months) :

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

the Nelson-Aalen looks like this



Exercise

- (1) Compute by hand the Nelson-Aalen estimator for the data given on slide 15. What is $\hat{\Lambda}(10)$? Compute also the Kaplan-Meier estimator for these data (by hand!).
- (2) Consider now the worm-data. Compute the Kaplan-Meier estimator using all data. Estimate the median waiting time and give a 95% confidence interval.
- (3) Compute the Kaplan-Meier estimator for each of the three treatment groups.
- (4) Compute the Kaplan-Meier estimator for each of the three treatment groups but now stratified by the variable `age.gr`. Plot these estimates in the same figure.
- (5) Estimate the median waiting times based on (4).