The next challenge: survival analysis

Gerspacher Thomas

Summary

Context

I. Survival Analysis

II. Challenge

Context

The objectives of the challenge are various:

- Assess participants’ performance.
- Teach participants’ a subject.
- Open a field of study to machine learning.

As a first step, the challenge is aimed at students of RPI, Rensselaer Polytechnic Institute, as part of a statistical course .

I.

Survival Analysis

A. Analysis

Objective:

Analyze the expected duration of time ***t*** until an event ***e*** appears.

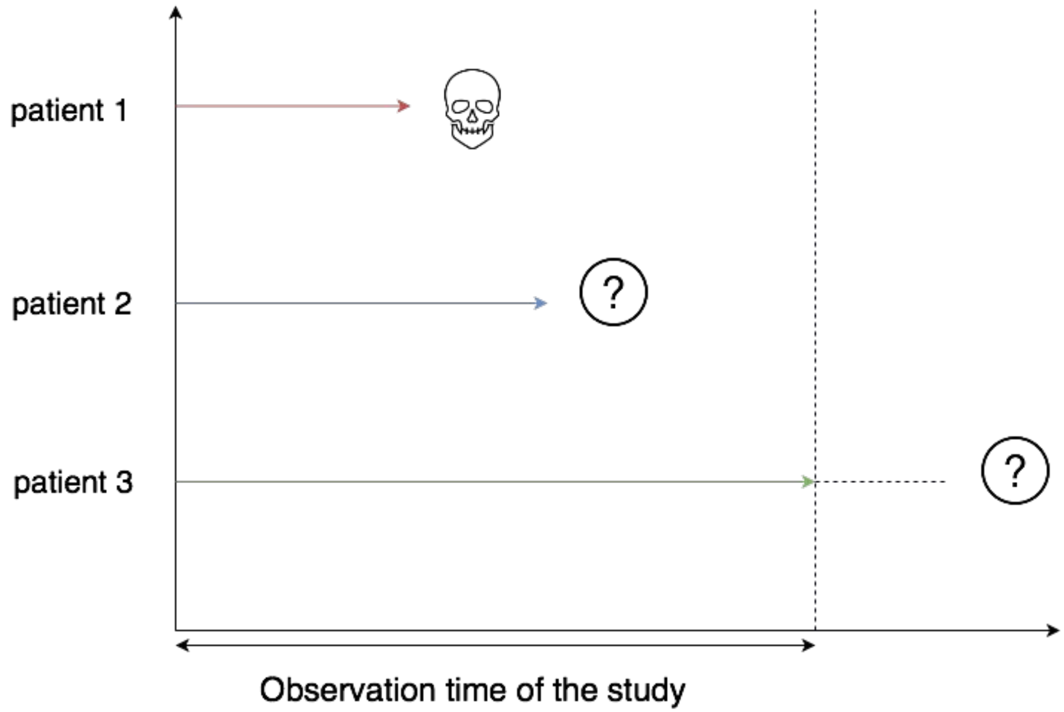
Examples:

- Effect of treatments on patients.
- Failures in mechanic systems.
- Relations between diseases and other factors.

A. Censoring

Definition:

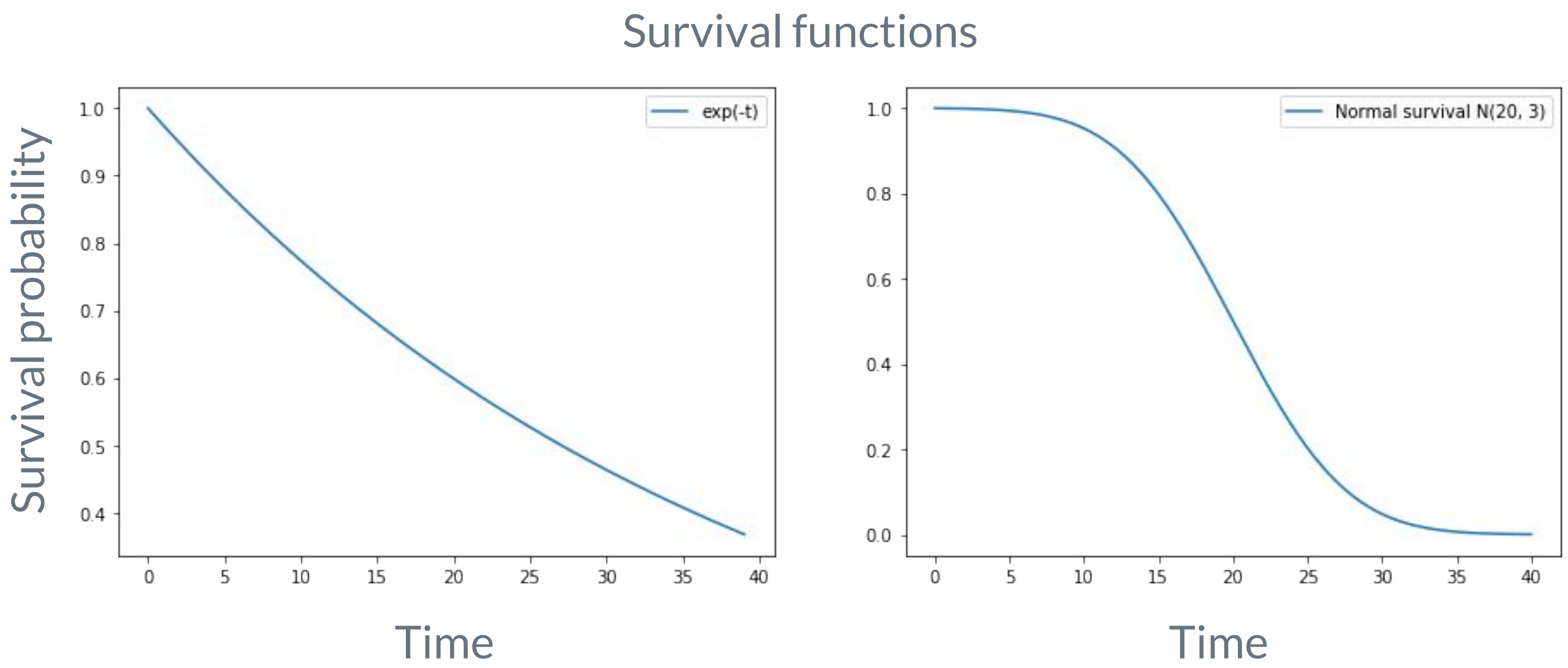
For some individuals, the event of interest never occurs during the observation time of the study.



B. Basics

- Survival function *S(t)*:

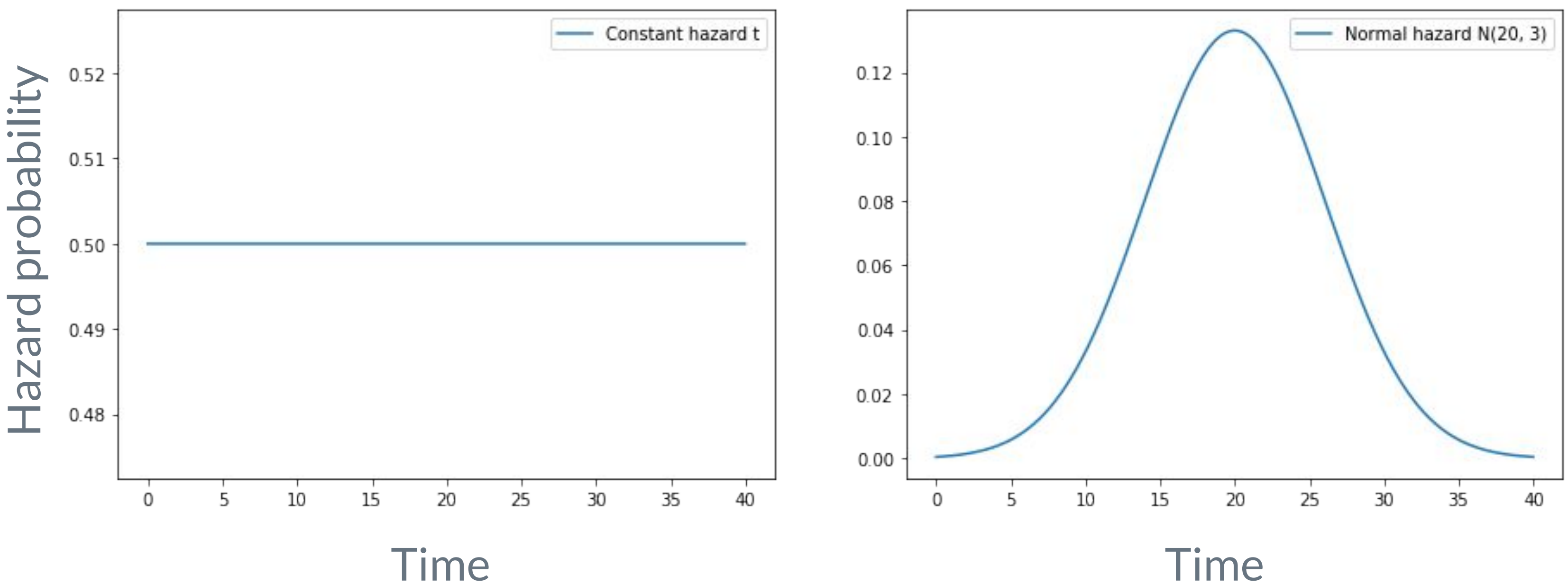
  Probability that a subject survives past time *t*



B. Basics

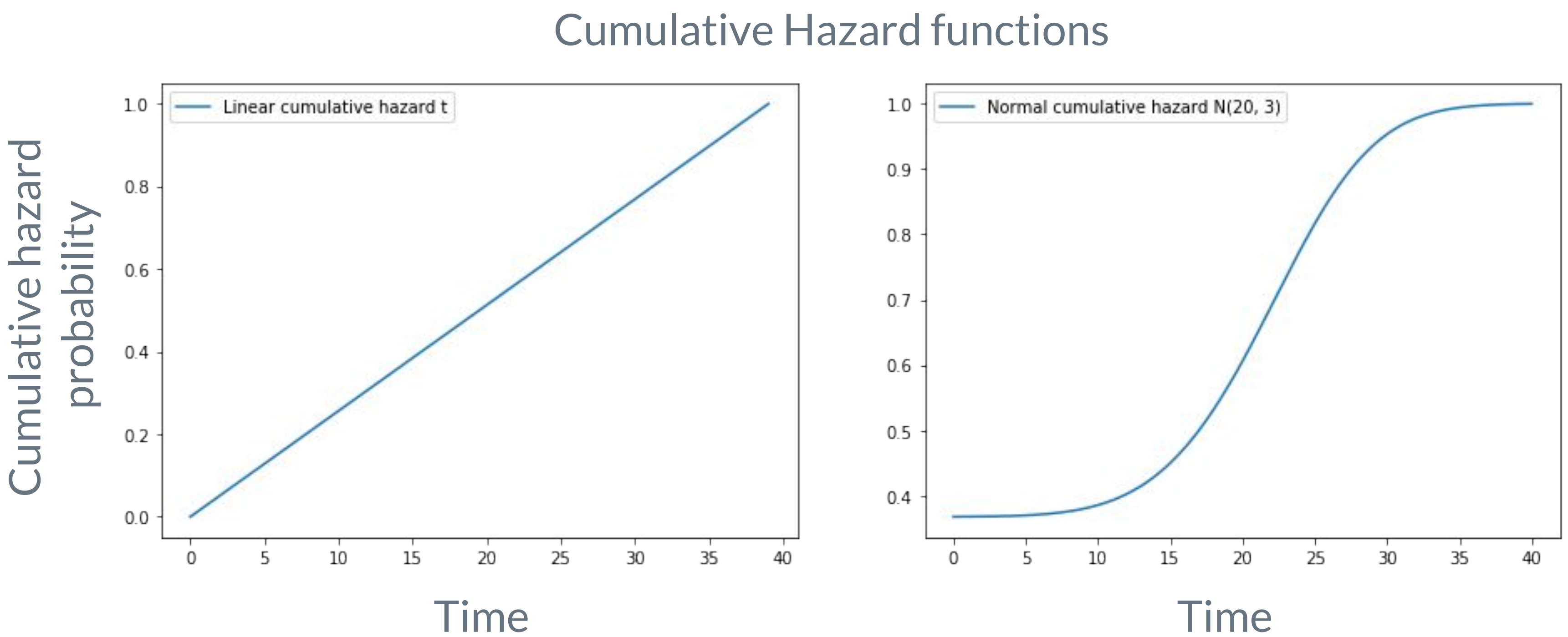
- Hazard function *h(t)*:
  Risk at time *t*

Hazard functions



B. Basics

- Cumulative hazard Function *H(t)*:
  Accumulated risk at time *t*



B. Basics

Relations between functions:

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

- $$h(t) = -\frac{\partial H(t)}{\partial t}$$

II.

Challenge

A. Progress

Steps:

1. Present a standard method used in survival analysis (ask the participant to implement it).

2. Use this baseline on real data and show how it performs.

3. Ask the participant to find a better solution than the standard method using, for example, machine learning algorithms

B. Application 1: Estimation

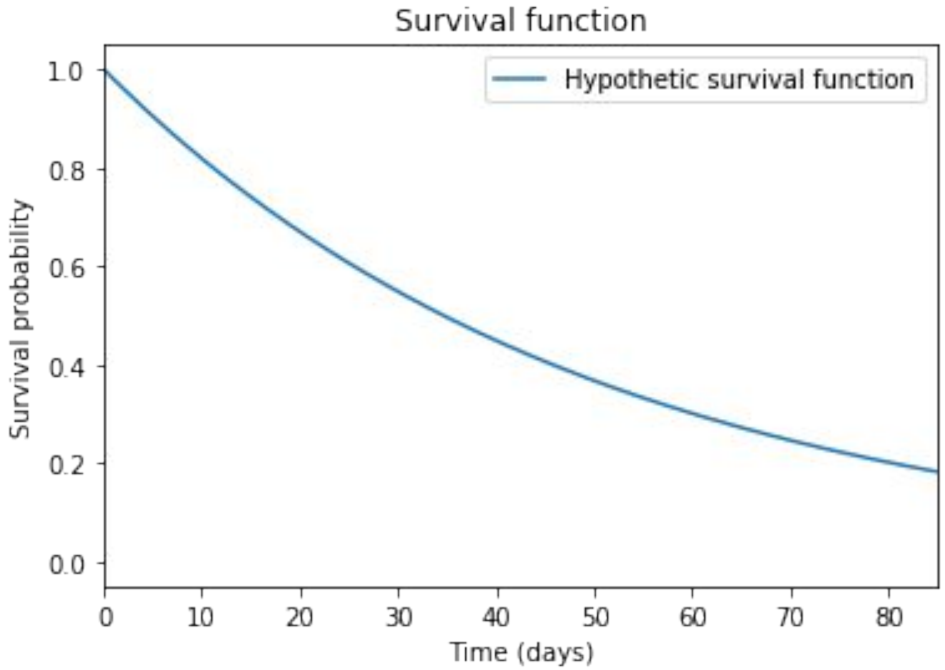
What is the probability that a patient dies during its stay in a hospital?

<u>Hypothesis:</u>

Let us say that each day, a patient as $\frac{1}{50}$ chance to survive.

$h(t)$	$H(t)$	$S(t)$
$\frac{1}{50}$	$\frac{t}{50}$	$e^{-\frac{t}{50}}$

B. Application 1: Estimation



$$S(t) = e^{-\frac{t}{50}}$$

B. Application 1: Estimation

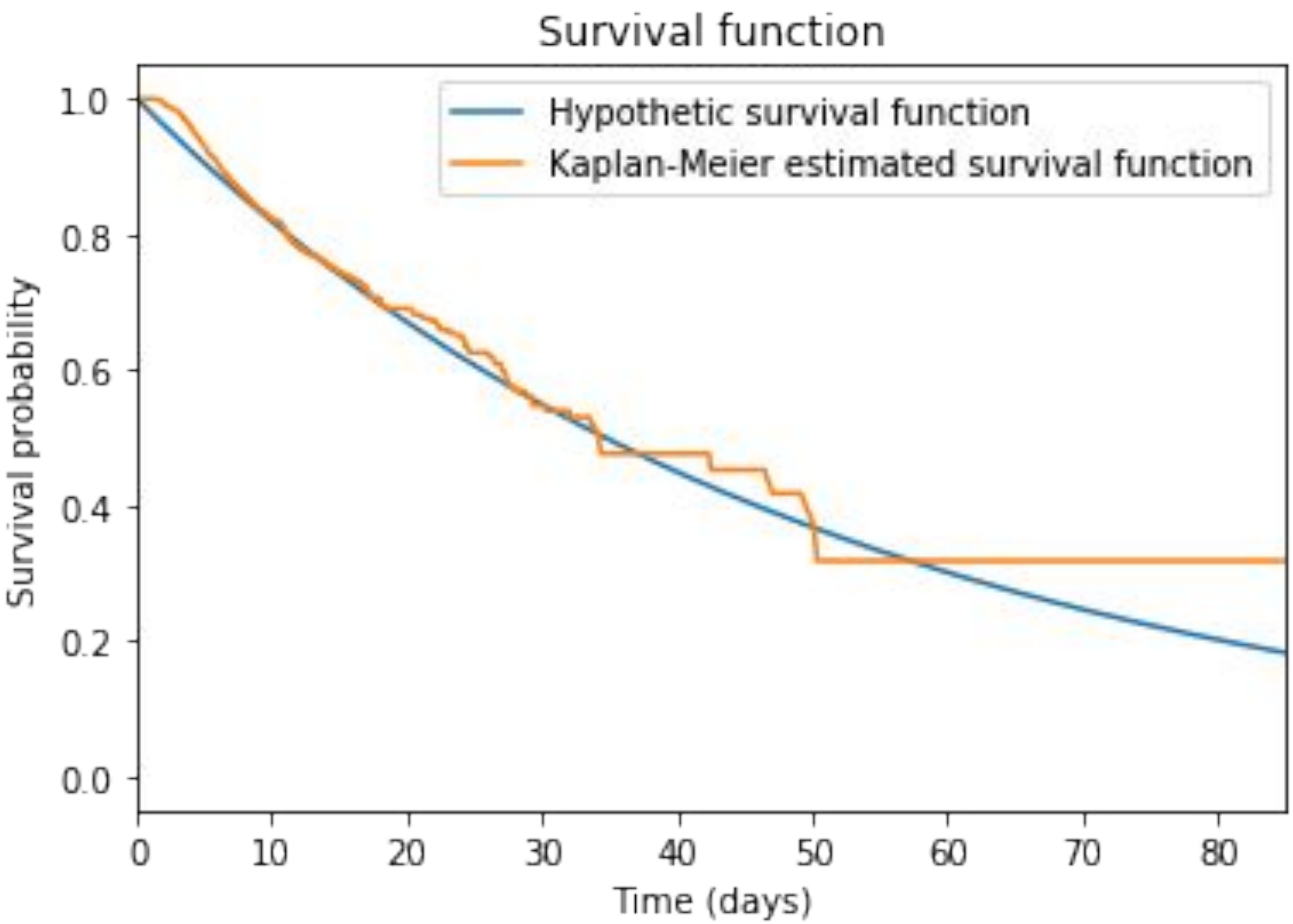
Kaplan-Meier estimator:

- non-parametric estimator of the survival function.
- constructed by using maximum-likelihood estimation.

$$\hat{S}(t) = \sum_{i:t_i<t} \frac{n_i - d_i}{n_i}$$

with n_i number of individuals who has not experienced the event at time t_i
and d_i number of individuals who experience the event at time t_i

B. Application 1: Estimation



C. Example & applications

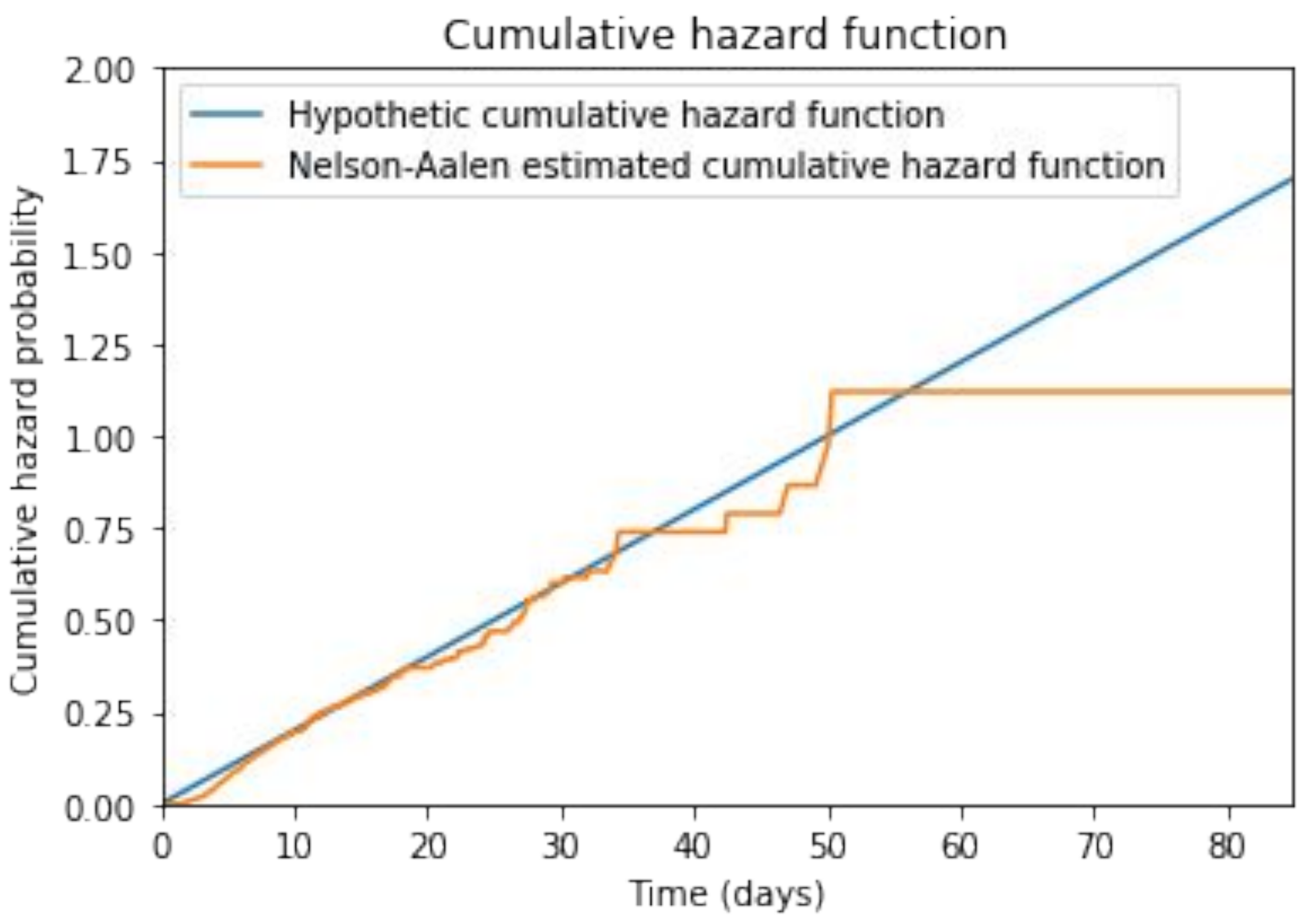
Nelson-Aalen estimator:

- non-parametric estimator of the cumulative hazard function.
- constructed by using maximum-likelihood estimation.

$$\hat{H}(t) = \sum_{i:t_i<t} \frac{d_i}{n_i}$$

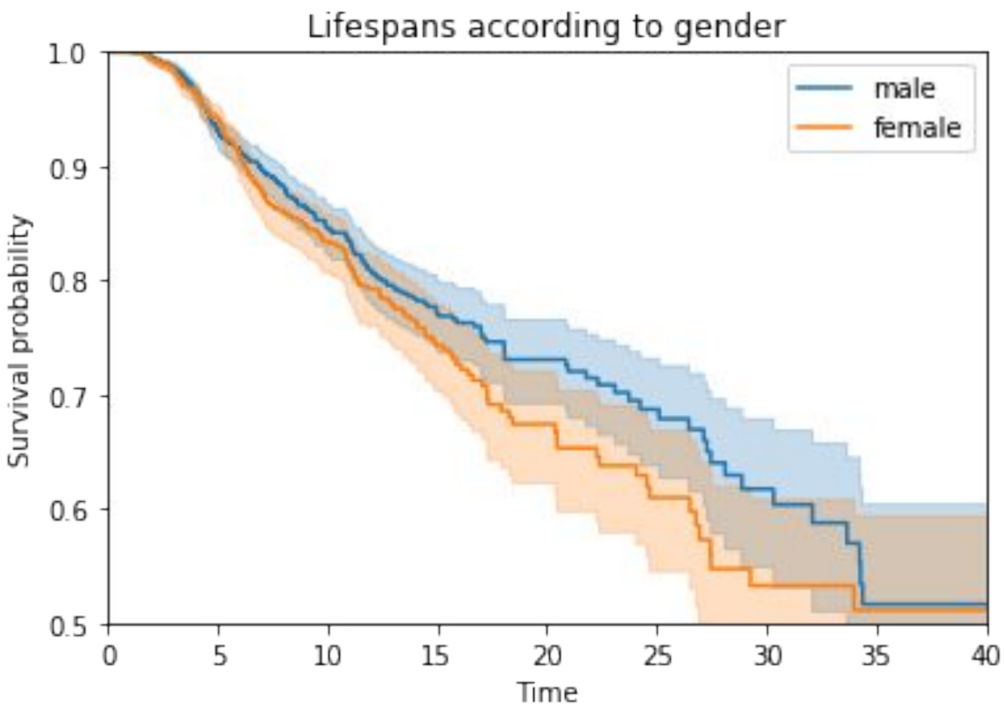
with n_i number of individuals who has not experienced the event at time t_i
and d_i number of individuals who experience the event at time t_i

C. Example & applications



C. Application 2: Comparison

Do men have the same survival probability than women?



C. Application 2: Comparison

Log-rank test:

- Non-parametric test
- Compare survival curves

Hypothesis 0 : Both survival curves are the same. $S_1(t) = S_2(t)$
Hypothesis 1 : Survival curves are different. $S_1(t) \neq S_2(t)$

```
null_distribution=chi squared, alpha=0.99, df=1, t_0=-1

test_statistic      p
        0.9006 0.3426
```

Survival curves of males and females are different (p-value significantly high).

C. Application 3: Regression

Cox Regression

Goal : Estimate the hazard function by using prior knowledge of some data X, i.e. λ(t\X).

$$\lambda(t|X) = b_0(t)e^{\sum_{i=1}^{d} b_i x_i}$$

Possible improvments: Feature selection

Presentation of the challenge

Survival Analysis - Jupyter Notebook

Conclusion: Objectives

- Data generation
  - Good quality
  - Ensure privacy (theoretically and practically)

- Challenge
  - Build fully functional challenges
  - Auto-grading system for the survival challenge
  - Check the viability of different tools such as Gitclass, Travis, Binder, ...

Thanks!

Any questions?