An Introduction to Survival Analysis

Chrysoula Tsismetzoglou

School of Medicine-Department of Mathematics
National and Kapodistrian University of Athens

March 18, 2025

ΔΠΜΣ **ΒΙΟΣΤΑΤΙΣΤΙΚΗ &**
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΥΓΕΙΑΣ

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ - ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

1. Motivation

2. Main functions

3. Basic concepts/assumptions

4. What is Survival Analysis?

5. Non-parametric methods

6. Semi-parametric methods

1 Motivation

2 Main functions

3 Basic concepts/assumptions

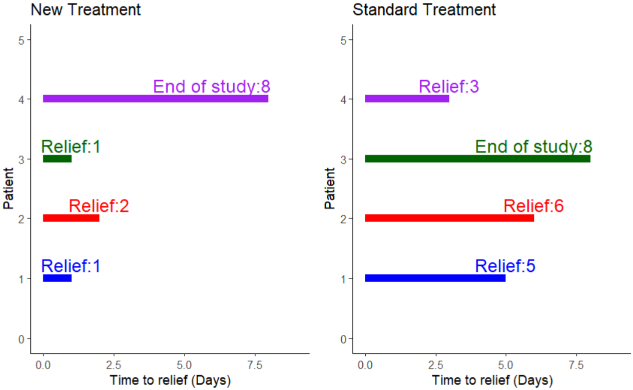
4 What is Survival Analysis?

5 Non-parametric methods

6 Semi-parametric methods

Motivation

Suppose we want to conduct a study to examine whether the new treatment for a disease is more effective than the standard one. For simplicity sake, suppose we have 4 patients in each treatment arm.



Motivation

Naive Analysis I

What can we do?
We can estimate (calculate) the **Cumulative Incidence** in each treatment arm as follows:

$$\textbf{Cumulative Incidence}_{Stand} = \frac{\#\text{Patients with relief in Standard treatment}}{\#\text{Total patients in Standard treatment}} = \frac{3}{4}(75\%)$$

$$\textbf{Cumulative Incidence}_{New} = \frac{\#\text{Patients with relief in New treatment}}{\#\text{Total patients in New treatment}} = \frac{3}{4}(75\%)$$

Does this mean that the two treatments have the same effect on patients?

Motivation

Naive Analysis II

The previous approach is not valid for this problem as ***it does not account for the patients' follow-up times and patients without an observed relief outcome until the end of the study***.

Solution 1!

We can use **Incidence Rates** to account for follow-up times!

$$\textbf{Incidence Rate}_{New} = \frac{\#\text{Patients with relief in New treatment}}{\text{Total follow-up times in New treatment}} = \frac{3 events}{12 \text{ person-days}} = \textbf{2.5 events per 10 person-days}$$

$$\textbf{Incidence Rate}_{Stand} = \frac{\#\text{Patients with relief in Standard treatment}}{\text{Total follow-up times in Standard treatment}} = \frac{3 \text{ events}}{22 \text{ person-days}} = \textbf{1.4 events per 10 person-days}$$

Motivation

Naive Analysis II-Notes

Incidence rates are **average rates**, there by useful when one can assume **relatively constant risks**[1].

So what do we do when instead?

The main idea is to introduce a risk measure that does not require constant risks over time and is defined **serially over shorter time intervals during which risk is reasonably constant**. This is called **Hazard Rate**.

[1] refers to the probability of experiencing the event per unit time

1 Motivation

2 Main functions

Hazard Rate

Survival function

Cumulative event function

Inter-relationships

3 Basic concepts/assumptions

4 What is Survival Analysis?

5 Non-parametric methods

6 Semi-parametric methods

1 Motivation

2 Main functions

Hazard Rate

Survival function

Cumulative event function

Inter-relationships

3 Basic concepts/assumptions

4 What is Survival Analysis?

5 Non-parametric methods

6 Semi-parametric methods

Hazard Rate

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

- This is **the instantaneous rate of the probability of event within the next small interval of time, given that the subject is still at risk for event at time t**(has “survived” the event up to t).

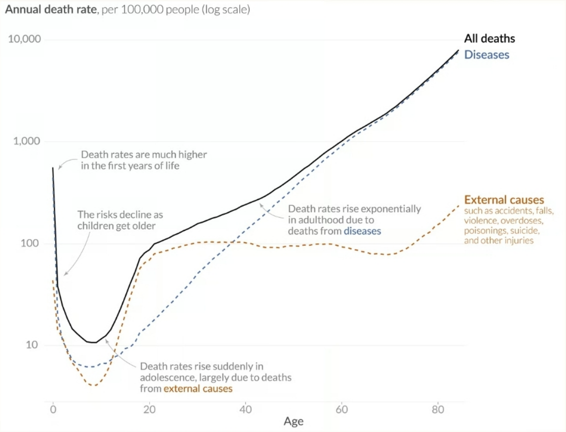
Hazard Rate

Example of death rates across ages

Death rates across ages

Our World in Data

National data from the United States between 2018 and 2021.



Note: Period death rates using ICD-10 categories. 'Diseases' includes all categories except 'external causes' and 'signs, symptoms and abnormal findings'.
Source: United States Centers for Disease Control and Prevention, via CDC Wonder database
OurWorldinData.org — Research and data to make progress against the world's largest problems. Licensed under CC-BY by the author Saloni Dattani

1 Motivation

2 Main functions

Hazard Rate

Survival function

Cumulative event function

Inter-relationships

3 Basic concepts/assumptions

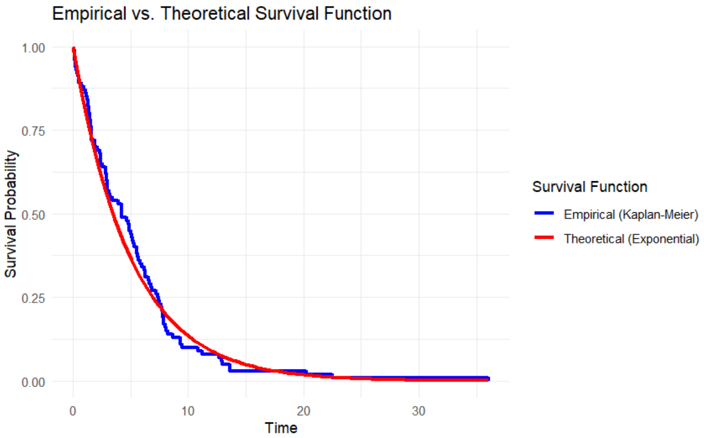
4 What is Survival Analysis?

5 Non-parametric methods

6 Semi-parametric methods

Survival function

$$\boxed{S(t) = P(T > t) = \textbf{probability of surviving beyond a specified time t}}$$



1 Motivation

2 Main functions

Hazard Rate

Survival function

Cumulative event function

Inter-relationships

3 Basic concepts/assumptions

4 What is Survival Analysis?

5 Non-parametric methods

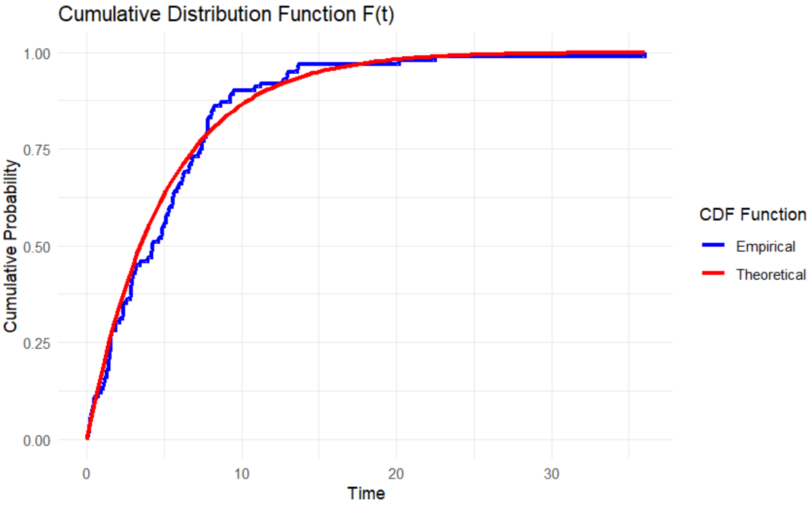
6 Semi-parametric methods

Cumulative event function

$$\boxed{F(t) = P(T \leq t) = 1 - S(t)}$$

- This is **the probability of having experienced the failure (or outcome event) by time t.**
- It gives **the cumulative probability of failure (or event occurrence) up to time** t.

Cumulative event function



1. Motivation
2. Main functions
   - Hazard Rate
   - Survival function
   - Cumulative event function
   - Inter-relationships
3. Basic concepts/assumptions
4. What is Survival Analysis?
5. Non-parametric methods
6. Semi-parametric methods

Inter-relationships

$$\begin{tikzcd}
 & f(t) \arrow[dl, shift right, "S(t)=\int_{t}^{+\infty} f(x)\,dx"'] \arrow[dr, shift left, "h(t)=\frac{f(t)}{S(t)}"'] & \\
S(t) \arrow[ur, shift right, "f(t)=-S'(t)"'] \arrow[rr, shift left, "h(t)=\frac{-S'(t)}{S(t)}"] & & h(t) \arrow[ul, shift left, "f(t)=h(t)\cdot e^{-\int_0^t h(x)dx}"'] \arrow[ll, shift left, "S(t)=e^{-\int_0^t h(x)dx}"]
\end{tikzcd}$$

Censored data

Censored times: when we don't know the exact survival times.

1 Motivation

2 Main functions

3 Basic concepts/assumptions

Right censoring

Left censoring

Interval censoring

Informative/Non-informative censoring

4 What is Survival Analysis?

5 Non-parametric methods

6 Semi-parametric methods

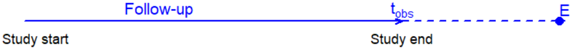
Censored data

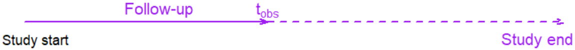
Right-censoring

This is the most common censoring type.

Right-censored data

When a lower bound for survival time is known ($T > t_{obs}$, where T is the true survival time and t_{obs} is the observed one.)





Censored data

Right-censoring

We have 2 types of right censoring:

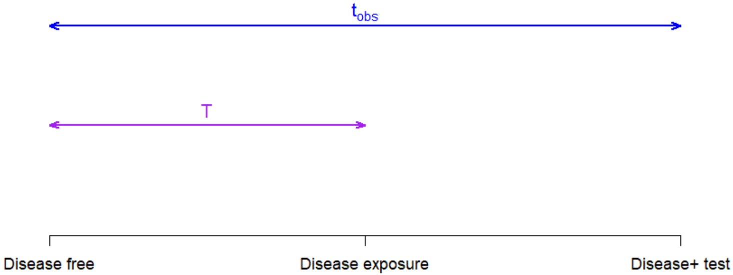
1. **Fixed-Administrative right-censoring**: it occurs when study ends and no event has been observed.
2. **Random right-censoring**: it occurs due to loss to follow-up, withdraws from study or competing events.

Censored data

Left-censoring

Left-censored data

When the true time event has been occurred before observing the event ($T \leq t_{obs}$, where T is the true survival time and t_{obs} is the observed one.) For example, when a patient living with AIDS has been infected by HIV before having a positive HIV test.



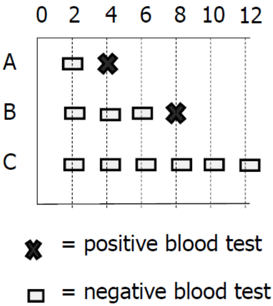
Censored data

Interval-censoring

Interval-censored data

The true survival time is known only to lie within an interval instead of being observed exactly.

This is more common in studies with periodic follow-up.



Censored data

Interval-censoring

Suppose we have a study with pre-scheduled observation times in which we take blood tests every 2 months for a year for 3 patients as shown above.

1. $2 < T_A < 4$ **Interval censored**
2. $6 < T_B < 8$ **Interval censored**
3. $12 < T_C$ **Right censored**

1. Motivation
2. Main functions
3. Basic concepts/assumptions
   - Right censoring
   - Left censoring
   - Interval censoring
   - Informative/Non-informative censoring
4. What is Survival Analysis?
5. Non-parametric methods
6. Semi-parametric methods

Censored data

Informative/Non-informative censoring

- Non-informative(Independent) censoring: **censoring is independent of survival time (but operates randomly)**.
  For example, an individual censored at t has similar prognosis as (is representative of) uncensored patients with equal follow-up (who survive up to t).
- Informative censoring: **censoring when patients drop-out for reasons related to the treatment allocation and the event rate of censored patients is different than those who are followed**.
  For example, patients withdrawn from study because of a deterioration in their physical condition-bias survival probabilities upwards because patients with better prognosis stay in the study.

1. Motivation
2. Main functions
3. Basic concepts/assumptions
4. **What is Survival Analysis?**
5. Non-parametric methods
6. Semi-parametric methods

What is Survival Analysis?

Survival Analysis

Survival Analysis is a branch of statistics that studies time-to-event data. It is a collection of statistical methods that aim to estimate and make general statistical inference on hazard rates and survival functions accounting,also,for censored data.

The term Survival Analysis is mainly used in medical research while there are equivalent terms and branches in other fields like:

- **Reliability Theory/Analysis** in engineering.
- **Duration Analysis/Modeling** in economics and political sciences.
- **Event History Analysis** in sociology.

1. Motivation
2. Main functions
3. Basic concepts/assumptions
4. What is Survival Analysis?
5. Non-parametric methods
6. Semi-parametric methods

Non-parametric methods

- In the previous slides, we saw a plot with the theoritical and the empirical survival function of exponential distribution survival times. We observed that the empirical one was a *step function*.
- Non-parametric methods are used primarily to estimate the *survival function* and subsequently the hazard,cumulative event function etc.
- There are 3 main non-parametric methods used:
  1. **Kaplan-Meier estimator**(mainly used).
  2. **Nelson-Aalen estimator**.
  3. **Life Table estimator**.
- In this presentation we will only see the first 2 ones.

Assumptions of the Kaplan-Meier estimator

The Kaplan-Meier survival (or hazard) curve is an **unbiased estimate** of the true survival (or hazard) curve if:

1. The patients are representative of the underlying population.
2. The events occurred at the times specified.
3. Censoring is “non-informative”.
4. Survival probabilities are the same for participants recruited early and late in the study.

Non-parametric methods

Kaplan-Meier estimator

- We observe survival times in a sample of n individuals with *right-censoring*.
- We denote $\tau_1, \tau_2, ..., \tau_\kappa$ the dinstict and ordered event (uncensored) times. For ties, censored time is taken to occur immediately after the event time.
- The **KM estimator of** $S(t)$ takes the form:

$$\hat{S}(t) = \prod_{j:\tau_j < t} \left(1 - \frac{d_j}{r_j}\right)$$

where r_j is the number of patients at risk (alive and not censored: in the risk set) just prior to time τ_j and d_j the number of events at time τ_j.

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

$$\widehat{Var} \approx [\hat{S}(t)] \cdot \sum_{j:\tau_j < t} \frac{d_j}{rj(r_j - d_j)}$$

- In large samples, $\hat{S}(t) \rightarrow$ Normal, so that a $100(1-a)\%$ CI is $\hat{S}(t) \mp 1.96 se[\hat{S}(t)]$. However, this estimation falls outside the bounds 0 and 1, so we use the log-minus-log transformation to restrict the interval between the desired bounds.

Non-parametric methods

Kaplan-Meier estimator-Notes

1. The **KM-type estimate of the hazard function** $h(t)$ is given by:

$$\hat{h}(t) = \frac{d_j}{r_j(\tau_{j+1} - \tau_j)}$$

2. We cannot estimate the last interval.
3. The **variance** of the $\hat{h}(t)$is $\widehat{var} = \hat{h}(t) \cdot \sqrt{\frac{r_j - d_i}{d_j r_j}}$. When d_j is small, this leads to imprecise confidence limits of limited use.
4. In practice, the KM-type estimates of $h(t)$ tend to be rather irregular, so we use some smoothing method to see the pattern more clearly.

Non-parametric methods

Nelson-Aalen estimator

- Unlike KM, Nelson-Aalen initially estimates the cumulative sum of hazard rates $H(t)$ (**Cumulative Hazard function** over time) at event times.
- The **Nelson-Aalen cumulative hazard function estimator of** $H(t)$ is given by: $\hat{H}(t)_{NA} = \sum_{j:\tau_j<t} \frac{d_j}{rj}$.
- The **Nelson-Aalen survival function estimator of** $S(t)$ is given by: $\hat{H}(t)_{NA} = \exp(-\hat{H}(t)_{NA}) = \sum_{j:\tau_j<t} \exp\left(\frac{d_j}{rj}\right)$.

Non-parametric methods

Comparing survival curves

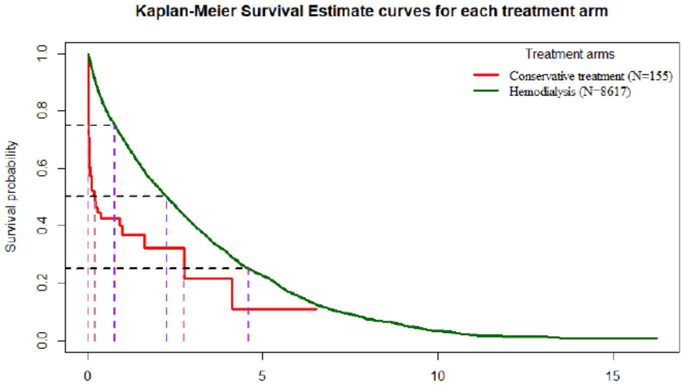
Another goal of the Survival Analysis is to **compare survival functions between 2 or more groups.**
For example, we want to answer questions like *Do patients live longer with a particular treatment? or Is there a difference in proportion of survival between different groups?*
There are several tools to compare survival curves:

1. Using **Kaplan-Meier curves**.
2. **Log-rank tests**.
3. **Wilcoxon test**.

Comparing survival curves

Using Kaplan-Meier

We can plot the KM survival curves of the groups (we can also plot the pointwise confidence intervals) and compute the median, the 25th and the 75th percentiles (if they exist) of each group.



Comparing survival curves

General hypothesis testing

Suppose we have *p* groups to compare:

$$H_0 : S_1(t) = S_2(t) = \dots = S_p(t) \textbf{ for every t}$$

vs.

$$H_A : \textbf{at least one of } S_j(t) \textbf{ is different}$$

Comparing survival curves

Log-rank and Wilcoxon tests

- The general idea of the **Log-rank test** and **Wilcoxon test** is to use 2×2 contingency tables in time intervals (the consecutive, ordered event times) to accommodate censoring.
- The difference between the 2 tests is that the Wilcoxon test is sensitive to **early differences** (it gives more weight to the early times), while the Log-rank test to later ones.
- Wilcoxon test cannot be used for more than 2 groups as it is. Generally, **Peto-Peto modification** is used for comparing more than 2 groups (this is what R do in *survival package*).
- Both tests follow χ^2 distribution with $p-1$ degrees of freedom, where p is the number of groups compared.
- Both tests have the right Type 1 power when testing the null hypothesis. The choice of which test to use may depend on the alternative hypothesis, which will drive the power of the test.
- The Log-rank test is most powerful under proportional hazards assumption (see Cox PH model).
- Wilcoxon test has high power when $T \sim Lognormal$ with equal variances in both groups but different means.

1. Motivation
2. Main functions
3. Basic concepts/assumptions
4. What is Survival Analysis?
5. Non-parametric methods
6. Semi-parametric methods

Semi-Parametric methods

The main semi-parametric used in Survival Analysis is the **Cox proportional hazards model.**

- The main idea is that we want to incorporate many covariates to our analysis and create a regression-type model adjusted for time-to-event data.
- The Cox PH model is the most commonly used model in Survival Analysis.
- With the Cox PH model we can estimate **hazard ratios**.

Cox Proportional Hazards model

Model specification

$$\boxed{h(t|X) = h_0(t) \cdot \exp\left(\beta_1 X_1 + b_2 X_2 + \ldots + \beta_p X_p\right)}$$

where:

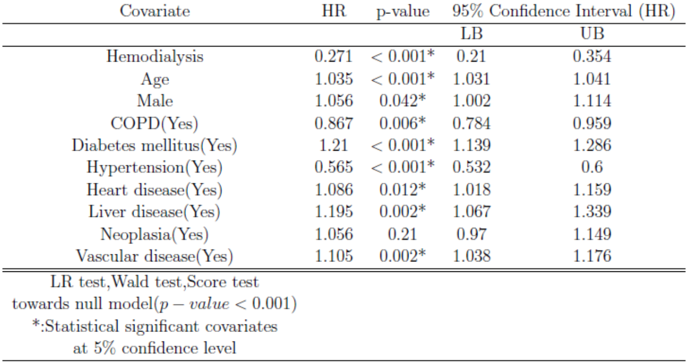
- $h_0(t)$ is the **baseline hazard** (for patients with all covariates equal to 0)
- X_i for $i = 1, \ldots, p$ are the covariates of the model. They can be either continuous (age,BMI etc) or categorical (sex,smoking status etc) that are recorded for each individual at the time origin.

- The right-hand side of the model ($\exp(\beta X)$) is called **linear predictor** and it shows the covariate effects.
- The **link function** of this model is the logarithm function.
- The main assumption of the model is the **Proportional Hazards assumption** $h_2(t) = \psi \cdot h_1(t) \longleftrightarrow S_2(t) = [S_1(t)]^{\psi}$. That means that survival curves do not cross.
- There is no distributional assumption for the baseline hazard, but the PH assumption makes the model semi-parametric.

Cox Proportional Hazards model

Interpretation of the model parameters e^{β_i}

change in the hazard ratio of Y per unit increase in X_i,adjusted for all other variables in the model.



Covariate	HR	p-value	95% Confidence Interval (HR)	
			LB	UB
Hemodialysis	0.271	< 0.001*	0.21	0.354
Age	1.035	< 0.001*	1.031	1.041
Male	1.056	0.042*	1.002	1.114
COPD(Yes)	0.867	0.006*	0.784	0.959
Diabetes mellitus(Yes)	1.21	< 0.001*	1.139	1.286
Hypertension(Yes)	0.565	< 0.001*	0.532	0.6
Heart disease(Yes)	1.086	0.012*	1.018	1.159
Liver disease(Yes)	1.195	0.002*	1.067	1.339
Neoplasia(Yes)	1.056	0.21	0.97	1.149
Vascular disease(Yes)	1.105	0.002*	1.038	1.176

LR test,Wald test,Score test
towards null model($p - value < 0.001$)
*:Statistical significant covariates
at 5% confidence level

Cox Proportional Hazards model

Notes 1

- To estimate the β coefficients we use the **Maximum Likelihood Estimation**.
- An estimation of the $h_0(t)$ is not required in the estimation of the coefficients.
- We use **partial likelihood** because it does not use the actual survival times but depends only on the rankings of event times. It can be shown that a partial likelihood satisfies all the regular likelihood properties.
- When dealing with *ties*, there is no problem when there are events and censored times but tied events cause comptutational complexity. **What can we do?**
  - **Breslow's Approximation**
  - **Efron's Approximation** (preferred)
- We can estimate $h_0(t)$ with **Breslow estimator**.

Cox Proportional Hazards model

Notes 2

- There are several *model selection* methods used in this framework that are similar to the standard ones we use in regression analysis.
- We can assess the **model fit**,mainly assessing ***linearity and proportional hazards assumption***. There are several tests and methods but here we demonstrate briefly the ones based on *residuals*.

  1. **Cox-Snell** $r_i's$:Assessing overall fit
  2. **Martingale** $r_i's$:Determing the functional form of variabales
  3. **Deviance** $r_i's$:Examining model accuracy and identifying outliers
  4. **Schoenfeld** $r_i's$:Testing PH assumption
  5. **Score** $r_i's$:Testing PH, identifying outliers

ANY QUESTIONS?