People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

MOHAMED KHIDER UNIVERSITY, BISKRA

FACULTY of EXACT SCIENCES

DEPARTMENT OF MATHEMATICS

Thesis Submitted in Partial Execution of the Requirements of the Degree of

Master In Mathematics

Option:STATISTICS

Presented by:

Azzouz fatiha sonia

Titled:

Simple and Mixed Censorship

Examination Committee Members:

Dr.	Sayah Abdallah	UMKB	Chairman
Dr.	Benatia Fatah	UMKB	Supervisor
Dr.	Touba Sonia	UMKB	Examiner

3^{th} Mai 2025

Dedication

I dedicate this humble work to

My dear mother.

My dear father .

My dear sisters and My dear brothers

and all my familly members.

To all my friends and coworkers. To everyone that helped me to finish this Master thesis.

REMERCIEMENTS

This work represents the conclusion of a long and challenging journey. Despite facing numerous obstacles in work, studies, I was able to complete it successfully.

First and foremost, I express my deepest gratitude to Almighty God for granting me the strength, courage, and patience to carry out this work and achieve success.

I am profoundly thankful to all those who supported and contributed to this accomplishment:

My heartfelt thanks go to Mr.Benatia Fattah, who was an outstanding advisor. His guidance, advice, and unwavering support were invaluable throughout this journey. I extend my sincere appreciation to him.

I also express my gratitude to the jury members Mr.Sayah Abdallah The Chairman and Ms.Touba Sonia The Examiner for accepting to review and discuss this Master thesis, offering their time and expertise.

Furthermore, I am deeply thankful to the professors in the Mathematics Department at Mohamed Khider University, whose dedication and knowledge shaped my academic journey during this Master's program.

Finally, I extend my heartfelt gratitude to my parents, siblings, and colleagues for their continuous encouragement and support throughout my university studies.

Thank you all.

Contents

List of Figures

List of Tables

Introduction

The assumptions needed to apply parametric methods are frequently not met in statistical studies, especially when the underlying distribution of the data is unclear or hard to ascertain. Researchers use what are referred to as non-parametric approaches in these circumstances. These statistical methods are adaptable and don't rely heavily on presumptions regarding the distribution of the data. When exploring the general behavior of data or estimating important functions, like the quantile function or survival function, without being limited by strict model structures, non-parametric approaches are particularly effective However, what occurs if we are unable to observe the phenomenon of interest in its entirety or if the data we gather is insufficient? This brings up the idea of censoring, which is a major problem in many real-world research projects. There are various types of censorship. One of the most prevalent is right-censoring, which happens, for instance, in studies on the length of unemployment: some people may still be unemployed at the end of the study period, giving us only a partial picture. Another type is left-censoring, which is seen in studies of chronic illnesses where the disease's actual onset occurs before the initial diagnosis is noted. Interval or double censoring is a more complicated situation that frequently occurs in engineering or technology testing, like when a machine is inspected on a regular basis and wenot precisely when a failure occurred, but only that it occurred between two inspection times.

and also this work is dedicated to leveraging non-parametric estimation techniques to handle doubly censored data, with a focus on estimating survival functions, density functions, and hazard rates. The methods and insights presented here aim to offer robust tools for analyzing incomplete data—an endeavor both statistically rich and practically essential.

Mixed censoring refers to a scenario where data is censored from both sides—left and right—making it a more complex form of censoring to handle. This type of censoring often arises in situations where both the starting and ending points of an event are unknown, and can be seen in

areas such as engineering, medical studies, and reliability testing. Addressing mixed censoring requires specialized non-parametric estimators that can provide accurate insights despite the incomplete nature of the data.

Chapter 1: In this opening chapter, we laid the foundation by reviewing the fundamental concepts of statistics that are essential for understanding the study. We covered key topics such as the cumulative distribution function, density function,probability space, Random Variable„ and probability distributions. These concepts form the theoretical basis for the estimation methods discussed in the following chapters.

Chapter 2: This chapter focuses on the concept of simple censoring, with particular emphasis on right-censoring, which is the most commonly encountered type in real-world applications. We introduced the main types of simple censoring, namely right-censoring, left-censoring, and interval censoring. Special attention was given to the Kaplan-Meier estimator, a widely used non-parametric method for estimating the survival function under right-censoring

Chapter 3: In this final chapter, we focus on mixed (or double) censoring, a more complex scenario where data is censored from both sides (left and right). We discuss the asymptotic properties of the non-parametric estimators used to analyze such data. Specifically, we examine the Patilea and Rolin Estimator and the Maximum Likelihood Estimator (MLE), both designed to handle mixed censoring

Chapter 1

A General Notion about Statistics

Statistics is the branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. It provides methods for making inferences and predictions based on limited information from data samples.

Statistics is broadly divided into two main branches:

1. Descriptive Statistics: Focuses on summarizing and describing data through measures like mean, median, standard deviation, and visualizations such as histograms.

2. Inferential Statistics: Involves drawing conclusions or making predictions about a population based on sample data, using techniques like hypothesis testing, confidence intervals, and regression analysis. [4] [9] [18]

This chapter is dedicated to a review of the basic notations in mathematical statistics, such as probability spaces, random variables, distributions, and the two types of estimation.

1.1 concepts and definitions:

1.1.1 probability space:

A probability space is a mathematical model used to represent random experiments or phenomena where the outcomes are uncertain. It consists of three componentsh.

Definition 1.1.1 *The sample space*(Ω)*:*

The set of all possible outcomes of the random experiment. Each outcome is considered an element of the sample space.

Definition 1.1.2 *The sigma-algebra ($\mathcal{F}$):*

A collection of events (subsets of the sample space) for which probabilities are defined. It includes the sample space itself, the empty set, and is closed under complement and countable unions and intersections.

Definition 1.1.3 *The probability function (P):*

A function that assigns a probability to each event in the sigma-algebra. It satisfies the following properties:

1. $P(\Omega) = 1$ (the probability of the sample space is 1),
2. $P(A) \geqslant 0$ for any event A(probabilities are non-negative),
3. P is countable additive, meaning that if A_1, A_2..... are disjoint events

, then $P(\sqcup_i A_i) = \sum_i P(A_i)$. [20]

The Sample Space:

The following definitions are from the book,Bernard Garel [2]

We will first discuss random experiments, which are defined by the fact that their outcomes cannot be precisely predicted and, when repeated under identical conditions, may yield different results. The set of all possible outcomes of such an experiment is denoted as

the set Ω is called the sample space (or fundamental space), representing all possible states or outcomes. An element ω of Ω is called an elementary event.

Sigma-Algebra:

let Ω be a sample space $\mathcal{F}$ sigma-algebra (or σ-algebra) is a non-empty collection $\mathcal{F}$ of subsets of Ω (i.e.,$\mathcal{F} \subseteq P(\Omega)$ where $P(\Omega)$ is the power set of Ω) that satisfies:

- $\Omega \in \mathcal{F}$ (the entire sample space is included).

- Closure under complementation: If $A \in \mathcal{F}$ then $A^C \in \mathcal{F}$
- Closure under countable unions: If $A_1, A_2 \ldots\ldots \in \mathcal{F}, A_n \in \mathcal{F}$.for all $n \in \mathbb{N}$ then $\cup_{n \in \mathbb{N}} A_n \in \mathcal{F}$

The pair $(\Omega, \mathcal{F})$ is called a probabilizable space.

Probability: As defined in Defnition 1.1.3

The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space.

1.1.2 Random Variable

A random variable is a mathematical function that assigns numerical values to the outcomes of a random experiment. It allows the transformation of abstract events into quantifiable data for probabilistic analysis.

In another definition, A real random variable X is a function defined on a measurable space, taking values in the set of real numbers $\mathbb{R}$ and it is measurable with respect to the Borel δ-algebra $\boldsymbol{B}(\mathbb{R})$ from Saporta [22], 2006

$$\begin{aligned} X: \quad (\Omega, F) &\to (\mathbb{R}, B_R) \\ A &\to X(A) \end{aligned}$$

Types of Random Variables

The following definitions are from the book Ross, S. M. (2014). [21]

Discrete Random Variables A random variable is discrete if it takes a finite or countable set of distinct values.

- Counting the number of heads obtained when flipping a coin three times.
- Common distributions include Binomial, Poisson, and Geometric distributions.

Continuous Random Variables A random variable is continuous if it can take on any value within an interval or across the real number line.

- Measuring the exact height of a person selected at random

- Frequently used distributions include Normal, Exponential, Uniform distributions.

1.1.3 statistical distribution:

from the book Krishnamoorthy[12] 2006

A statistical distribution describes how the values of a random variable are spread or distributed across different outcomes. It provides a detailed representation of the probability of various values or ranges of values that a random variable can take. This distribution can be discrete or continuous and is characterized by its probability mass function (PMF) in the case of discrete variables or probability density function (PDF) for continuous variables. Statistical distributions are essential for modeling, analyzing, and making inferences from data, as they help to quantify uncertainty and predict future outcomes based on observed patterns.

probability density function (PDF) The density distribution, is a function that describes the likelihood of a continuous random variable taking on a particular value. Unlike discrete distributions where probabilities are assigned to specific outcomes, a continuous distribution assigns probabilities over intervals. The probability that the random variable takes a value within a certain range is given by the area under the PDF curve within that range. The total area under the PDF curve is always equal to 1, reflecting the certainty that the random variable will take some value within its possible range.

For a continuous random variable X the probability that X lies within the interval $[a,b]$ is given by:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

where $f_X(x)$ is the probability density function of X and x represents a possible value of the variable.

cumulative distribution function (CDF): from Saporta [22], 2006

let X be a random variable. The cumulative distribution function (CDF) of X is defined as a function from $(\mathbb{R}, B_R)$ to the interval [0,1] , given for every $x \in \mathbb{R}$ by

$$F(x) = P(X \leq x)$$

1. $F(x)$ is a non-decreasing function on $\mathbb{R}$
2. $F(x)$ is continuous from the right at every point in $\mathbb{R}$
3. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$

Survival Function The survival function, denoted by $S(x)$ or sometimes written as $\overline{F}\ (x)$ is defined on $\mathbb{R}+$ (the set of non-negative real numbers) as:

$$S(x) = 1 - F(x) = P(X > x)$$

For a time t this function represents the probability of "surviving" or not experiencing the event until time t

1. $\overline{F}\ (x)$ is also called the tail function
2. $\overline{F}\ (x)$ is a monotonically decreasing function.
3. $\overline{F}\ (x)$ is a left-continuous function.

Estimation from [5] cour yahya djabrane 2024

Parametric Statistics: Parametric statistics is the "classical" approach to statistics. We have a sample $X_1, X_2, ..., X_n$ of observations drawn from a population X The goal is to estimate a function or quantity related to this population (such as mean, variance, density, distribution, etc.) based on the sample $X_1, ...X_n$In this approach, it is assumed that the function to be estimated is known, except for a vector of parameters.

Example 1.1.1 *Consider a sample* $(X_1, ..., X_n)$*of i.i.d. observations from a normal distribution* $\boldsymbol{N(m, \sigma^2)}$*Estimating the mean* $\boldsymbol{m}$ *corresponds to estimating the probability density function:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \quad \exp(-\frac{(x-m)^2}{2\sigma^2})$$

This is an example of parametric estimation.

However, often:

- We do not assume a specific parametric form for the function to estimate. For example, if we want to study the average size of a dwelling Y as a function of salary X : $m(x) = E[Y \mid X = x]$ we don't assume any specific distribution or functional form for this relationship.
- Nonparametric statistics (SNP) involves cases where we do not make any assumptions about the form of the distribution of random variables. For example, if

  $F = \{ f : [0,1] \rightarrow \mathbb{R}$ is increasing $\}$we are engaging in nonparametric estimation, where we estimate the density function f without assuming any specific model for it.

When to use nonparametric methods (SNP):

- When it is difficult to fit the observations to a parametric distribution.
- When there is no clear model to use or if you prefer not to impose a prior assumption on the model.
- When it is unclear how many components should be included in the model.
- When dealing with a high-dimensional problem where parametric models are impractical due to the large number of parameters

  Advantages and Disadvantages:

- Less prior information is required about the observations.
- More general models that are more robust and flexible
- Slower convergence rates: nonparametric methods generally require more data to achieve the same level of precision as parametric methods.

Empirical Distribution Function (EDF): Suppose we observe a sample $X_1, X_2, ..., X_n$from a real-valued random variable with a cumulative distribution function (CDF) F :

$$F(x) = P(X \leq x)$$

The natural estimator for F , called the empirical distribution function F_n is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

where $I(X_i \leq x)$is an indicator function that equals 1if $X_i \leq x$ and 0 otherwise.

Formally, the EDF is given by:

Let n be the sample size, and let the random sample be:

$$X_1, X_2, X_3 \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot$$

We sort these observations in ascending order:

$$X_1 \leq X_2 \leq X_3 \leq \cdot \cdot \cdot \cdot \cdot \cdot \leq X_n$$

The i-th order statistic, denoted$X_{(i)}$is the value that occupies the i-th position in this sorted list, i.e., the i-th smallest value in the sample.

$$F_n(x) = \begin{cases} 0 & \text{If } X(1) > x \\ \frac{i}{n} & \text{If } X(i) \leq x \leq X(i+1) \quad \text{for } i = 1, 2, ...n \\ 1 & \text{If } X(n) < x \end{cases}$$

where $X(1)$ is the minimum and $X(n)$ is the maximum of the sample $X_1, X_2, ..., X_n$

It is clear that F_n is a nonparametric estimator of the CDF F

Propertie 1.1.1

1. Bias of $F_n(x)$:

   The bias of the empirical distribution function $F_n(x)$is defined as the difference between the expected value of $F_n(x)$and the true CDF$F(x)$To calculate the bias:

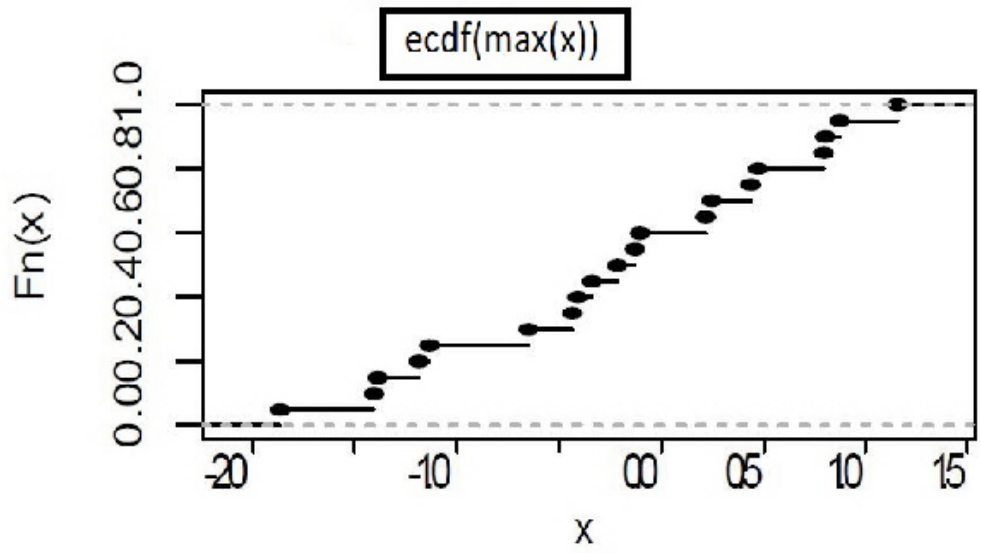

Figure 1: Distribution Empirique d'une va $N(0,1)$

Figure 1.1: ghraph 1

$$E(F_n(x)) = E\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right) = \frac{1}{n}\sum_{i=1}^{n} E(I_{\{X_i \leq x\}}) =$$
$$= \frac{1}{n}\sum_{i=1}^{n} p(X_i \leq x) = \frac{1}{n} \times n\ F(x)$$
$$= F(x)$$

This is equal to the probability $P(X \leq x) = F(x)$ so:

$$Bias(F_n(x)) = E\left(F_n(x) - F(x)\right) = 0$$

Thus, the bias of $F_n(x)$ is zero, meaning$F_n(x)$ is an unbiased estimator of $F(x)$

2. Variance of $F_n(x)$

Next, the variance of $F_n(x)$is computed. We start with the squared expected value:

$$E\left(F_n^2(x)\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right)^2\right)$$

This expands into:

$$\frac{1}{n^2}\left[\sum_{i=1}^{n} E(I(X_i \leq x))^2 + \sum_{i \neq j} E(I(X_i \leq x)(X_j \leq x))\right]$$

Evaluating each part:

$$E(I(X_i \leq x)) = F(x)$$

Thus, the variance is:

$$Var\left(F_n(x)\right) = \frac{1}{n}F(x)(1 - F(x))$$

This shows that the variance of $F_n(x)$ decreases as n increases.

3. Convergence of $F_n(x)$(Probability Convergence):

   as $Var\left(F_n(x)\right) \to 0$ (i.e., the variance becomes smaller as n grows), by Chebyshev's inequality, we can conclude that:

$$F_n(x) \to F(x) \text{ in probability as } n \to \infty$$

4. Convergence Distribution:

   For all x the empirical distribution function $F_n(x)$converges to the true CDF $F(x)$ almost surely (with probability 1), which is a consequence of the strong law of large numbers (SLLN).

5. Central Limit Theorem (CLT):

   According to the Central Limit Theorem, we have:

$$\sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{d} \boldsymbol{N(0,}F(x)(1 - F(x)))$$

   This means that the difference between $F_n(x)$ and $F(x)$ scaled by $\sqrt{n}$ converges in distribution to a normal distribution with mean 0 and variance$F(x)(1 - F(x)$

6. Glivenko-Cantelli Theorem:

The Glivenko-Cantelli Theorem provides the following result:

$$\sup_x |F_n(x) - F(x)| \xrightarrow{p} 0 \text{ as } n \to \infty$$

This means that $F_n(x)$ converges uniformly to $F(x)$almost surely.

7. Kolmogorov's Theorem:

   Kolmogorov's theorem on uniform convergence implies that:

$$P\left(\sqrt{n}\sup_x |F_n(x) - F(x)| \leq z\right) \to e^{-2z^2} \text{ as } n \to \infty$$

   This result provides a probabilistic bound on the uniform convergence of $F_n(x)$ to $F(x)$

Application The empirical survival function: $\overline{F}_n(t)$ is defined as:

$$\overline{F}_n(t) = \frac{1}{n}\sum_{j=1}^{n} I(X_j > t)$$

where t$\in \mathbb{R}$

Key Properties of $\overline{F}_n(t)$:

- $\overline{F}_n(t)$ serves as the empirical estimate of the survival function. The indicator function $I(X_j > t)$ is 1 if $X_j > t$ and 0 otherwise.

  where each term $I(X_j > t)$ follows a Bernoulli distribution with parameter$p = P(X_j > t) = \overline{F}(t)$ Therefore, the sum of these Bernoulli random variables:

$$n\overline{F}_n(t) = \sum_{j=1}^{n} I(X_j > t)$$

  follows a Binomial distribution with parameters(n, p)

  Expectation of

$$E(\overline{F}_n(t)) = \frac{1}{n}E\left(\sum_{j=1}^{n} I(X_j > t)\right) = \frac{1}{n} \times n \times \overline{F}(t) = \overline{F}(t)$$

  Thus,$\overline{F}_n(t)$is an unbiased estimator of $\overline{F}(t)$

- Variance of $\overline{F}_n(t)$:

  The variance of $\overline{F}_n(t)$is given by:

$$Var(\overline{F}_n(t)) = \frac{1}{n^2} Var\left(\sum_{j=1}^{n} I(X_j > t)\right)$$

  Since the random variables$I(X_j > t)$are Bernoulli with parameter $p = \overline{F}(t)$we have:

$$Var(I(X_j > t)) = \overline{F}(t)(1 - \overline{F}(t))$$

  Thus, the variance of $\overline{F}_n(t)$becomes:

$$Var(\overline{F}_n(t)) = \frac{1}{n}\overline{F}(t)(1 - \overline{F}(t))$$

The empirical quantile :

The empirical quantile is an estimate of the quantile of a random variable X based on a sample of size n it is defined as follows:

The theoretical Quantile:

Let $Q(p)$be the quantile function associated with X

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, \text{for } p \in (0, 1)$$

where $F(x)$ is the cumulative distribution function (CDF) of X

the Empirical Quantile:

Given a sample $X_1, X_2, ...X_n$sorted in ascending order $\left(X_{(1)}, X_{(2)}, ...X_{(n)}\right)$ the empirical quantile $Q_n(p)$ is defined as:

$$Q_n(p) = F_n^{-1}(p) = X_{(j)} \qquad \text{for } \frac{j-1}{n} \leq p \leq \frac{j}{n}$$

where$F_n(x)$ is the empirical cumulative distribution function.

Expectation Estimation Using the Quantile Function:

The mathematical expectation of X can be expressed using quantiles:

$$E(x) = \int_0^1 Q(p)dq$$

Using the empirical quantile estimator, we obtain an estimate of

$$\hat{E}(x) = \frac{1}{n}\sum_{j=1}^{n} X_{(J)}$$

This estimation is simply the empirical mean of the sample.

Kernel Density Estimation (KDE): from the book Density Estimation for Statistics and Data Analysis" by B. W. Silverman [7]

"Kernel density estimation (KDE) is a fundamental non-parametric method for estimating the probability density function (PDF) of a random variable. Unlike parametric methods, which assume a specific form for the distribution, KDE constructs an estimate directly from the observed data, providing a flexible and smooth approximation of the underlying distribution.

The mathematical formulation of the KDE estimator is given by:

$$\hat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where:

- $\hat{f_h}(x)$ is the estimated probability density function at point x
- n is the number of observed data points
- $X_1, X_2, ..., X_n$ are the observed data points
- $K(\cdot)$ is the kernel function, which determines the shape of the local weighting
- h is the bandwidth parameter, which controls the smoothness of the estimated density.

Explanation of KDE Components:

1. Kernel Function$K(\cdot)$
    - The kernel function assigns weights to data points relative to the estimation point x
    - Common kernel choices include:

- Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

- Epanechnikov kernel:

$$K(u) = \frac{3}{4}(1 - u^2) \qquad \text{for } |u| \leq 1$$

- Other kernels include uniform, biweight, and triweight functions.

2. Bandwidth h

- The bandwidth is a crucial parameter that controls the smoothness of the density estimate.
- Small h →The estimate is too sensitive to fluctuations (overfitting, too much noise).
- Large$h$ → The estimate is oversmoothed (important details may be lost).
- .An optimal choice of h can be determined using methods like Silverman's rule of thumb or cross-validation.

Chapter 2

Simple Censorship

This Master thesis focuses on estimation in a censored model, starting with an introduction to the concept of censoring. In survival analysis and reliability studies, the interest lies in the time until a specific event occurs. This time is referred to as failure time, lifetime, survival time, or simply duration. It is a positive random variable, often assumed to have an upper bound. This could represent, for example, the lifespan of a patient after receiving treatment, the duration of unemployment, the time until a machine breaks down, or the age at which a child learns to perform a specific task. Often, for various reasons, the time of interest cannot be fully observed. This might occur due to losing contact with a patient, the beginning or end of the study period, and so on. These values are considered censored. While the censored values are unknown, they must still be accounted for to ensure accurate estimates and valid conclusions. Depending on the specific context, the statistical literature offers many methods for handling censored observations. Different types of censoring exist.

Definition 2.0.4 *(Definition of Censoring)*

Censoring: in statistics refers to a situation where the full data on an event is not fully observed due to some limitations in the data collection process. It occurs when the exact time of an event is unknown, but there is enough information to determine that the event either occurred before or after a certain time point. Censoring is commonly encountered in survival analysis and time-to-event studies.

The primary types of censoring are:

Right Censoring: The event of interest has not occurred by the end of the observation period or the individual left the observation before the end of the experiment.

Left Censoring: The event has already occurred before the subject was observed, but the exact time is unknown.

Interval Censoring: The event is known to have occurred within a specific time interval, but the exact time remains unknown.

In statistical modeling, censoring is addressed by utilizing the available information to estimate the probability of the event happening over time, even with incomplete data.[1] [14] [6]

In this section, we introduce the concept of censorship and its types, such as left-censoring, right-censoring, and interval-censoring.

The censoring variable, denoted by Y refers to the fact that the event of interest is not fully observed $\Leftrightarrow$Instead of observing the actual event time$X$we observe $Y$and we only know that:

- if $X > Y$, we have right censoring
- if $X < Y$, it's called left censoring
- If $Y_1 < X < Y_2$, it's referred to as interval censoring

  For a given individual J , we consider:

- X_J : the individual's true survival time
- Y_J : the censoring time, i.e., the time when observation ends
- Z_j : the actual observed duration, which is either the event time or the censoring time, depending on which came first

2.1 right censoring:

The lifetime is said to be right-censored if the individual has not experienced the event by their last observation. In the presence of right censoring, not all lifetimes(y)are fully observed; for some of them, we only know that they are greater than a certain known value. Let R be a random censoring variable. Instead of observing the variable Y, which we are interested in, we observe a pair of variables$(Z, \delta)$where $Z = \min(Y, R)$ and $\delta = I_{\{Y \leq R\}}$. δ is called the censoring indicator because its values inform us whether the observation is complete(if $\delta = 1$)or right-censored

Figure 2.1: RIGHT CENSORED

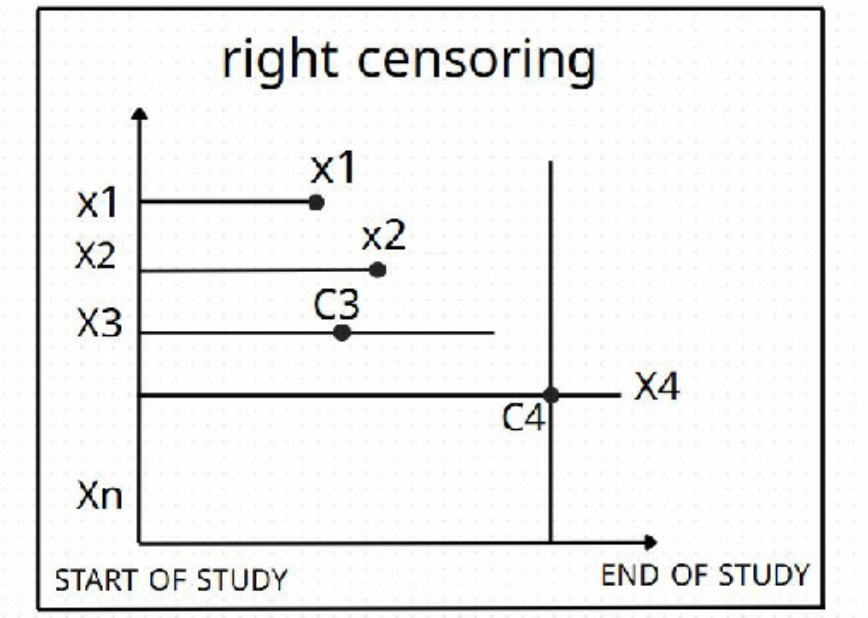

Figure 2.2: graph 2

(if $\delta = 0$)An illustrative example is when we are interested in the lifetime of a specific type of machine, but these machines break down if there is an electrical surge. Here, the machine's lifetime is right-censored by the moment at which the surge occurs." [10]

Example 2.1.1 *Consider a study designed to monitor the post-operative survival time of a group of patients. Each patient is followed for a maximum of five years or until death, whichever occurs first, starting from January 1, 2020. ,Suppose that, as of January 1, 2025—the end of the follow-up period—one patient remains alive and has not experienced the event of interest (death).In this scenario, we only know that the patient survived at least five years, but the exact time of death is unknown. Thus, the patient's survival time is considered right-censored.[25]*

2.2 Left Censoring:

refers to the situation where the individual has already experienced the event before they are observed. In this case, we only know that the event of interest occurred before a certain known value, represented by a random variable$L$ For each individual, we can associate a pair of random variables(Z, ∂) such that$Z = \max(Y, L)$and $\partial = I_{\{Y \geq L\}}$One of the first examples of left censoring found in the literature involves researchers observing baboons and their behavior of descending

Figure 2.3: left cnsored

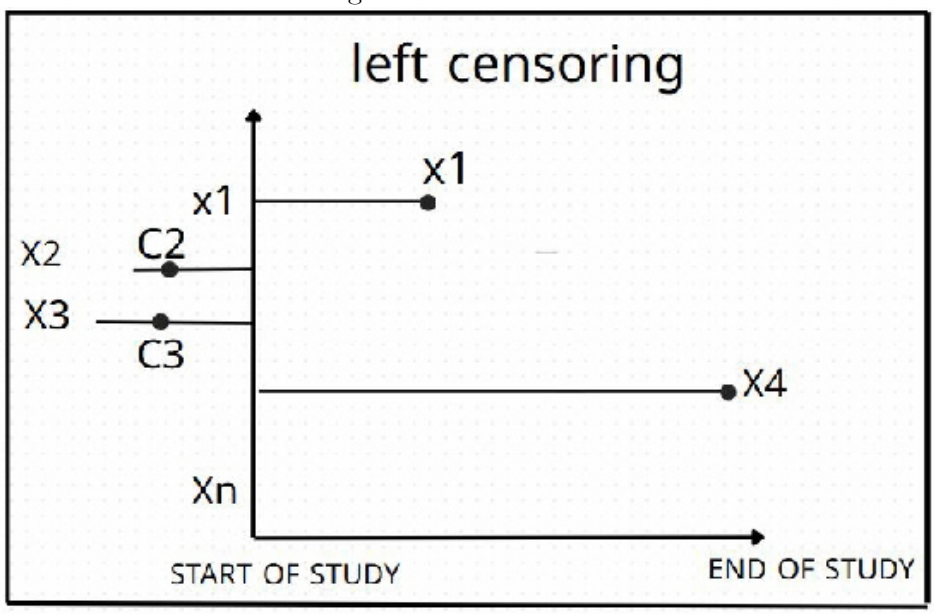

Figure 2.4: graph 3

from trees to eat (baboons spend the night in trees). The event of interest here is the time when the baboon descends from the tree. This time is observed if the baboon descends after the arrival of the observers. However, the data is censored if the baboon had already descended before the observers arrived. In this case, we only know that the time the baboon descended was before the observers arrived. Therefore, we observe the maximum of the baboon's descent time and the observers' arrival time.

The goal of a study on monkey behavior in a forest was to track how long the monkeys spent in trees over a given time frame. Although the study began on January 1, 2020, it was noted that some monkeys had already started to spend time in the trees before that date.

In this case, it is unclear exactly when the monkeys began to remain in the trees, but it is evident that they were there prior to the start of the study period. Their time spent in the trees is therefore regarded as left-censored because it took place before the observation period began[15].

2.3 Interval censoring:

Occurs when the exact time of an event is unknown, but it is known to lie within a specific time interval. This situation arises when data are collected at discrete time points, and the event

of interest is only observed between these points. For example, in medical studies, a patient may be examined periodically, and the event (e.g., disease progression) is only known to have occurred between two visits, but the precise time of the event is not observed.

This concept is central to survival analysis methods dealing with censored data, where the exact event times are not directly observed but are constrained within known intervals

In early childhood education, it's common to assess the age at which children acquire specific skills. This "time-to-event" refers to when a child successfully performs a particular task. However, some children may already have these skills upon entering a study, making it difficult to determine the exact age they acquired them. This situation is known as left censoring in survival analysis, where the event of interest has occurred before observation begins, but the precise timing is unknown.

Survival analysis deals with situations where the exact timing of events isn't fully known. Left censoring happens when an event has already occurred before a subject joins the study, but the exact time is uncertain. This contrasts with right censoring, where the event hasn't occurred by the study's end, and interval censoring, where the event's timing is known to fall within a specific range.

To handle left-censored data effectively, specialized statistical methods are needed to accurately estimate event time distributions. Ignoring left censoring can lead to biased results, as it overlooks instances where the event has already occurred before observation[16]

2.4 Double censoring

For example, a study focused on the age at which children in an African community learn to perform certain tasks. At the start of the study, some children already knew how to perform the tasks being studied. In this case, we only know that the age at which they learned is earlier than their age at the start of the study. By the end of the study, some children had still not learned these tasks, and we only know that the age at which they will eventually learn the tasks is later than their age at the end of the study. The age at the start of the study (left-censoring variable$L$)

s obviously less than the age at the end of the study (right-censoring variable R)The age of interest is observed if it falls within the study period. We observe$Z = \max(\min(Y, R), L)$ with a

censoring indicator. This model was studied by Turnbull, who introduced an implicit estimator for the survival function of$Y$ which is the solution to a self-consistency equation.

Type I Censoring: Fixed The experimenter sets a fixed (non-random) end date for the experiment. The maximum duration of participation is then fixed (non-random) and, for each observation, it is the difference between the experiment's end date and the date the patient enters the study. The number of observed events, however, is random. This model is commonly used in epidemiological studies.

The number of observed events is random.

This model is commonly used in epidemiological studies.

Type II Censoring: Waiting The experimenter pre-defines the number of events to observe. In this case, the end date of the experiment becomes random, while the number of events remains non-random. This model is often used in reliability studies.

This model is frequently used in reliability studies.

Type III Censoring: Random This is typically the model used in therapeutic trials. In this type of experiment, the inclusion date of the patient into the study is fixed, but the end observation date is unknown (for example, it may correspond to the patient's hospital stay duration). Suppose $\{X_1, ...X_n\}$ is a sample from a positive random variable X, we say that there is random censoring of this sample if there exists another positive random variable Y with a sample$Y_1.....Y_n$ In this case, instead of observing the values ofX_{J5}, we observe a pair of random variables(Z_J, δ_J), where:

$Z_J = \min(X_J, Y_J)$

$\delta_J = I(X_J \leq Y_J)$

Here, δ_J is the censoring indicator, which determines whether X was censored or not:

If $\delta_J = 1$ the event duration is observed $(Z_J = X_J)$

if $\delta_J = 0$ the event is censored $(Z_J = X_J)$

In this case, the observed durations are incomplete.

This chapter focuses exclusively on random right censoring, which is widely used in real-world applications and thus deserves particular attention.[15]

The hazard rate

The hazard rate (or risk function) is defined as the instantaneous probability that an individual will experience the event of interest within a very short time interval, given that the individual has survived up to the start of the interval.

Formally, the hazard rate $\lambda(t)$ is expressed as:

$$\lambda(t) = \lim_{h \to 0} \frac{P(X \leq t + h \mid X \geq t)}{h} = \lim_{h \to 0} \frac{P(X \leq t + h \mid X \geq t)/P(X \geq t)}{h} = \frac{1}{P(X \geq t)} \frac{P(X \leq t + h \mid X \geq t)}{h}$$
$$= \lim_{h \to 0} \frac{P(t \leq X \leq t + h \mid X \geq t)}{h}$$

In other words, $\lambda(t)$ represents the probability that an individual will experience the event within a small time interval $[t, t + h]$ conditional on surviving up to time

This can also be written as:

$$\lambda(t) = \frac{f(x(t))}{S(x(t))}$$

$$\frac{f(x(t))}{1 - f(x(t))}$$

where

- $f(x(t))$ is the probability density function of the event time,t
- $S(x(t))$ is the survival function, representing the probability of surviving beyond time t [10]

Cumulative Hazard Rate:[23]

The cumulative hazard rate, also known as the cumulative hazard function at time t is obtained by integrating the hazard function from 0 to t

$$H(t) = \int_0^t \lambda(u)\ du = -\log S(t)$$

where $H(t)$ is the cumulative hazard at time$t$,and $\lambda(u)$ is the hazard function at time u,

The survival function can be derived from the cumulative hazard rate using the following relation:

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t \lambda(u)\ du\right)$$

2.5 Kaplan-Meier Estimator:

From the books [3] [8] [24]

let X_1.....X_n be a sample representing the durations of interest (assumed to be positive), with a distribution function F and C_1.....C_n be a sample representing the censoring times, assumed to be independent of the durations of interest, with a distribution function G In the random right-censorship model, instead of observing the duration of interest$X_i$we observe the minimum of the two values,$Z_i = \min(X_i, C_i)$

Censoring Indicatorδ_i : The censoring indicator δ_i takes the value of 1 if the event of interest is observed and 0 if it is censored. It is defined as: $\delta_i = I\{X_i \leq C_i\}$where X_i

is the observed event time, and$C_i$is the censoring time.

Survival or Reliability Data: In the context of survival analysis or reliability data, where the goal is to estimate the time until a specific event occurs, the distribution function F

is estimated using the Kaplan-Meier estimator, introduced by Kaplan and Meier (1958).This estimator is defined for values

$$F_n(z) = 1 - \prod_{i=Z_i \leq z} \left(\frac{N_i(Z_i) - 1}{N_i(Z_i)}\right)^{\delta_i}$$

Here $N_n(x) = \sum_{i=1}^n I\{Z_i \geq x\}$ represents the number of subjects at risk at time x for $z \geq Z_n$ there are several conventions to define $F_n(z)$

1. It can be defined as $F_n(z_n)$but this might not make F_n a proper distribution function if$Z_n$is censored.
2. It can be defined as 0
3. It can be left undefined.

Some properties of the Kaplan-Meier estimator

In survival analysis $\tilde{S}_n$ serves a similar purpose for incomplete data as the empirical distribution function does for standard data.

Bias and Convergence

The Kaplan-Meier estimator is slightly biased; generally

$$E(\tilde{S}_n(t)) < \bar{S}(t)$$

where $E(\cdot)$ represents the expectation. However, it is a consistent estimator For all $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\tilde{S}_n(t) - S(t)\right| \geq \varepsilon\right) = 0$$

Thus, it is asymptotically unbiased:

$$\lim_{n\to\infty} E(\tilde{S}_n(t) - S(t)) = 0$$

Auto-coherence: In the absence of censoring, an estimator for $S(t)$ is

$$\check{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(t_i>t)}$$

In the presence of censoring, we can still write:

$$\check{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\delta_i \Pi_{t_i>t} + (1-\delta_i)\, I_{t_i>t}\right)$$

However, the value of $I_{t_i>t}$ is not known for censored data.

If an estimator$\check{S}_n(t)$ of $S(t)$ is known, we can estimate the expectation of $I_{t_i>t}$ given that $\delta_i = 0$ and $t_i > s_i$

we have

$$E\left(I_{t_i>t} | \delta_i = 0 \text{ and } t_i > s_i\right) = \frac{P(t_i > t)}{P(t_i > s_i)}$$

Thus,

$$\check{E}\left(I_{t_i>t}\delta_i = 0 \text{and } t_i > s_i\right) = \frac{\check{S}_n(t)}{\check{S}(s_i)}$$

The Kaplan-Meier estimator has the property of being auto-coherent, i.e.,

$$\hat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i I_{t_i>t} + (1-\delta_i)\,\frac{\hat{S}_n(t)}{\hat{S}_n(s_i)}\right.$$

(Constraints on a Survival Function)

Given $\check{S}_n^0$we can iteratively calculate an estimate $\check{S}_n^k$by:

$$\check{S}_n^k(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i I_{t_i>t} + (1-\delta_i)\,\frac{\check{S}_n^{(k-1)}(t)}{\check{S}_n^{(k-1)}(s_i)}\right.$$

And we have:

$$\lim_{k\to n}\check{S}_n^k = \hat{S}_n$$

Limitations of the Kaplan-Meier Estimator: The Kaplan-Meier estimator has a discontinuous nature. In some situations, it may be necessary to smooth the estimator by applying a kernel function. It also doesn't account for uncertainties in the observed values of t_i (event times) and s_i (censoring times).

In medical statistics, these uncertainties are usually small, as event times like death or discharge are typically known precisely. However, when combined with the inherent variability of the data (which can be described by the standard deviation around the mean), this can affect the estimated variability from the Kaplan-Meier estimator. To correct for these effects, simulations, such as bootstrap methods, can help model these uncertainties and adjust the results accordingly.

The Kaplan-Meier estimator

The Kaplan-Meier estimator (EKM) is the most widely used method for estimating the survival function without making any assumptions about the distribution of survival times. It is also known as the Product Limit (PL) estimator because it is derived from the limit of a product.

How the Kaplan-Meier Estimator (EKM) is Constructed:

The Kaplan-Meier estimator is based on progressively calculating the probability of survival at each time point, considering the events that occur and the individuals who remain "at risk" at each moment .The mathematical formula for the estimated survival function $S(t)$at time t such

that $t_0 < t$

$$\begin{aligned} S(t) &= P(X > t, X > t_0) \\ &= P(X > t/X > t_0)S(t_0) \end{aligned}$$

We repeat the operation by choosing $t'' \leq t_0$ yielding:

$$S(t_0) = P(X > t_0 \ \mid \ X > t'')S(t'')$$

Therefore:

$$S(t) = P(X > t, X > t_0)P(X > t_0 \ \mid \ X > t'')S(t'')$$

When selecting the dates for conditioning, we choose those where an event (death or censoring) has occurred, i.e.,$T_{(i)}$We then estimate quantities of the form:

$$Pi = P(X > \ T_{(i)} \mid \ X > T_{(i-1)}$$

where P_i represents the probability of surviving during the interval $I_i = [T_{(i-1)}, \ T_{(i)}]$ given survival at the start of this interval. Let R_i denote the number of subjects at risk at time $T_{(i)}$ and M_i the number of observed deaths at this time. The probability $q_i = 1 - p_i$ represents the chance of dying during interval I_i conditional on being alive at the start of the interval. A natural estimator for qi is

$$\widehat{q_i} = \frac{M_i}{R_i} = \frac{d_j}{D_j} = \frac{number \ of \ deaths \ at \ time \ \ T_{(i)}}{number \ of \ \ subjects \ at \ risk}$$

Assuming no ties (i.e.$T_{(i)}$ are distinct), if $\delta(i) = 0$(indicating censoring at time$T_{(i)}$) then $M_i = 0$ In this case, we have:

$$\widehat{q_i} = \begin{cases} \frac{1}{R_i} \ \ if & \delta(i) = 1 \\ 0 \quad if & \delta(i) = 0 \end{cases}$$

Consequently $\widehat{p_i} = 1 - \widehat{q_i}$ becomes:

$$\widehat{p_i} = \begin{cases} 1 - \frac{1}{R_i} & if \quad \delta(i) = 1 \\ 1 & if \quad \delta(i) = 0 \end{cases}$$

$$\widehat{p_i} = (1 - \frac{1}{R_i})^{\,\delta(i)}$$

It is evident that $R_i = n - i + 1$ Thus, the Kaplan-Meier estimator (KME) for the survival function of the lifetime variable X is obtained:

$$\hat{S}_{km}(t) = 1 - \widehat{F}_{km}(t) = \begin{cases} \prod\limits_{i=T(i)\leq t} (1 - \frac{1}{R_i})^{\,\delta(i)} & if \quad t < T(n) \\ 0 & if \quad t \geq T(n) \end{cases}$$

Similarly, the KME for the survival function of the censoring variable$C$ is

$$\overline{G}_n(t) = 1 - \hat{G}_{km}(t) = \begin{cases} \prod\limits_{i=T(i)\leq t} (1 - \frac{1}{n-i+1})^{\,1-\delta(i)} & if \quad t < T(n) \\ 0 & if \quad t \geq T(n) \end{cases}$$

here $T(i)$ and $\delta(i)$ for $i = 1 \cdot \cdot \cdot \cdot \cdot n$ are such that $T(1) \leq T(2) \leq T(3) \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \leq T(n)$ and $\delta(i)$ are the corresponding indicator variables.

The Kaplan-Meier estimator can also be expressed in the following form:

$$\hat{S}_{km}(t) = \begin{cases} \prod\limits_{i=1}^{n} (1 - \frac{1}{n-i+1})^{\mathbb{I}\{T(i)\leq t\}} & if \quad t < T(n) \\ 0 & if \quad t \geq T(n) \end{cases}$$

$$\overline{G}_n(t) = \begin{cases} \prod\limits_{i=1}^{n} (1 - \frac{1}{n-i+1})^{\mathbb{I}\{T(i)\leq t\}} & if \quad t < T(n) \\ 0 & if \quad t \geq T(n) \end{cases}$$

2.6 Simulation

2.6.1 Kaplan-Meier estimator of the survival function

Clinical Study Example:

Title: Clinical Trial of HAART (Highly Active Antiretroviral Therapy) in HIV Patients

Study Objective: The goal of the study was to evaluate the effectiveness of Highly Active Antiretroviral Therapy (HAART) in improving CD4+ T cell counts and survival duration among HIV patients.

Study Design: A total of 200 HIV-infected patients were divided into two groups:

Group 1: 100 patients receiving Highly Active Antiretroviral Therapy (HAART).

Group 2: 100 patients receiving traditional antiretroviral treatment (ART) with lower efficacy.

Results:

Group 1: Showed a 45% increase in CD4+ T cell count after 6 months of treatment, and a 30% improvement in survival time compared to Group 2.

Group 2: Showed no significant improvement in CD4+ T cell counts and had a lower survival duration compared to Group 1.

Data for Group 1 (HAART):

8, 15, 24, 34, 45, 50, 60, 72, 84*, 100, 120*, 135, 150, 180*, 200*, 210, 225*, 240, 265*, 280, 300, 320*, 340*, 365*

(* indicates that these patients are still alive after the specified days in the trial).

Data for Group 2 (Traditional ART):

12, 18, 23, 31, 40, 52, 60, 70, 90*, 110, 130*, 145, 155, 170*, 190*, 210, 220*, 235, 250*, 270, 290, 310*, 330*, 350*

(* indicates that these patients are still alive after the specified days in the trial)[26].

Methodology

1. Sorting the Time Points: Arrange the event (or censoring) times t_1, $t_2, \cdots\cdots\cdots, t_n$ in ascending order such that $t_1 < t_2 < \cdots\cdots\cdots < t_n$

2. Calculating Events and Risk Set:

   - For each time point t_j , determine:
     - d_j :The number of events (such as relapses or deaths) that occur at time t_j
     - D_j The number of individuals at risk at time t_j (i.e., individuals who were alive or in remission up to time t_j)

3. Calculating the Conditional Probability: For each time point t_j calculate the conditional probability p_j of surviving between $t_{j-1} and\ t_j$

   where $p_j = 1 - \frac{d_j}{D_j}$ where $\frac{d_j}{D_j}$ is the event rate at time t_j

4. Calculating the Cumulative Survival Function: The cumulative survival function at time t_j is calculated as:

   $\hat{S}(t_j) = \prod_{k=1}^{j} p_k$ where $\hat{S}(t_j)$ represents the cumulative survival probability up to time t_j

HAART Group Data (Group 1) CODE IN R

```
library(survival)
df_haart <- data.frame(
time = c(8, 15, 24, 34, 45, 50, 60, 72, 84, 100, 120, 135, 150, 180, 200, 210, 225, 240, 265, 280, 300, 320, 340, 365),
status = c(1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0)
)
event_times <- sort(unique(df_haart$time[df_haart$status == 1]))
n_at_risk <- sapply(event_times, function(t) sum(df_haart$time >= t))
n_died <- sapply(event_times, function(t) sum(df_haart$time == t & df_haart$status == 1))
death_prob <- n_died / n_at_risk
surv_step <- 1 - death_prob
cum_surv <- cumprod(surv_step)
result <- data.frame(
Time = event_times,
Died = n_died,
AtRisk = n_at_risk,
‘d/n‘ = round(death_prob, 3),
‘1 - d/n‘ = round(surv_step, 3),
‘Survival Probability‘ = round(cum_surv, 4)
```

```
)
print(result)
km_fit <- survfit(Surv(time, status) ~1, data = df_haart)
plot(km_fit,
xlab = "Days",
ylab = "Survival Probability",
main = "Kaplan-Meier Survival Curve - HAART Group",
col = "blue",
lwd = 2)
grid()
```

Even in these conditions we can calculate the **Kaplan-Meier** estimates as summarized in this Table1:

	Time to event (t)	Died (d)	AtRisk (n)	P(death) $=\frac{d}{n}$	P(survival)$=1-\frac{d}{n}$	P(Survival at L)
1	8	1	24	0,042	0,958	0,9583
2	15	1	23	0,043	0,957	0,9167
3	24	1	22	0,045	0,955	0,8750
4	34	1	21	0,048	0,952	0,8333
5	45	1	20	0,050	0,950	0,7917
6	50	1	19	0,053	0,947	0,7500
7	60	1	18	0,056	0,944	0,7083
8	72	1	17	0,059	0,941	0,6667
9	100	1	15	0,067	0,933	0,6222
10	135	1	13	0,077	0,923	0,5744
11	150	1	12	0,083	0,917	0,5265
12	210	1	9	0,111	0,889	0,4680
13	240	1	7	0,143	0,857	0,4011
14	280	1	5	0,200	0,800	0,3209
15	300	1	4	0,250	0,750	0,2407

Table 2.1: Kaplan Miere estimate group 1

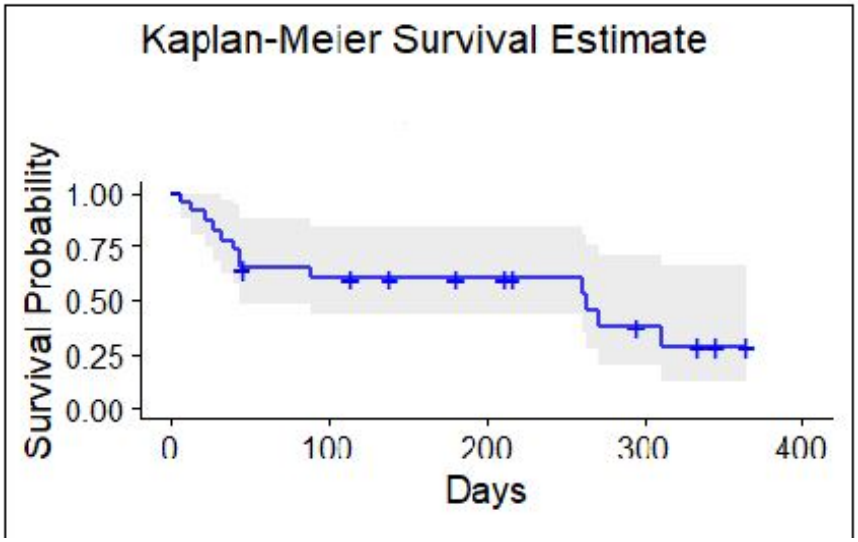

Figure 2.5: KPLAN MEIER HAART

Group 2 (Conventional Treatment - ART) CODE IN R

library(survival)

df_art <- data.frame(

time = c(12, 18, 23, 31, 40, 52, 60, 70, 90, 110, 130, 145, 155, 170, 190, 210, 220, 235, 250, 270, 290, 310, 330, 350),

status = c(1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0)

)

event_times <- sort(unique(df_art$time[df_art$status == 1]))

n_at_risk <- sapply(event_times, function(t) sum(df_art$time >= t))

n_died <- sapply(event_times, function(t) sum(df_art$time == t & df_art$status == 1))

death_prob <- n_died / n_at_risk

surv_step <- 1 - death_prob

cum_surv <- cumprod(surv_step)

result <- data.frame(

Time = event_times,

Died = n_died,

AtRisk = n_at_risk,

‘d/n‘ = round(death_prob, 3),

‘1 - d/n‘ = round(surv_step, 3),

‘Survival Probability‘ = round(cum_surv, 4)

)

print(result)

km_fit <- survfit(Surv(time, status) ~1, data = df_art)

plot(km_fit,

xlab = "Days",

ylab = "Survival Probability",

main = "Kaplan-Meier Survival Curve",

col = "red",

lwd = 2)

grid()

Even in these conditions we can calculate the **Kaplan-Meier** estimates as summarized in this Table 2

	Time to event (t)	Died (d)	AtRisk (n)	P(death) $=\frac{d}{n}$	P(survival) $=1-\frac{d}{n}$	P(Survival at L)
1	12	1	24	0,042	0,958	0,9583
2	18	1	23	0,043	0,957	0,9167
3	23	1	22	0,045	0,955	0,8750
4	31	1	21	0,048	0,952	0,8333
5	40	1	20	0,050	0,950	0,7917
6	52	1	19	0,053	0,947	0,7500
7	60	1	18	0,056	0,944	0,7083
8	70	1	17	0,059	0,941	0,6667
9	110	1	15	0,067	0,933	0,6222
10	145	1	13	0,077	0,923	0,5744
11	155	1	12	0,083	0,917	0,5265
12	210	1	9	0,111	0,889	0,4680
13	235	1	7	0,143	0,857	0,4011
14	270	1	5	0,200	0,800	0,3209
15	290	1	4	0,250	0,750	0,2407

Table 2.2: Kaplan mier group 2

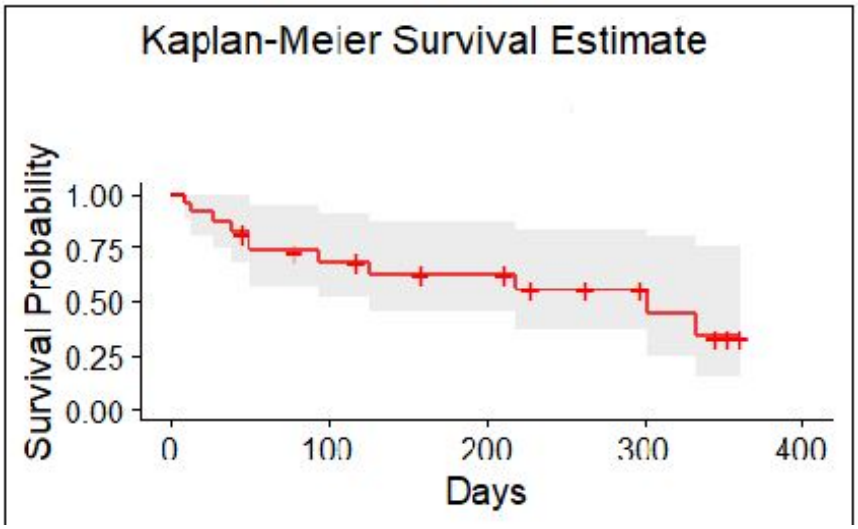

Figure 2.6: KPLAN MIEIER ART

Comparative Survival Analysis: HAART vs. ART We will compare two groups of patients—those receiving highly active antiretroviral therapy (HAART) and those receiving conventional antiretroviral therapy (ART)—using Kaplan-Meier survival curves. The goal is to assess whether there are significant differences in survival duration between the two treatment approaches.

CODE IN R

```
group1 <- c(8, 15, 24, 34, 45, 50, 60, 72, 84, 100, 120, 135, 150, 180, 200, 210, 225, 240, 265, 280, 300, 320, 340, 365)

status1 <- c(1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0)

group2 <- c(12, 18, 23, 31, 40, 52, 60, 70, 90, 110, 130, 145, 155, 170, 190, 210, 220, 235, 250, 270, 290, 310, 330, 350)

status2 <- c(1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0)

X <- c(group1, group2)

D <- c(status1, status2)

t <- c(rep("HAART", length(group1)), rep("ART", length(group2)))

f <- data.frame(X, D, t)

library(survival)

s <- survfit(Surv(X, D) ~t, data = f)

plot(s, lty = c(1, 2), col = c("blue", "red"), xlab = "Time", ylab = "Survival Probability")

legend("topright", legend = c("HAART", "ART"), lty = c(1, 2), col = c("blue", "red"))
```

This R code performs survival analysis on a dataset that includes survival times, censoring indicators, and treatment types. The dataset is created by combining three variables: X, D, and t.

X contains the survival times for each observation, where each observation represents a patient. D includes the censoring indicators, which indicate whether an observation was censored or not. A value of 1 means the observation was censored, while 0 means it was not.

t is a categorical variable indicating which treatment each observation received. In this case, there are two treatments: HAART and ART. The first group of patients received HAART, while the second group received ART.

The code then loads the 'survival' library and uses the 'survfit' function to fit a Kaplan-Meier survival curve to the data. The formula specifies that the survival object should be modeled based on the treatment variable t.

Finally, the 'plot' function is used to generate a plot of the survival curves, and the 'legend' function is used to add a legend to the plot.

The resulting plot shows the estimated survival probabilities for each treatment over time, and the legend indicates which line corresponds to each treatment. Overall, the code is used to compare the survival properties of two different treatments and visually display the results.

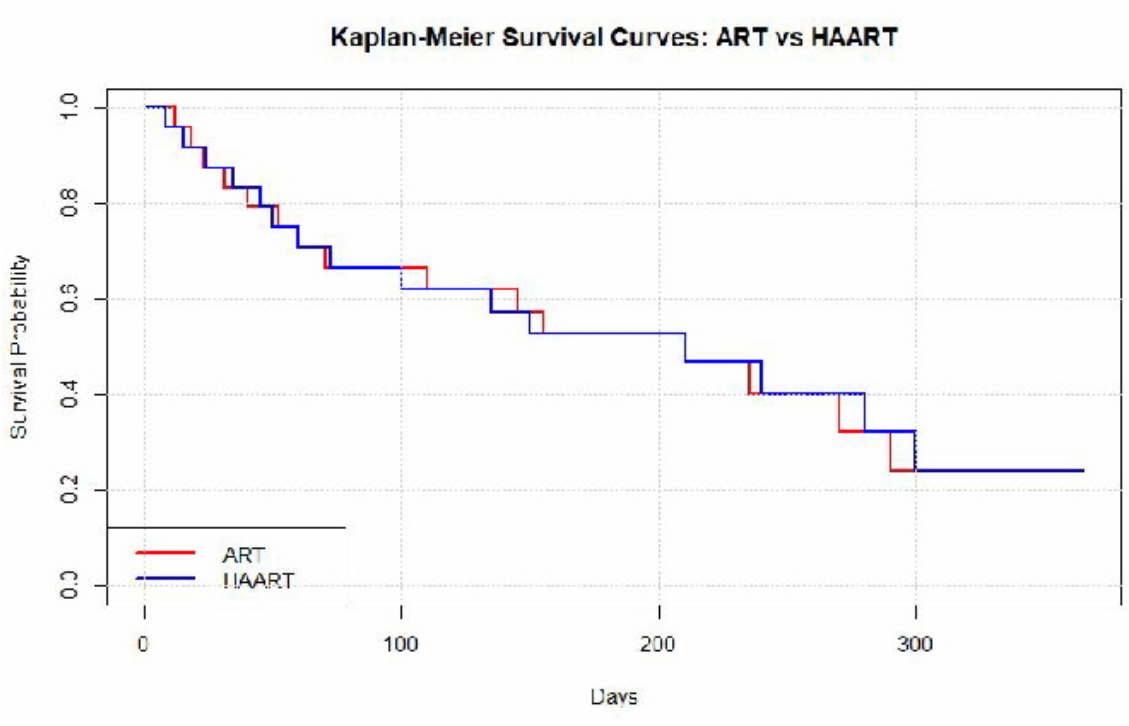

Figure 2.7: KAPLAN MIEIER FOR 2 GROUP

Survival analysis using Kaplan-Meier curves showed a clear difference between the two groups: patients who received Highly Active Antiretroviral Therapy (HAART) and those who received traditional Antiretroviral Therapy (ART).

The survival curves for the HAART group remained consistently higher throughout the follow-up period, indicating a greater chance of survival compared to the ART group.

This difference was evident through a vertical gap between the two curves at various time points, where the proportion of surviving patients was higher in the HAART group. A horizontal gap was also observed, showing that HAART patients took longer to reach the same mortality rate

seen in the other group.

These findings suggest that HAART not only increases the likelihood of survival but also delays death compared to traditional ART, supporting its effectiveness as a better treatment option for HIV patients.

Chapter 3

Mixed Censorship

Survival analysis experienced significant development in the second half of the twentieth century after Kaplan and Meier introduced their famous estimator of the survival function for right-censored data. This estimator generalizes the complement to one of the empirical distribution function. Later, in 2006, Patilea and Rolin expanded this framework to include more complex cases such as mixed censoring, where censoring occurs from both the left and the right. These are situations that the Kaplan-Meier estimator does not adequately address. In this context, they introduced new statistical tools and more sophisticated models to extend the scope of survival data analysis to cover these more complex censoring scenarios.

in this section, we will address all aspects related to the concept of mixed censoring in survival analysis, such as the properties of mixed censoring and the estimator of the mixed censoring distribution (Patilea and Rolin). Below is a simplified definition of mixed censoring:

3.1 Definiton of Mixed censoring :

Mixed censoring refers to a situation in survival analysis where one or two types of censoring occur simultaneously within the same dataset. Typically, this includes right censoring—where the exact time of the event is unknown because the event did not occur during the observation period—and left censoring—where the event has already occurred before the observation begins. This dual occurrence complicates the analysis of time-to-event data, as it introduces additional uncertainty in the estimation of survival functions and hazard rates. Specialized statistical methods are required to properly handle these varying censoring mechanisms, ensuring that the

resulting estimates are both accurate and reliable.[25]

With another definition [13]

It is said that there is mixed censoring when two types of censorship (one on the left and the other on the right) can prevent the observation of the phenomenon of interest, without necessarily being able to determine the interval to which it belongs. Instead of observing a sample of the variable of interest$Y$ we observe a sample of the pair $(Z; A)$

$$Z = \max(\min(Y; R); L)$$

and A write As in the model described in the article by [Patilea and Rolin (2006)].[19]

$$A = \begin{cases} 0 \; if \quad L < Y < R \\ 1 \; if \quad L < R < Y \\ 2 \; if \quad \min(Y; R) \leq L \end{cases}$$

Here L and R denote the left-censoring and right-censoring times, respectively, and A is the censoring indicator that distinguishes among three types of observations

- $A = 0$ indicates an exact observation (with Y observed within the interval $(L; R)$)
- $A = 1$ corresponds to right censoring (where we know only that $Y > R$)
- $A = 2$ corresponds to left censoring (where we know only that $Y \leq L$)

To highlight the difference between one-sided censoring, such as right or left censoring, and mixed censoring, the following table presents the key characteristics of mixed censoring

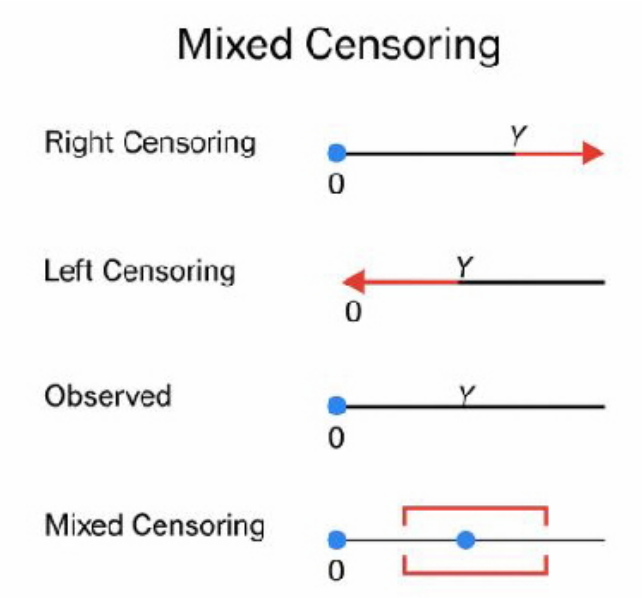

Figure 3.1: graph 4

Property	Explications
Unbiasedness	Estimators like the Kaplan–Meier are unbiased when dealing with right -censored data, but they may become biased when applied to mixed or doubly censored datasets
Consistency	Turnbull's nonparametric estimator is known to be consistent even under complex censoring structures, including mixed or double censoring
Non-parametric flexibility	Survival functions can be estimated without needing to assume a, particular distributionusing methods such as Turnbull or Patilea–Rolin.
Bias risk	When left censoring is heavy or censoring types are unbalanced, bias may be introduced into the estimation process.
Use of iterative algorithms	For complex censoring scenarios, iterative methods like the EM algorithm are employed to achieve accurate estimates.

Table 3.1: The properties of mixed censoring

1. Strong Uniform Consistency:

$$\sup_{t\in[a,b]} \left|\hat{S}_n(t) - S(t)\right| \rightarrow^{a,s} 0 \qquad as\ n \rightarrow \infty$$

where $[a, b]$is a condensed subset of T, is support, and the convergence is almost certain.

2. Asymptotic Normality: In the presence of regularity

$$\sqrt{n}\left(\hat{S}_n(t) - S(t)\right) \rightarrow^{d} N(0, \delta^2(t))$$

where $\delta^2(t)$ depends on the underlying censoring and survival distributions.

3. Turnbull's estimator bounds are provided by the Patilea-Rolin estimator.

$$\hat{S}_n^{lower}(t) \leq \hat{S}_n^{Turnbull}(t) \leq \hat{S}_n^{upper}(t)$$

The modified product-limit formulas are applied to intervals that are defined by the observed censoring patterns in order to construct these bounds.

4. Bootstrap Validity: Given bootstrap samples $\{(X_i^*, \delta_i^*)\}$, the estimator $\hat{S}_n^*(t)$ satisfies:

$$\sqrt{n}\left(\hat{S}_n^*(t) - \hat{S}_n(t)\right) \rightarrow^{d} N(0, \hat{\delta}^2(t))$$

where convergence is conditional on the sample, or in the bootstrap sense

.

Non-parametric estimation for mixed censoring model

For situations in which a censoring mechanism is applied to the observations T_i a new class of estimators is presented. Patilea and Rolin (2006) discussed this model, which is based on non-parametric estimation.

this is due to the fact that the estimator of the distribution of kaplan-Meier is no longer valid in this case hence the use of the new Patilea and Rolin estimator

3.2 Patilea and Rolin Estimator:

Examine a sample (Z_i, δ_i) for $1 < i < n$ of the pair (Z, δ) where Z is defined as follows:

$$Z = (T \wedge C) \vee L = \max(\min(T, C), L)$$
$$for \quad X = (T \wedge C \ \ )$$

where T, C, L are independent positive random variables representing the variable of interest, the left-censored variable, and the right-censored variable, respectively

Assume that H is the distribution function of Z and that $H^{(0)}$ represents the sub-distribution for the uncensored observations. The following expressions supply these.

$$H(t) = P(Z \leq t) = F_L(t)F_X(t)F_C(t)(1 - S_T(t)S_C(t))$$
$$Uncensored\ data\ \ (\delta = 0)\ \rightarrow\ \ H^{(0)}(t) = P(Z \leq t,\ \delta = 0)\ = \int_0^t F_L(x)S_C(x) \quad dF_T(x)$$

Right-censored data ($\delta = 1$)

$$H^{(1)}(t) = P(Z \leq t,\ \delta = 1)\ = \int_0^t F_L(x)F_T(x) \quad dF_C(x)$$

Left-censored data ($\delta = 2$)

$$H^{(2)}(t) = P(Z \leq t,\ \delta = 2)\ = \int_0^t S_T(x)F_C(x) \quad dF_L(x)$$

The following are their corresponding empirical versions:

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq t\}}$$

$$H_n^{(0)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq t, \delta_i = 0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq t, \quad T_i - C_i \leq 0, \quad L_i - T_i \leq 0\}}$$

$$H_n^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \le t, \delta_i = 1\}}$$

$$H_n^{(2)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \le t, \delta_i = 2\}}$$

Let Z_J for $(1 \le j \le M)$ represent the different observed values of Z_i arranged in ascending order. For each $k \in \{0, 1, 2\}$ define:

$$D_{kj} = \sum_{i=1}^{n} \mathbb{I}\left\{Z_i = Z'_J\ ,\ \delta_i = k\right\}$$

The following form represents n, the bounded product estimator put forth by Patilea and Rolin (2006)

$$\overline{S}_n(t) = 1 - \widetilde{F}_n(t)$$
$$\overline{S}_n(t) = \prod_{j/Z'_J \le t} (1 - \frac{D_{0J}}{\dot{F}(Z'_{J-1}) - nH_n(Z'_{J-1})})$$

$\dot{F}_n$ stands for the Kaplan–Meier estimator of the distribution functionF_L, which is created by time reversal, or by inverting time in the manner described below:

$$\dot{F}_n(t) = \prod_{j/Z'_J \le t} (1 - \frac{D_{0J}}{nH_n(Z'_J)})$$

Non parametric Estimation of Copulas under Double Censoring: [11]

Let $X = (X_1, X_2)$ be a pair of positive random variables with support $X = X_1 \times X_2$and joint distribution function$F$ let $R = (R_1, R_2)$ respectively $L = (L_1, L_2)$ be a pair of random variables for right censoring (respectively, left censoring). We consider X, L and R to be independent variables. When double censoring is used, we see independent copies $(Z_{1i}, Z_{2i}, A_{1i,}A_{2i},)\quad 1 \le i \le n$

of the vector $(Z_1, Z_2, A_{1,}A_{2,})$ where for each $K \in \{1, 2\}, Z_K = \max(\min(X_K, R_K), L_K)$ and A_K is the censoring indicator given by:

$$A_K = \begin{cases} 0 \; if \quad L_K < X < R_K \\ 1 \; if \quad L_K < R_K < X_K \\ 2 \; if \quad \min(X_K; R_K) \leq L_K \end{cases}$$

For any random variable $V, F_V, S_V, I_v T_v$,the following represent its distribution function, survival function, lower endpoint, and upper endpoint of the support of V , respectively Furthermore, we define $\varphi(t^-) = \lim_{\varepsilon \to 0^+} \varphi(t - \varepsilon)$ for any right-continuous function $\varphi = R \to R$ When it exists, the left-hand limit of φ at t when it exists Moreover, for any differentiable function $\psi = R^2 \to R$ we represent the partial derivative of $\psi$with respect to the first (or second) variable by $\partial_1 \psi (or \; \partial_2 \psi)$ respectively

We assume that the functions $F_{X_K}, F_{R_K}$and F_{L_K} (for $K \in \{1, 2\}$)are continuous.

The following notations must be introduced in order to define the empirical copula$C_n(U, V)$for$(U, V) \in [0, 1]^2$we need to introduce the following notations.for$K \in \{1, 2\}$and $j \in \{0, 1, 2\}$ denote by $H_K^{(j)}(t) = P(Z_K \leq t, A_K = j)$

when $A_K = j$ we take into account the sub-distribution function of Z_K which is represented by $H_K^{(j)}$ which is represented by $I_{H_K^{(j)}} = \inf\{t \in \boldsymbol{R} \; / \; H_K^{(j)}(t) > 0\}$This is the lowest value of t for which the sub-distribution function is larger than zero.The empirical representations of the overall distribution function and the sub-distribution function $H_K^{(j)}, F_{Z_K}$ are described as follows:

$$H_{nK}^{(j)}(t) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{I}_{\{Z_{Ki} \leq t \; , \; A_{Ki} = j\}}$$

$$\hat{F}_{Z_K} = \sum_{i=1}^{n} \boldsymbol{I}_{\{Z_{Ki} \leq t \; \}}$$

The indicator function in this case, represented by $\boldsymbol{I}_{\{\cdot\}}$is equal to 1 when the condition inside is true and 0 otherwise.let$(Z'_{Kj} \;)$ $1 \leq J \leq m$ where $m \leq n$ represent the unique observed values of

$(Z_{Ki})1 \leq i \leq n$ These are used to define the product-limit estimator $\hat{F}_{L_K}(t)$ of the function $F_{L_K}(t)$ as

$$\hat{F}_{L_K}(t) = \prod_{j/Z'_{Kj}>t} \left(1 - \frac{\sum_{i=1}^{n} \boldsymbol{I}_{\{Z_{Ki}=Z'_{Kj}\ ,\ A_{Ki}=2\}}}{n\hat{F}_{Z_K}(Z'_{Kj})}\right)$$

This estimator is conceptually similar to the Kaplan-Meier estimator, but it is constructed by reversing time.

Additionally, the product-limit estimator for S_{Rk} is given in reference [25]

$$S_{Rk}(t) = \prod_{i/Z_{ki}\leq t} \left(1 - \frac{\boldsymbol{I}_{\{A_{ki}=1\}}}{n\left(\hat{F}_{L_K}(Z_{\bar{K}i}) - \hat{F}_{Z_K}(Z_{\bar{K}i})\right)}\right)$$

The empirical distribution function of X cannot be directly constructed from the data due to its unobservability.

$$\tilde{F}_n(x_1x_2) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}_{\{X_{1i}\leq x_1\ ,\ X_{2i}\leq x_2\}}$$

The empirical distribution function cannot be used to estimate$F(x_1, x_2)$ since X is unobserved. Thus, in accordance with[27] and noting that.

$$\begin{aligned} &E\left[g(Z_1Z_2)\boldsymbol{I}_{\{A_i=0\}}\boldsymbol{I}_{\left\{Z_{1\leq x_1}, Z_{2\leq x_2}\right\}}\right] \\ &= E\left[\boldsymbol{I}_{\{X_1\leq x\ ,\ X_2\leq x_2\}}\right] \\ &= F(x_1\ ,\ x_2) \end{aligned}$$

$$\hat{g}(Z_{1i}Z_{2i})\boldsymbol{I}_{\{A_i=0\}}\boldsymbol{I}_{\{A_{2i}=0\}}\boldsymbol{I}_{\left\{Z_{1i\leq x_1}, Z_{2i\leq x_2}\right\}}$$

where

$$F_{n1}(x_1) = \lim_{x_2\to\infty} F(x_1, x_2)\ and\ F_{n2}(x_2) = \lim_{x_1\to\infty} F(x_1, x_2)$$

Conclusion

We looked at a number of nonparametric estimation topics in the context of censored data in my graduation Master thesis.

We covered some fundamental and basic ideas in Chapter One, including random variables and the distribution function. Simple censoring techniques, such as left- and right-censoring, were covered in the second section, along with how to use the Kaplan-Meier estimator to estimate these censored parameters. Alongside ideas like risk rate estimation

We showed in the last section that another nonparametric estimator, like the Patilea and Rolin estimator, can be applied when there is mixed censoring, or combined right and left censoring

Bibliography

[1] Applied Survival Analysis: Regression Modeling of Time-to-Event Data" by David W. Hosmer Jr. and Stanley Lemeshow

[2] Bernard Garel · 2002 .Modélisation probabiliste et statistique

[3] Breslow and Crowley, 1974; Gill, 1983

[4] Casella, G., & Berger, R. L. (2002). Statistical Inference (2nd ed.). Duxbury

[5] COUR YAHYA 2024

[6] Cox, D.R., & Oakes, D. (1984). Analysis of Survival Data

[7] Density Estimation for Statistics and Data Analysis" by B. W. Silverman (1986)

[8] Földes and Rejtö, 1981

[9] Freedman, D., Pisani, R., & Purves, R. (2007).

[10] Gannoun et al. (2003)

[11] Gribkova, S. and Lopez, O., Non-parametric Copula estimation under bivariate cen soring, Scandinavian Journal of Statistics 42, 925-946, (2015)

[12] Handbook of Statistical Distributions with Applications" by K. Krishnamoorthy (published in 2006),

[13] Idiou, N. and Benatia, F., Survival copula parameters estimation for Archimedean family under singly censoring, Advances in Mathematics: Scientific Journal 10, 1-4,(2021)

[14] Kalbfleisch, J.D., & Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data.

[15] [Khardani et al. (2011) and Collomb et al. (1987)

[16] Klein, J. P., & Moeschberger, M. L. (2003), "Survival Analysis: Techniques for Censored and Truncated Data

[17] Lopez, O. and Saint-Pierre, P., Bivariate censored regression relying on a new esti mator of the joint distribution function, Journal of Statistical Planning and Inference142, 2440-2453, (2012)

[18] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). Introduction to the Practice of Statistics (9th ed.). W.H. Freeman and Company

[19] Patilea & Rolin (2006)

[20] Ross, S. (2014). A First Course in Probability. 9th Edition. Pearson.

[21] Ross, S. M. (2014). Introduction to Probability Models (11th ed.). Academic Press.

[22] Saporta, G. (2006). ProbabilitÈs, analyse des donnÈes et statistique. Editions Technip.

[23] "Statistical Methods in Medical Research" by P. Armitage and T. Colton

[24] Stute and Wang, 1993; Winter et al., 1978

[25] Survival Analysis: Techniques for Censored and Truncated Data" by John P. Klein and Melvin L. Moeschberger

[26] Zhang F, Dou Z, Ma Y, Zhao Y, Liu Z, Bulterys M, Chen RY. Effect of earlier initiation of antiretroviral treatment and increased treatment coverage on HIV-related mortality in China: a national observational cohort study. Lancet Infect Dis. 2011 Jul;11(7):516–524. doi:10.1016/S1473-3099(11)70097-4

Appendixe A: Abbreviations and Notations

$v.a.r$:	$Real\ random\ variable(s)$
$i.i.d$:	$Independent\ and\ identically\ distributed$
$p.s$:	$Almost\ surely$
$\bar{F}$:	$S(t) = 1 - F(t) \qquad survival\ function$
F :	$Cumulative\ distribution\ function\ (CDF)$
$\mathbb{R}$:	$Set\ of\ real\ numbers$
$X \xrightarrow{d} Y$:	$Convergence\ in\ \ distribution$
S_X :	$survival\ function\ of\ \ X$
f :	$Probability\ density\ function$
F_n :	$Empirical\ cumulative\ distribution\ function$
I_A :	$Indicator\ function\ of\ \ the\ set\ A$
Ω :	$Sample\ space$
λ_T :	$Hazard\ function\ of\ T$
$\mathbb{N}$	$The\ set\ of\ natural\ numbers$

ملخص

في هذا البحث، تم التطرق إلى موضوع المراقبة في البيانات الإحصائية، حيث تم تقديم دالة البقاء كأداة أساسية في تحليل البيانات التي تحتوي على مراقبة، خصوصًا في حالة المراقبة اليمنى. تم استعراض مقدر كابلان-ماير الذي يُستخدم بشكل واسع لتقدير دالة البقاء في البيانات المراقبة يمينيًا. ومع ذلك، تبين أن هذا المقدر غير صالح في حالة المراقبة المختلطة، حيث يتم مراقبة البيانات من كلا الجانبين (اليسار واليمين). من هنا كانت الحاجة لتقديم مقدر باتيليا ورولين (2006)، الذي يتميز بقدرته على التعامل مع البيانات المراقبة المختلطة، ويحل مشكلة عدم صلاحية المقدر السابق في هذا السياق

الكلمات المفتاحية:.دالة البقاء, المراقبة اليمنى, المراقبة المختلطة . مقدر كابلان-ماير , مقدر باتيليا ورولين

Résumé

Dans cette étude, nous avons abordé la problématique de la censure dans les données statistiques, en présentant la fonction de survie comme un outil fondamental pour l'analyse des données censurées, notamment dans le cas de la censure à droite. L'estimateur de Kaplan-Meier, largement utilisé pour estimer la fonction de survie dans ce type de données, a été examiné. Cependant, il s'avère que cet estimateur n'est pas adapté dans le cas de la censure mixte, où les données sont censurées à la fois à gauche et à droite. Cela a donc motivé la présentation de l'estimateur de Patilea et Rolin (2006), qui est spécifiquement conçu pour traiter ce type de censure et pallie les limites de l'estimateur précédent dans ce contexte.

Mots-clés : Fonction de survie, censure à droite, censure mixte, estimateur de Kaplan-Meier, estimateur de Patilea et Rolin.

Abstract

In this study, the topic of censoring in statistical data was addressed, with the survival function presented as a fundamental tool for analyzing data affected by censoring, particularly in the case of right-censoring. The Kaplan-Meier estimator, widely used for estimating the survival function in right-censored data, was reviewed. However, it was found that this estimator is not suitable for mixed censoring, where data is censored from both sides (left and right). This highlighted the need to introduce the Patilea and Rolin (2006) estimator, which is specifically designed to handle mixed censored data and addresses the limitations of the previous estimator in this context.

Keywords: Survival function, right censoring, mixed censoring, Kaplan-Meier estimator, Patilea and Rolin estimator.