République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

*Université Mohamed Khider, Biskra*

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : statistique

**Par :**

Alimi Ferial

**Titre :**

# Test d'une différence de survie entre plusieurs échantillons.

Devant le Jury :

| | | | |
|---|---|---|---|
| Dr | DHIABI SAMRA | U. Biskra | Encadreur |
| Dr | ABDELLI JIHANE | U. Biskra | Président |
| Dr | OUANOUGHI YASMINA | U. Biskra | Examinatrice |

**Soutenu Publiquement le 26/06/2022**

# Dédicace

$A^{'}$ *Mes Chers Parents*,

$A'$ *ma tante Nadjah*,

$A^{'}$ *Mes Soeurs et Mon Frère*,

$A'$ *Ma Grand* − *Mère khadidja*.

# Remerciements

Avant tout, je tiens à remercier ALLAH Le Tout Puissant qui m'a aidée et donnée la santé, la patience et le courage durant ces longues années d'études.

Je tiens remercié sincèrement mon encadreur, Dr. Daibi Samra, qui s'est toujours montré l'écoute et très disponible tout au long de la réalisation de ce mémoire.

Je remercie également les membres du Jury pour avoir accepté d'évaluer et de juger ce modeste travail. Leurs questions et propositions ont permis de l'améliorer et l'enrichir.

Mes remerciements s'adressent également à tous les enseignants du département de mathématiques, qui m'ont aidée tout au long des années de ma scolarité, avec une mention spéciale au Fateh Benatia.

Enfin, je remercie du fond du cœur et avec grand amour mes chère parents (Sayeh, Mebarka) qui n'ont jamais cessé de croire en moi pendant toutes mes années d'études. Merci pour les sacrifices consentis à mon éducation, pour le soutien et surtout pour la patience. Merci aussi à ma sœur (Riham et Asma) qui m'ont toujours encouragé et a toutes ma familles et aussi à mon frère (Larbi), et ma tante (Nadjah), a mes amies surtout (Hana Legouirah).

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction

L'analyse de survie est une branche des statistiques qui consiste à analyser et etulier durée de vie attendue (de personne, d'appareil,...) d'analyser la durée. Attendue jusqu'à ce qu'un événement étudié (communément appelé "décès","panne") est le passage irréversible entre deux états (communément nommé "vivant" / "décès", ou fonction panne).

Ce sujet est appelé théorie de la fiabilité en ingénierie. Analyse de la durée ou modélisation de la durée en économie et analyse de l'histoire des événements en sociologie.

L'analyse des données de survie a pour première particularité de ne concerner que des variables aléatoires positive. Une deuxième particularité de cette analyse est l'existence par fois d'observations incomplète. En effet, les données sont souvant recueillies partiellement, notamment. à cause des proses de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information, ce qu'on appelle "données manquants".

La censure à droite est le phénomène le plus couramment rencontré lors du recueil de données de survie. On cherche alors à estimer la distribution des temps de survie,de plusieurs maniéres connues, dant en particulier la comparer les fonctions de survie de plusieurs groupes, ceci represente notre travail ici présente.

Le mémoire se Compares essenrillement de trois chapitres :

• Au premier chapitre on présente une introduction sur la distribution de survie et les résultats concernant la théorie des données manquauter particuliément la censure à droite.

• Dans le deuxième chapitre nous présentons les principaux résultat, concernant l'estimateur de Kaplan-Meier (1958) ; ainsi que l'espérance, la variance et les propriétés asymptotiques de cet estimateur.

• Dans le troisième et dernier chapitre ; on s'intéresse à une approche non-paramétrique. Qui consiste principalement a comparer le nombre de décès observés dans chacun de deux groupes (voir plusieurs) étudier comparativement au nombre de décès attendus (calculés sous l'hypothèse d'égalité). Nous avons utilisées pour l'extension du Log-Rank pour sa performance établie dans la litterature du domaine.

# Introduction

L'a plupart des recherches statistiques reposent généralement sur l'étude du comportement de différents phénomènes pendant des périodes précises et sur l'utilisation des résultats de ces études pour élaborer des recommandations appropriées et prendre des décisions adéquates. Notre travail est basé sur Analyse de la durée de survie.

En mathématique, la durée de vie est assimilée à une variable aléatoire non négative dont l'étude a reçu une attention particulière de la part des statisticiens. La base de toute analyse statistique est l'échantillon à qui il arrive parfois d'être censuré. Il existe plusieurs mécanismes de censure dont la plus couramment rencontrée est la censure aléatoire à droite.

L'analyse de survie à été généralisée à un ajustement pour des événements non uniques mais fréquents telle que les rechutes en cas de maladie. Les objectifs de l'analyse de survie sont : estimer, interpréter et comparer la fonction de survie à des distributions classiques.

Dans ce travail, on s'intéresse à l'estimation statistique non paramétrique en présence de données censurées. Dans ce cadre, l'estimateur introduit en 1958 par Kaplan-Meier voir [8] est très utile et constitue l'élément principal dans l'estimation de la fonction de survie. Le mémoire est organisé comme suit :

• Le premier chapitre se regroupe en trois sections. La section 1, porte sur quelques rappels et définitions sur la fonction de survie. Dans la section 2, on présente une petite introduction sur la fonction de survie. Dans la section 3, on présente un schéma des données incomplètes : données de censure et ses types (à droite, à gauche).

• Le deuxième chapitre est réservé à une synthèse sur les principaux estimateurs non-paramétriques de la fonction de survie et notamment celui de l'estimateur de Kaplan-Meier.

• Dans le troisième et dernier chapitre ; on s'intéresse à une approche non-paramétrique. Qui consiste principalement a comparer le nombre de décès observés dans chacun de deux groupes (voir plusieurs) étudier comparativement au nombre de décès attendus (calculés sous l'hypothèse d'égalité). Nous avons utilisées pour l'extension du Log-Rank pour sa performance établie dans la littérature du domaine.

# Chapitre 1

# Distribution de survie

L'analyse de survie est usuellement définie comme l'étude du délai de la survenue au cours du temps d'un événement d'intérêt, comme une panne de machine dans le domaine de la fiabilité, une rechute ou une rémission dans le domaine médical. Nous donnons ci-dessous les définitions des principaux outils utilisés en analyse de la survie (1.1). Nous allons définir les fonctions permettant de décrire une distribution de survie.

## 1.1 Définitions et carractéristiques

### Évènement d'intérêt

Le temps d'évènement est le temps correspondant à l'évènement se produisant le premier dans l'ordre chronologique, comme par exemple "rechute ou décès".

### Cohorte

On appelle cohorte un groupe de sujets inclus dans une étude en même temps, suivis dans des conditions standardisées pendant une durée prédéterminée.

### Durée de survie

C'est le délai entre la date d'origine , la date de survenue et la date des dernières nouvelles.

▷Date d'origine

La date correspondant au point de départ de la surveillance. Elle peut être différente pour chaque sujet en fonction de la façon dont le sujet est répertorié. Dans certains cas la date d'origine peut être antérieure à l'inclusion dans l'étude.

▷Date de point

C'est la date choisie pour l'évaluation.

▷Date des dernières nouvelles

Il s'agit de la date la plus récente à laquelle les informations sur le patient ont été collectées, y compris si l'événement d'intérêt s'est produit ou non.

### Perdu de vue

Une personne est perdue de vue lorsque son suivi est interrompu avant la date du point et que l'événement d'intérêt ne s'est pas encore produit.

### Temps de participation

Le temps de participation est voir simplement la durée de surveillance pour chaque sujet utilisée dans l'estimation de la survie. Trois cas :

▷L'événement a lieu au cours de la surveillance ce qui implique que le temps

de participation est égale à la date de survenue de l'évènement moins la date d'origine.

▷Le sujet est vivant à la date de point ce qui implique que le temps de participation est égale à la date de point moins la date d'origine.

▷Le sujet est perdu de vue ce qui implique que le temps de participation est égale à la date de dernière nouvelle moins la date d'origine.

### Temps de recul

C'est le temps entre la date d'origine et la date de point, c'est-à-dire (*c.à.d*) le temps maximum possible pour suivre l'événement. Ainsi, les renversements minimum et maximum d'une série de sujets déterminent l'ancienneté de cette série.

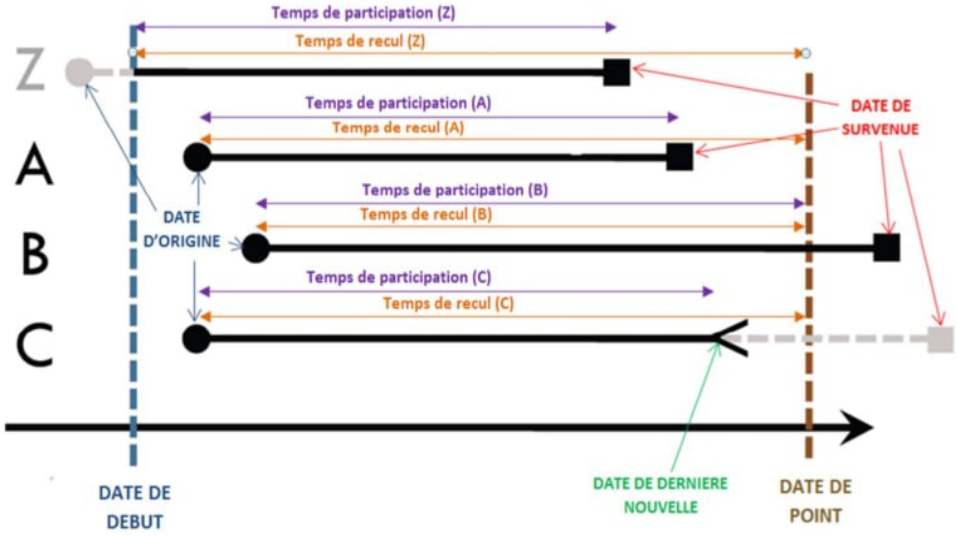


FIG. 1.1 – Schéma représentant les principales définitions relatives à l'analyse de la durée de survie

▷Le patient A est suivi entièrement durant la période de l'étude.

▷Le patient B est encore vivant après la date de point. Son décès surviendra peut

être plus tard, mais hors la période d'étude.

▷Le patient C n'est pas suivi durant toute la période d'étude. Il était vivant jusqu'au moment où on l'a perdu de vue (date de dernière nouvelle).

▷Le patient Z, sa maladie est apparue avant la date de début d'étude.

## 1.2 Distribution de la durée de survie

La durée de survie à laquelle on s'intéresse est une variable aléatoire continue positive, qu'on note par $T$.

### 1.2.1 Densité et fonction de répartition de fonction de survie

**Fonction de survie**

La fonction de survie $S(t)$ est égale à la probabilité de survivre jusqu'à l'instant $t$ *c.à.d* pour $t$ fixé et positive $t \geq 0$

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t). \end{aligned} \tag{1.1}$$

$S(t) \in [0; 1]$, c'est une probabilité. Elle est décroissante de 1 en 0. La probabilité d'être en vie à l'origine est égale à 1. La probabilité d'être en vie après en temps infini est nulle.

**Exemple 1.2.1** *Prenons le cas de la vie humaine. Quelle est la fonction de survie d'un être humain avec l'âge, une personne grandit puis meurt à un moment*

*donné ?*

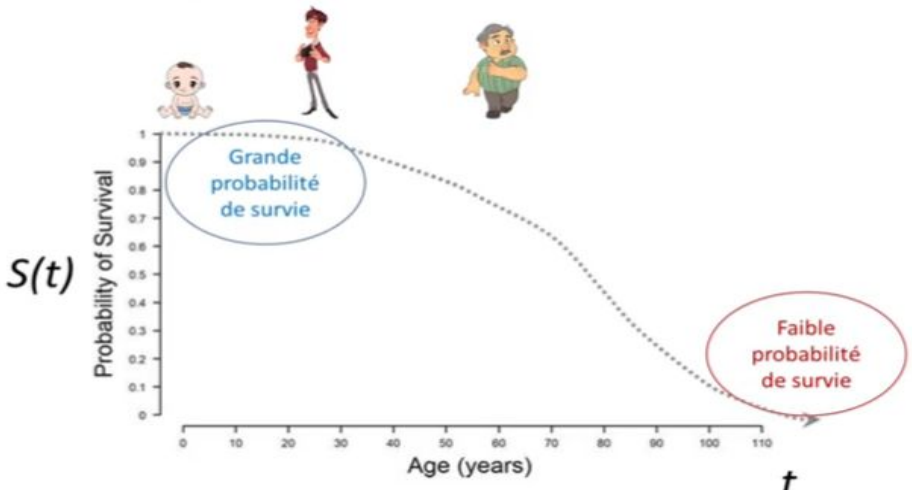


FIG. 1.2 – la fonction de survie d'un être humaine avec l'age

**Exemple 1.2.2** *(En biostatistique)*

•*Survie d'un individu après l'apparition d'une tumeur.*

•*Âge de décès chez des patients atteints de diabète...*

**Exemple 1.2.3** *(D'autres domaines)*

•*Fiabilité : durée entre deux pannes d'un matériel, d'un logiciel, ...*

•*Economie : durées des périodes de travail ou de chômage, temps avant faillite,...*

•*Assurances : durée de cotisation avant le premier remboursement.*

**Fonction de répartition**

telle que : $F(t) \in [0, 1]$, c'est une probabilité d'avoir eu l'événement avant l'origine est égale à "0", la probabilité d'avoir eu l'événement avant en temps de suivi très long est certaine est égale à "1".

**Densité de probabilité**

C'est la fonction $f(t) > 0$ telle que pour tout $t > 0$, $f(t)$ représente la limite de probabilité que l'événement se produise au temps $t$

$$\begin{aligned} f(t) &= \lim_{h \longrightarrow 0} \frac{P(t \leq T < t+h)}{h} \\ &= (1 - S(t))' \\ &= -S'(t). \end{aligned} \tag{1.2}$$

### 1.2.2 Fonction risque instantanéet et de risque cumulé

**Fonction risque instantanéet (ou taux de hasard)**

La fonction de risque instantané $h(t)$ représente la limite de probabilité que l'événement se produise à l'instant $t$ sachant qu'il ne s'est pas produit avant (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\begin{aligned} h(t) &= \lim_{h \longrightarrow 0} \frac{P(t \leq T < t+h \setminus T \geq t)}{h} \qquad h > 0 \\ &= \lim_{h \longrightarrow 0} \frac{1}{h} \frac{P(t \leq T < t+h)}{P(T \geq t)} \\ &= \lim_{h \longrightarrow 0} \frac{1}{h} \frac{P(t \leq T < t+h)}{1 - P(T < t)}, \end{aligned}$$

$$\begin{aligned} &= \lim_{h \longrightarrow 0} \frac{1}{h} \frac{P(t \leq T < t+h)}{1 - P(T \leq t)} \\ &= \frac{f(t)}{S(t)} \\ &= \frac{-S'(t)}{S(t)} \\ &= -\ln(S(t))'. \end{aligned}$$

**Exemple 1.2.4** *supposons que le temps de survie d'une population a la fonction de densité suivante*

$$f(t) = e^{-t} \qquad t \geq 0.$$

*alors la fonction de répartition de $F$ est*

$$F(t) = \int_0^t f(x)dx = \int_0^t e^{-x}dx = -e^{-x} \mid_0^t = 1 - e^{-t}.$$

*et de là*

$$S(t) = 1 - F(t) = 1 - 1 + e^{-t} = e^{-t},$$

*-la fonction de risque instantané $h(t)$ est définie par*

$$h(t) = \frac{f(t)}{S(t)} = \frac{e^{-t}}{e^{-t}} = 1.$$

**Fonction de risque cumulé**

La fonction de risque cumulé $H(t)$ est la somme des risques instantanés jusqu'à l'instant $t$

$$H(t) = \int_0^t h(x)dx = \int_0^t -\ln(S(x))'dx = -\ln(S(t)).$$

**Propriété 1.2.1** *Supposons que $T$ admette une densité $f$ continue sur $\mathbb{R}^+$. Soit $A = \{t > 0; S(t) \neq 0\}$, alors, pour tout $t \in A$, les propriétés suivantes sont équivalentes :*

1)$h(t) = \frac{f(t)}{S(t)}$.

2)$h(t) = (-\ln S(t))\prime$.

3)$S(t) = exp(-H(t))$.

4)$f(t) = h(t)exp(-H(t))$.

**Preuve.**

1)$\Longrightarrow$2)

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\ln(S(t))'.$$

2)$\Longrightarrow$3)

$$h(t) = -\ln(S(t))' = -\frac{S'(t)}{S(t)},$$

D'oú

$$S(t) = exp(-\int_0^t h(x)dx) = exp(-H(t)).$$

3)$\Longrightarrow$4) on a

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{exp(-H(t))},$$

D'oú

$$f(t) = h(t)exp(-H(t)).$$

4)$\Longrightarrow$1) on a

$$f(t) = h(t)exp(-H(t)),$$

$$h(t) = \frac{f(t)}{exp(-H(t))},$$

et on a

$$S(t) = exp(-H(t)),$$

donc

$$h(t) = \frac{f(t)}{S(t)}.$$

D'oú le résultat ■

### 1.2.3 Quantités associées à la distribution de survie

**Moyenne et variance de la durée de survie**

Le temps moyenne de survie $E(T)$ et la variance de survie $V(T)$ sont définies par :

•le temps moyenne de survie $E(T)$

$$E(T) = \int_0^{+\infty} tf(t)dt = \int_0^{+\infty} t(-S^{'}(t))dt,$$

en utilisant l'intégration par parties $(I.P.P)$

$$\begin{aligned} E(T) &= -tS(t) \mid_0^{+\infty} + \int_0^{+\infty} S(t)dt \\ &= \int_0^{+\infty} S(t)dt. \end{aligned} \tag{1.3}$$

•la variance de la durée de survie $V(T)$

$$V(T) = E(T^2) - E(T)^2,$$

alors on calcul $E(T^2)$ :

$$E(T^2) = \int_0^{+\infty} t^2 f(t)dt = \int_0^{+\infty} - t^2 S^{'}(t)dt$$

en utilisant l'intégration par parties $(I.P.P)$ :

$$\begin{aligned} E(T^2) &= -t^2 S(t) \mid_0^{+\infty} + 2\int_0^{+\infty} tS(t)dt \\ &= 2\int_0^{+\infty} tS(t)dx. \end{aligned}$$

donc

$$V(T) = 2\int_0^{+\infty} tS(t)dt - (E(T))^2. \tag{1.4}$$

**Remarque 1.2.1** *on peut déduire l'espérance et la variance à partir de n'importe la quelle des fonctions $F, S, f, h, H$ (mais pas l'inverse).*

### Quantiles de la durée de survie

Le quantile $q(p)$ est le temps où une proportion $p$ de la population a disparu, ainsi la fonction quantile de la durée de survie est définie par

$$\begin{aligned} q(p) &= \inf(t : F(t) \geq p) \qquad\qquad 0 < p < 1 \\ &= \inf(t : S(t) \leq 1 - p). \end{aligned}$$

lorsque la fonction de répartition $F$ est strictement croissante et continue alors

$$\begin{aligned} q(p) &= F^{-1}(p), \qquad\qquad 0 < p < 1 \\ &= S^{-1}(1 - p). \end{aligned} \tag{1.5}$$

Le quantile particulier d'ordre 0.5 est appelé médiane de la durée de survie. C'est le temps $t_{0.5}$ pour lequel la probabilité de survie $S(t)$ est égale à 0.5 ; *c.à.d* $S(t_{0.5}) = F(t_{0.5}) = 0.5$. Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[b_i; b_s]$ un intervalle de confiance de niveau de $S(t_{0.5})$, alors un intervalle de confiance de niveau du temps médian $t_{0.5}$ est $[S^{-1}(b_s); S^{-1}(b_i)]$.

## 1.3 Censure

Dans l'analyse de survie les données ne sont pas toujours complètement observées, parce que, pour certains individus l'évènement du début et / ou de la fin n'est pas observé. C'est-à-dire privées d'une partie de l'information. Dans ce cas les données sont censurées, Il n'est pas rare, mais elles sont plutôt incomplètes.

pour un individu $i$, on a donc :

-son temps de survie $T_i$.

-son temps de censure $C_i$.

-la durée réellement observée $X_i$.

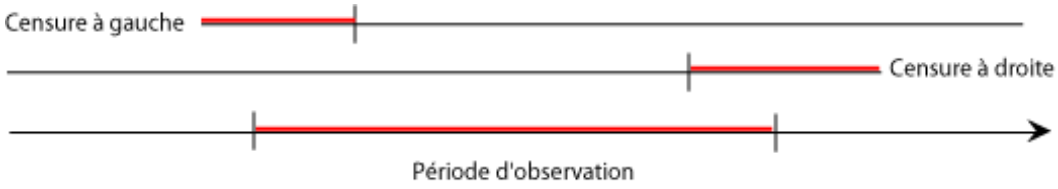


FIG. 1.3 – Différents types de censure des durées

**Remarque 1.3.1** *Pour une présentation détaillée des différents types de censure voir comme un exemple le livre de Klein, M [7] et Elisa, T [4]*

### 1.3.1 Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation (1.3). En présence de censure à droite, les durées de vie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue [12].

**Censure de type I : censure fixée**

Soit $C$ une valeur fixée, au lieu d'observer les variables $T_1, ..., T_n$ qui nous intéressent, on n'observe $T_i$ uniquement lorsque $T_i \leq C$

$$X_i = T_i \wedge C = min(T_i; C).$$

Par exemple, on peut tester la durée de vie de n objet identiques (ampoules) sur un intervalle d'observation fixé $[0; u]$.

**Censure de type II : attente**

On décide d'observer les durées de survie des $n$ patients jusqu'à ce que $r$ $(1 \leq r \leq n)$ d'entre eux soient décédés et d'arrêter l'étude à ce moment-là, Si

$$T_{1;n} \leq T_{2;n} \leq ... \leq T_{n;n}.$$

désignent la statistique d'ordre associée à l'échantillon $(T_1, ..., T_n)$, alors la date de censure est $T_{r;n}$ et on observe(voir [8] page 17)

$$\begin{cases} X_{i,n} = T_{i,n} & si \quad i \leq r, \\ X_{i,n} = T_{r;n} & si \quad i \geq r. \end{cases}$$

**Censure de type III (ou censure aléatoire de type II)**

Soient $C_1, ..., C_n$ des variables aléatoires indépendantes et identiquement distribué. $(i.i.d)$. On observe les variables

$$X_i = T_i \wedge C_i, \qquad i = 1, ..., n.$$

L'information disponible peut être résumée par :

▷la durée réellement observée $X_i$

▷un indicateur $\delta_i = I_{\{T_i \leq C_i\}}$,$tq$

$*$ $\delta_i = 1$ si l'événement est observé (d'où $X_i = T_i$). On observe les vraies durées ou les durées complètes.

$*$ $\delta_i = 0$ si l'individu est censuré (d'où $X_i = C_i$). On observe des durées incomplètes (censurées).

La censure aléatoire est la plus courante. Par exemple : lors d'un essai thérapeutique, elle peut être engendrée par

▷Perte de vision : le patient quitte l'étude en cours et n'est plus revu (en raison d'un mouvement, le patient décide de se faire soigner ailleurs). Ce sont des patients perdus de vue.

▷Arrêt ou changement de traitement : Des effets indésirables ou l'inefficacité du traitement peuvent conduire à un changement ou à l'arrêt du traitement. Ces patients sont exclus de l'étude.

▷Fin de l'étude : L'étude se termine alors que certains patients sont encore en vie (ils n'ont pas vécu cet événement). Ce sont des exclus vivants. Les perdus de vue (et exceptions) et la serviabilité sont identiques aux observations censurées, mais les deux mécanismes sont de nature différente (la censure peut être bénéfique pour les perdus de vue).

### 1.3.2 Censure à gauche

Il y a censure à gauche lorsque l'individu á déjà subi l'événement avant qu'il soit observé(1.3). On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue.

Par exemple : la censure à gauche peut s'appliquer à une mesure ou une observation d'un phénomène dont on ignore le moment de l'apparition.

En présence de censure à gauche, l'information disponible pour chaque individu est $\{T_i; \delta_i\}$ avec

$\triangleright T_i$, la durée réellement observée : $T_i = X_i \vee C_i = max(X_i, C_i)$.

$\triangleright \delta_i = I_{\{X_i \geq C_i\}}$, un indicateur qui vaut 1 si l'événement est observé et 0 si l'individu est censuré.

# Chapitre 2

# Estimateur de Kaplan-Meier de la distribution de survie

Dans ce chapitre nous nous placerons dans le cadre le plus fréquent d'une censure à droite aléatoire de type $I$. Si aucun modèle n'est pas supposé, on estime la fonction de survie avec l'utilisation de l'estimateur de Kaplan-Meier.

## 2.1 Estimation de Kaplan-Meier de la survie

### 2.1.1 Fonctions empiriques de répartition et de survie

Soit $T_1, ..., T_n$ un échantillon de taille $n \geq 1$ d'une $v.a$ positive $T$ de fonction de répartition $F$ et de fonction de survie $S$. Les fonctions empiriques de répartition et de survie (Figure 2.2), $F_n$ et $S_n$ sont respectivement définies par

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i \leq t\}} \qquad \forall t \geq 0.$$

et

$$S_n(t) := 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{T>t\}} \qquad \forall t \geq 0.$$

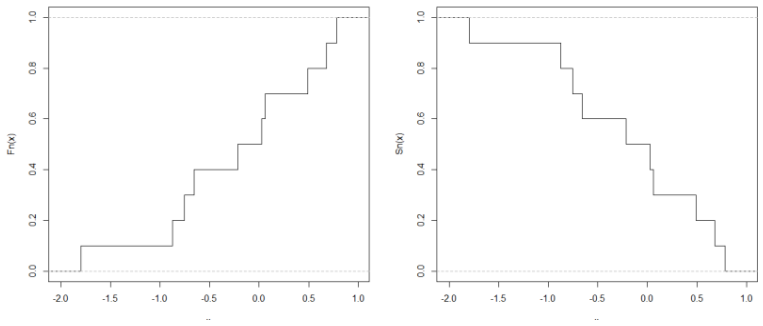


FIG. 2.1 – Fonctions empiriques de répartition (gauche) et de survie (droite) d'un échantillon Gaussien de taille 10

L'estimateur $S_n(t)$ jouit de bonnes propriétés de convergence. Il vérifie les lois (faible et forte) des grands nombres et d'après le théorème de Glivenko-Cantelli, on a :

$$\sup_{t\geq 0} \mid S_n(t) - S(t) \mid \xrightarrow{p.s} 0,$$

de plus, on a la propriété de normalité asymptotique

$$mpkjnnn\sqrt{n}(S_n(t) - S(t)) \xrightarrow{D} N(0; \sigma^2(t)) \qquad ; t \geq 0 \; quand \; n \longrightarrow \infty,$$

où

$$\sigma^2(t) = \frac{S(t)(1 - S(t))}{n}.$$

### 2.1.2 Estimateur de Kaplan-Meier

#### Définition

En pratique on estime la fonction de survie à partir de données collectées, cette approximation s'appelle la courbe de Kaplan-Meier.

L'estimateur de Kaplan-Meier (Kaplan-et Meier, 1958) aussi dénommé estimateur "produit-limite", est l'estimateur du maximum de vraisemblance non-paramétrique de la fonction de survie. Il permet d'estimer la fonction de survie $S(t)$ à partir d'un échantillon de $n$ sujets avec des durées de survie qui peuvent être censurées à droite.

Pour cela on utilise les notations définies précédemment $(\tilde{T}_i, \delta_i)$ $iid$, $i = 1, ..., n$. La définition de l'estimateur découle du raisonnement simple suivant : ne pas encore avoir subi l'événement à l'instant $t$, c'est ne pas l'avoir encore subi juste avant $t$ et ne pas le subire. Ce qui permet de décomposer le calcul connaissant $(\tilde{T}_i, \delta_i)$ on a $t_1 < t_2 < ... < t_j < ... < t_k$ les différents temps d'évènements observés, alors on peut définir

$\triangleright d_i$ : le nombre de décès en $T_i$.

$\triangleright n_i$ : le nombre d'individus à risque de subir l'événement juste avant le temps $T_i$ (effectif à risque).

Avec le raisonnement précèdent calculer la probabilité de ne pas encore avoir subi l'évènement en $t_i$ $(p(T > t_i))$ revient à calculer la probabilité de ne pas encore avoir subi l'évènement en $t_{i-1}$ et celle de ne pas le subir en $t_i$ sachant que l'évènement n'à pas en lieu jusqu'en $t_{i-1}$. Ainsi, la survie à un instant quelconque est le produit

de probabilités conditionnelles de survie de chacun des instants précédents

$$\begin{aligned}\hat{S}(t_i) &= p(T > t_i) \\ &= p(T > t_i / T > t_{i-1}) \times p(T > t_{i-1}) \\ &= p(T > t_i / T > t_{i-1}) \times ... \times p(T > t_2 / T > t_1) \times p(T > t_1) \\ &= p_i \times p_{i-1} \times ... \times p_1. \end{aligned} \tag{2.1}$$

où $p(T > t_i/T > t_{i-1})$ on estime $p_i$ par $p_i = (n_i - d_i)/n_i$

L'estimateur de Kaplan-Meier de la probabilité de survivre au moins jusqu'en $t$ est donc

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}),$$

on a donc

$$\hat{S}(t_{i+1}) = \hat{S}(t) \times (1 - \frac{d_i}{n_i}).$$

par définition $\hat{S}(0) = 1$. la fonction $\hat{S}(t)$ est une fonction en escalier décroissante constante entre deux temps d'évènement observé.

**Remarque 2.1.1** *comme les temps d'évènements sont supposés distincts, on a*
$\begin{cases} d_i = 0 \text{ en cas de censure en } T_i \text{ i.e quand } \delta_i = 0, \\ d_i = 1 \text{ en cas de décès en } T_i \text{ i.e quand } \delta_i = 1. \end{cases}$
*On obtient alors l'estimateur de Kaplan-Meier*

$$\begin{aligned}\hat{S}(t) &= \prod_{\substack{i=1,...,n \\ T_i \leq t}} \left(1 - \frac{\delta_i}{n_i}\right) \\ &= \prod_{\substack{i=1,...,n \\ T_{(i)} \leq t}} \left(1 - \frac{\delta_i}{n - (i-1)}\right) \\ &= \prod_{\substack{i=1,...,n \\ T_{(i)} \leq t}} \left(1 - \frac{n-i}{n-i+1}\right)^{\delta_i}. \end{aligned} \tag{2.2}$$

### Moyenne et médiane de survie

Quand on décrit des données de survie, on doit préciser le nombre total d'événements observés sur la période de survie et on peut donner une représentation graphique de la fonction de survie estimée. Il est intéressant de préciser aussi le temps de survie médian ou moyenne des sujets. En ce qui concerne la durée de survie on préfèrera indiquer la médiane de survie estimée plutôt que la moyenne. La médiane estimée de survie est le délai $(t_M)$ tq $S(t_M) = 0,5$.

En effet, la moyenne que l'on pouvuait calculer à partir des temps d'évènements observés n'estime pas bien le délai de survie moyenne car les délais censurés à droites ne sont pas pris en compte.

On purrait estimer l'espérance à partir de l'estimateur de Kaplan-Meier si le temps d'observation le plus élevé correspondait à une observation non censurée ; cependant il arrive souvent qu'il corresponde à une observation censurée, auquel cas l'espérance n'est pas définie.

La médiane de survie, estimée par la méthode de Kaplan-Meier est souvent bien définie, mais il peut aussi arriver qu'elle ne puisse pas être estimée si le recul de l'étude n'est pas assez important par raport à l'incidence de l'évènement.

### Estimation de la variance

L'estimateur de Kaplan-Meier est une statistique et certains estimateurs sont utilisés pour approcher sa variance. Un de ces estimateurs les plus courants est la formule de Greenwood

$$\hat{V}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{i:T_{i;n} \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{2.3}$$

**Biais**

De nombreux auteurs se sont intéressés au calcul du biais de l'estimateur de Kaplan-Meier. Un premier résultat dû à Gill (1980) dit qu'à moins qu'aucune observation ne soit censurée, l'estimateur de Kaplan-Meier de la fonction de survie présente un biais positif, $c'.à.d$

$$E\left[\hat{S}(t) - S(t)\right] \geq 0.$$

Une expression explicite de ce biais est détaillé dans [6], page 118. Pour tout t tel que $S(t) > 0$, on a

$$E\left[\hat{S}(t) - S(t)\right] = E\left[I_{\{T_{n;n}<t\}} \frac{\hat{S}(T_{n;n})(S(T_{n;n}) - S(t))}{S(T_{n;n})}\right].$$

On déduit que $\hat{S}(t)$ est sans biais pour tout $T_{n;n} < t$ ou si la dernière observation n'est pas censurée (puisque dans ce cas là, $\hat{S}(T_{n;n}) = 0$). Supposons que dans notre échantillon, $T_{n;n}$ est censuré. On remarque alors que pour tout $t$, $T_{n;n}$, $\hat{S}(t)$ a un biais positif et ce biais est d'autant plus grand (en valeur absolue) que l'on a de données censurées et/ou que $t$ est grand. En effet, plus l'écart entre $t$ et la dernière observation $T_{n;n}$ est important et plus $S(T_{n;n}) - S(t)$ est grand. De même, plus il y a de données censurées et plus $S(T_{n;n})$ est grand. Donc l'estimateur de Kaplan-Meier est généralement sans biais sauf dans les queues de distribution.

**Estimation par l'intervalle de confiance**

On peut utiliser la formule suivante pour donner un intervalle de confiance ($IC$) contenant la vrai du paramètre

$$IC_{95°/_\circ} = [\hat{S}(t) \pm 1,96 \times \hat{\sigma}_i]. \tag{2.4}$$

cependant, cet intervalle est symétrique autour de $\hat{S}(t)$, et fournit donc, pour des valeurs de $\hat{S}(t)$ proches de 0 et de 1, des bornes pouvant dépasser 0 ou 1.

Il peut être préférable d'utiliser l'intervalle de confiance de Rathman, dont les bornes sont toujours comprises entre 0 et 1.

$$IC_{95°/_\circ} := \frac{M}{M + (1.96)^2} \left[ \hat{S}(t) + \frac{(1,96)^2}{2M} \pm 1.96 \sqrt{\hat{\sigma}_t^2 + \frac{(1.96)^2}{4M^2}} \right].$$

où $M = \frac{\hat{S}(t)[1-\hat{S}(t)]}{\hat{\sigma}_t^2}$

**Cohérence**

Un estimateur $\hat{S}$ d'une fonction $S$ est dit cohérent si, pour tout $t$, on a

$$\hat{S}(t) := \frac{1}{n} \left[ \sum_{i=1}^{n} I_{\{T_i > t\}} + \sum_{i=}^{n} I_{\{T_i \leq t, \delta_i = 0\}} \frac{\hat{S}(t)}{\hat{S}(T_i)} \right].$$

Cette formule signifie que les survivants au-delà de $t$ sont la somme : des individus ni décédés, ni censurés à la date $t$ et des individus qui, censurés en $T_i$ avant $t$, survivent après $t$ avec la probabilité conditionnelle $\hat{S}(t)/\hat{S}(T_i)$.

D'après Fermanian[5], l'estimateur de Kaplan-Meier est l'unique estimateur cohérent de la fonction de survie.

**Exemple 2.1.1** *On observe les durées de vie de 10 diodes exprimées en mois (+*

*si censurées)*

$$1 \quad 3 \quad 4^+ \quad 5 \quad 7^+ \quad 8 \quad 9 \quad 10^+ \quad 11 \quad 13^+$$

*L'estimateur de Kaplan-Meier de la fonction de survie $S$ des durées de vie de 10 diodes exprimées en mois dans cet exemple se calcule par la formule*

$$\hat{S}(t) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{n-i}{n-i+1}\right)^{\delta_i}.$$

*et on a $Y_i = n - i + 1$, et donne le tableau suivant 2.1 :*

| $Temps$ | $n_i$ | $d_i$ | $\hat{S}(t_i)$ | $Intervalle$ |
|---|---|---|---|---|
| 0 | 10 | 0 | 1 | $[0;1[$ |
| 1 | 10 | 1 | $(1-1/10)\hat{S}(0) = 0.900$ | $[1;3[$ |
| 3 | 9 | 1 | $(1-1/9)\hat{S}(1) = 0.800$ | $[3;5[$ |
| 5 | 7 | 1 | $(1-1/7)\hat{S}(3) = 0.686$ | $[5;8[$ |
| 8 | 5 | 1 | $(1-1/5)\hat{S}(5) = 0.549$ | $[8;9[$ |
| 9 | 4 | 1 | $(1-1/4)\hat{S}(8) = 0.411$ | $[9;11[$ |
| 11 | 2 | 1 | $(1-1/2)\hat{S}(9) = 0.205$ | $[11;+\infty[$ |

TAB. 2.1 – L'estimateur de Kaplan-Meier de la fonction de survie S des durées de vie de 10 diodes

***Représentation graphique :*** *le graphe suivant présente l'estimation de la fonction de survie par Kaplan-Meier et les intervalles de confiance à un niveau de* 95% *par exemple $\hat{S}(10) = 0.4$ (Figure 2.2)*

**Exemple 2.1.2** *Calculons l'estimateur de Kaplan-Meir sur les données de Frechet (1963) : "étude des durées de rémission, exprimées en semaines, des sujets atteints de leucémie, selon qu'ils ont reçu de 6-mercaptopurnie ou un placébo". Les* 21 *observations se présentent de la manière suivante (une signe + indiquant*

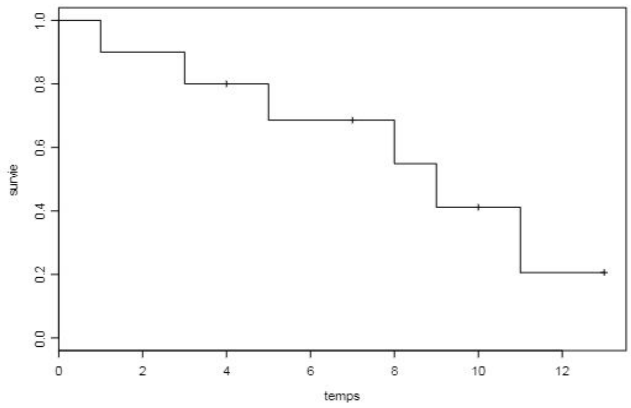


FIG. 2.2 – L'estimateur de K.M de la fonction de survie des durées de vie de 10 diodes

*une donnée censurées à droite) :*

$$6 \quad 6 \quad 6 \quad 6^+ \quad 7 \quad 9 \quad 10 \quad 10^+ \quad 11^+ \quad 13 \quad 16 \quad 17^+$$

$$19^+ \quad 20^+ \quad 22 \quad 23 \quad 25^+ \quad 32^+ \quad 32^+ \quad 34^+ \quad 35^+$$

*L'estimateur de Kaplan-Meier de la survie du groupe de 21 malades traité par le traitement 6MP (existence d'ex-aequo) se calcule par la formule :*

$$\hat{S}_{6MP}(t) = \prod_{i:t_i \leq t}(1 - \frac{d_i}{n_i}).$$

*et on a* $Y_i = n - i + 1$*, et donne le tableau suivant 2.2*

***Représentations graphiques :*** *Le graphe suivant présente les estimateurs de la fonction de survie par Kaplan-Meier pour les deux traitements. Nous remarquons que la courbe de survie 6MP est nettement supérieure à celle du Placébo. D'où l'efficacité du traitement 6MP.*

| $temps$ | $n_i$ | $d_i$ | $\hat{S}_{6MP}(t)$ | $Intervalle$ |
|---|---|---|---|---|
| 0 | 21 | 0 | 1 | $[0, 6[$ |
| 6 | 21 | 3 | $(1 - \frac{3}{20}) \times 1 = 0.875$ | $[6, 7[$ |
| 7 | 17 | 1 | $(1 - \frac{1}{17}) \times 0.875 = 0.807$ | $[7, 10[$ |
| 10 | 15 | 1 | $(1 - \frac{1}{15}) \times 0.807 = 0.753$ | $[10, 13[$ |
| 13 | 12 | 1 | $(1 - \frac{1}{12}) \times 0.753 = 0.690$ | $[13, 16[$ |
| 16 | 11 | 1 | $(1 - \frac{1}{11}) \times 0.690 = 0.627$ | $[16, 22[$ |
| 22 | 7 | 1 | $(1 - \frac{1}{7}) \times 0.627 = 0.538$ | $[22, 23[$ |
| 23 | 6 | 1 | $(1 - \frac{1}{6}) \times 0.538 = 0.448$ | $[23, +\infty[$ |

TAB. 2.2 – Les estimateurs de K.M de la fonction de survie du groupe de 21 malades traité par le traitement 6MP

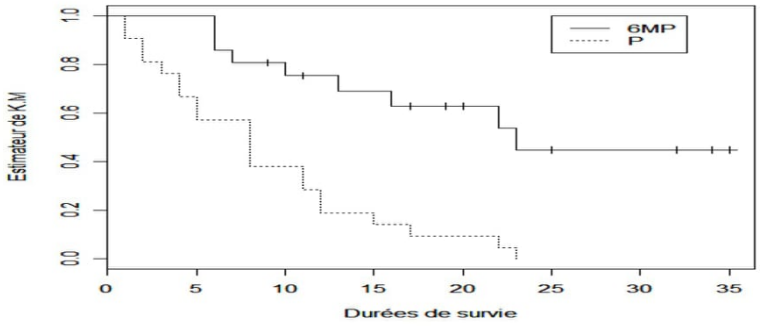


FIG. 2.3 – Les estimateurs de K.M de la fonction de survie du groupe de 21 malades traité par le traitement 6MP

# Chapitre 3

# Comparaison de deux ou plusieurs fonctions de survie

Il existe plusieurs tests pour comparer les fonctions de survie $S$ de deux ou plusieurs échantillons : Log-Rank, Gehan-Wilcoxon, Test de Gehan, Prentice-Wilcoxon ou Peto-Wilcoxon. Dans ce chapitre nous nous concentrerons sur le test du Log-Rank.

## 3.1 Comparaison de deux groupes

Dans un premier temps, l'objectif est de comparer les durées de survie $S_A$ et $S_B$ de deux groupes notés "A" et "B" :

Il est possible de comparer les survies de deux groupes à un instant $t$ donné. La survie au temps $t$ est une proportion, ainsi, en utilisant l'approximation de la loi

binomiale par la loi normale on montre que la statistique suivante

$$\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\sqrt{\hat{V}(\hat{S}_A(t)) + \hat{V}(\hat{S}_B(t))}}.$$

suit asymptotiquement une loi $N(0;1)$ sous l'hypothèse $H_0 : S_A(t) = S_B(t)$. Néanmoins, cela ne permet pas de tester (globalement) l'égalité des distributions de survie ce qui limite l'intérêt de la méthode.

**Notations**

Dans ce chapitre, nous nous intéressons à une approche non enseignante. Le principe des tests est de comparer le nombre de décès observés dans chaque groupe avec le nombre de décès attendus (calculés sur l'hypothèse d'égales distributions de survie).

Considérons les notations suivantes

$\trianglerighteq T_1 < ... < T_N$ les temps de décès ordonnés des deux échantillons réunis,

$\trianglerighteq d_{A_i}$ et $d_{B_i}$ le nombre de décès observés au temps $T_i$ dans chacun des groupes A et B,

$\trianglerighteq d_i = d_{A_i} + d_{B_i}$ le nombre total de décès observés en $T_i$,

$\trianglerighteq n_{A_i}$ et $n_{Bi}$ le nombre de sujets à risques en $T_i$ dans les groupes A et B,

$\trianglerighteq n_i = n_{A_i} + n_{B_i}$ le nombre total de sujets à risques en $T_i$,

$\trianglerighteq w_i$ le poids associé au temps $T_i$.

Pour chaque temps d'événement $T_i$, l'information peut être résumée sous forme de tableau

| | Décès en $T_i$ | Vivant après $T_i$ | Total |
|---|---|---|---|
| Groupe A | $d_{A_i}$ | $n_{A_i} - d_{A_i}$ | $n_{A_i}$ |
| Groupe B | $d_{B_i}$ | $n_{B_i} - d_{B_i}$ | $n_{B_i}$ |
| Ensemble | $d_i$ | $n_i - d_i$ | $n_i$ |

TAB. 3.1 – tableau d'information peut être résumée sous forme au temps T dans chacun des groupes A et B

### 3.1.1 Statistiques de test

Nous cherchons à tester l'hypothèse $H_0 : S_A(t) = S_B(t)$ qui est l'égalité des fonctions de survie dans les deux groupes. Ainsi, sous l'hypothèse $H_0$, Le taux de mortalité attendu (chez les personnes à risque) est identique dans les deux groupes pour tous les instants de décès $T_i$ : Pour chaque temps $T_i$ on peut comparer les pourcentages de décès parmi les sujets à risque dans chacun des groupes en utilisant le test du $Chi - 2$.

Soit $D_{A_i}$ ($D_{B_i}$ et $D_i$) la variable dont la valeur est $d_{A_i}$ ($d_{B_i}$ et $d_i$) ; on peut montrer que $D_{A_i}$ suit une loi hypergéométrique (cf : test du $Chi - 2$) d'espérance

$$E(D_{A_i}) = \frac{n_{A_i} \times d_i}{n_i}. \tag{3.1}$$

et de variance

$$V(D_{A_i}) = \frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{A_i} \times n_{B_i}}{n_i^2}. \tag{3.2}$$

où $E(D_{A_i})$ correspond au nombre de décès attendus dans le groupe A : Sous $H_0$, on montre que les variables $D_{A_i} - E(D_{A_i})$ suivent asymptotiquement des lois $N(0; V(D_{A_i}))$ ($\frac{[D_{A_i} - E(D_{A_i})]^2}{V(D_{A_i})}$ suivent asymptotiquement des loi de $\chi^2$).

Considérons des pondérations $w_i = 1, ..., N$, alors par indépendance entre les

variables $D_{A_i}$ et $D_{A_j}$ (associées aux $T_i$ et $T_j$), les variables

$$\sum_{i=1}^{N} w_i(D_{A_i} - E(D_{A_i})) = \sum_{i=1}^{N} w_i(D_{A_i} - \frac{n_{A_i} \times d_i}{n_i}).$$

Suivre l'approximation des distributions normales avec des moyennes et des variances nulles $w_i^2 V(D_{A_i})$ Par conséquent, sous $H_0$, les statistiques suivantes

$$\chi_0^2 = \frac{\left[\sum_{i=1}^{N} w_i(D_{A_i} - \frac{n_{A_i} \times d_i}{n_i})\right]^2}{\sum_{i=1}^{N} w_i^2 \frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{A_i} \times n_{A_i}}{n_i^2}}.$$

suivent asymptotiquement des lois de $\chi^2$ à 1 degré de liberté.

**Remarque 3.1.1** *La logique précédente s'applique à toutes les cases de la matrice* $2 \times 2$*, ce qui produit d'autres statistiques de test équivalentes.*

**Remarque 3.1.2** *Les tests conditionnels sont définis aux marges du tableau et supposent que les temps d'événement sont constants. Les tableaux peuvent alors être traités comme des matrices indépendantes.*

### 3.1.2 Test du Log-Rank

Le test du Log-Rank est le test standard pour comparer la courbe de survie. Lorsqu'elle compte, elle permet de rejeter l'hypothèse de superposition des deux courbes. Il analyse si, au niveau de chaque décès, la distance entre les deux courbes est supérieure à ce qui peut être expliqué par le hasard. C'est-à-dire si la différence cumulée entre les courbes mesurées à chaque décès est supérieure à la valeur attendue uniquement en raison du hasard. Ainsi, le test du Log-Rank analyse les courbes dans leur ensemble. Il pourrait être significatif même si les deux courbes se rencontraient à la fin du suivi, entraînant le même nombre de décès dans chaque

groupe. Cependant, le test perd de son efficacité lorsque les deux courbes n'évoluent pas proportionnellement, notamment lorsque les deux courbes se croisent. Par conséquent, l'analyse visuelle des courbes doit toujours accompagner l'interprétation du test Log-Rank.

D'autres tests, comme le test de Jihan (également appelé test de Wilcoxon) ou le test de Prentice, sont plus sensibles aux décès précoces qu'aux décès tardifs. Le test du Log-Rank est équivalent au test de Mantel Haenszel pour l'intégration des données stratifiées. C'est une épreuve non pédagogique. Le test du Log-Rank est plus puissant que le test du risque relatif, qui ne tient pas compte de la censure.

**Principe du test du Log-Rang**

Les individus sont classés par ordre croissant selon les temps observés $T_i$ dans les deux groupes A et B réunis. On a $T_1 < ... < T_l$ avec $l \leq N$. On note :

$E_{B_i}$ : nombre de décès attendus ($i.e$ sous $(H_0)$) au temps $T_i$ dans le groupe B

$$E_{B_i} = \frac{d_{A_i} + d_{B_i}}{n_{A_i} + n_{B_i}} \times n_{B_i}.$$

Ce test généralise au cas de données censurées le test de Savage. On peut noter que sous l'hypothèse nulle $d_A + d_B = E_A + E_B$, en d'autres termes la valeur de la statistique de test ne dépend pas du groupe sur laquelle on l'évalue. La forme de la statistique suggère la formule approchée suivante

$$\chi^2 = \frac{(d_A - E_A)^2}{E_A} + \frac{(d_B - E_B)^2}{E_B}, \tag{3.3}$$

d'une autre manière :

$$\chi^2 = \frac{(d_A - E_A)^2}{V(E_A)} + \frac{(d_B - E_B)^2}{V(E_B)}.$$

dont on peut montrer qu'elle est inférieure à celle du log-rank. Sa forme évoque celle d'un $Khi - 2$ d'ajustement usuel. Le test du log-rank est le test le plus couramment employé.

**Remarque 3.1.3** *a) Le test du log-rank peut-être utiliser pour comparer les courbes de survie de $k$ groupes avec $k \geq 2$.*

*b) Le critère statistique $\chi^2$ suit alors une loi du $Chi - 2$ à $k - 1$ degrés de liberté.*

**Exemple d'application de calcul du test du Log-Rank 01**

Nous reprenons l'exemple des courbes de survies établies par la méthode de Kaplan-Meier des courbes de survie.

Dans une première étape il faut regrouper les temps de participation de l'ensemble des sujets des groupes A et B. Puis, pour chaque $t_i$ où se produisent des décès il faut calculer les décès attendus.

Les résultats sont présentés dans le tableau suivant :

$\triangleright d_A = \sum_{i=1}^{20} d_{A_i} = 8.$

$\triangleright d_B = \sum_{i=1}^{20} d_{B_i} = 20.$

$\triangleright E_A = \sum_{i=1}^{20} E_{A_i} = 13,333.$

$\triangleright E_B = \sum_{i=1}^{20} E_{B_i} = 14,667.$

| *Donnée en jours* | $d_{A_i}$ | $d_{B_i}$ | $n_{A_i}$ | $n_{B_i}$ | $E_{A_i}$ | $E_{Bi}$ |
|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 370 | 340 | $1,042$ | $0,958$ |
| 5 | 0 | 1 | 369 | 339 | $0,521$ | $0,479$ |
| 24 | 1 | 0 | 368 | 338 | $0,521$ | $0,479$ |
| 28 | 0 | 1 | 367 | 337 | $0,521$ | $0,479$ |
| 31 | 1 | 0 | 367 | 336 | $0,521$ | $0,479$ |
| 36 | 1 | 0 | 366 | 335 | $0,522$ | $0,479$ |
| 40 | 0 | 1 | 365 | 335 | $1,043$ | 0.957 |
| 55 | 2 | 1 | 364 | 333 | $1,567$ | $1,133$ |
| 72 | 0 | 1 | 362 | 332 | $0,522$ | $0,478$ |
| 85 | 0 | 2 | 362 | 331 | $0,522$ | $0,478$ |
| 97 | 1 | 1 | 362 | 330 | $1,576$ | $0,431$ |
| 110 | 0 | 1 | 361 | 328 | $0,524$ | $0,476$ |
| 125 | 1 | 1 | 361 | 326 | $1,567$ | $1,424$ |
| 140 | 0 | 1 | 359 | 324 | $0,526$ | $0,474$ |
| 174 | 0 | 1 | 359 | 322 | $0,527$ | $0,473$ |
| 192 | 0 | 1 | 357 | 321 | $0,526$ | $0,474$ |
| 250 | 0 | 1 | 357 | 319 | $0,528$ | $0,472$ |
| 310 | 0 | 1 | 357 | 318 | $0,529$ | $0,471$ |
| 345 | 0 | 1 | 356 | 316 | $0,530$ | $0,470$ |
| 360 | 0 | 1 | 355 | 315 | $0,530$ | $0,470$ |

TAB. 3.2 – Les résultats de calcul du test du Log-Rank

▷Exemple : Calcul des $E_{A_i}$ et $E_{B_i}$ au temps $t_i = 55$ jours

$$E_{B_i} = \frac{n_{A_i}}{n_{A_i} + n_{B_i}} \times (d_{A_i} + d_{B_i}) = \frac{n_{B_i}}{n_i} \times d_i = \frac{333}{333 + 364}(1 + 2) = 1,133,$$
$$E_{A_i} = \frac{n_{B_i}}{n_{A_i} + n_{B_i}} \times (d_{A_i} + d_{B_i}) = \frac{n_{A_i}}{n_i} \times d_i = \frac{364}{333 + 364}(1 + 2) = 1,567.$$

On vérifie que : $d_A + d_B = E_A + E_B = 8 + 20 = 13,333 + 14,667 = 28$

Le test du Log-Rank est le suivant

$$\begin{aligned}\chi^2 &= \frac{(d_A - E_A)^2}{E_A} + \frac{(d_B - E_B)^2}{E_B} \\ &= \frac{(8 - 13,333)^2}{13,333} + \frac{(20 - 14,667)^2}{14,667} \\ &= 6,364.\end{aligned}$$

La valeur du $\chi^2$ est supérieure à $\chi^2_{1,1-\frac{\alpha}{2}} = 3,84$ (Se référer à la table des quantiles du $Chi-2$) On rejette $H_0$ et on conclut que les deux courbes de survie sont statistiquement différentes.

**Exemple d'application de calcul du test du Log-Rank 02**

Nous suivons respectivement deux échantillons A et B de patients atteints de cancer et nous enregistrons les résultats dans le tableau suivant

| $i$ | $t_i$ | $d_{A_i}$ | $d_{B_i}$ | $n_{A_i}$ | $n_{B_i}$ | $E_{A_i}$ | $E_{B_i}$ | $\sigma^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 2 | 21 | 21 | 1 | 1 | 0,49 |
| 2 | 2 | 0 | 2 | 21 | 19 | 1,05 | 0,95 | 0,49 |
| 3 | 3 | 0 | 1 | 21 | 17 | 0,55 | 0,45 | 0,25 |
| 4 | 4 | 0 | 2 | 21 | 16 | 1,14 | 0,86 | 0,48 |
| 5 | 5 | 0 | 2 | 21 | 14 | 1,2 | 0,8 | 0,47 |
| 6 | 6 | 3 | 0 | 21 | 12 | 1,91 | 1,01 | 0,65 |
| 7 | 7 | 1 | 0 | 17 | 12 | 0,59 | 1,41 | 0,24 |
| 8 | 8 | 0 | 4 | 16 | 12 | 2,29 | 1,71 | 0,87 |
| 9 | 10 | 1 | 0 | 15 | 8 | 0,65 | 0,35 | 0,23 |
| 10 | 11 | 0 | 2 | 13 | 8 | 1,24 | 0,76 | 0,45 |
| 11 | 12 | 0 | 2 | 12 | 6 | 1,33 | 0,67 | 0,42 |
| 12 | 13 | 1 | 0 | 12 | 4 | 0,75 | 0,25 | 0,19 |
| 13 | 15 | 0 | 1 | 11 | 4 | 0,73 | 0,27 | 0,2 |
| 14 | 16 | 1 | 0 | 11 | 3 | 0,79 | 0,21 | 0,17 |
| 15 | 17 | 0 | 1 | 10 | 3 | 0,77 | 0,23 | 0,18 |
| 16 | 22 | 1 | 1 | 7 | 2 | 1,56 | 0,44 | 0,3 |
| 17 | 23 | 1 | 1 | 6 | 1 | 1,71 | 0,29 | 0,2 |

TAB. 3.3 – Les résultats de calcul du test du Log-Rank

▷Exemple : Calcul $\sigma^2(E_{jk})$ au temps $t_1, t_2, t_3$

$$V(D_{A_i}) = \frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{A_i} \times n_{B_i}}{n_i^2}.$$

alors :

$$\sigma^2(E_{1k}) = \frac{21 \times 21 \times 2 \times (42-2)}{42^2 \times (42-1)} = 0,49.$$

$$\sigma^2(E_{2k}) = \frac{21 \times 19 \times 2 \times (40-2)}{40^2 \times (40-1)} = 0,49.$$

$$\sigma^2(E_{3k}) = \frac{21 \times 17 \times 1 \times (38-1)}{38^2 \times (38-1)} = 0,25.$$

$tq : n_i = n_{A_i} + n_{B_i} et d_i = d_{A_i} + d_{B_i}$

Le test du Log-Rank est le suivant

$$\begin{aligned}\chi^2 &= \frac{(d_A - E_A)^2}{V(E_A)} + \frac{(d_B - E_B)^2}{V(E_B)} \\ &= \frac{(9 - 19,25)^2}{6,26} + \frac{(21 - 10,75)^2}{6,26} \\ &= 33.58.\end{aligned}$$

La valeur du $\chi^2$ est supérieure à $\chi^2_{1,1-\frac{\alpha}{2}} = 3,84$ (Se référer à la table des quantiles du $Chi-2$) On rejette $H_0$ et on conclut que les deux courbes de survie sont statistiquement différentes.

## 3.2 Comparaison de plusieurs groupes

Les tests de la section précédente sont généralisés au cas de la comparaison des fonctions de survie de deux échantillons. Dans cette section, seule l'extension du test de connexion sera considérée (car c'est le test le plus utilisé).

Considérons le cas des trois groupes A, B et C, Le tableau suivant résume les notations

| | Décès en $T_i$ | Vivant après $T_i$ | Total |
|---|---|---|---|
| Groupe A | $d_{A_i}$ | $n_{A_i} - d_{A_i}$ | $n_{A_i}$ |
| Groupe B | $d_{B_i}$ | $n_{B_i} - d_{B_i}$ | $n_{B_i}$ |
| Groupe C | $d_{C_i}$ | $n_{C_i} - d_{C_i}$ | $n_{C_i}$ |
| Ensemble | $d_i$ | $n_i$ | $n_i$ |

TAB. 3.4 – Les résultats de calcul du test du Log-Rank le cas des trois groupes

En suivant la même démarche que dans le cas de deux échantillons, on montre que le vecteur suivant

$$V = \begin{pmatrix} \sum_{i=1}^{N}(D_{A_i} - E(D_{A_i})) \\ \sum_{i=1}^{N}(D_{B_i} - E(D_{B_i})) \end{pmatrix}. \tag{3.4}$$

avec

$$E(D_{A_i}) = \frac{n_{A_i} \times d_i}{n_i}.$$

$$E(D_{B_i}) = \frac{n_{B_i} \times d_i}{n_i}.$$

Il suit asymptotiquement une loi normale dans $\mathbb{R}^2$ avec une espérance nulle. Nous concluons alors que la statistique suivante

$$\chi_0^2 = V' \begin{pmatrix} \sum_{i=1}^{N} V(D_{A_i})) & \sum_{i=1}^{N} cov(D_{A_i}, D_{B_i}) \\ \sum_{i=1}^{N} cov(D_{A_i}, D_{B_i}) & \sum_{i=1}^{N} V(D_{B_i})) \end{pmatrix}_V^{-1}.$$

avec

$$V(D_{A_i}) = \frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{A_i}(n_i - n_{A_i})}{n_i^2}.$$

$$V(D_{B_i}) = \frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{B_i}(n_i - n_{B_i})}{n_i^2}.$$

$$cov(D_{A_i}, D_{B_i}) = -\frac{(n_i - d_i)}{(n_i - 1)} \times \frac{d_i \times n_{A_i} \times n_{B_i}}{n_i^2}.$$

Suit asymptotiquement une distribution du $\chi^2$ à deux degrés de liberté. Le raisonnement ci-dessus avec le couple $(A, B)$, est aussi possible avec des couples $(A, C)$ ou $(B, C)$ permettant d'obtenir d'autres statistiques de test d'équation.

Dans le cadre général de la comparaison de $k$ groupes, la statistique de test s'obtient de la même façon et suit asymptotiquement une loi de $\chi^2$ à $k - 1$ degrés de liberté.

**Exemple d'application : cas de la comparaison des fonctions de survie de deux échantillons**

Pour comparer les survies entre les trois groupes A, B et C statistiquement. On fait la construction du test Log-Rank d'égalité des 3 courbes. On voit que le test du $Chi - 2$ (3.3) est $20,7$ de 2 degrés de liberté et la valeur de probabilité est $0,00003$. Donc, on serait conduit à rejeter l'hypothèse nulle (l'hypothèse d'égalité des survies) et donc il n'y a pas une équivalence en matière d'efficacité des trois traitements.

| | N | Observes | censures | Attendee | $(D_A - E_A)^2/E_A$ | $(D_B - E_B)^2/E_B$ |
|---|---|---|---|---|---|---|
| Group : A | 50 | 36 | 14 | 56.4 | 7.406 | 15.29 |
| Group : B | 50 | 44 | 6 | 41.2 | 6.192 | 0.32 |
| Group : C | 50 | 50 | 0 | 32.4 | 9.608 | 15.49 |

TAB. 3.5 – Tests d'égalité des 3 survies : résultats de la statistique du Log-Rank.

| temps $t_i$ | $n_{A_i}$ | $n_{B_i}$ | $n_{C_i}$ | $d_{A_i}$ | $d_{B_i}$ | $d_{C_i}$ | $E_{A_i}$ | $E_{B_i}$ | $\hat{S}_{A_i}(t_i)$ | $\hat{S}_{B_i}(t_i)$ | $\hat{S}_{C_i}(t_i)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 50 | 50 | 8 | 12 | 19 | 13.50 | 15.5 | 0.84 | 0.76 | 0.62 |
| 2 | 42 | 38 | 31 | 1 | 7 | 3 | 2.301 | 5.254 | 0.82 | 0.62 | 0.56 |
| 3 | 41 | 31 | 28 | 5 | 2 | 1 | 3.565 | 1.576 | 0.72 | 0.58 | 0.54 |
| 4 | 36 | 29 | 27 | 1 | 3 | 4 | 2.857 | 3.625 | 0.70 | 0.52 | 0.46 |
| 5 | 35 | 26 | 23 | 1 | 5 | 2 | 1.666 | 3.714 | 0.68 | 0.42 | 0.42 |
| 7 | 34 | 21 | 21 | 2 | 1 | 3 | 3.090 | 2.000 | 0.64 | 0.40 | 0.36 |
| 8 | 32 | 20 | 18 | 1 | 2 | 2 | 1.920 | 2.105 | 0.62 | 0.36 | 0.32 |
| 11 | 31 | 18 | 15 | 1 | 1 | 1 | 1.347 | 1.090 | 0.60 | 0.34 | 0.30 |
| 12 | 30 | 17 | 15 | 1 | 1 | 4 | 3.333 | 2.656 | 0.58 | 0.32 | 0.22 |
| 17 | 29 | 16 | 11 | 1 | 2 | 1 | 1.450 | 1.777 | 0.56 | 0.28 | 0.20 |
| 19 | 28 | 14 | 10 | 1 | 1 | 1 | 1.473 | 1.166 | 0.52 | 0.26 | 0.18 |
| 21 | 26 | 13 | 9 | 2 | 2 | 1 | 2.228 | 1.772 | 0.50 | 0.22 | 0.16 |
| 38 | 25 | 11 | 8 | 1 | 1 | 3 | 3.030 | 2.315 | 0.48 | 0.20 | 0.08 |
| 40 | 24 | 10 | 5 | 1 | 1 | 1 | 1.655 | 1.333 | 0.46 | 0.18 | 0.08 |
| 41 | 23 | 9 | 4 | 1 | 1 | 1 | 1.703 | 1.384 | 0.32 | 0.16 | 0.06 |
| 46 | 22 | 8 | 3 | 1 | 2 | 1 | 1.760 | 2.181 | 0.28 | 0.12 | 0.04 |

TAB. 3.6 – Résultats de l'estimation de la fonction de survie du groupe A,B et C traité au bolus avec une forte dose de tueur de vers.

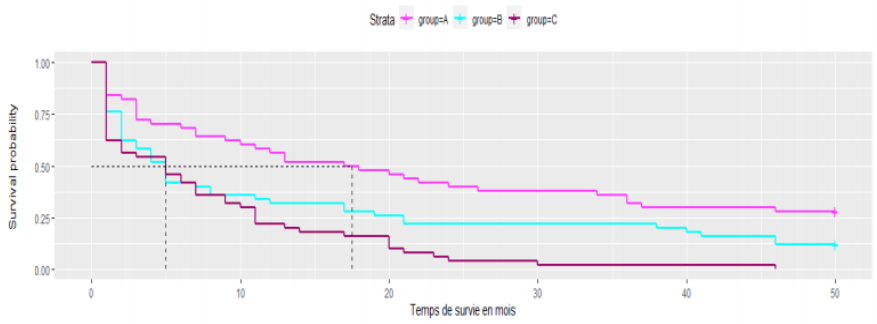


FIG. 3.1 – la représentation des courbes de survie pour les groupes A,B et C

# Conclusion

L'estimateur non paramétrique de la fonction de survie de Kaplan-Meier représente l'outil fondamental de l'estimation non paramétrique sous censure. Il a historiquement été utilisé dans l'analyse de la durée de vie dans le domaine médical en particulier.

Dans ce travail, on s'est concentré sur l'étude de cet estimateur qui, dans le cas ou les données incomplètes, se réduit à l'estimateur non paramétrique empirique usuel et fait un test pour comparer les groupes des données de survie (la statistique du Log-Rank), ainsi que les conclusions tirées.

Pour conclure, on signale que ce mémoire n'est qu'un point de départ pour mieux connaitre ce monde immense des Test d'une déférence de survie.

# Notations et symbols

Les diférentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées à dessous

| | | |
|---|---|---|
| $v.a$ | : | *variable aleatoire.* |
| $i.i.d$ | : | *Indépendantes et identiquement distribué.* |
| $c.á.d$ | : | *c'est á dire.* |
| $E(T)$ | : | *Esperance mathematique ou moyenne de v.a* $T$. |
| $V(T)$ | : | *varianc emathematique ou moyenne de v.a* $T$. |
| $tq$ | : | *telle que.* |
| $I.P.P$ | : | *Integrale par parti.* |
| $I_{x\in\{A\}}$ | : | *la fonction indicatrice de l'ensemble* $A$. |
| $p.s$ | : | *presque sûrement.* |
| $I.C$ | : | *Intervalle de confiance.* |
| $x \wedge y$ | : | $\min(x, y)$. |
| $\xrightarrow{D}$ | : | *convergence en distribution.* |
| $\mathbb{R}$ | : | *Ensemble des valeurs réelles.* |
| $x \vee y$ | : | $\max(x, y)$. |
| $:=$ | : | *Égalité par définition.* |
| $N(0, \sigma^2(t))$ | : | *Loi normale d'éspérance* $0$ *et variance* $\sigma^2(t)$. |

# Bibliographie


[1] Analyses de Survie Jonathan Lenoir (MCU), jonathan.Unité "Écologie et Dynamique des Systèmes Anthropisés"

[2] Commenges, D., & Jacqmin-Gadda, H. (2015). Modèles biostatistiques pour l'épidémiologie. De Boeck Superieur.

[3] Courgeau, D., & Lelièvre, E. (1989). Analyse démographique des biographies. Ined.

[4] Elisa, T. Lee. (2003). Statistical methods for survival data analysis. JohnWiley & Sons.

[5] Fermanian, J.D. Modèles de durées. Cours ENSAE 3ème année.http ://www.crest.fr/ckfinder/userfiles/pageperso/fermania/JDF_durée3.pdf.

[6] Fleming, T. R., & Harrington, D. P. (1991). Counting processes and survival analysis john wiley & sons. Inc. New York.

[7] Klein,M. (2003). Survival analysis techniques for censored and truncated data. Spring.

[8] Kaplan, E.L., Meier, P. 1958. Nonparamétric estimation from incomplete observations. J.A.Statist. Assoc. 53, 457-481.

[9] Lee, E. T., & Wang, J. (2003). Statistical methods for survival data analysis (Vol. 476). John Wiley & Sons.

[10] Lafont-Terranova, J. (2009). Se construire, à l'école, comme sujet-écrivant : l'apport des ateliers d'écriture (Vol. 15). Presses universitaires de Namur.

[11] Lorino, T. (2002). Modèles statistiques pour des données de survie corrélée(Doctoral dissertation, Institut national agronomique paris-grignon-INA PG).

[12] M'ziou, I. (2014). Mémoire de master. Estimation non paramétrique, Université Mohamed Khider, Biskra

[13] Saint Pierre, P. (2015). Introduction à l'analyse des durées de survie. Université Pierre et Marie Curie, France.

[14] Semmari, M. (2020). l'Analyse de Survie et Applications.

[15] Sibide, I. D. B. (2014). Analyse non-paramétrique des politiques de maintenance basée sur des données des durées de vie hétérogènes (Doctoral dissertation, Université de Lorraine).

[16] Soltane, L. (2017). Analyse des valeurs extrêmes en présence de censure (Doctoral dissertation, Université Mohamed Khider-Biskra).

[17] Planchet, F., & Thérond, P. (2006). Modèles de durée. Economica.

[18] Viallon, V. (2006). Processus empiriques, estimation non paramétrique et données censurées (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

[19] Zerroukhi, H. (2019). Mémoire de master. Estimation de l'Indice des valeurs extrémes sous censure. Université Mohamed Khider, Biskra, Algerie.