République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE MOHAMED KHIDER, BISKRA

FACULTE des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DEPARTEMENT DE MATHEMATIQUES

Mémoire présenté par

Ben Terki Nousseiba

Pour l'obtention du Dipmlôme de

MASTER en Mathématiques

Option : **Statistique**

Titre

Introduction à l'Analyse de Survie

Membres du comité d'examen

Pr.	Abdelhakim Necir	UMKB	Président
Pr.	Djamel Meraghni	UMKB	Encadreur
Dr.	Nour Elhouda Zouaoui	UMKB	Examinateur

Juin 2024

Dédicace

À ma chère famille et à mes amis,

À vous qui avez été mes piliers tout au long de ce parcours académique, je vous dédie humblement cette recherche. Votre soutien indéfectible, vos encouragements chaleureux et vos précieux conseils ont été une source de force et d'inspiration pour moi à chaque étape de ce voyage. Cette réussite est autant la vôtre que la mienne, car c'est grâce à votre amour et à votre soutien que j'ai pu atteindre ce point. Que cette dédicace soit un témoignage de ma profonde gratitude envers vous tous.

Avec toute ma reconnaissance et mon amour,

Nousseiba Ben Terki

REMERCIEMENTS

Je remercie Dieu Le Tout-Puissant de m'avoir donné la volonté, la force et le courage pour bien mener et finir mon travail de mémoire de master.

Tout d'abord, je voudrais remercier sincèrement mon directeur de recherche, Pr. MERAGHNI Djamel, pour son soutien constant, ses conseils éclairés et son expertise inestimable tout au long de ce projet. Son mentorat a été essentiel pour orienter mes efforts et pour m'aider à atteindre mes objectifs académiques.

Je suis également reconnaissante envers mes parents et ma famille pour leur amour inconditionnel, leur soutien indéfectible et leur encouragement constant tout au long de ce voyage.

Je tiens également à remercier tous mes enseignants qui m'ont guidée, inspirée et partagé leur expertise avec moi tout au long de mon cursus universitaire. Leurs enseignements ont enrichi mes connaissances et ont contribué à façonner ma pensée critique.

Enfin, je suis reconnaissante envers toutes les autres personnes qui ont apporté leur soutien, leur expertise et leur encouragement à différents moments de ce voyage académique. Votre soutien et vos contributions ont été essentiels pour que cette recherche voie le jour et pour que j'atteigne ce jalon important dans ma vie académique. Je vous en suis profondément reconnaissante.

Avec toute ma gratitude.

Nousseiba Ben Terki

Table des matières

Table des figures

Liste des tableaux

Introduction

Dans de nombreux domaines tels que la médecine, l'économie et les sciences sociales, comprendre le temps écoulé avant qu'un événement ne se produise est crucial pour la prise de décision et la planification stratégique. L'analyse de survie fournit un cadre statistique pour étudier ces événements et les facteurs qui les affectent. Une introduction à l'analyse de survie est essentielle pour permettre aux chercheurs et aux praticiens de comprendre et d'appliquer efficacement cette méthode statistique importante dans divers domaines de recherche et d'application.

L'objectif de l'analyse de survie est de modéliser l'espérance de vie des organismes biologiques ou le temps restant avant l'échec ou la panne dans les systèmes artificiels, que l'on représente graphiquement sous la forme d'une courbe de survie. Pour mieux comprendre, on se pose les questions suivantes : qu'est-ce que l'analyse de survie et quels sont les principaux modèles et méthode utilisés ?

Dans ce travail, on s'intéresse à trouver une estimation appropriée de la fonction de survie basée sur certaines données dans le cas non paramétrique. Ce mémoire se compose de deux chapitres :

- Le premier chapitre contient les notions de base de l'analyse de survie et les deux phénomènes causant des données incomplètes : censure et troncature. Les informations précédentes sont obtenues des références [9], [3], [12] et [6].
- Le second chapitre traite l'estimation non paramétrique de la fonction de survie sous des données incomplètes. Les estimateurs les plus répandus sont l'estimateur de Kaplan-

Meier pour les données censurées et celui de Lynden-Bell dans le cas de données tronquées. On suggère les références [10], [4], [11] et [7]. Des exemples d'application sur des données réelles et simulées sont traités à la fin du chapitre.

Enfin, il y a lieu de noter que les calculs numériques et les représentations graphiques sont réalisés à l'aide des packages "cluster","coin" et "survival" du logiciel d'analyse statistique R.

Chapitre 1

Eléments d'analyse de survie

On trouve ci-dessous un aperçu des outils et des concepts permettant de comprendre l'analyse de survie.

1.1 Concepts de base

1.1.1 Qu'est-ce que l'analyse de survie ?

Il s'agit d'une méthode statistique qui traite de l'analyse des données de temps (en général) durant la période d'une étude statistique. Elle sert à estimer la probabilité qu'un événement particulier se produise à un instant donné. L'étude concerne une population d'individus qui peuvent être des personnes ou autres. L'analyse de survie est largement utilisée dans la recherche médicale, l'ingénierie et les sciences sociales. a titre d'exemple, l'événement d'intérêt peut être un décès, une maladie ou une défaillance d'un organe pour une personne (médecine), une panne pour une machine (mécanique) ou un divorce pour un couple (sociologie).

Le but de l'analyse de survie est d'estimer la fonction de survie, qui décrit la probabilité qu'un événement ne se produit pas, et d'identifier les facteurs associés au risque de l'événement.

1.1.2 Données incomplètes

Il arrive que lors d'une étude statistique, on ne dispose pas de la totalité des données. On dit qu'on est en présence de données incomplètes dont le traitement nécessite l'analyse de survie. Les données incomplètes sont dues à deux phénomènes différents : la censure et la troncature.

1.2 Censure

Une observation est dite censurée si l'événement d'intérêt ne s'est pas produit pour l'individu en question ou le temps auquel l'événement s'est produit n'est pas connu. La censure représente un défi dans l'analyse de survie car elle conduit souvent à des données incomplètes, ce qui peut biaiser l'estimation et affecter la validité de l'analyse statistique. Selon l'étude statistique, il existe plusieurs catégories de censure, qu'on décrit ci-dessous.

1.2.1 Censure à droite

On dit que la variable d'intérêt est censurée à droite si on n'a aucune information sur la dernière observation en relation avec l'individu concerné. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées.

Exemple 1.2.1

On transplante 4 nanopuces dans les cœurs de 4 bébés tortues marines numérotés de 1 à 4 dont les œufs ont éclos tous en même temps puis on les lâche dans l'océan. Ces nanopuces mesurent les battements de cœur de ces bébés tortues et sont connectées de façon continue à un outil de mesure dans lequel on reçoit :

- le nombre de battements de cœur du bébé tortue tant que celui ci est vivant.
- le message «Décès» au moment où le coeur du bébé tortue s'arrête de battre.
- le message «Erreur» quand le signal avec la nanopuce est perdu.

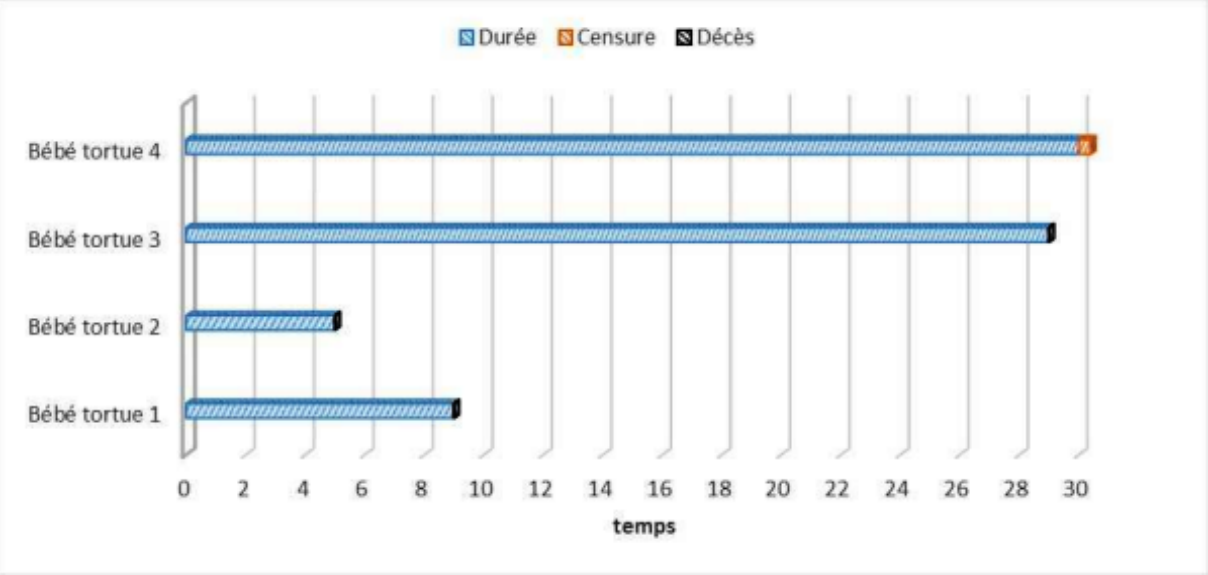

FIG. 1.1 – Censure à droite

On désigne par X la durée de vie d'un bébé tortue. On dispose d'un échantillon, de taille 4, $(X_1, ..., X_4)$ des durées de vie des bébés tortues transplantés. On a suivi cette cohorte et au bout d'un mois on a constaté ce qui suit :

- le bébé tortue n°1est décédé au bout de 9 jours.
- on a perdu tout contact avec la nanopuce du bébé tortue n° 2 au bout de 5jours.
- le bébé tortue n°3 est décédé au bout de 29 jours.
- le bébé tortue n°4 est toujours en vie.

Ainsi, ce que l'on sait sur les durées de vie des ces tortues est comme suit :

– pour le bébé tortue n^{o}1 : on observe $X_1 = 9$ jours, qui est sa durée de vie.

– pour le bébé tortue n^o 2 : on observe $C_2 = 5$ jours qui est la durée pendant laquelle il a survécu jusqu'à perte de tout signal avec la nanopuce. Même si l'on ne la connait pas avec exactitude, on sait que sa durée de vie X_2 est forcément plus grande que C_2.

– Pour le bébé tortue n^{o}3 : on observe $X_3 = 29$ jours.

– Pour le bébé tortue n^{o}4 : on reçoit toujours des signaux de la nanopuce. Donc, on observe $C_4 = 30$ jours et on sait que forcément $X_4 > C_4$.

1.2.2 Censure à gauche

Il y a censure à gauche lorsque l'individu a déjà subi l'événement d'intérêt avant qu'il ne soit observé, c'est à dire avant le début de l'étude. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue.

Exemple 1.2.2 ([9]) *On veut étudie en fiabilité un certain composant électronique monté en paralléle avec un ou plusieurs autres composants. Une panne de ce composant n'entraine pas nécessairement l'arrêt du système :le système peut continuer à fonctionner jusqu à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du systéme). La durée observée pour ce composant est alors censurée à gauche.*

1.2.3 Censure par un intervalle

Dans ce cas, comme le nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. En général, on retrouve ce modèle dans les études de suivi médical où les patients sont contrôlés et se présentent ensuite après que l'événement d'intérêt se soit produit. On a aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type de censure est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles de la forme $[C, \infty[$ ou $[0, C]$ respectivement.

Exemple 1.2.3 ([3]) *Cet exemple concerne une étude de suivi de patients prenant des biothérapies. L'apparition d'anticorps, anti-biothérapies (ADA) chez un patient ne peut être constaté que lors d'une visite médicale. Sur le graphique de la Figure 1.2, l'individu 2 a effectué deux visites : la première au temps C_1 et la seconde en C_3. Si lors de la seconde visite, il est déclaré ADA-positif, la seule information sur la date de positivité du patient est un intervalle de temps entre les deux visites, soit $X_2 \in [C_1, C_3]$.*

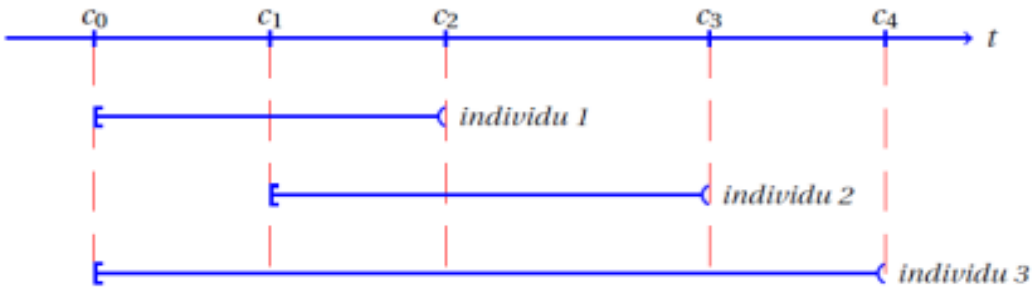

FIG. 1.2 – Censure par intervalle([3])

Censure double (mixte)

Ce type de censure est un mélange entre la censure à droite et la censure à gauche, dans le même échantillon.

Exemple 1.2.4 *Un ethnologue étudie la durée d'apprentissage d'une tâche par un enfant. On désigne par X et C les v.a's représentant la durée d'apprentissage et l'âge de l'enfant respectivement. Pour un enfant qui sait déjà accomplir la tâche, C censure X à gauche car peut eux X est inconnu mais inférieur à C c.à.d $X < C$. Cet exemple comporte aussi des censures droites. En effet, la durée d'apprentissage d'un enfant qui ne sait pas encore accomplir la tâche en question lors du départ de l'ethnologue est censuré à droite par la durée d'apprentissage C' observée par l'ethnologue : on a $X > C'$.*

Remarque 1.2.1 *Il est à noter que chacune des quatre catégories de censure décrites ci-dessus peuvent se présenter en fonction du mode ou mécanisme de censure suivant :*

1. ***Censure de type 1 (fixé) :*** *dans ce cas la durée de l'expérience est fixée d'avance. Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Soit C une valeur (déteministe : durée, par exemple) fixée. En censure à droite, au lieu d'observer toutes les variables $X_1, ..., X_n$ qui nous intéressent, n'on observe X_j que lorsqu' elle est inférieure ou égale à la valeur C. On observe donc une variable Z_j telle que $Z_i := \min(X_i, C)$, $i = 1, ...n$.*

2. ***Censure de type 2 (attente) :*** *dans ce cas l'expérimentateur fixe le nombre d'événements à observer. Alors, la date de fin d'expérience devient aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie,... L'inconvénient de ce type est que la durée de l'expérience peut être longue.*

3. ***Censure type 3 (aléatoire) :*** *dans ce type d'expérience, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue. Ce modèle est utilisé pour les essais thérapeutiques. Soit $X_1, ..., X_n$ un échantillon, de taille $n \geq 1$, d'une v.a positive X. On dit qu'il y a censure aléatoire (à droite) de cet échantillon s'il existe une autre v.a positive C, d'échantillon $C_1, ..., C_n$, appelée variable de censure. Dans ce cas, au lieu d'observer les X_i, on observe un couple de v.a's (Z_i, δ_i) avec*

$$Z_i := \min(X_i; C_i) \text{ et } \delta_i := \mathbb{1}\{X_i \leq C_i\}, \ i = 1, ..., n, \tag{1.1}$$

où δ_i est l'indicateur de censure, qui détermine si X a été censurée ou non. On a :

$$\delta_i = \begin{cases} 1, & X_i \text{ est observée} \quad Z_i = X_i \\ 0, & X_i \text{ est censurée} \quad Z_i = C_i \end{cases}$$

1.3 Troncature

Une observation est dite tronquée si elle est conditionnelle à un autre événement. On dit qu'une v.a (durée de vie, par exemple) est tronquée si elle n'est observable que sous une certaine condition dépendant de sa valeur. La troncature se produit lorsque les observations sont exclues en raison du délai d'apparition de l'événement d'intérêt. Contrairement à la censure, la troncature influe directement sur la composition de l'échantillon dont la taille est réduite (par la troncature). Par exemple, une étude s'intéresse uniquement aux salariés qui gagnent plus de 25000 $ par an. Ainsi, toute personne gagnant moins de cette somme

est simplement supprimée de l'ensemble des données. Il y a plusieurs formes de troncature, chacune présentant ses propres défis et implications pour l'analyse statistique. Les plus courantes sont décrite ci-dessous. Pour cela, on dispose d'une v.a. X et d'une autre v.a. T indépendante de X.

1.3.1 Troncature à gauche

Définition 1.3.1 *On dit que X est tronquée à gauche par T lorsque X n'est observable que si $X \geq T$.*

Remarque 1.3.1 *Les valeurs de X qui sont inférieures ou égales à T ne sont pas observées et ne sont pas incluses dans l'échantillon.*

Exemple 1.3.1 *On désigne par X la durée de vie d'un individu. On étudie cette variable à partir d'une cohorte tirée au sort dans une une ceratine population. Seule la survie des sujets vivants à la date T d'inclusion dans la cohorte pourra être étudiée. Il y a troncature à gauche de X par T, car seuls les sujets ayant survécu jusqu'à et après T sont observables.*

1.3.2 Troncature à droite

Définition 1.3.2 *On dit qu il y a troncature à droite de X par T lorsque X n'est observable que si $X < T$.*

Remarque 1.3.2 *Les valeurs de X qui sont supérieures ou égales à T ne sont pas observées et ne sont pas incluses dans l'échantillon.*

Exemple 1.3.2 ([1]) *.Dans le problème relatif au SIDA transmis par transfusion, la variable d'intérêt est ici la durée d'induction X de la maladie, durée qui s'écoule entre la date d'infection$Y$ et la date $(Y + X)$ de déclaration de la maladie. On suppose que l'observation a lieu entre deux dates fixes c (la date de transfusion) et b(fixé), voir la figure 1.3.*

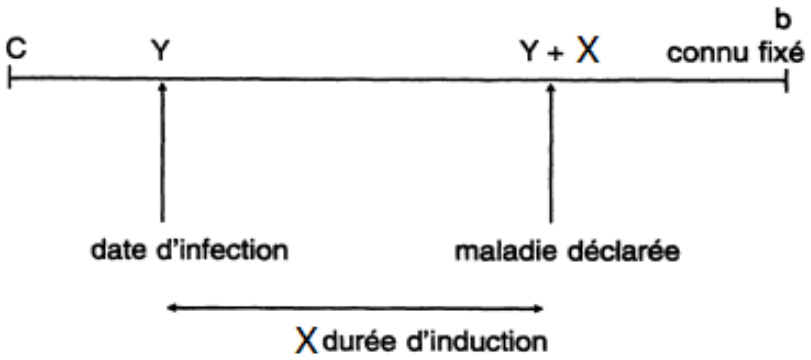

FIG. 1.3 – Schéma correspondant au SIDA ([1])

1.3.3 Troncature par intervalle

Définition 1.3.3 *On dit qu'il y a troncature par intervalle lorsque X est tronqué à gauche par une variable T_1 et à droite par une autre variable T_2.*

Exemple 1.3.3 *On rencontre ce type de troncature lors de l'étude des patients d'un registre : les patients diagnostiqués avant la mise en place du registre (tronqués à gauche) ou répertoriés après la consultation du registre (tronqués à droite) ne seront pas inclus dans l'étude.*

1.4 Distributions de la durée de survie

On désigne le temps de survie par la v.a. X (continue, positive). La distribution de X est entièrement caractérisée par l'une des cinq fonctions suivantes.

1.4.1 Fonction de répartition

Définition 1.4.1 *La fonction de répartition (f.d.r) de X (ou de sa loi P_X) est une fonction F définie sur $\mathbb{R}_+$par :*

$$F(t) := P(X \leq t), \ t \geq 0. \tag{1.2}$$

Pour t fixé, $F(t)$ représente la probabilité de mourir avant l'instant t.

Propriété 1.4.1

1. *La fonction F est une fonction croissante sur $\mathbb{R}_+$.*
2. *F est continue à droite sur $\mathbb{R}_+$.*
3. *On a :*

$$\lim_{t\to 0_+} F(t) = 0 \text{ et } \lim_{t\to\infty} F(t) = 1.$$

1.4.2 Fonction de survie

Définition 1.4.2 ([12]) *La fonction de survie S est, pour t fixé, la probabilité de survivre au delà de l'instant t, c'est à-dire*

$$S(t) := P(X > t),\ t \geq 0. \tag{1.3}$$

La fonction de survie d'une va X est décroissante, continue à gauche sur $\mathbb{R}_+$ et vérifie :

$$\lim_{t\to 0_+} S(t) = 1 \text{ et } \lim_{t\to\infty} S(t) = 0.$$

- Les fonctions F et S sont aussi appellées fonction de distribution et queue de distribution repectivement.
- Elles sont liées par la relation :

$$S(t) = 1 - F(t),\ t \geq 0. \tag{1.4}$$

1.4.3 Fonction de densité de probabilité

Comme toute autre v.a. continue, X a une fonction de densité de probabilité f définie ci-dessous.

Définition 1.4.3 *La densité de probabilité de X est égale à la dérivée de sa f.d.r :*

$$f(t) = F'(t),$$

où t est un point de $\mathbb{R}_+$ où la dérivée existe.

Remarque 1.4.1

1. *On a :*
$$f(t) \geq 0,\ t \geq 0 \text{ et } \int_0^\infty f(t)dt = 1.$$

2. *La fonction de répartion est égale à l'intégrale de la densité :*
$$F(t) = \int_0^t f(u)du,\ t \geq 0.$$

3. *La densité de probabilité de X est aussi définie en termes de la fonction de survie :*
$$f(t) = -S'(t).$$

4. *Pour $t_0 \geq 0$ fixé, $f(t_0)$ représente la probabilité de mourir dans un petit intervalle $[t_0, t_0 + h]$ de temps après l'instant t_0 :*
$$f(t_0) = \lim_{h \to 0} \frac{F(t_0 + h) - F(t_0)}{h} = \lim_{h \to 0} \frac{P(t_0 \leq X \leq t_0 + h)}{h}. \tag{1.5}$$

1.4.4 Fonctions de risque et de risque cumulé

Définition 1.4.4 (risque) *La fonction du risque, notée par λ, est défini par :*

$$\lambda(t) := \lim_{h \to 0} \frac{P(t \leq X \leq t + h/X > t)}{h},\ t \geq 0. \tag{1.6}$$

On l'appelle aussi, selon les domaines d'application, "taux instantané de défaillance", "taux de risque", "taux de hasard" ou encore, "quotient de mortalité". Pour ces différentes appellations, on renvoie au livre [11], page 19.

Cette fonction, qui est une probabilité conditionnelle, donne habituellement plus d'information, que la fonction de survie, sur le mécanisme de décés. C'est pourquoi elle est souvent employée pour résumer les données de survie.

Remarque 1.4.2

1. *La fonction 1.6 peut également être définie en termes de la f.d.r F (ou de la fonction de survie) et de la densité de probabilité f :*

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}, \ t > 0.$$

2. *La fonction λ peut aussi être écrite sous la forme d'une équation différentielle :*

$$\lambda(t) = -\frac{d}{dt} \ln S(t), \ t > 0. \tag{1.7}$$

Définition 1.4.5 (risque cumulé) *La fonction du risque cumulé, qu'on note par Λ, est égale à l'intégrale de fonction du risque :*

$$\Lambda(t) := \int_0^t \lambda(u)du, \ t > 0. \tag{1.8}$$

La fonction Λ permet (aussi) de caractériser la loi de la v.a. X. En effet, on a le résultat suivant :

Proposition 1.4.1 *On exprime $F(t)$ à partir de $\Lambda(t)$:*

$$F(t) = 1 - \exp\{-\Lambda(t)\}, \ t > 0.$$

Preuve. L'équation différentielle (1.7) a pour solution :

$$\Lambda(t) = \int_0^t \lambda(u)du = -\int_0^t \frac{d}{du} \ln S(u) = -\ln S(t).$$

En multipliant les deux membres par -1, puis en prenant l'exponentielle, on obtient :

$$S(t) = \exp\{-\Lambda(t)\}.$$

à partir de 1.4, on obtient la formule désirée. ■

Enfin, il est à noter que toutes les fonctions définies par (1.2) , (1.3) , (1.5) , (1.6) et (1.8) sont liées entre elles. En d'autres termes, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également définies. Chacune d'entre elles permet de caractériser la distribution de la variable d'intérêt X.

Chapitre 2

Estimation sous données incomplètes

Il existe plusieurs façons d'analyser la survie. Le choix de la méthode dépend de la nature des données et de la question de recherche. Dans ce qui suit, on discute l'estimation non paramétrique de la fonction de survie (à partir de laquelle on déduit les estimations des quantités qui lui sont liées). On choisit les cas de censure aléatoire à droite et de troncature aléatoire à gauche.

2.1 Estimation sous censure

Soient $(\Omega, \mathcal{A}, P)$ un espace probabilisé et $X_1, ..., X_n$ une suite de v.a's $i.i.d$ positives, de f.d.r commune F et $C_1, ..., C_n$ une suite de v.a's de censure $i.i.d$ positives. On suppose aussi que C_i sont indépendantes des X_i. Soit $\{(Z_i, \delta_i), i = 1, ..., n\}$ l'échantillon réellement observé défini par (1.1) dans le cadre le plus fréquent d'une censure à droite de type 3.
On rappelle qu'en l'absence de censure, la f.d.r F est estimée de manière très simple en utilisant la fonction de répartition empirique usuelle F_n definie par

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}, \ t \geq 0.$$

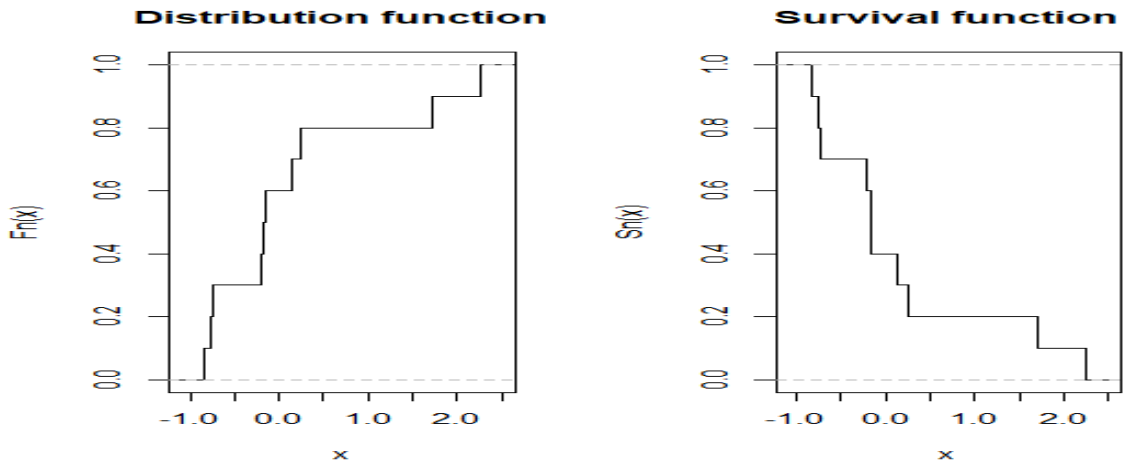

FIG. 2.1 – Fonctions de répartition (à gauche) et de survie (à droite) d'un échantillon de taille 20 d'une variable exponentielle standard.

La fonction de survie empirique S_n correspondante est donc

$$S_n(t) := 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > t\}},\ t \geq 0.$$

Exemple 2.1.1 *On génére un échantillon de taille* 20 *d'une variable exponentielle standard (de paramètre* 1*). Les fonctions de répartition et de survie empiriques sont représentées dans la figure 2.1.*

Mais, dans le cas où les données sont censurées, il est impossible d'utiliser la fonction S_n puisqu'elle comprend des quantités non observées (tous les X_i censurés ne sont pas observés). Il est donc nécessaire de construire un estimateur de la fonction de survie 1.3 en présence de données censurées.

2.1.1 Estimateur de Kaplan-Meier

En 1958, Kaplan et Meier [4]ont proposé un estimateur non paramétrique, appelé estimateur de Kaplan-Meier.

Construction de l'estimateur

L'idée de la construction est la suivante : survivre après un instant t c'est être en vie juste avant t et ne pas mourir au temps t. Pour $0 \leq t^{''} \leq t^{'} \leq t$, on a :

$$\begin{aligned} S(t) &= P(X > t) \\ &= P(X > t^{'}, X > t) \\ &= P(X > t / X > t^{'}) \times P(X > t^{'}) \\ &= P(X > t \setminus X > t^{'}) \times P(X > t^{'} / X > t^{''}) \times P(X > t^{''}). \end{aligned}$$

On considère les temps d'événements (décès et censure) distincts $t = Z_{i,n}$, $i = 1, ..., n$, avec $Z_{0,n} = 0$, ordonnés par ordre croissant, on obtient :

$$P(X > Z_{i,n}) = \prod_{k=1}^{i} P(X > Z_{k,n} / X > Z_{k-1,n}). \tag{2.1}$$

On considère les deux notations suivantes :

$*$ y_i le nombre d'individus à risque de subir l'événement (mourir) juste avant $Z_{i,n}$.

$*$ d_i le nombre de décès en $Z_{i,n}$.

Alors la probabilité p_i de mourir dans l'intervalle $]Z_{i-1,n}, Z_{i,n}]$ sachant qu'on était vivant en $Z_{i-1,n}$, c.à.d $p_i = P(X \leq Z_{i,n} / X > Z_{i-1,n})$, peut être estimée par :

$$\hat{p}_i = \frac{d_i}{y_i}. \tag{2.2}$$

Remarque 2.1.1 *Comme les temps d'événements sont supposés distincts, on a :*

$$d_i = \begin{cases} 0, & \textit{censure en } Z_{i,n} \ \ (\delta_i = 0), \\ 1, & \textit{décès en } Z_{i,n} \ \ (\delta_i = 1) \end{cases}$$

Dans le cas où il y a des ex-aequo :

- *Si ce sont des événements de nature différente, on considère que les observations non censurées ont lieu avant celles qui sont censurées.*
- *S'il y a plusieurs décès au même temp $Z_{i,n}$, alors $d_i > 1$.*

En combinant les équations (2.1) et (2.2) , on obtient la définition (usuelle) suivante de l'estimateur de Kaplan-Meier du temps de survie à l'instant t.

Définition 2.1.1 *L'estimateur de Kaplan-Meier de la fonction de survie S à l'instant t est défini par :*

$$S_n^{KM}(t) := \prod_{Z_{i,n} \leq t} (1 - \frac{d_i}{y_i}),\ t \geq 0. \tag{2.3}$$

Remarque 2.1.2

1. *L'estimateur S_n^{KM} est parfois appelé estimateur du produit limite (P-L) car ils'obtient comme la limite d'un produit.*
2. *L'estimateur de Kaplan-Meier est une fonction en escalier décroissante qui fait des sauts à chaque instant t_i. La valeur du saut dépend du nombre d'événements au temps t_i et aussi du nombre de censures à ce temps là.*
3. *L'estimateur de Kaplan-Meier est continue à droite par construction.*
4. *Le nombre y_i est égal à $n-i+1$. Si on note par $\delta_{[i,n]}$ le concomitant de la $i^{ème}$ statistique d'ordre $Z_{i,n}$ $\left(\delta_{[i,n]} = \delta_j \text{ si } Z_{i,n} = Z_j, 1 \leq j \leq n\right)$, alors on peut écrire $S_n^{KM}(t)$ sous la forme suivante :*

$$S_n^{KM}(t) = \prod_{Z_{i,n} \leq t} (1 - \frac{\delta_{[i,n]}}{n-i+1}),\ t < Z_{n,n}.$$

5. *D'autres formes sont obtenues à partir de la formule ci-dessus :*

$$S_n^{KM}(t) = \prod_{Z_{i,n} \leq t} (\frac{n-i}{n-i+1})^{\delta_{[i,n]}} = \prod_{i=1}^{n} (1 - \frac{\delta_{[i,n]}}{n-i+1})^{\mathbf{1}_{\{Z_{i,n} \leq t\}}},$$

qui, d'après [10] page 122, *peut réécrite sous la forme d'une somme :*

$$S_n^{KM}(t) = \sum_{i=1}^{n} \frac{\delta_{[i,n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j,n]}} \mathbb{1}_{\{Z_{i,n} \leq t\}}$$

Propriétés

Les propriétés asymptotiques de l'estimateur de Kaplan-Meier sont résumées dans la proposition suivante.

1. Absence de biais ,pour tout t ,on a :

$$\lim_{n\to\infty} E\left[S_n^{KM}(t)\right] = S(t)$$

2. Consistance uniforme :soit $x_h = H^{-1}(1) := \inf\{t : H(t) = 1\} \leq \infty$,où $H(t)$ est la f.d.r de Z. Alors :

$$\sup_{0\leq t<x_h} \left|S_n^{KM}(t) - S(t)\right| \xrightarrow{p.s} 0, \text{ quand } n \to \infty.$$

3. Normalité asymptotique : pour tout $t \geq 0$, on a

$$\sqrt{n}\left(S_n^{KM}(t) - S(t)\right) \xrightarrow{D} \mathbf{X}_t.$$

où $\mathbf{X}_t$ est un processus Gaussien centré.

Intervalle de confiance

L'estimateur de Kaplan Meier est asymptotiquement normal de moyenne $S(t)$. Ainsi,

$$\frac{S_n^{KM}(t) - E\left[S_n^{KM}(t)\right]}{\sqrt{Var\left(S_n^{KM}(t)\right)}} \sim \mathcal{N}(0,1).$$

Donc, l'intervalle de confiance asymptotique à 95% de $S(t)$ est :

$$\left[\hat{S}_n(t) \pm 1,96 \times \sqrt{Var\left(S_n^{KM}(t)\right)}\right],$$

où $Var\left(S_n^{KM}(t)\right)$ désigne la variance de $S_n^{KM}(t)$ pour laquelle il existe un estimateur appelé estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier et defini par :

$$\widehat{Var(S_n^{KM}(t))} := \left(S_n^{KM}(t)\right)^2 \sum_{Z_{i,n} \leq t} \frac{d_i}{y_i\left(y_i - d_i\right)}$$

2.1.2 Test du Log-Rank

C'est un test statistique qui permet de comparer les courbes de Kaplan-Meier. Cela revient à comparer les durées de vies entre deux ou plusieurs groupes en fonction du sexe, âge, etc. Pour des ensembles de données complètes, il est courant d'utiliser le test de Kolmogorov-Smirnov et/ou le test de Mann-Withney. Par contre, la présence des données censurées nécessite l'utilisation d'autres types de tests tels que le test de Wilcoxon généralisé (test de Gehan), le test du Log-Rank. Ce dernier est le plus populaire pour comparer plusieurs courbes de survie. C'est un test non-paramétrique, dont la démarche est la suivante (voir [5]) :

On considère deux groupes A et B et on désigne par $S_A(t)$ et $S_B(t)$ leur fonctions de survie respectives. Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 : S_A(t) = S_B(t) \text{ (survie identique entre les groupes)} \\ H_1 : S_A(t) \neq S_B(t) \text{ (survie différente entre les groupes)} \end{cases}$$

Principe du test Log- Rank

Soient $t_1, t_2, ..., t_k$ les temps de décès observés dans les deux groupes A et B, tels que :

– d_{A_i} : nombre de décès observés dans A en t_i.

– d_{B_i} : nombre de décès observés dans B en t_i.

– d_i : nombre global de décès observés en t_i.

– y_{A_i} : nombre de sujets exposés au risque dans A en t_i.

– y_{B_i} : nombre de sujets exposés au risque dans B en t_i

– y_i : nombre global de sujets exposés au risque en t_i.

On calcule, en chaque temps de décés observé, les quantités e_{A_i} et e_{B_i} représentant les nombres de décès attendus en t_i, sous l'hypothèse nulle H_0, dans les groupes A et B. Ces nombres sont respectivement définis par :

$$e_{A_i} = d_i \frac{y_{A_i}}{y_i} \text{ et } e_{B_i} = d_i \frac{y_{B_i}}{y_i},\ i = 1, ..., k.$$

Puisque ce n'est pas evident de faire le test pour chaque instant t_i $(i = 1, ..., k)$, alors on calcule les quantités précédentes pour tous les temps t_i. On obtient alors :

– $E_A = \sum_{i=1}^{k} e_{A_i}$: nombre total de décès attendus dans A sous H_0.

– $E_B = \sum_{i=1}^{k} e_{B_i}$: nombre total de décès attendus dans B sous H_0.

– $O_A = \sum_{i=1}^{k} d_{A_i}$: nombre total de décès observés dans A.

– $O_B = \sum_{i=1}^{k} d_{B_i}$: nombre total de décès observés dans B.

L'écart (au carré) entre le nombre O_A observé (empirique) et le nombre E_A attendu (théorique) sous l'hypothèse nulle H_0, relativement à E_A, est distribué selon la loi du khi-deux à un degré de liberté (ddl). Il en est de même pour l'écart entre O_B et E_B :

$$\frac{(O_A - E_A)^2}{E_A} \rightsquigarrow \chi_1^2 \text{ et } \frac{(O_B - E_B)^2}{E_B} \rightsquigarrow \chi_1^2.$$

On définit la somme des deux écarts ci-dessus par :

$$\chi^2 := \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B},$$

et on l'appelle valeur du khi-deux calculée. Sous l'hypothèse H_0 (relation entre les deux

groupes), cette dernière est de distribtion du khi-deux à $(2-1)=1$ ddl.

La décision est :

– Si χ^2 est supérieure au khi-deux tabulé, on rejette H_0.

– Si χ^2 est inférieure au khi-deux tabulé, on accepte H_0.

Remarque 2.1.3 *Le test du log rank peut-être utilisé pour comparer les courbes de survie de q groupes, avec $q \geq 2$. Le critère statistique χ^2suit alors la loi du khi-deux à $(q-1)$ ddl.*

2.2 Estimation sous troncature

Compte tenu de la présence de données tronquées, on choisit le type d'estimation qui correspond à la nature de ces données. L'estimateur le plus connu dans le cas de la troncature aléatoire à gauche, introduite dand la définition 1.3.1, est dû à de Lynden-Bell.

2.2.1 Estimateur de Lynden-Bell

Soient $X_1, ..., X_N$ une suite de v.a's réelles (d'intérêt) i.i.d, de f.d.r commune F, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. Soient $T_1, ..., T_N$ une suite de v.a's (de troncatures) i.i.d de f.d.r continue L. On suppose aussi que ces variables sont indépendantes des X_i. La fonction de répartition conjointe de X et T est :

$$H(x,t) = P(X \leq x, T \leq t) = F(x)L(t).$$

La taille de l'échantillon $N \geq 1$ est déterministe.

Soit $\{(X_i^*, T_i^*),\ i = 1, ..., n\}$ l'échantillon réellement observé $(X_i \geq T_i)$. Une première conséquence de la troncature est que la taille de l'échantillon diminue, c.à.d $n \leq N$, et devient indéterminée. En d'autres termes, la taille n de l'échantillon vraiment observé est une v.a. Elle est distrubée selon la loi Binomiale de paramètres N et $\mu := P(X_i \geq T_i) > 0$. En

vertu de la loi forte des grands nombres on a :

$$\mu_n = \frac{n}{N} \xrightarrow{p.s} \mu, \text{ quand } N \to \infty.$$

Conditionnellement à la valeur de n, les données observées (X_i^*, T_i^*) sont encore i.i.d, mais leur f.d.r conjointe diffère de celle des données initiales (X_i, T_i) :

$$\begin{aligned} H^*(x,t) &= P(X \leq x, T \leq t \ / \ X \geq T) \\ &= \mu^{-1} \int_{-\infty}^{x} L(t \wedge z) dF(z). \end{aligned}$$

Pour définir les estimateurs de F et L, on a besoin d'introduire la fonction :

$$C(x) := P(T \leq x \leq X \ / \ X \geq T),$$

qui peut être estimée empiriquement par :

$$C_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{T_i \leq x \leq X_i\}.$$

En 1971, Lynden-Bell [7] introduisit les estimateurs de maximum de vraisemblance non paramétriques de F et L, donnés par les estimateurs produit-limite suivants :

$$F_n(x) = 1 - \prod_{i:X_i \leq x} \left(1 - \frac{1}{nC_n(X_i)}\right) \quad \text{et} \quad L_n(x) = \prod_{i:X_i > x} \left(1 - \frac{1}{nC_n(T_i)}\right).$$

Remarque 2.2.1 *Les propriétés asymptotiques de ces estimateurs ont été étudiées par* ***Woodroofe*** ([13]).

2.3 Exemples d'application

Cette section est consacrée à l'application de l'estimation de Kaplan-Meier sur des données réelles et simulées. Pour cela, on utilise les packages **survival, cluster, coin** du logiciel d'analyse statistique R [2]

2.3.1 Données simulées

Les temps de survie X et de censure C sont supposés de loi exponentielle de paramètres respectifs 0.2 et 0.1. On extrait deux échantillons, de taille $n = 20$, de ces variables. Les résultats de simulation sont résumés dans le tableau 2.1.

$\mathbf{t}_i$	$\mathbf{n}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t}_i)$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
0.065	20	1	0.950	0.049	$0.859 - 1.000$
0.211	19	1	0.900	0.067	$0.777 - 1.000$
0.652	17	1	0.847	0.081	$0.702 - 1.000$
1.136	15	1	0.791	0.094	$0.6270 - 0.997$
1.308	14	1	0.734	0.103	$0.558 - 0.965$
1.857	12	1	0.673	0.112	$0.488 - 0.929$
3.235	9	1	0.598	0.121	$0.402 - 0.889$
4.319	7	1	0.513	0.131	$0.311 - 0.844$
4.695	4	1	0.385	0.148	$0.181 - 0.818$
5.250	3	1	0.256	0.144	$0.085 - 0.770$
5.763	2	1	0.128	0.116	$0.022 - 0.752$

TAB. 2.1 – Estimation de Kaplan-Meier (avec intervalle de confiance IC) du temps de survie de données exponentielles de paramètre 0.2 censurées par une variable exponentielle deparamètre 0.1. La valeur s représente l'écart-type de l'estimation.

2.3.2 Etude de cas

Données de Freireich

Freireich en 1963 a mené une expérience thérapeutique, pour comparer les durées de rémission, en semaines, de 21 patients atteints de leucémie selon qu'ils ont pris ou non un médicament appelé 6-mercaptopurnie (6-MP). Le groupe témoin a reçu un placebo. Cet

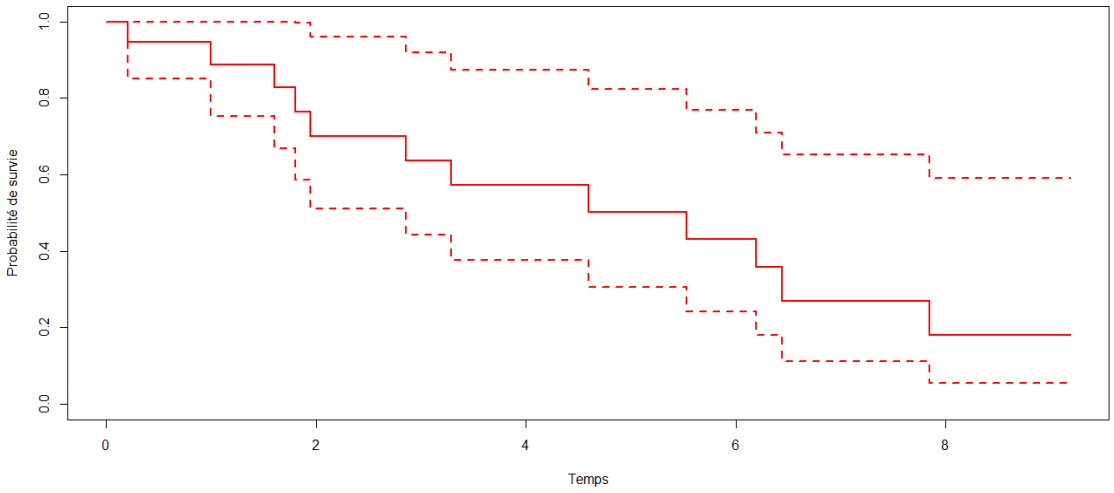

FIG. 2.2 – Fonction de survie de données exponentielles de paramètre 0.2 censurées par une variable exponentielle de paramètre 0.1. Les lignes pointillées représentent les bornes de confiance de niveau 95%.

6-MP	$6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$
placebo	$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$

TAB. 2.2 – Données de Freireich

exemple est célèbre dans l'illustration et l'application des outils de l'analyse de survie. Les données peuvent être trouvées dans plusieurs documents comme par exemple. [6], page 2.

Essai thérapeutique : étude contrôlée visant à préciser, sur une population sélectionnée et particulièrement surveillée, les effets d'un médicament sur une maladie bien précisée.

Placebo : préparation dépourvue de tout principe actif, utilisée à la place d'un médicament pour son effet psychologique, dit "effet placebo".

Les nombres suivis du signe + correspondent à des données censurées. Par exemple : le 9ème patient est perdu de vue au bout de 11 semaines de traitement avec le 6-MP : il a donc une durée de rémission supérieure à 11 semaines. Dans le groupe traité par le 6-MP, 9 seulement parmi les 21 valeurs sont observées.

Dans le groupe placebo, il n'y a aucune donnée censurée mais. Ceci entraine que, dans ce groupe, l'estimateur de Kaplan-Meier coïncide avec l'estimateur empirique usuel de la

$\mathbf{t}_i$	$\mathbf{Y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
1	21	2	0.905	0.064	$0.779 - 1.000$
2	19	2	0.810	0.086	$0.642 - 0.977$
3	17	1	0.762	0.093	$0.580 - 0.944$
4	16	2	0.667	0.103	$0.465 - 0.868$
5	14	2	0.571	0.108	$0.360 - 0.783$
8	12	4	0.381	0.106	$0.173 - 0.589$
11	8	2	0.286	0.099	$0.093 - 0.479$
12	6	2	0.191	0.086	$0.023 - 0.358$
15	4	1	0.143	0.076	$0.000 - 0.293$
17	3	1	0.095	0.064	$0.000 - 0.221$
22	2	1	0.048	0.047	$0.000 - 0.193$
23	1	1	0.000	/	/

TAB. 2.3 – Rèsultats de l'estimation de la fonction de survie des 21 patients traitès au placebo

$\mathbf{t}_i$	$\mathbf{y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
6	21	3	0.857	0.076	$0.707 - 1.000$
7	17	1	0.807	0.087	$0.636 - 0.977$
10	15	1	0.753	0.096	$0.564 - 0.942$
13	12	1	0.690	0.107	$0.481 - 0.900$
16	11	1	0.627	0.114	$0.404 - 0.851$
22	7	1	0.538	0.128	$0.286 - 0.789$
23	6	1	0.448	0.135	$0.184 - 0.712$

TAB. 2.4 – Rèsultats de l'estimation de la fonction de survie des 21 patients traitès à la 6-mercaptopurnie

fonction de survie.

Les résultats numériques de l'estimation des fonctions de survie relatives aux deux groupes de patients sont résumés dans les tableaux (2.3 et 2.4). Sur ce dernier, on contate que la probabilité de survie correspondant à la dernière valeur observée est non nulle $\left(S_n^{KM}(23) = 0.448 \neq 0\right)$. Ceci implique qu'il existe des observations censurées au delà de cette valeur.

Les représentations graphiques des estimateurs des deux fonctions de survie sont présentées dans la figure (2.3). On voit que la courbe de survie relative au groupe de patients traités par la 6-MP est en tout temps, au dessus de celle du groupe témoin. Ceci prouve l'efficacité de la 6-mercaptopurnie dans le traitement de la leucémie.

Test Log-Rank

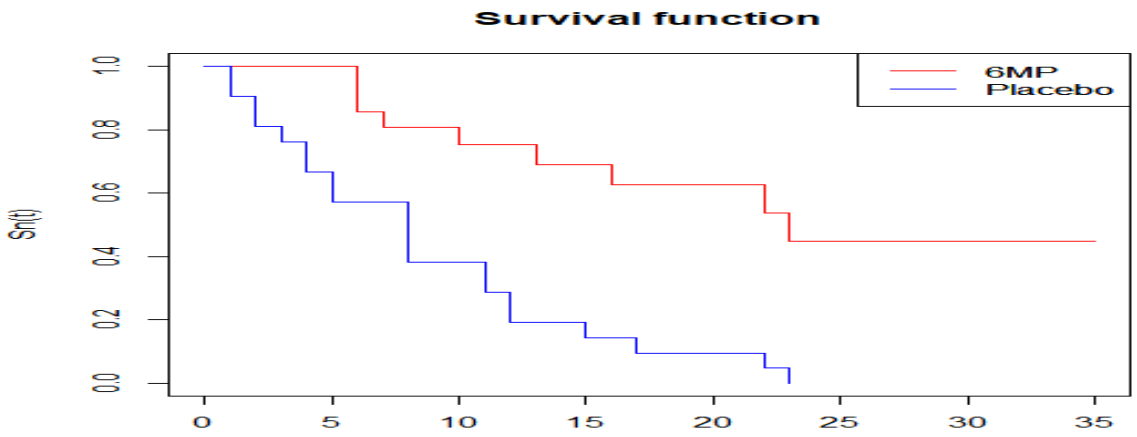

FIG. 2.3 – Estimation de Kaplan-Meier de la fonction de survie (en semaines) pour les 21 données de Freireich

	n	**O**	**E**	$(\mathbf{O}-\mathbf{E})^2/\mathbf{E}$	$(\mathbf{O}-\mathbf{E})^2/\mathbf{V}$
6-MP	21	9	19.3	5.46	16.8
placebo	21	21	10.7	9.77	16.8
ddl= 1, khi-deux= 16.8, p-valeur= 4.17×10^{-5}.					

TAB. 2.5 – résultats de test de Log-Rank

Le tableau 2.5 affiche les résultats obtenus lors du test Log-Rank :

$$\begin{cases} H_0 : & \text{absence d'effet du 6-MP} \\ H_1 : & \text{existence d'effet du 6-MP} \end{cases}$$

La valeur $\chi^2 = 16.8$, du khi-deux observé, correspond à une très faible p-valeur égale à 4.17×10^{-5}. Cette denière étant inférieure à n'importe quel niveau de signification usuel ($\alpha = 0.01, 0.05, 0.10$) conduit au rejet de l'hypothèse nulle H_0. Par conséquent, on conclut que le médicament 6-MP a un effet significatif dans le traitement de la leucémie.

Gliome maligne([8])

Les données de cet exemple, concernent les résultats d'un essai clinique, qui consiste en une nouvelle radio-immunothérapie (RIT), sur 37 personnes atteintes de gliome maligne. Les données relatives à cette étude sont présentées dans les tableaux 2.6 et 2.7. La survie

globale, exprimée en mois, (temps ecoulé du début de thérapie à la mort du patient, causée par la maladie) est comparée pour deux groupes de patients : un groupe de contrôle subissant un traitement standard et un autre traité par la RIT. L'intérêt principal est de déterminer si les personnes traitées à la radio-immunothérapie survivent pendant un temps plus long, par rapport aux patients du groupe de contrôle. En d'autres termes, on veut s'assurer de l'efficacité du nouveau mode de traitment.

Essai clinique : essai clinique dans lequel les sujets sont répartis entre un groupe expérimental et un groupe témoin.

RIT : méthode de radiothérapie qui consiste à irradier des petites tumeurs disséminées dans l'organisme après injection intraveineuse d'un anticorps porteur d'une forte radioactivité et ayant la propriété de se fixer sur les cellules tumorales.

Les colonnes 3 et 5 représentent le type de cancer (grade 3 ou grade 4 connu sous le nom de glioblastome) et l'état de la donnée (observée "Vrai" ou censurée "Faux") respectivement. Les résultats numériques de l'estimation des fonctions de survie relatives aux deux groupes sont résumés dans les Tableaux 2.8, 2.10, 2.9 et 2.11. On constate que, pour le groupe de contrôle, la plus grande des valeurs non censurées a pour estimation $S_n^{KM}(34) = 0.333 \neq 0$, ceci témoigne de l'existence de données censurées supérieures à 34. Dans le cas du glioblastome, les données du groupe de contrôle sont complètes, d'où une estimation nulle pour la survie à l'observation maximale 25 et au delà.

Les fonctions de survie aux deux types de cancer dans chacun des deux groupes sont repréentées graphiquement dans la figure 2.4, les graphiques représentent les estimations des quatre fonctions de survie. On remarque que les courbes de survie des patients traités par radio-immunothérapie (RIT) sont constamment supérieures à celles des deux groupes de contrôle. Cette observation suggère que la RIT a un impact favorable sur la durée de survie des individus souffrant de gliome.

Test Log-Rank

Les résultats du test de log-rank pour les Grade 3 et Grade 4 sont présentés dans les

Age	**Sexe**	**Histologie**	**Groupe**	**évènement**	**temps**
41	Femme	Grade 3	RIT	Vrai	53
45	Femme	Grade 3	RIT	Faux	28
48	Homme	Grade 3	RIT	Faux	69
54	Homme	Grade 3	RIT	Faux	58
40	Femme	Grade 3	RIT	Faux	54
31	Homme	Grade 3	RIT	Vrai	25
53	Homme	Grade 3	RIT	Faux	51
49	Homme	Grade 3	RIT	Faux	61
36	Homme	Grade 3	RIT	Faux	57
52	Homme	Grade 3	RIT	Faux	57
57	Homme	Grade 3	RIT	Faux	50
55	Femme	Grade 4	RIT	Vrai	43
70	Homme	Grade 4	RIT	Vrai	20
39	Femme	Grade4	RIT	Faux	14
40	Femme	Grade 4	RIT	Faux	36
47	Femme	Grade 4	RIT	Vrai	59
58	Homme	Grade 4	RIT	Vrai	31
40	Femme	Grade4	RIT	Vrai	14
36	Homme	Grade 4	RIT	Vrai	36
27	Homme	Grade 3	Contrôle	Vrai	34
32	Homme	Grade 3	Contrôle	Vrai	32
53	Femme	Grade 3	Contrôle	Vrai	9
46	Homme	Grade 3	Contrôle	Vrai	19
33	Femme	Grade 3	Contrôle	Faux	50
19	Femme	Grade 3	Contrôle	Faux	48
32	Femme	Grade 4	Contrôle	Vrai	8
70	Homme	Grade 4	Contrôle	Vrai	8
72	Homme	Grade 4	Contrôle	Vrai	11
46	Homme	Grade 4	Contrôle	Vrai	12

TAB. 2.6 – Observations sur 37 patients atteints de deux types de gliome

Age	**Sexe**	**Histologie**	**Groupe**	**évènement**	**tepms**
44	Homme	Grade 4	Contrôle	Vrai	15
83	Femme	Grade 4	Contrôle	Vrai	5
57	Femme	Grade 4	Contrôle	Vrai	8
71	Femme	Grade 4	Contrôle	Vrai	8
61	Homme	Grade 4	Contrôle	Vrai	6
65	Homme	Grade 4	Contrôle	Vrai	14
50	Homme	Grade 4	Contrôle	Vrai	13
42	Homme	Grade 4	Contrôle	Vrai	25

TAB. 2.7 – Observations sur 37 patients atteints de deux types de gliome (suite)

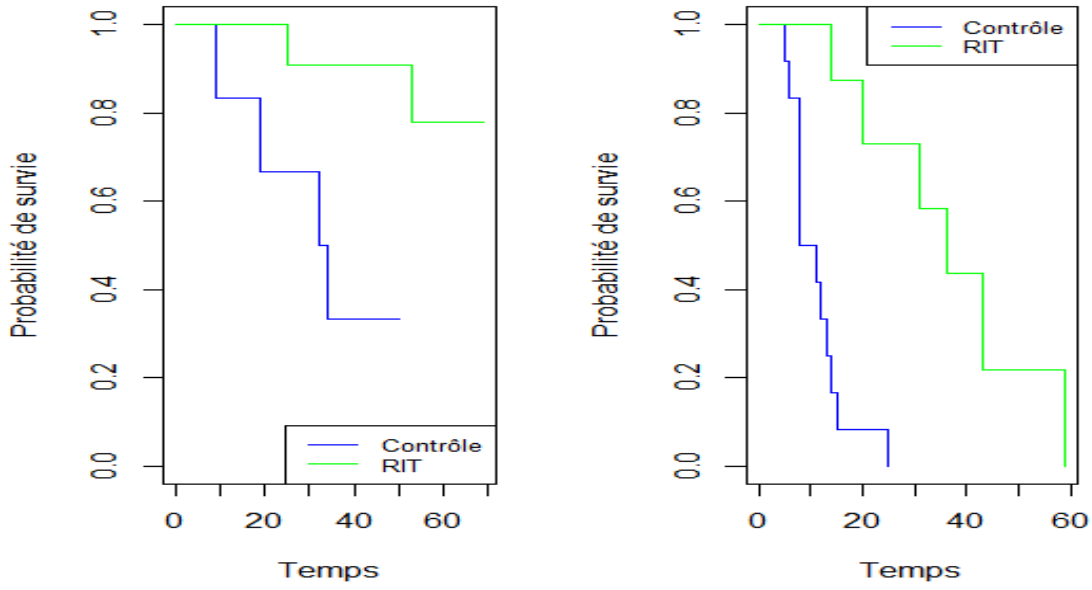

FIG. 2.4 – Fonctions de survies pour les données du gliome de grade 3 (panneau de gauche) et de grade 4 (panneau de droite)

$\mathbf{t}_i$	$\mathbf{y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
9	6	1	0.833	0.152	$0.583 - 1.000$
19	5	1	0.667	0.192	$0.379 - 1.000$
32	4	1	0.500	0.204	$0.225 - 1.000$
34	3	1	0.333	0.192	$0.108 - 1.000$

TAB. 2.8 – Rèsultats de l'estimation de la fonction de survie de 6 patients atteints de gliome de grade 3 subissant un traitement standard

$\mathbf{t}_i$	$\mathbf{y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
25	11	1	0.909	0.087	$0.754 - 1.000$
53	7	1	0.779	0.141	$0.546 - 1.000$

TAB. 2.9 – Résultats de l'estimation de la fonction de survie de 11 patients atteints de gliome de grade 3 traitès à la radio-immunothérapie

$\mathbf{t}_i$	$\mathbf{y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
5	12	1	0.917	0.080	$0.773 - 1.000$
6	11	1	0.833	0.108	$0.647 - 1.000$
8	10	4	0.500	0.144	$0.284 - 0.880$
11	6	1	0.417	0.142	$0.213 - 0.814$
12	5	1	0.333	0.136	$0.150 - 0.742$
13	4	1	0.250	0.125	$0.094 - 0.666$
14	3	1	0.167	0.108	$0.047 - 0.591$
15	2	1	0.083	0.0789	$0.013 - 0.544$
25	1	1	0.000	/	/

TAB. 2.10 – Résultats de l'estimation de la fonction de survie de 12 patients atteints de gliome de grade 4 subissant un traitement standard

$\mathbf{t}_i$	$\mathbf{y}_i$	$\mathbf{d}_i$	$\mathbf{S}_n^{KM}(\mathbf{t})$	$\hat{\sigma}$	$\mathbf{95\% - IC}$
14	8	1	0.875	0.117	$0.673 - 1.000$
20	6	1	0.729	0.165	$0.468 - 1.000$
31	5	1	0.583	0.186	$0.313 - 1.000$
36	4	1	0.438	0.188	$0.189 - 0.917$
43	2	1	0.219	0.181	$0.043 - 1$
59	1	1	0.000	/	/

TAB. 2.11 – Résultats de l'estimation de la fonction de survie de 8 patients atteints de gliome de grade 4 traités à la radio-immunthérapie.

	n	**O**	**E**	$(\mathbf{O}-\mathbf{E})^2/\mathbf{E}$	$(\mathbf{O}-\mathbf{E})^2/\mathbf{V}$
Contrôle	6	4	1.49	4.23	6.06
RIT	11	2	4.51	1.4	6.06

TAB. 2.12 – résultats de test de Log-Rank(Grade 3)

	n	**O**	**E**	$(\mathbf{O}-\mathbf{E})^2/\mathbf{E}$	$(\mathbf{O}-\mathbf{E})^2/\mathbf{V}$
Contrôle	12	12	5.65	7.13	14.4
RIT	8	6	12.35	3.26	14.4

TAB. 2.13 – résultats de test de Log-Rank(Grade 4)

tableaux 2.12 et 2.13 respectivement.

Pour le Grade 3, la p-valeur est de 0.01, indiquant une signification statistique à un seuil de 0.05 (ou 0.10). Pour le Grade 4, la p-valeur, égale à 2×10^{-4}, est très faible. Ceci prouve qu'il existe une différence statistiquement significative, aux seuils usuels $\alpha = 0.01, 0.05$ et 0.10, entre les fonctions de survie des groupes «Contrôle» et «RIT» pour les cas d'histologie 4. Au vu de ce qui précède, on conclut que la RIT a un effet significatif dans le traitement de la gliome (Grade 3 et Grade 4).

Conclusion

Introduction à l'analyse de survie explore un domaine important de la statistique, fournissant des outils puissants pour étudier le temps écoulé jusqu'à ce qu'un événement d'intérêt se produise. Dans ce travaill, on s'est intéressé à l'estimation non paramétrique de la fonction de survie sous des données incomplètes (censurées et tronquées) ainsi qu'au test Log-Rank. Ce dernier est un test de comparaison des fonctions de survie entre deux ou plusieurs groupes en fonction du nombre de catégories de la variable qualitative.

Dans le cas de censure, la méthode non paramétrique de Kaplan-Meier permet d'estimer la valeur de la fonction de survie, en tenant compte du fait que les données soient observées ou censurées et pour la troncature, c'est l'estimateur non prarmétrique de Lynden-Bell qui est le plus populaire.

Enfin, on note qu'il existe des méthodes d'estimation semi paramétriques particulèrement efficaces dans le cas où on possède une information partielle sur la distribution de probabilité de la variable d'intérêt. Dans ce contexte, on réfère le lecteur aux travaux du Professeur Abdelhakim Necir et son équipe.

Bibliographie

[1] Huber-Carol, C. (1994) . *Durées de survie tronquées et censurées.* Journal de la Société Statistique de Paris, 135 (4) , 3 − 23.

[2] Ihaka, R. et Gentleman, R. (1996). R : A language for data analysis and graphics. *Journal of computational and graphical statistics* **5**,299 − 314

[3] Jonas, S.F. (2018) . *Méthodes de comparaisons de deux ou plusieurs groupes de données censurées par intervalle avec application en immunologie clinique* (Doctora dissertation, Université Paris Saclay (COmUE)).

[4] Kaplan, E.L. et Meier, P. (1958). *Nonparametric estimation from incompete observations.* Journal of the American association,.53 (282), 457 − 481.

[5] Kankoé,S. (2016) *Méthode actuarielle d'estimation des courbes de survie, principe,différences avec la méthode de kaplan-meier.*

[6] Klein, J.P. et Moeschberger, M.L. (2003). *Survival analysis : techniques for censored and truncated data* (Vol. 1230). New York : Springe.

[7] Lynden-Bell, D.(1971). *A method of allowing for known observational selection in small samples applied to* $3C$ R *quasars.* Monthly Notices of the Royal Astronomical Society, 155(1), 95 − 118.

[8] M'ziou, I. (2014) . *Estimation non paramétrique avec données censurées* (mémoire de Master). L'université Mohamed Khider, Biskra.

[9] Ndao, P (2015). *Modélisation de valeurs extrêmes conditionnelles en présence de censure* (Doctoral dissertation, PhD thesis, Université Gaston Berger de Saint-Louis).

[10] Reiss, R. D, Thomas, M, et Reiss, R. D. (1997). *Statistical analysis of extreme values* (Vol. 2). Basel : Birkhauser..

[11] Saporta, G. (2006). Probabilités, analyse des données et statistique. Technip, Paris.

[12] Sissaoui, A, et Abdlaziz, S. (2019 – 2020). *Estimation pour les données censurées* (mémoire de Master). L'université Mohamed Seddik Ben Yahia, Jijel.

[13] Woodroofe, M. (1985). *Estimating a distribution function with truncated data.* Annals of Statistics, 13(1), 163 – 177.

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Notation	**Signification**
c-à-d	C'est-à-dire.
ddl	Degré de liberté.
f.d.r	Fonction de répartition.
i.i.d	Indépendantes et identiquement distribuées.
v.a	Variable aléatoire.
$\mathbb{1}\{A\}$	Indicatrice de l'ensemble A.
F	Fonctionde répartition.
S	Fonction de survie.
λ	Fonction du risque.
Λ	Fonction du risque cumulé.
(Ω, A, P)	Espace probabilisé.
F_n	Fonction de répartition empirique.
$\mathbf{S}_n^{KM}$	Estimateur de Kaplan-Meier.
$E(X)$	Espérance de X.
$\xrightarrow{p.s}$	Convergence presque sûre.
$\xrightarrow{D}$	Convergence en distribution.
$:=$	Egalité par définition

الملخص

تحليل البقاء هو أسلوب إحصائي يستخدم لدراسة الوقت حتى وقوع حدث معين. يعد التقدير اللامعلمي لدالة البقاء كابلن-ماير وليندن بيل من الأدوات الأساسية للتحليل الإحصائي في ظل رقابة واقتطاع البيانات على التوالي.

الكلمات المفتاحية:البيانات الخاضعة للرقابة، البيانات المقتطعة، كابلن ماير، ليندن-بيل، دالة البقاء.

Abstract

Survival analysis is a statistical method used to study the time until the occurrence of a specific event. Kaplan-Meier and Lynden-Bell non parametric estimators of the survival function are fundamental tools in the statistic alanalysis of data under censoring and truncation respectively.

Keywords: Censored Data, Kaplan-Meier, Lynden-Bell, Survival function , Truncated Data.

Résumé

L'analyse de survie est une method statistique utilisée pour étudier le temps jusqu'à l'occurrence d'un événement spécifique. Les estimateurs non paramétriques de la fonction de surviede Kaplan-Meier et de Lynden-bell sont des outils fondamentaux dans l'analyse statistique des données sous censure et troncature respectivement.

Mots clés : Données censurées, Données tronquées, Fonction de survie, Kaplan-Meier, Lynden-Bell.