
République Algérienne 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 en vue de l'obtention du
DIPLÔME DE MASTER EN MATHÉMATIQUES

Option : **Statistique**

Par
Semmari Manal

Thème

Sur l'Analyse de Survie et Applications

Membres du Comité d'Examen

Pr. Djamel Meraghni	UMKB	Président
Dr. Louiza Soltane	UMKB	Encadreur
Dr. Jihane Abdelli	UMKB	Examineur

Septembre 2020

*À mes chers Parents,
À mes soeurs et mes frères,
À ma famille et mes amis,
À tous mes enseignants.*

REMERCIEMENTS

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. Louiza Soltane, qui s'est toujours montré l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi que pour l'inspiration, l'aide et le temps qu'il ont bien voulu me consacrer.

Je remercie également tous les enseignants et employés du département de Mathématiques présidé par Monsieur Mokhtar Hafayedh, avec une mention spéciale à Messieurs les membres du jury "le Pr. Djamel Meraghni et le Dr. Jihane Abdelli" qui nous ont faites l'honneur de participer à notre soutenance.

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience.

Enfin, j'adresse mes plus sincères remerciements ma famille, mes proches, amis et tous ceux qui, de près ou de loin, m'ont soutenu et encouragé au cours de la réalisation de ce mémoire.

Merci toutes et à tous.

S. MANAL

TABLE DES MATIÈRES

Dédicaces	2
Remerciements	3
Table des Matières	4
Liste des Figures	6
Liste des Tableaux	7
Introduction	1
1 Modélisation de l'analyse de survie	3
1.1 Définition de l'analyse de survie	3
1.2 Donnée de survie	4
1.2.1 Date d'origine	5
1.2.2 Date de point	5
1.2.3 Date des dernières nouvelles	5
1.2.4 Recul	5
1.2.5 Temps de participation	5
1.3 Censure	5
1.3.1 Types de censure	6
1.4 Distributions de la durée de survie	9
1.4.1 Fonction de répartition	9

1.4.2	Fonction de survie	10
1.4.3	Fonction de Densité	10
1.4.4	Fonctions du Risque et du Risque Cumulé	11
1.4.5	Fonctions Empiriques de Répartition et de Survie	12
1.5	Méthodes non paramétriques	13
1.5.1	Méthode de Kaplan-Meier pour la fonction de survie	13
1.5.2	Méthode de Nelson-Aalen pour la fonction du risque cumulé	18
1.6	Comparaison de deux groupes de données de survie	20
1.6.1	Test du log-rank	20
2	Exemple d'application	23
2.1	Présentation des données	23
2.2	Statistiques descriptives	25
2.3	Statistique inférentielle	27
	Conclusion	34
	Bibliographie	35
	Annexe B : Abréviations et Notations	37

TABLE DES FIGURES

1.1	Schéma représentant les principales définitions relatives à l'analyse de la durée de survie.(source [7])	6
1.2	Schema Representant les Cas de Censure Aléatoire (Source [8]).	9
1.3	Fonction Empirique de Répartition (Droite) et de Survie (Gauche) d'un Echantillon Gaussien Standard de Taille 200.	13
2.1	Distribution du poids des moutons mâles sauvages atteints par parasites intestinaux.	25
2.2	Distribution du groupe en fonction du poids.	26
2.3	Courbe de survie estimée pour les 150 données de moutons.	27
2.4	Courbes de survie estimées au sein de chaque groupe.	28
2.5	Courbes du risque cumulé estimé au sein de chaque groupe.	29

LISTE DES TABLEAUX

2.1	sheep data (package survival). Observations sur la mortalité chez 150 moutons mâles sauvages atteints par parasites intestinaux. Seules les données des 20 premiers moutons sont présentées ici.	24
2.2	Résumé statistiques descriptives élémentaires des moutons mâles sauvages atteints par parasites intestinaux.	25
2.3	Effectifs et proportions des 150 moutons mâles sauvages atteints par parasites intestinaux répartis selon le poids et le groupe	26
2.9	Tests d'égalité des 3 survies : résultats de la statistique du log-rank	29
2.4	Estimateur de Kaplan-Meier de la fonction de survie (en mois) pour les 150 moutons mâles sauvages atteints par parasites intestinaux	30
2.5	Résultats de l'estimation de la fonction de survie du groupe A traité au bolus avec une forte dose de tueur de vers.	31
2.6	Résultats de l'estimation de la fonction de survie du groupe B traité au bolus avec une faible dose de tueur de vers.	32
2.7	Résultats de l'estimation de la fonction de survie du groupe C traité au bolus sans contenu destructeur de vers.	33
2.8	La médiane et l'intervalle de confiance estimées par KM pour chaque groupe.	33

Introduction

La 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 éclairées. Cette recherche a été basée sur le concept de variables aléatoires et de distributions de probabilités, où les variables aléatoires sont adaptées aux applications probabilistes appropriées en observant le comportement de ces variables.

Le développement des statistiques a conduit au développement du concept de distributions pour inclure un autre facteur est le temps. Il est appliqué dans de nombreux domaines différents tels que la médecine, l'économie, la population, etc ...

L'analyse de survie est basée sur l'estimation de la fonction de survie et la fonction de risque et grâce à la fonction de survie on peut être estimée la médiane du temps de survie pour les patients actuels ou futurs, et les résultats estimés sont particulièrement utiles pour concevoir un système de traitement ou conseiller le patient lors du diagnostic. L'application d'une telle étude aide à identifier les conditions et les caractéristiques qui augmentent ou diminuent la probabilité de survie, l'absence de modèles de risque entraîne la perte de toutes les caractéristiques qui mènent à la création d'un bon système thérapeutique.

Dans ce travail, on s'intéresse à la comparaison de courbes de survie estimées par Kaplan-Meier, la comparaison de courbes de risque cumulée et estimée la médiane du temps de survie.

Le mémoire est composé de deux chapitres :

- Le premier chapitre contient le cadre théorique du mémoire et comprend l'analyse de la survie, la définition des données censurées, les raisons de la censure, la fonction de

survie, la fonction de risque et la comparaison de deux groupes de données de survie. Ces définitions et résultats proviennent, entre autres, des références [4], [8], [11], [14] et [19].

- Le second chapitre contient les concepts théoriques du premier chapitre (vus au [Chapter 1](#)), où il est consacré un exemple d'application sur données réelles : les moutons mâles sauvages atteints par parasites intestinaux, dans lequel les données de ce mémoire ont été décrites et trouvé la fonction de survie, la fonction de risque et fait un test pour comparer les groupes des données de survie (la statistique du Log-Rank), ainsi que les conclusions tirées. Pour une présentation détaillée de ces techniques : d'analyse de survie à l'aide du logiciel d'analyse statistique R ([The R Project for Statistical Computing](#)), on suggère les références [3], [5], [11] et [13].

Enfin, il y a lieu de noter que les calculs numériques et les représentations graphiques sont réalisés à l'aide des packages "survival", "survminer", "dplyr" et "ggplot2" du R.

CHAPITRE 1

Modélisation de l'analyse de survie

Dans ce chapitre il est adressé toutes les méthodes statistiques qui va utiliser dans la partie pratique de cette mémoire, et la partie principale de ce chapitre est la section sur la recherche de la fonction de survie, la fonction de risque et la comparaison de deux ou plusieurs ensembles de données de survie. Pour plus de détail sur ce thème on peut refère aux [6], [11] et [5].

1.1 Définition de l'analyse de survie

Définition 1.1.1 (Analyse de survie)

L'analyse de survie est le terme utilisé pour décrire l'analyse des données qu'ils sont sous forme de temps (times) allant de l'origine du temps (time origin) à la survenance d'un événement ou d'un point final spécifique (end point).

Dans la recherche médicale, l'origine du temps est souvent la date d'enregistrement de l'élément (l'individu) dans une étude, comme les essais médicaux pour comparer deux types de médicaments ou plus (tester l'efficacité d'un médicament) si le point final est la mort du patient, les données résultantes sont les temps de survie (*Survival Time*), mais si le point final n'est pas la mort, alors les données résultantes sont appelées données de temps sur événement (*Time to Event Data*).

Remarque 1.1.1

L'analyse de survie est un ensemble de procédures statistiques pour analyser les données lorsque la variable adoptée est la durée avant qu'un évènement précis se produise, et cette durée peut être des jours, des semaines, des mois, des années, depuis le début de l'individu et que l'évènement, il peut être la mort, le début de la maladie, la rechute du patient.

On remarque qu'il existe des nombreux autres noms comme :

- Dans les statistiques médicales et épidémiologiques, il est connu sous le nom d'analyse de survie ou analyse de risque (Hazard Analysis).
- Dans les études d'ingénierie, elle est connue sous le nom d'analyse du temps d'échec (Failure Time Analysis).
- Dans les études de psychologie, il est connu comme analyse d'historique d'évènement (Event History Analysis) (voir par exemple le livre de [Göran Broström \[9\]](#)).
- En économie connue sous le nom d'analyse de transition (Transition Analysis).

Ici quelques exemples :

Exemple 1.1.1 (En biostatistique)

- *Survie d'un individu après l'apparition d'une tumeur.*
- *Âge de décès chez des patients atteints de diabète...*
- *temps de rémission après opération chirurgicale.*
- *Durée de vie d'une ampoule, d'un pièce mécanique,...*

Exemple 1.1.2 (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.*

1.2 Donnée de survie

Donnée de survie (Survival Data) est une expression utilisée pour décrire les données qui mesurent le temps jusqu'à un évènement d'intérêt et que la variable résultante est le temps jusqu'à un évènement d'intérêt, qui est connu par le temps de survie, qui est toujours une variable réelle des valeurs positives. Il y a trois conditions de base pour la durée de survie qui doivent être définies précisément qui sont la data d'origine, la date de point et la date des dernières nouvelles (l'évènement), elles sont présentées dans la [Figure 1.1](#), qu'on abordera en détail comme suit :

1.2.1 Date d'origine

L'origine de la durée étudiée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente, ce qui n'est pas important car c'est la durée qui nous intéresse.

1.2.2 Date de point

C'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

1.2.3 Date des dernières nouvelles

C'est la date à laquelle on a eu pour la dernière fois des nouvelles du sujet : cela peut être la date du décès ou la date de survenue de l'évènement étudié (guérison, première rechute, première apparition d'un évènement indésirable, ...), mais aussi la date de la dernière consultation si le sujet est perdu de vue ou n'a pas présenté l'évènement étudié (décès, guérison, ...) (source : [M'ziou Imane \[14\]](#), mémoire de master en Statistiques).

1.2.4 Recul

C'est le délai écoulé entre la date d'origine et la date de point.

1.2.5 Temps de participation

C'est le temps écoulé entre :

- la date d'origine et la date de dernières nouvelles, si cette dernière est antérieure à la date du point (décédé d'avant la date de point).
- la date d'origine et la date de point, si celle-ci est antérieure à la date de dernières nouvelles (vivant aux dernières nouvelles). (source : [M'ziou Imane \[14\]](#)).

1.3 Censure

Une des caractéristique des données de survie est l'existence d'observations incomplètes, en effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de

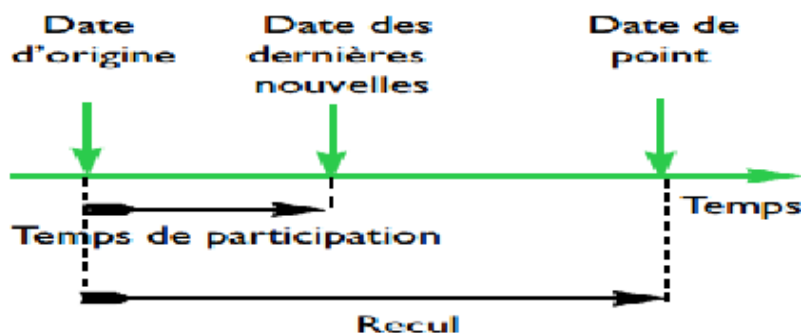


FIG. 1.1 – Schéma représentant les principales définitions relatives à l'analyse de la durée de survie.(source [7])

censure et troncature, les données censurées proviennent du fait qu'on n'a pas accès à toute l'information, au lieu d'observer des réalisations indépendantes et identiquement distribuées (iid) de durée X , on observe la réalisation diverses perturbations indépendantes ou non de l'événement étudié. (Source : Soltane [20], thèse de doctorat de Probabilité et Statistique).

Définition 1.3.1 (Variable censure)

la variable de censure Y est définie par la non-observation de l'événement étudié si au lieu d'observer X , on observe Y , et que l'on sait que $X > Y$ (respectivement $X < Y, Y_1 < X < Y_2$), on dit qu'il ya censure à droite (respectivement censure à gauche, censure par intervalle)

pour un individu donné j , on va considérer

- son temps de survie X_j , de fonction de répartition F .
- son temps de censure Y_j , de fonction de répartition G .
- la durée rèlement observée Z_j , de fonction de répartition H .

Pour une présentation détaillée des différents types de censure voir comme un exemple le livres de Elisa [6] et Klein et Moeschberger [11].

1.3.1 Types de censure

Censure à droite :

La variable d'intérêt est dite censure à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées.

Censure à gauche :

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

Censure par intervalle :

Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. On retrouve ce modèle en général dans des études suivi médical ou les patients sont contrôlés et se présente ensuite après que l'événement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite au à gauche par des intervalles du type $[y, +\infty[$ et $[0, y]$ respectivement. (Source : [Soltane \[20\]](#)).

Censure double :

La censure double (ou mixte) ce type de censure c'est un mélange entre les deux censure, la censure à droite et la censure à gauche, dans le même échantillon.

Ces quatre catégories de censure décrites ci-dessus peuvent se présenter en fonction du mode ou mécanisme de censure suivantes :

Censure de type 1 : fixée L'expérimentateur fixe une valeur (une date par exemple non aléatoire de fin d'expérience). Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles.

Soit Y une valeur fixée. Par exemple en censure à droite, au lieu d'observer les variables X_1, \dots, X_n qui nous intéressent, on observe X_j que lorsqu'elle est inférieure ou égale à une durée fixée Y . On observe donc une variable Z_j telle que $Z_j := \min(X_j, Y)$, $j = 1, \dots, n$.

Censure de type 2 : attente L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie.

Par exemple en épidémiologie 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. Soient $X_{j:n}$ et $Z_{j:n}$ les statistiques d'ordre des variables X_j et Z_j . La date de censure est donc $X_{r:n}$

et on observe

$$\begin{cases} Z_{j:n} = X_{j:n} & \text{si } j \leq r, \\ Z_{j:n} = X_{j:n} & \text{si } j \geq r. \end{cases}$$

Censure de type 3 : aléatoire C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient, voir l'[Exemple 1.3.1](#)).

Soit X_1, \dots, X_n un échantillon d'une va positive X , on dit qu'il y a censure aléatoire de cet échantillon s'il existe une autre va positive elle aussi Y d'échantillon Y_1, \dots, Y_n dans ce cas au lieu d'observer les X_i 's, on observe un couple de va's (Z_i, δ_i) avec

$$Z_i := \min(X_i, Y_i) \quad \text{et} \quad \delta_i := \mathbb{1}\{X_i \leq Y_i\} \quad \text{pour } i = 1, \dots, n, \quad (1.1)$$

où δ_i l'indicateur de censure, qui détermine si X a été censuré ou non :

- si $\delta_i = 1$, la durée d'intérêt est observée ($Z_i = X_i$).
- si $\delta_i = 0$, elle est censurée ($Z_i = Y_i$). On observe des durées incomplètes.

Exemple 1.3.1

On considère une étude relative à la durée de survie de patients soumis à un traitement particulier. L'évènement d'intérêt est la mort de la patient. Tous les individus sont suivis pendant les 10 semaines suivant la première administration du traitement. On considère plus particulièrement 3 sujets qui vont permettre d'illustrer certaines des caractéristiques les plus fréquentes des données de survie et notamment deux cas possibles de censure à droite. On peut citer certaines causes entraînant la censure aléatoire :

1. "Perdu de vue" : le patient peut décider de se faire soigner ailleurs et on ne le revoit plus.
2. Arrêt du traitement : suite à des effets secondaires le traitement est arrêté.
3. Fin de l'étude : l'étude se termine et certains patients sont toujours vivants.

Dans la [Figure 1.2](#) on représente le suivi de trois patients. Le premier patient est décédé 4 semaines après le début du traitement. Il s'agit d'une observation non censurée. Le deuxième patient est vivant au terme des 14 semaines d'observation. L'information de ce patient n'est pas connue lorsque la constitution de la base de données est arrêtée (en $t = 10$). il est donc censuré. Quand au troisième patient, il a été perdu de vue à $t = 7$, donc il est censuré à $t = 7$.

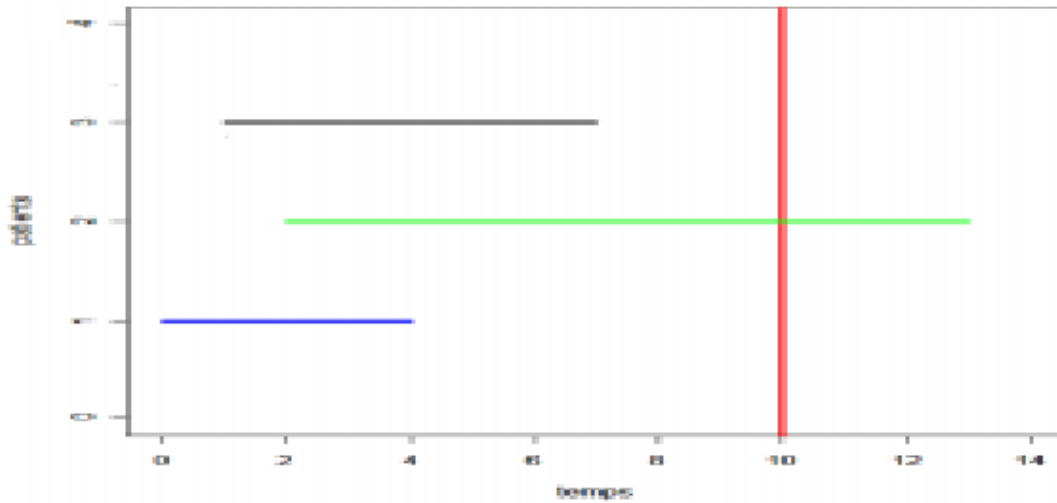


FIG. 1.2 – Schema Representant les Cas de Censure Aléatoire (Source [8]).

Pour des détails complets sur les types de censure avec des exemples, on réfère aux livres de ([6], [11] et on peut voir aussi Pierre [16], page 7) et (Vivian [22], page 14). Dans ce mémoire, on s'intéresse uniquement au cas des censures à droite du type aléatoire. Celui-ci correspond à un modèle fréquemment utilisé dans la pratique, ce qui justifie amplement qu'on y attache de l'intérêt.

1.4 Distributions de la durée de survie

Soit X le temps de survie La distribution de X est entièrement caractérisée par l'une des cinq fonctions suivantes :

1.4.1 Fonction de répartition

Définition 1.4.1 (Fonction de répartition)

La fonction de répartition (fdr ou fd) de X ou de sa loi P_X est la fonction sur \mathbb{R}_+ définie par

$$\begin{aligned} F : \mathbb{R}_+ &\longrightarrow [0, 1] \\ t &\longrightarrow \mathbb{P}(X \leq t), \end{aligned}$$

Pour t fixé, c'est la probabilité de mourir avant l'instant t .

Propriété 1.4.1

La fonction F est une fonction croissante monotone continue à droite telle que

$$\lim_{t \rightarrow 0} F(t) = 0 \quad \text{et} \quad \lim_{t \rightarrow \infty} F(t) = 1.$$

1.4.2 Fonction de survie**Définition 1.4.2 (Fonction de survie)**

La fonction de survie qu'on note par $S(t)$ ou $\bar{F}(t)$ est définie sur \mathbb{R}_+ par

$$S(t) = \bar{F}(t) := \mathbb{P}(X > t). \quad (1.2)$$

Pour t fixé c'est la probabilités de survivre jusqu'à l'instant t .

Propriété 1.4.2

La fonction de survie d'une va X est décroissante monotone continue à gauche et vérifie

$$\lim_{t \rightarrow 0} S(t) = 1 \quad \text{et} \quad \lim_{t \rightarrow \infty} S(t) = 0.$$

Remarque 1.4.1

1. F est aussi appelé fonction de distribution et $S(t)$ est la queue de distribution.
2. On peut définir $F(t) := 1 - S(t)$, pour tout $t \geq 0$,

1.4.3 Fonction de Densité

Comme toute autre va continue, la durée de survie X a une fonction de densité de probabilité.

Définition 1.4.3 (Fonction de densité)

Si F admet une dérivée par rapport à la mesure de Lebesgue sur \mathbb{R}_+ , la fonction de densité de probabilité existe, elle est définie pour tout $t \geq 0$, par

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} = \lim_{dx \rightarrow \infty} \frac{\mathbb{P}(t \leq X \leq t + dt)}{dt}. \quad (1.3)$$

Pour t fixée, la densité de probabilité caractérise la probabilité de mourir dans un petit intervalle de temps après l'instant t .

1.4.4 Fonctions du Risque et du Risque Cumulé

Appelée selon les domaines d'application : "*taux instantané de défaillance*", "*taux de risque*", "*le taux de hasad*" ou encor, "*quotient de mortalité*". Pour plus de détails, on renvoie au livre de Saporta [19], page 19.

Définition 1.4.4 (*Fonction du risque*)

Si X est une variable continue positive représentant une durée. La fonction du risque, notée par $h(t)$, est définie par

$$h(t) = \lim_{dx \rightarrow 0} \frac{\mathbb{P}(t \leq X < t + dx / X \geq t)}{dx} = \frac{f(t)}{S(t)}. \quad (1.4)$$

Cette fonction donne habituellement plus d'informations sur le mécanisme de décès que la fonction de survie. C'est pourquoi cette fonction est souvent employée pour résumer les données de survie.

Définition 1.4.5 (*Fonction du risque cumulé*)

la fonction du risque cumulé, qu'on note par $H(t)$, c'est l'intégral de fonction du risque

$$H(t) := \int_0^t h(u) du = \int_0^t \frac{dF(u)}{S(u)}. \quad (1.5)$$

$h(t)$ caractérise la loi de X car on peut retrouver $F(t)$ à partir de $h(t)$

$$h(t) = -\frac{d}{dt} \ln S(t). \quad (1.6)$$

On note que, en vertu de (1.5), on peut réécrire

$$H(t) := -\log \{S(t)\} \quad (1.7)$$

Il est facile de deduire de (1.6) que

$$F(t) = 1 - \exp \left\{ -\int_0^t \frac{dF(u)}{S(u)} \right\} = 1 - \exp(-H(t)). \quad (1.8)$$

Cette égalité présente une caractérisation de distribution et une fonction de survie par l'intermédiaire de fonction du risque. Toutes ces fonctions (1.2), (1.3), (1.4) et (1.5) sont donc 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. Pour plus d'illustrer sur les relations d'équivalence entre ces fonctions précédentes on réfère à, par exemple, (Lee et Wang [12], Exemple 2.2, page 17), (Wienke [23], Exemple 2.1, page 17).et (Gilbert [8], page 21).

1.4.5 Fonctions Empiriques de Répartition et de Survie

Définition 1.4.6 (*Fonctions empiriques de répartition et de survie*)

Soit X_1, \dots, X_n un échantillon de taille $n \geq 1$ d'une va positive X de fdr F et de fonction de survie S . Les fonctions empiriques de répartition et de survie, F_n et S_n sont respectivement définies par

$$F_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq t\}, \quad \forall t \geq 0, \quad (1.9)$$

et

$$S_n(t) = 1 - F_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i > t\}, \quad \forall t \geq 0, \quad (1.10)$$

où $\mathbb{I}\{A\}$ est la fonction indicatrice de l'ensemble A .

On peut écrire (1.9) et (1.10) en termes des valeurs des statistiques d'ordre¹ comme suite F

$$F_n(t) = \begin{cases} 0 & \text{si } t < X_{1:n}, \\ \frac{i}{n} & \text{si } X_{i:n} \leq t < X_{i+1:n}, \\ 1 & \text{si } t \geq X_{n:n}, \end{cases} \quad \text{et } S_n(t) = \begin{cases} 1 & \text{si } t < X_{1:n}, \\ 1 - \frac{i}{n} & \text{si } X_{i:n} \leq t < X_{i+1:n}, \\ 0 & \text{si } t \geq X_{n:n}. \end{cases} \quad (1.11)$$

Pour une représentation graphique de ces deux fonctions, voir la [Figure 1.3](#).

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

$$\sup_{t \geq 0} |S_n(t) - S(t)| \xrightarrow{p.s.} 0, \quad \text{quand } n \longrightarrow \infty. \quad (1.12)$$

De plus, on a la propriété de normalité asymptotique, on a

$$\sqrt{n}(S_n(t) - S(t)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, S(t)(1 - S(t))/n), \quad t \geq 0, \quad \text{quand } n \longrightarrow \infty. \quad (1.13)$$

¹Les statistiques d'ordre associées à l'échantillon X_1, \dots, X_n , sont obtenues en classant ces va's par ordre croissant $X_{1:n} \leq \dots \leq X_{n:n}$.

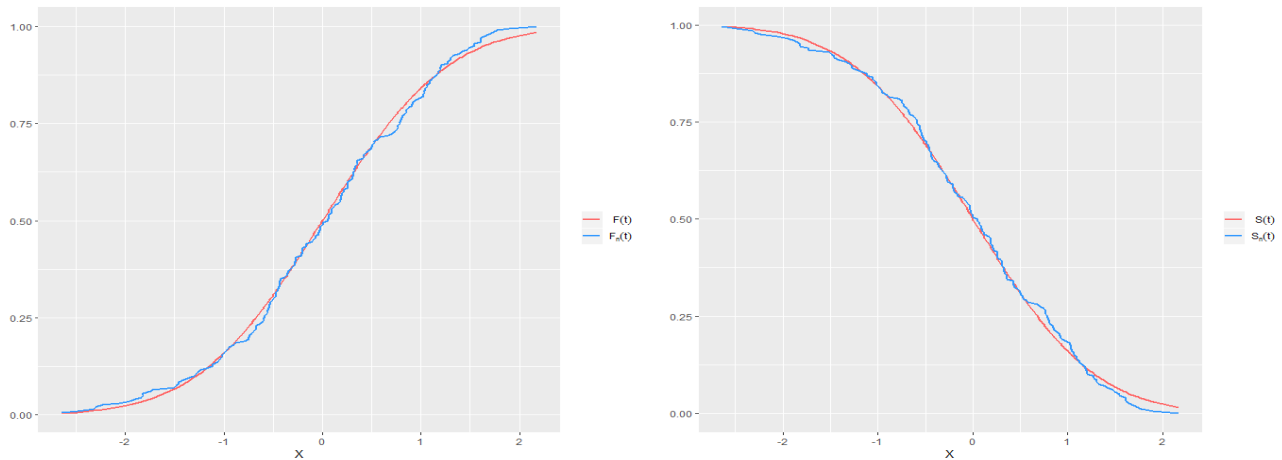


FIG. 1.3 – Fonction Empirique de Répartition (Droite) et de Survie (Gauche) d'un Echantillon Gaussien Standard de Taille 200.

1.5 Méthodes non paramétriques

Dans l'analyse de survie, il est toujours utile de trouver des résumés numériques ou graphiques du temps de survie des individus, généralement les estimations de la fonction de survie (1.2) et de la fonction de risque (1.6) sont souvent appropriées qui sont des méthodes non-paramétrées, parce qu'il ne nécessite pas aucune hypothèse sur la distribution du temps de survie et pour comparer les temps de survie de deux groupes ou plus, on utilise le test du (Log-rank).

1.5.1 Méthode de Kaplan-Meier pour la fonction de survie

Lorsque les données sont censurées il est impossible d'utiliser (1.10) puis qu'il fait intervenir des quantités non observées. Dans ce cas, Kaplan et Meier [10] en 1958 ont proposé un estimateur, appelé estimateur de Kaplan-Meier ou estimateur du produit limite (P-L) car il s'obtient comme la limite d'un produit. L'idée de la construction de cet estimateur est donnée par la probabilité conditionnelle. Soient $0 < t'' < t' < t$, on a :

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

On considère les temps d'événements (décès et censure) distincts $Z_{i:n}, (i = 1, \dots, n)$ ordonnés par ordre croissant, on obtient

$$P(X > Z_{i:n}) = \prod_{k=1}^i P(X > Z_{k:n} \setminus X > Z_{k-1:n}), \quad i = 1, \dots, n,$$

avec $Z_{0:n} = 0$. Si on note par n_i le nombre d'individus à risque de subir l'événement (mourir) juste avant le temps $Z_{i:n}$, et par d_k 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 que l'on était vivant en $Z_{i-1:n}$ (c'est à dire $p_i = P(X > Z_{i:n} \setminus X > Z_{i-1:n})$, peut être estimée par

$$\hat{p}_i := \frac{d_i}{n_i}. \quad (1.14)$$

Comme les temps d'événements sont supposés distincts, on a $d_i = 0$ en cas de censure en $Z_{i:n}$, et $d_k = 1$ en cas de décès en $Z_{i:n}$, Si on désigne par $\delta_{[i:n]}$ le concomitant de la $i^{\text{ème}}$ statistique d'ordre $Z_{i:n}$ défini par $\delta_{[i,n]} = \delta_j$ si $Z_{i:n} = Z_j$ alors on a

- $\delta_{[i,n]} = 0$ en cas de censure en $Z_{i:n}$.
- $\delta_{[i,n]} = 1$ en cas de décès en $Z_{i:n}$.

Définition 1.5.1

L'estimateur de Kaplan-Meier, pour $t < Z_{n:n}$, $i = 1 \dots n$, est défini par

$$S_n^{KM}(t) = \bar{F}_n^{KM}(t) := \prod_{Z_{i:n} \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{Z_{i:n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n_i}\right) = \prod_{Z_{i:n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n - i + 1}\right), \quad (1.15)$$

où $Z_{1:n} \leq \dots \leq Z_{n:n}$ les statistiques d'ordre associées à Z_1, \dots, Z_n .

Remarque 1.5.1

Il existe d'autres formes pour l'estimateur de Kaplan-Meier :

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

Une écriture sous la forme d'une somme peut être trouvée dans le livre de Reiss et Thomas [18], page 122

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

Remarque 1.5.2

L'estimateur de Kaplan-Meier est très utile dans l'estimation, sous censure, de quantités statistiques définies en termes de la fonction de répartition ou de survie, telles la moyenne, la variance, les quantiles, les mesure de risque, (Source : [14]).

Remarque 1.5.3

Dans le cas où il y a des ex-aequo, si ces ex-aequo sont tous des morts, la seule différence tient à ce que d_i n'est plus égal à 1 mais au nombre des morts et l'estimateur de Kaplan Meier devient

$$S_n^{KM}(t) = \prod_{i=1}^n \left(1 - \frac{d_i}{n_i}\right) = \prod_{i=1}^n \hat{p}_i \quad (1.16)$$

Erreur standard de l'estimateur de Kaplan-Meier

Une aide essentielle à l'interprétation d'une estimation de toute quantité est la précision de l'estimation, qui se reflète dans l'erreur standard de l'estimation. Ceci est défini comme étant de la racine carrée de la variance estimée de l'estimateur et il est utilisé dans la construction d'une estimation d'intervalle de confiance pour une quantité d'intérêt. Dans cette partie, on va donner l'erreur standard de l'estimateur de Kaplan-Meier, parce que ce dernier est le plus important et le plus utilisé pour la fonction de survie. L'erreur standard de $S_n^{KM}(t)$ il est présenté en détail dans la Section 2.2, page 25, Collett [4].

L'estimateur de Kaplan-Meier pour la fonction de survie, pour toute valeur de t dans l'intervalle de t_k à t_{k+1} , on peut s'écrire :

$$S_n^{KM}(t) = \prod_{j=1}^k \hat{p}_j, \quad k = 1, \dots, n.$$

Où $\hat{p}_j = (n_j - d_j)/n_j$ est la probabilité estimée qu'un l'individu survit à travers l'intervalle du temps qui commence à t_j , $j = 1, \dots, n$. Prendre des logarithmes,

$$\log S_n^{KM}(t) = \sum_{j=1}^k \log \hat{p}_j,$$

et donc la variance de $\log S_n^{KM}(t)$ est donnée par :

$$\text{var} \{ \log S_n^{KM}(t) \} = \sum_{j=1}^k \text{var} \{ \log \hat{p}_j \}. \quad (1.17)$$

Maintenant, le nombre d'individus qui survivent au début de l'intervalle à t_j peut être supposé avoir une distribution binomiale avec les paramètres n_j et p_j , où p_j est la vraie probabilité de survie à travers cet intervalle. Le nombre observé qui survit est $n_j - d_j$, et en utilisant le résultat que la variance d'une variable aléatoire binomiale avec les paramètres n, p est $np(1 - p)$, la variance de $n_j - d_j$ est donné par

$$\text{var}(n_j - d_j) = n_j p_j (1 - p_j).$$

Puisque $\hat{p}_j = (n_j - d_j)/n_j$, la variance de $\hat{p}_j = (n_j - d_j)/n_j^2$, c'est-à-dire $p_j(1 - p_j)/n_j$. La variance de \hat{p}_j peut alors être estimée par

$$\hat{p}_j(1 - \hat{p}_j)n_j. \quad (1.18)$$

Pour obtenir la variance de $\log \hat{p}_j$, on utilise un résultat général pour la variance approximative d'une fonction d'une variable aléatoire. Selon ce résultat, la variance d'une fonction $g(X)$ de la variable aléatoire X est donnée par

$$\text{var}\{g(X)\} \approx \left\{ \frac{dg(X)}{dX} \right\}^2 \text{var}(X). \quad (1.19)$$

C'est ce qu'on appelle l'approximation de la série de Taylor à la variance d'une fonction d'une variable aléatoire. En utilisant l'équation (1.19), la variance approximative de $\log \hat{p}_j$ est $\text{var}(\hat{p}_j)/\hat{p}_j^2$, et en utilisant l'expression (1.18), la variance estimée approximative de $\log \hat{p}_j$ est $(1 - \hat{p}_j)/(n_j \hat{p}_j)$, qui, par substitution à \hat{p}_j , se réduit à

$$\frac{d_j}{n_j(n_j - d_j)}. \quad (1.20)$$

En suite, à partir de l'équation (1.17)

$$\text{var}\{\log S_n^{KM}(t)\} \approx \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}, \quad (1.21)$$

et une nouvelle application du résultat de l'équation (1.19) donne

$$\text{var}\{\log S_n^{KM}(t)\} \approx \frac{1}{[S_n^{KM}(t)]^2} \text{var}\{S_n^{KM}(t)\}.$$

Donc

$$\text{var} \{S_n^{KM}(t)\} \approx [S_n^{KM}(t)]^2 \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}. \quad (1.22)$$

Enfin, l'erreur standard de l'estimateur de Kaplan-Meier est donné par :

$$S_n^{KM}(t) \approx S_n^{KM}(t) \left\{ \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}}, \quad \text{pour } t_k \leq t < t_{k+1}. \quad (1.23)$$

Ce résultat est connu sous le nom de formule de Greenwood. S'il n'y a pas du temps de survie censuré, $n_j - d_j = n_{j+1}$ et l'expression (1.20) devient $(n_j - n_{j+1})/n_j n_{j+1}$. Maintenant

$$\sum_{j=1}^k \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^k \left(\frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \sum_{j=1}^k \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

qui peut s'écrire

$$\frac{1 - S_n^{KM}(t)}{n_1 S_n^{KM}(t)}.$$

Puisque $S_n^{KM}(t) = n_{k+1}/n_1$ pour $t_k \leq t < t_{k+1}$, $k = 1, 2, \dots, n-1$. En l'absence de censure. Par conséquent, d'après l'équation (1.23), la variance estimée de $S_n^{KM}(t)$ est $S_n^{KM}(t) [1 - S_n^{KM}(t)]/n_1$. Il s'agit d'une estimation de la variance de la fonction de survie empirique, donnée dans l'équation (1.23), en supposant que le nombre des individus à risque au temps t a une distribution binomiale avec des paramètres $n_1, S(t)$.

Normalité asymptotique

Il est montré que, sous certaines conditions, l'estimateur de Kaplan-Meier est uniformément consistant, asymptotiquement normal. Pour une discussion détaillée sur le comportement asymptotique de cet estimateur, se référer à l'article de [Stute \[21\]](#) et les références qui y sont mentionnées. La normalité asymptotique de $S_n^{KM}(t)$ est établie dans le théorème suivant.

Théoreme 1.5.1

En tout point de continuité $0 \leq t_0 \leq \tau$ de S , où $\tau > 0$ avec $S(\tau) > 0$, on a

$$\sqrt{n}(\hat{S}(t_0) - S(t_0)) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N(0, \text{var}^2(t_0)),$$

où $\text{var}^2(t_0) := -S^2(t) \int_0^{t_0} S(du)/S^2(u)G(u)$ avec G désignant la fonction de survie de la variable Y .

Preuve. voir [21]. ■

Intervalles de confiance

Sous les conditions du [Théoreme 1.5.1](#) . On peut construire, pour un niveau de confiance $0 < \alpha < 1$ fixé, un intervalle d'estimation $IC(\alpha)$ pour $S(t)$ de la manière suivante

$$IC(\alpha) := \left[S_n^{KM}(t) \pm Z_{\alpha/2} \sqrt{\widehat{var}(S_n^{KM}(t))} \right],$$

où $Z_{\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de loi normale standard $N(0, 1)$, c'est-à-dire $Z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

Remarque 1.5.4

d'après [16], cet intervalle ne peut être utilisé quand $S_n^{KM}(t)$, est proche de 0 ou de 1. En effet l'intervalle étant symétrique autour de $S_n^{KM}(t)$, les bornes peuvent dépasser les valeurs 0 ou de 1.

Pour remédier à ce problème, on utilise l'intervalle de confiance de rothman défini comme suit

$$IC(\alpha) := \frac{K}{K + (Z_{\alpha/2})^2} \left[S_n^{KM}(t) \pm \frac{Z_{\alpha/2}}{2K} \pm Z_{\alpha/2} \sqrt{\widehat{var}(S_n^{KM}(t)) + \frac{(Z_{\alpha/2})^2}{4K^2}} \right].$$

Avec

$$K = \frac{S_n^{KM}(t)(1 - S_n^{KM}(t))}{\widehat{var}(S_n^{KM}(t))}.$$

(source [14]).

1.5.2 Méthode de Nelson-Aalen pour la fonction du risque cumulé

L'estimateur de Kaplan-Meier est un estimateur non-paramétrique de la fonction de survie. L'estimateur de Nelso-Alen aussi est un estimateur non-parametrique, mais pour la fonction du risque cumulé, il est intoduit par Nelson [15] en 1972 et Aalen [1] en 1978. D'après la propriété de l'indépendance entre X et Y , on peut écrit $N(t)$ comme suit

$$N(t) = 1 - (1 - F(t))(1 - G(t)) = N^0(t) + N^1(t),$$

où

$$N^0(t) = P(Z \leq t; \delta = 0) = \int_0^t \bar{F}(x) dG(x) \quad \text{et} \quad N^1(t) = P(Z \leq t; \delta = 1) = \int_0^t \bar{G}(x) dF(x).$$

La fonction du risque cumulé (1.5) peut s'exprimer de la forme suivante

$$H(t) = \int_0^t \frac{\bar{G}(x) dF(x)}{\bar{N}(x)} = \int_0^t \frac{dN^1(t)}{\bar{N}(t)}.$$

Définition 1.5.2 (Estimateur de Nelson-Aalen)

L'estimateur de Nelson-Aalen $H_n(t)$ de H basé sur l'échantillon $\{(Z_i, \delta_i)\}, 1 \leq i \leq n$ donné par

$$H_n(t) = \int_0^t \frac{dN_n^1(n)}{\bar{N}(t)} = \begin{cases} \sum_{Z_{i:n} \leq t} \frac{\delta_{[i:n]}}{n-i+1} & \text{si } t < Z_{i:n} \\ 1 & \text{si } t \geq Z_{i:n}, \end{cases}$$

où

$$N_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{Z_i \leq t\} \quad \text{et} \quad N_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^n \delta_i \mathbb{I}\{Z_i \leq t\}.$$

représentent respectivement la fdr empirique de $N(t)$ et version empirique de $N^1(t)$ de l'échantillon Z_1, \dots, Z_n .

Remarque 1.5.5

En remplaçant $H(t)$ par $H_n(t)$ dans (1.8) on obtient un nouvel estimateur de la fonction de survie S

$$S_n^{NA}(t) := \begin{cases} \prod_{Z_{i:n} < t} \exp\left\{-\frac{\delta_{[i:n]}}{n-i+1}\right\} & \text{si } t < Z_{n:n} \\ 0 & \text{si non} \end{cases}.$$

Se référer à l'article de [Brahimi et al \[2\]](#) et la thèse de [Soltane \[20\]](#).

Remarque 1.5.6

Une première procédure d'estimation du risque cumulé consiste à exploiter la relation (1.7), en remplaçant $S(t)$ par $S_n^{KM}(t)$. D'où un premier estimateur de la forme

$$H_n(t) = - \sum_{i:Z_{i:n} \leq t} \log\left(1 - \frac{\delta_{[i:n]}}{n-i+1}\right).$$

On note que si on utilise la formule (1.16) pour $S_n^{KM}(t)$, on obtient

$$H_n(t) = -\sum_{i=1}^n \log\left(1 - \frac{d_i}{n_i}\right).$$

1.6 Comparaison de deux groupes de données de survie

Après avoir estimé les courbes de survie sur deux groupes d'individus ou plus, on est souvent amené à vouloir les comparer. L'exemple le plus simple est celui de deux échantillons d'individus issus initialement de la même population mais dont l'un a été soumis à une intervention (par exemple un traitement, une action marketing, ...). La question est alors de tester l'efficacité de cette intervention. Il est extrêmement important de s'assurer qu'à l'exception de l'intervention en question, les deux ensembles d'individus possèdent les mêmes autres caractéristiques. Dans le cas contraire, une divergence des courbes de survie ne peut pas être attribuable à la seule intervention dont on cherche à apprécier l'impact. Par exemple que les caractéristiques d'âges, de sexe, de poids, de catégories socio-professionnelles, ... sont les mêmes dans les deux groupes dont l'un est la cible d'une action marketing et l'autre pas. La plupart des statistiques utilisées sont fondées sur des tableaux de contingences construits sur l'ensemble des temps d'évènements.

Pour la comparaison de deux ou plusieurs groupes de données de survie, il existe un certain nombre de méthodes qui peuvent être utilisées pour quantifier l'étendue des différences entre les groupes (méthode graphique, statistiquement, ...). Donc, statistiquement : on a les tests les plus couramment utilisés sont ceux du log-rank et de Wilcoxon-Gehan. Sera maintenant considérée uniquement à la procédure non paramétrique du test du log-rank.

1.6.1 Test du log-rank

Le test du log-rank est la méthode la plus largement utilisée pour comparer deux ou plusieurs courbes de survie. L'hypothèse nulle est qu'il n'y a aucune différence de survie entre les deux groupes. Le test du log-rank est un test non paramétrique, qui ne fait aucune hypothèse sur les distributions de survie. Essentiellement, le test du log-rank compare le nombre d'évènements observés dans chaque groupe à ce qui serait attendu si l'hypothèse nulle était vraie (c-à-d si les courbes de survie étaient identiques). La statistique du log-rank est approximativement distribuée sous forme de statistique de test du Chi^2 .

Aussi appelée statistique de Mantel-Haenzel, elle doit son nom à Peto et Peto (1972) qui

la dérive en considérant les estimateurs du logarithme des fonctions de survie. Pour simplifier nous l'exposerons relativement en détail pour 2 seulement groupes d'individus, avant d'aborder l'extension à plus de deux groupes.

Comparaison de deux fonctions de survie

Soient 2 groupes d'individus indicés 1 et 2 et les effectifs observés au temps t_i :

Groupe	1	2	total
Individus ayant connu l'évènement	d_{1i}	d_{2i}	d_i
Individus n'ayant pas connu l'évènement	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
Individus risqués	n_{1i}	n_{2i}	n_i

L'hypothèse nulle est l'égalité des courbes de survie. Sous H_0 la proportion attendue d'évènements à un temps t_i quelconque est donnée par d_i/n_i et le nombre espéré d'évènements au sein des groupes est obtenu en appliquant cette proportion à l'effectif observé de chacun des groupes. Soit :

$$\begin{aligned}
 e_{1i} &= n_{1i} \frac{d_i}{n_i}, \quad \text{pour le 1}^{er} \text{ groupe} \\
 e_{2i} &= n_{2i} \frac{d_i}{n_i} \\
 &= (n_i - n_{1i}) \frac{d_i}{n_i} \\
 &= d_i - e_{1i}, \quad \text{pour le seconde groupe.}
 \end{aligned}$$

Les règles de construction des données nécessaires à ces calculs sont celles utilisées dans les calculs de la statistique KM. Ainsi, le nombre d'individus risqués au temps t_i est égal au nombre d'individus risqués en t_{i-1} diminué de l'effectif des individus ayant connu l'évènement en t_{i-1} et de celui des individus censurés entre t_{i-1} inclus et t_i exclu.

Une fois ces tables construites pour l'ensemble des temps d'évènement, ensemble obtenu par l'union des deux sous-ensembles de temps d'évènement afférents à chacun des groupes (les temps considérés sont donc soit observés dans le groupe 1 soit observés dans le groupe 2), on calcule 4 quantités :

- Le nombre total d'évènements observés dans le premier groupe, c-à-d la somme des d_{1i} , noté O_1 .
- Le nombre total d'évènements observés dans le second groupe, c-à-d la somme des d_{2i} , noté O_2 .

- Le nombre total d'évènements espérés sous H_0 pour le premier groupe, c-à-d la somme des e_{1i} , noté E_1 .
- Le nombre total d'évènements espérés sous H_0 pour le second groupe, c-à-d la somme des e_{2i} , noté E_2 .

On note au passage la relation d'égalité $O_1 + O_2 = E_1 + E_2$.

La statistique $O_1 - E_1$ est statistique de log-rank ou de Mantel-Haenzel. Si on imagine qu'une action marketing a été effectuée auprès des individus du groupe 2 et que l'évènement est la rupture de la relation client, alors une valeur positive de la statistique signifie que le nombre d'évènements observés dans le groupe 1 est supérieur à celui attendu sous H_0 où, de façon équivalente puisque $O_1 - E_1 = E_2 - O_2$, que le nombre d'évènements observés dans le groupe 2 est inférieur à celui attendu sous H_0 et que donc l'action marketing en question affecte globalement favorablement la courbe de survie et donc la fidélisation du client. Nous reviendrons ultérieurement sur la signification du terme globalement.

Il reste à dériver la distribution de cette statistique pour être en mesure de réaliser un test de significativité de l'intervention. Pour cela on retrouve une statistique de *Chi2* de Pearson couramment employée dans l'étude des tableaux de contingence (Cf. cours de statistique non paramétrique) :

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = (O_1 - E_1)^2 \left(\frac{1}{E_1} + \frac{1}{E_2} \right).$$

Sous l'hypothèse nulle d'égalité des courbes de survie, cette quantité est asymptotiquement la réalisation d'un *Chi2* à un degré de liberté.

CHAPITRE 2

Exemple d'application

Dans ce chapitre, on applique quelques unes des méthodes vues aux [Chapter 1](#) sur un exemple de données réelles. Les résultats numériques et les représentations graphiques sont obtenus à l'aide du package "survival", "survminer", "dplyr" et "ggplot2" du logiciel d'analyse statistique R.

2.1 Présentation des données

Les données de l'exemple suivant, présentées dans le [Table 2.1](#), sont prises du package "survival" du logiciel R. Ces données, très célèbres dans l'illustration des outils de l'analyse de survie, peuvent être trouvées dans plusieurs documents comme par exemple, le livre de [Michael J. Carwley \[13\]](#), la page 265. Elles concernent les résultats d'un essai d'étude sur la mortalité chez 150 moutons mâles sauvages atteints par parasites intestinaux. La survie globale, exprimée en mois, (temps écoulé du début de thérapie à la mort du mouton, causée par la maladie), tous poids de 0.100 à 50 *kg*, certains de ces moutons n'étaient pas morts au moment où l'étude a été terminée, et ces moutons contribuent donc à des temps de survie censurés à droite.

Il y avait trois groupes expérimentaux, où les moutons ont été suivis pendant 50 mois. Les groupes ont été traités avec trois médicaments différents contre leurs parasites intestinaux : Le groupe *A* a reçu un bolus avec une forte dose de tueur de vers, le groupe *B* a reçu une faible dose et le groupe *C* a reçu le placebo (un bolus sans contenu destructeur de vers).

La masse corporelle initiale de chaque individu (poids) a été enregistrée comme covariable. Le mois au cours duquel chaque mouton est mort a été enregistré et les moutons qui ont survécus jusqu'au 50 mois (fin de l'étude) ont été enregistrés comme étant censurés (pour eux, l'indicateur de censure **événement**= 0, tandis que les moutons qui sont morts ont tous un **événement**= 1). Un sous-ensemble des données de mouton est présenté dans le [Table 2.1](#).

L'intérêt principal est de déterminer si les moutons du groupe *A* qui a reçu un bolus avec une forte dose de tueur de vers survivent pendant un temps plus long, par rapport aux moutons du groupe *B* et groupe *C*. En d'autres termes, on veut s'assurer de l'efficacité du nouveau mode de traitement.

Temps	Événement	Poids	Groupe
20	1	5.385	<i>A</i>
34	1	7.413	<i>A</i>
1	1	9.266	<i>A</i>
2	1	6.228	<i>A</i>
3	1	5.229	<i>A</i>
3	1	9.699	<i>A</i>
50	0	1.973	<i>A</i>
26	1	5.838	<i>A</i>
1	1	2.088	<i>A</i>
50	0	0.237	<i>A</i>
21	1	6.814	<i>A</i>
3	1	5.502	<i>A</i>
13	1	1.137	<i>A</i>
11	1	6.323	<i>A</i>
22	1	7.384	<i>A</i>
50	0	8.713	<i>A</i>
50	0	7.458	<i>A</i>
1	1	1.424	<i>A</i>
50	0	1.312	<i>A</i>
9	1	5.162	<i>A</i>
⋮	⋮	⋮	⋮

TAB. 2.1 – sheep data (package survival). Observations sur la mortalité chez 150 moutons mâles sauvages atteints par parasites intestinaux. Seules les données des 20 premiers moutons sont présentées ici.

2.2 Statistiques descriptives

Ici, on va utiliser l'approche descriptive pour décrire les données d'étude à l'aide de tableaux ou de représentations graphiques qui décrivent les caractéristiques générales des données. Cette approche, elle représente une étape nécessaire et constitue un premier pas dans l'analyse et l'interprétation des données. Pour ce là, les statistiques descriptives élémentaires des données sont résumées dans le [Table 2.2](#).

	Temps	Poids (kg)	groupe
Minimum	1.00	0.055	<i>A</i> : 50
1er Q_1	1.00	2.459	<i>B</i> : 50
Médiane	7.00	6.316	<i>C</i> : 50
Moyenne	15.17	9.390	
Écart-type	17.542	9.483	
3ème Q_3	21.00	11.955	
Maximum	50.00	42.090	

TAB. 2.2 – Résumé statistiques descriptives élémentaires des moutons mâles sauvages atteints par parasites intestinaux.

On constate que la plus petite poids est $Min = 0.055$, la plus grande poids est $Max = 42.090$. On sait qu'il y a 150 moutons et on note à partir le tableau au-dessus il y a 50 moutons pour chaque groupe, alors la 1^{er} quartile : 25% de 150 moutons ont une poids inférieure ou égale à $Q_1 = 2.459$. Pour la 3^{ème} quartile : 25% de 150 moutons ont une poids supérieure ou égale à $Q_3 = 11.955$. On a $l'ecart - type = 9.483$, il est fort, donc les valeurs de poids ne sont pas dispersées autour de la moyenne cela signifie que les poids des moutons hétérogène.

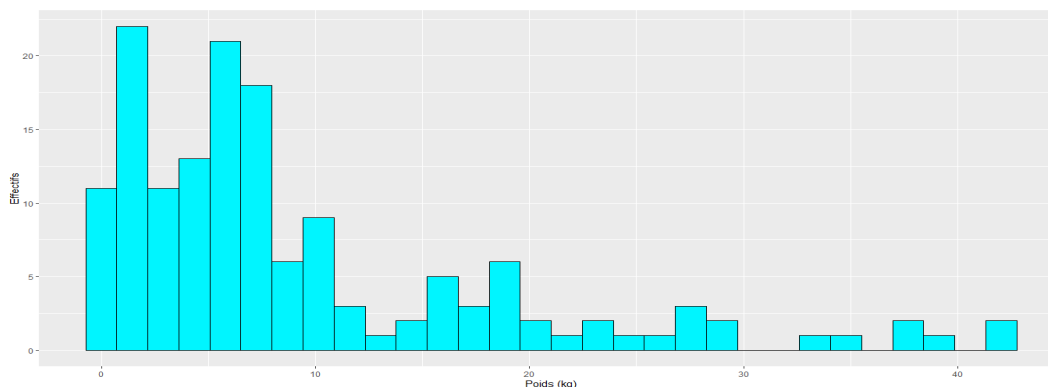


FIG. 2.1 – Distribution du poids des moutons mâles sauvages atteints par parasites intestinaux.

La [Figure 2.1](#) représente la distribution du poids des moutons mâles sauvages atteints par parasites intestinaux, en vertu des résultats du [Table 2.2](#) et du fait que le graphe de l'histogramme [2.1](#), nous ont permis de séparer les poids en deux groupes : le premier constitué des moutons qui lui sont supérieurs à 10 kg et les autres moutons, qui lui sont inférieurs à 10 kg . Par conséquent, les 150 moutons sont répartis selon le poids et le groupe comme le montrent le [Table 2.3](#) et on obtiendra de représentation graphique presque parfaite pour chaque groupe en fonction du poids comme le montre la [Figure 2.2](#).

Poids	Groupe					
	A		B		C	
	Effectifs	Proportions	Effectifs	Proportions	Effectifs	Proportions
$\geq 10\text{ kg}$	0	0%	25	50%	18	36%
$< 10\text{ kg}$	50	100%	25	50%	32	64%
Total	50	100%	50	100%	50	100%

TAB. 2.3 – Effectifs et proportions des 150 moutons mâles sauvages atteints par parasites intestinaux répartis selon le poids et le groupe

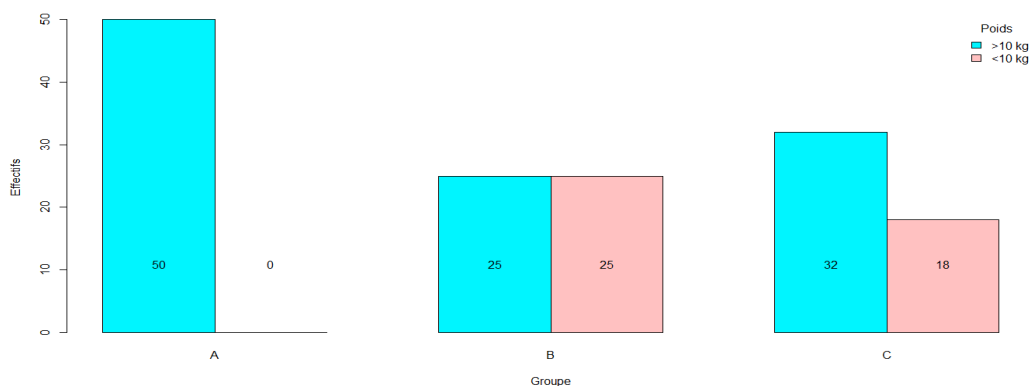


FIG. 2.2 – Distribution du groupe en fonction du poids.

Il est clairement que du [Table 2.3](#) et la [Figure 2.2](#), qu'il y a 50 moutons du pourcentage 100% leurs poids sont inférieurs à 10 kg et 0 mouton du pourcentage 0% leurs poids sont supérieurs à 10 kg dans le groupe A. Alors que, dans le groupe B c'est égal. L'effectif des moutons dans le groupe C qui lui sont inférieurs à 10 kg sont 32 moutons du pourcentage 64%, tans que lui sont supérieurs à 10 kg sont 18 moutons du pourcentage 36%.

2.3 Statistique inférentielle

Ici, on va utiliser l'approche inférentielle (inductive), elle a pour objectif de mettre en place des règles de décision afin de réaliser de l'inférence statistique, qui réside dans l'étudier de l'analyse de survie et son utilisation pour trouver la fonction de survie et la fonction de risque et connaître l'importance de la variable "poids" sur le risque de décès.

La [Définition 1.5.1](#) permet de calculer l'estimation de la probabilité de survie par la méthode de KM au [Table 2.1](#). Les résultats obtenus sont donnés dans le [Table 2.4](#), où la dernière colonne contient, pour chaque instant t_i (Temps), l'intervalle d'estimation à 95% de confiance. La représentation graphique associée au [Table 2.4](#) est présentée dans la [Figure 2.3](#), où la courbe de survie estimée pour les 150 données de moutons en escaliers décroissants (avec les bornes de confiance à 95%), sur lequel on constate que la plus grande des valeurs non censurées a pour estimation $\mathbf{S}_n^{KM}(46) = 0.133 \neq 0$. Ceci témoigne de l'existence de données censurées au delà de 46.

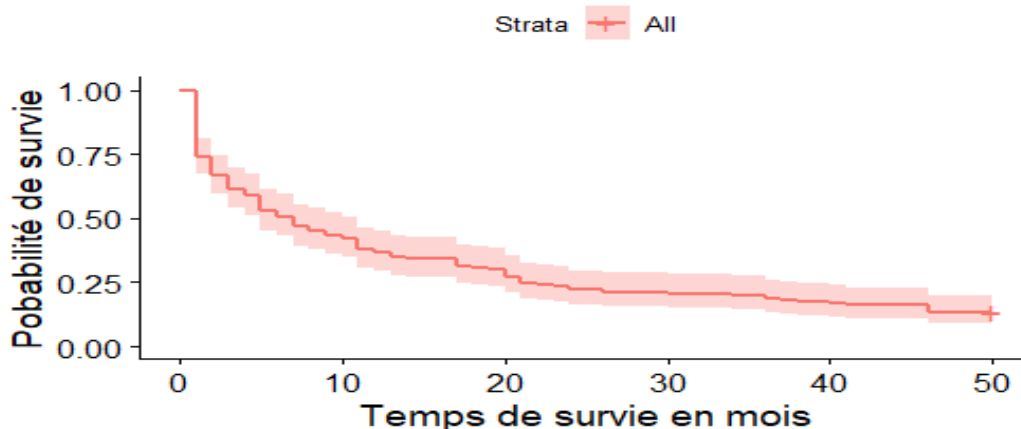


FIG. 2.3 – Courbe de survie estimée pour les 150 données de moutons.

Les résultats numériques de l'estimation des fonctions de survie relatives aux groupes de A , B et C sont respectivement résumés dans les [Table 2.5](#), [Table 2.6](#) et [Table 2.7](#). La dernière colonne contient, pour chaque instant t_i , l'intervalle d'estimation 95% de confiance, sur lequel on constate que pour le groupe A , la plus grande des valeurs non censurées a pour estimation $\mathbf{S}^{KM}(46) = 0.28 \neq 0$. Ceci témoigne de l'existence de données censurées supérieures à 46. Il en est de même pour le groupe B , où l'observation maximale non censurée a pour estimation $\mathbf{S}^{KM}(46) = 0.12 \neq 0$. Dans le cas du groupe C , les données sont complètes, d'où une estimation nulle pour la survie à l'observation maximale 46 et au delà.

La Figure 2.4 contient la représentation des courbes de survie avec la médiane estimées au sein de chaque groupe associée au Table 2.5, Table 2.6, Table 2.7 et Table 2.8, où on voit que les courbes de survie relatives au groupe A sont, en tout temps, au-dessus de celles des deux groupes de B et C. Ceci permet de conclure que le traitement au bolus avec une forte dose de tueur de vers a un effet positif sur la durée de survie de moutons atteints par parasites intestinaux. Autrement dit, la probabilité de survie estimée est plus élevée pour le groupe A.

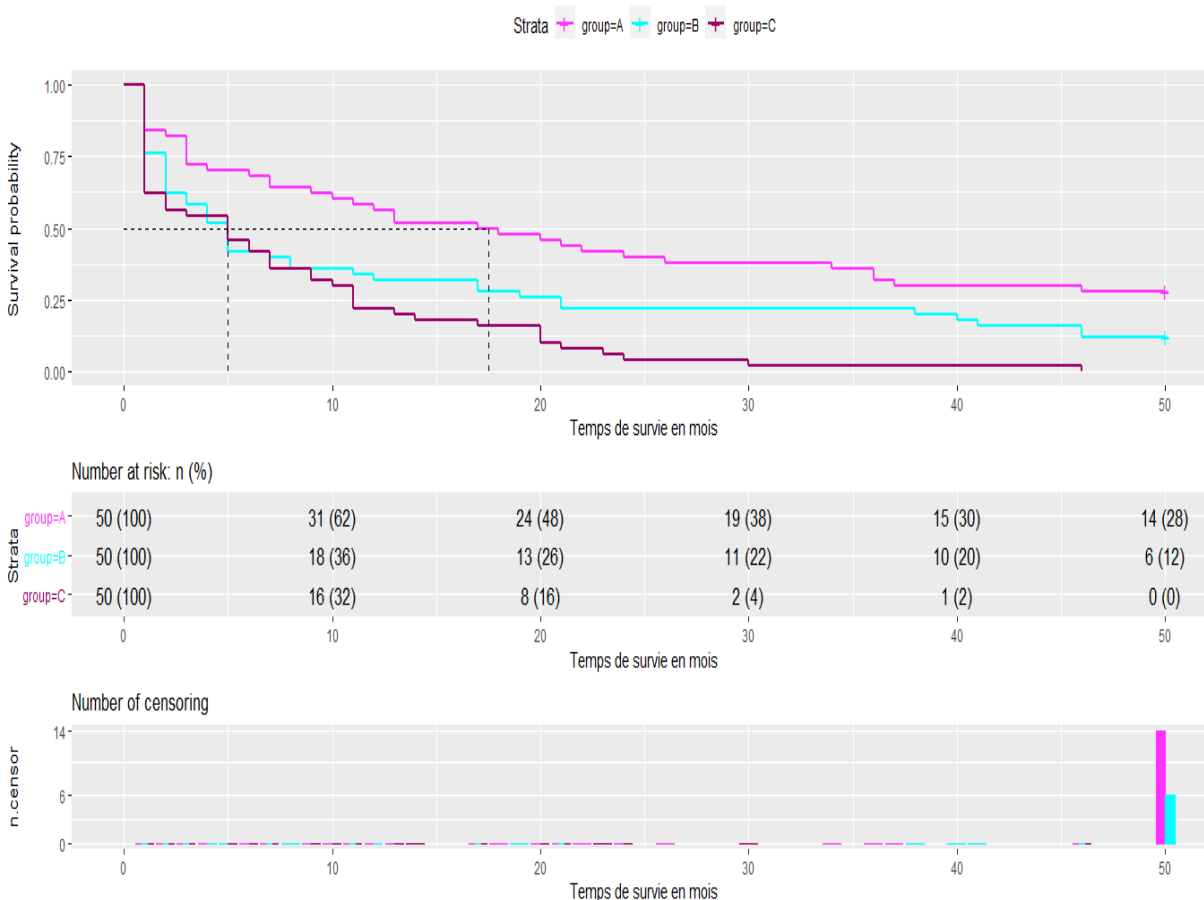


FIG. 2.4 – Courbes de survie estimées au sein de chaque groupe.

Le Table 2.8, fournit un résumé utile des données qui peuvent être utilisées pour estimer des mesures comme la médiane du temps de survie. On constate que la médiane du temps de survie pour le groupe A est 17.5 mois, contre 5 mois pour les groupes B et C. Il semble y avoir un avantage de survie pour le groupe A par rapport aux groupes B et C. Cependant, pour déterminer si cette différence est statistiquement significative, il faut un test statistique officiel, c'est un sujet qui est abordé dans la partie suivante.

Le risque cumulé, donné dans la relation (1.7), peut être interprété comme la force cumulative de la mortalité. En d'autres termes, il correspond au nombre d'événements qui seraient attendus pour chaque individu au temps t si l'événement était un processus répétable. On peut illustrer cela par une représentation graphique (voir la Figure 2.5) du risque cumulé estimé au sein de chaque groupe.

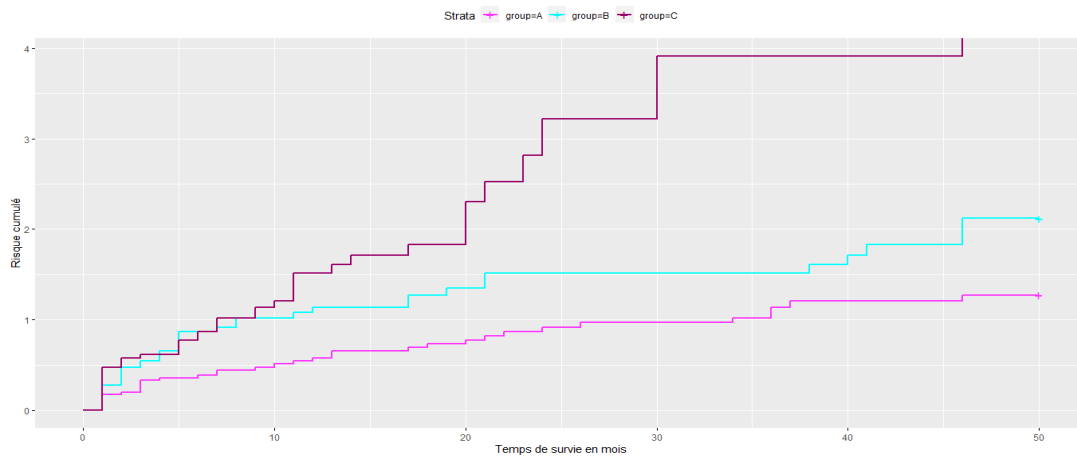


FIG. 2.5 – Courbes du risque cumulé estimé au sein de chaque groupe.

La figure ci-dessus montre qu'il y a des courbes croissantes qui augmentent au temps de décès observés.

Pour comparer les survies entre les trois groupes A , B et C statistiquement. On fait la construction du test log-rank (refer à Section 1.6) d'égalité des 3 courbes. Le résultat du test demandé est présenté dans le Table 2.9, on voit que le test du $Chi2$ 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	Observées	Censurées	Attendue	$(O_1 - E_1)^2 / E_1$	$(O_2 - E_2)^2 / E_2$
group= A	50	36	14	56.4	7.406	15.29
group= B	50	44	6	41.2	0.192	0.32
group= C	50	50	0	32.4	9.608	15.49

$Chi2 = 20.7$ de 2 degrés de liberté, $p = 0.00003$

TAB. 2.9 – Tests d'égalité des 3 survies : résultats de la statistique du log-rank

Temps mois (t_i)	n.risque n_i	n.évén. d_i	pro.survie $S_n^{KM}(t)$	err.std.	IC (95%)
1	150	39	0.740	0.0358	0.673 – 0.814
2	111	11	0.667	0.0385	0.595 – 0.747
3	100	8	0.613	0.0398	0.540 – 0.696
4	92	4	0.587	0.0402	0.513 – 0.671
5	88	9	0.527	0.0408	0.452 – 0.613
6	79	3	0.507	0.0408	0.433 – 0.593
7	76	6	0.467	0.0407	0.393 – 0.554
8	70	2	0.453	0.0406	0.380 – 0.540
9	68	3	0.433	0.0405	0.361 – 0.520
10	65	2	0.420	0.0403	0.348 – 0.507
11	63	6	0.380	0.0396	0.310 – 0.466
12	57	2	0.367	0.0393	0.297 – 0.452
13	55	3	0.347	0.0389	0.278 – 0.432
14	52	1	0.340	0.0387	0.272 – 0.425
17	51	4	0.313	0.0379	0.247 – 0.397
18	47	1	0.307	0.0376	0.241 – 0.390
19	46	1	0.300	0.0374	0.235 – 0.383
20	45	4	0.273	0.0364	0.211 – 0.355
21	41	4	0.247	0.0352	0.186 – 0.326
22	37	1	0.240	0.0349	0.180 – 0.319
23	36	1	0.233	0.0345	0.175 – 0.312
24	35	2	0.220	0.0338	0.163 – 0.297
26	33	1	0.213	0.0334	0.157 – 0.290
30	32	1	0.207	0.0331	0.151 – 0.283
34	31	1	0.200	0.0327	0.145 – 0.275
36	30	2	0.187	0.0318	0.134 – 0.261
37	28	1	0.180	0.0314	0.128 – 0.253
38	27	1	0.173	0.0309	0.122 – 0.246
40	26	1	0.167	0.0304	0.116 – 0.238
41	25	1	0.160	0.0299	0.111 – 0.231
46	24	4	0.133	0.0278	0.089 – 0.201

TAB. 2.4 – Estimateur de Kaplan-Meier de la fonction de survie (en mois) pour les 150 moutons mâles sauvages atteints par parasites intestinaux

Temps mois (t_i)	n.risque n_i	n.évén. d_i	pro.survie $S_n^{KM}(t)$	err.std.	IC (95%)
1	50	8	0.84	0.0518	0.744 – 0.948
2	42	1	0.82	0.0543	0.720 – 0.934
3	41	5	0.72	0.0635	0.606 – 0.856
4	36	1	0.70	0.0648	0.584 – 0.839
6	35	1	0.68	0.0660	0.562 – 0.822
7	34	2	0.64	0.0679	0.520 – 0.788
9	32	1	0.62	0.0686	0.499 – 0.770
10	31	1	0.60	0.0693	0.478 – 0.752
11	30	1	0.58	0.0698	0.458 – 0.734
12	29	1	0.56	0.0702	0.438 – 0.716
13	28	2	0.52	0.0707	0.398 – 0.679
17	26	1	0.50	0.0707	0.379 – 0.660
18	25	1	0.48	0.0707	0.360 – 0.641
20	24	1	0.46	0.0705	0.341 – 0.621
21	23	1	0.44	0.0702	0.322 – 0.602
22	22	1	0.42	0.0698	0.303 – 0.582
24	21	1	0.40	0.0693	0.285 – 0.562
26	20	1	0.38	0.0686	0.267 – 0.541
34	19	1	0.36	0.0679	0.249 – 0.521
36	18	2	0.32	0.0660	0.214 – 0.479
37	16	1	0.30	0.0648	0.196 – 0.458
46	15	1	0.28	0.0635	0.180 – 0.437

TAB. 2.5 – Résultats de l'estimation de la fonction de survie du groupe A traité au bolus avec une forte dose de tueur de vers.

Temps mois (t_i)	n.risque n_i	n.évén. d_i	pro.survie $S_n^{KM}(t)$	err.std.	IC (95%)
1	50	12	0.76	0.0604	0.650 – 0.888
2	38	7	0.62	0.0686	0.499 – 0.770
3	31	2	0.58	0.0698	0.458 – 0.734
4	29	3	0.52	0.0707	0.398 – 0.679
5	26	5	0.42	0.0698	0.303 – 0.582
7	21	1	0.40	0.0693	0.285 – 0.562
8	20	2	0.36	0.0679	0.249 – 0.521
11	18	1	0.34	0.0670	0.231 – 0.500
12	17	1	0.32	0.0660	0.214 – 0.479
17	16	2	0.28	0.0635	0.180 – 0.437
19	14	1	0.26	0.0620	0.163 – 0.415
21	13	2	0.22	0.0586	0.130 – 0.371
38	11	1	0.20	0.0566	0.115 – 0.348
40	10	1	0.18	0.0543	0.100 – 0.325
41	9	1	0.16	0.0518	0.085 – 0.302
46	8	2	0.12	0.0460	0.057 – 0.254

TAB. 2.6 – Résultats de l'estimation de la fonction de survie du groupe B traité au bolus avec une faible dose de tueur de vers.

Temps mois (t_i)	n.risque n_i	n.évén. d_i	pro.survie $S^{KM}(t)$	err.std.	IC (95%)
1	50	19	0.62	0.0686	0.499 – 0.770
2	31	3	0.56	0.0702	0.438 – 0.716
3	28	1	0.54	0.0705	0.418 – 0.697
5	27	4	0.46	0.0705	0.340 – 0.621
6	23	2	0.42	0.0698	0.303 – 0.582
7	21	3	0.36	0.0679	0.249 – 0.521
9	18	2	0.32	0.0660	0.214 – 0.479
10	16	1	0.30	0.0648	0.196 – 0.458
11	15	4	0.22	0.0586	0.130 – 0.371
13	11	1	0.20	0.0566	0.115 – 0.348
14	10	1	0.18	0.0543	0.100 – 0.325
17	9	1	0.16	0.0518	0.085 – 0.302
20	8	3	0.08	0.0424	0.043 – 0.230
21	5	1	0.08	0.0384	0.031 – 0.205
23	4	1	0.06	0.0336	0.020 – 0.180
24	3	1	0.04	0.0277	0.010 – 0.156
30	2	1	0.02	0.0198	0.003 – 0.139
46	1	1	0.00	/	/

TAB. 2.7 – Résultats de l'estimation de la fonction de survie du groupe C traité au bolus sans contenu destructeur de vers.

	N	Médiane	IC (95%)
group= A	50	17.5	9 – 36
group= B	50	5.0	2 – 12
group= C	50	5.0	1 – 9

TAB. 2.8 – La médiane et l'intervalle de confiance estimées par KM pour chaque groupe.

Conclusion

L'analyse de survie est basée sur l'estimation de la fonction de survie (Kaplan-Meier) et la fonction de risque cumulé (Nelson-Aalen) et grâce à la fonction de survie on peut être estimée la médiane du temps de survie pour les patients actuels ou futurs. Dans ce travail, on s'est concentré sur l'étude des deux estimateurs qui, dans le cas où les données sont censurés. Les résultats estimés par l'estimateur de Kaplan-Meier sont particulièrement utiles pour concevoir un système de traitement ou conseiller le patient lors du diagnostic. Il arrive assez souvent qu'une série statistique, ayant des données censurées, présente des valeurs extrêmes. Il serait intéressant de faire le même travail, par ce que ce cas mérite d'être considéré attentivement vu.

BIBLIOGRAPHIE

- [1] Aalen, O. (1978). Non parametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist*, 534-545.
- [2] Brahim, D. Meraghni, A. Necir and L. Soltane. (2018). Nelson-Aalen tail product-limit process and extreme value index estimation under random censorship. Unpublished manuscript, available on the ArXiv archive : <https://arxiv.org/abs/1502.03955>.
- [3] Brian S. E. and Torsten H. (2006). A handbook of statistical analyses using R. Chapman & Hall/CRC. Taylor & Francis Group, LLC.
- [4] Collett, D. (2015). Modelling survival data in medical research. Third Edition. CRC press. aylor & Francis Group, A chapman & hall book.
- [5] Dirk, F. (2016). Applied survival analysis using R. New York, NY : Springer.
- [6] Elisa, T. Lee. (2003). Statistical methods for survival data analysis. John Wiley & Sons.
- [7] Fovea, Groupe. L'analyse de survie des études cliniques. Département Biométrie de Fovea. <http://www.Fovea-group.com/cro/pdf/fms8>. Pdf.
- [8] Gilbert, C. (2019). Économétrie des durées de survie. Note de Cours Master 2 ESA. Voies professionnelle et recherche.
- [9] Göran. B, (2012). Event history analysis with R. (Chapman & Hall/CRC The R Series). Taylor & francis group, LLC.
- [10] Kaplan, E.L, and Meier, P. (1958). Non parametric estimation from incomplete observations. *J. Amer. Statist. Assoc*, 53(282) ; 457481.
- [11] Klein, M. (2003). Survival analysis techniques for censored and truncated data. Spring

-
- [12] Lee, E. T and Wang, J. (2003). Statistical methods for survival data analysis. John Wiley.
- [13] Michael J, C. (2005). Statistics. An introduction using R. John wiley and sons, Ltd.
- [14] M'ziou, I. (2014). Mémoire de master. Estimation non paramétrique, Université Mohamed Khider, Biskra.
- [15] Nelson, W. (1972). A short life test for comparing a sample with previous accelerated. Test Results. Technometrics, 14(1), 175-185.
- [16] Saint, P. P. (2012). Introduction à l'analyse des durées de survie. [Http ://www. lsta. Upmc. Fr/psp/cours survie 1. Pdf.](http://www.lsta.upmc.fr/psp/cours/survie/1.pdf)
- [17] Ramdani, H. (2017) Mémoire de master. Construction d'un test paramétrique de l'estimateur de l'indice des valeurs extrêmes censurées, Université Kasdi Merbah, Ouargla, Algerie.
- [18] Reiss, R. D, Thomas, M. (2006). Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. Birkhauser, Basel.
- [19] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.
- [20] Soltane, L. (2017). Analyse des Valeurs Extrême en Presence de Censure. Thèse de doctorat. Université de Biskra, Algerie.
- [21] Stute, W. (1995). The central limit théorème under random censorship. Ann. Status 2, 422-439.
- [22] Vivian, V. (2006). Processuse empirique, estimation non parametrique et donnée censurées. Paris.
- [23] Wienke, A. (2010). Frailty models in survival analysis. CRC Press.
- [24] Zerroukhi, H. (2019). Mémoire de master. Estimation de l'Indice des valeurs extrêmes sous censure. Université Mohamed Khider, Biskra, Algerie.

Annexe B : Abréviations et Notations

Les différentes abréviations et notation utilisées tout au long de cette thèse sont expliquées ci-dessous.

$err.std.$: Erreur type de l'estimation de Kaplan-Meier.
KM	: Kaplan Meier.
iid	: Indépendantes et identiquement distribué.
IC	: Intervalle de confiance.
NA	: Nelson-Aalen.
n.évén.	: Le nombre d'individus à risque de subir l'événement.
n.risque	: Le nombre de décès.
pro.survie	: Probabilité de survie.
$E(X)$: Espérance ou moyenne de X.
f	: Fonction de densité.
F	: Fonction de répartition.
S	: Fonction de survie.
F_n	: Fonction de répartition empirique.
S_n	: Fonction de survie empirique.
S_n^{KM}	: Estimateur de Kaplan-Meier.
S_n^{NA}	: Estimateur de Nelson Aalen.
F^{-1}	: Inverse généralisé de F .
$var(X)$: Variance de X.
$\mathbb{1}_A$: Fonction indicatrice de l'ensemble A .

Q	:	Fonction quantile.
H	:	Fonction de hazard cumulé.
h	:	Fonction de hazard.
$Z_{1:n}, \dots, Z_{n:n}$:	Statistique d'ordre associées à Z_1, \dots, Z_n .
$\xrightarrow{\mathcal{D}}$:	Convergence en distribution.
$:=$:	Egalité par définition.
$\xrightarrow{p.s}$:	Convergence presque sûre.
$N(0, 1)$:	Loi normale d'espérance 0 et variance 1.
Φ^{-1}	:	Fonction quantil de $N(0, 1)$.

مُلخَص

المقدّر اللامعلمي لدالة البقاء كابلان-ماير هو الأداة الأساسية لتقدير اللامعلمي الخاضع للرقابة. تاريخياً تم استخدام تحليل البقاء على قيد الحياة في المجال الطبي على وجه الخصوص. يساعد تطبيق مثل هذا المقدّر على تحديد الظروف والخصائص التي تزيد أو تقلل من احتمالية البقاء على قيد الحياة للفرد حيث يؤدي غياب نماذج المخاطر إلى فقدان جميع الخصائص التي تؤدي إلى إنشاء نظام علاجي جيد. نقترح مثال لتطبيق هذا الإجراء على ذكور الأغنام البرية المصابة بالطفيليات المعوية.

الكلمات المفتاحية: تحليل البقاء، وظيفة البقاء، دالة المخاطر التراكمية، الرقابة اليمنى، مقدر كابلان-ماير، مقدر نيلسون-آلين.

Résumé

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 survie dans le domaine médical en particulier. L'application d'un tel estimateur aide à identifier les conditions et les caractéristiques qui augmentent ou diminuent la probabilité de survie d'individu où l'absence des modèles de risque entraîne la perte de toutes les caractéristiques qui mènent à la création d'un bon système thérapeutique. On propose un exemple d'application de cette procédure sur moutons mâles sauvages atteints par parasites intestinaux.

Les mots clés : Analyse de survie, Fonction de survie, Fonction de risque cumulé, Censure à droite, Estimateur de Kaplan-Meier, Estimateur de Nelson-Aalen.

Abstract

The nonparametric estimator of the Kaplan-Meier survival function is the fundamental tool of nonparametric estimation under censorship. It has historically been used in survival analysis in the medical field in particular. The application of such an estimator helps to identify the conditions and characteristics that increase or decrease the probability of survival of individuals where the absence of the risk models results in the loss of all the characteristics that lead to the creation of a good system therapeutic. An example of the application of this procedure to wild male sheep affected by intestinal parasites is proposed.

Key words: Survival analysis, Survival function, Cumulative risk function, Right censoring, Kaplan-Meier estimator, Nelson-Aalen estimator.