République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahmane MIRA-Bejaia

Faculté des Sciences Exactes

Département Mathématique/MI

Mémoire de fin d'étude

En vue d'obtention du diplôme de Master en Mathématiques

***O*ption : *P*robabilités *S*tatistique et *A*pplication**

Thème :

Application du modèle de Cox dans l'analyse des données de survie.

Réalisé par :

SAHLI SAIDA

Setenu devant le jury composé de :

Présidante :	*M^me TIMERIDJINE.K*	**M.C.A**	**U.A/Mira BEJAIA**
Rapporteur :	*M^me LAGHA.K*	**M.C.A**	**U.A/Mira BEJAIA**
Examinateur :	*M^r BOURAINE.M*	**M.A.A**	**U.A/Mira BEJAIA**
Examinatrice :	*M^me TABTI.H*	**M.C.B**	**U.A/Mira BEJAIA**

Promotion : 2018/2019

REMERCIEMENTS

Je remercie tout d'abord Dieu tout puissant de m'avoir donné le courage, la force et la patience d'achever ce modeste travail.

Mes premiers remerciements s'adressent à ma promotrice, pour son indispensable soutien, sa grande disponibilité, sa générosité, son encouragement ses remarques constructives, ses aides inestimables et son suivi attentif tout au long de la réalisation de mon mémoire. J'ai beaucoup apprécié travailler sous sa direction, d'autant plus que ceci m'a permis d'approfondir mes connaissances gràce à ses judicieux conseils.

Mes remerciements s'adressent à mon encadreur au sein de l'hôpital d'Amizour, l'assistant médical Mr CHELHABE.L et aussi à Mr TENKHI.H, Dr HOCINI et au personnel du service oncologie, pour leurs accueil chaleureux. Je tiens à remercier particulièrement Mr BRIKH.M l'archiviste de l'hôpital pour l'encouragement et les aides précieuses qu'il m'a fournis, j'ai beaucoup appris de lui.

Je remercie aussi tous les enseignants du département de Mathématiques qui ont contribué à ma formation.

Mes remerciements s'adressent aussi à la présidente du jury Mme TIMREDJINE.K, l'examinatrice Mme TABTI.H et l'examinateur BOURAINE.M pour l'immense honneur qu'ils m'ont fait en acceptant d'évaluer ce travail.

Enfin, mes profonds remerciements vont à mes chers parents, frères et sœurs, pour leur soutien et leur confiance en moi, sans oublier mes amis (es) et mes camarades du groupe, de tous les niveaux, et tous ceux qui ont contribué, de près ou de loin à ma formation.

Merci à tous

DÉDICACES

Je dédie ce modeste travail aux personnes les plus chères à mes yeux mes

parents.

A ces deux grands cœurs qui m'entourent toujours par leur tendresse et leur affection.

A ceux qui m'ont toujours encouragé et soutenu dans mes études et m'ont éclairé et

ouvert la vie de l'avenir.

Je vous dédie le fruit de mes efforts, comme un symbol de gratitude.

Que Dieu vous garde pour moi

Soyez toujours fières de moi.

A mes chères sœurs que j'aime plus que tout au monde.

A mes frères.

A mes chères nièces.

A toute ma famille.

A tous mes amis (es), pour leur amitié, leur soutien moral, et leur conseils

en particulier

A ma trés chère amie : Nouara .

en souvenir de nos éclats de rire et des bons moments

en souvenir de tout ce qu'on a vécu ensemble, j'éspère de tout mon cœur que notre amitié

durera éternellement.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

L'analyse des données de survie voit le jour au $XVII^e$ siècle, dans le domaine de la démographie. L'objectif des analystes de ce siècle est l'estimation, à partir des registres des décés, de diverses caractéristiques de la population : son effectif, sa longévité, etc. Ces analyses, très générales, ne sont affinées qu'à partir du XIX^e siècle, avec l'apparition de catégorisations suivant des **variables éxogènes** : sexe, nationalité, catégories socio-professionnelles,.... Durant ce siècle sont apparues également les premières modélisations de la probabilité de la mortalité par âge, probabilité qui sera par la suite désignée sous le terme de **fonction de risque**. Enfin, l'analyse des données de survie commence à déborder le cadre stricte de la démographie pour investir, au XX^e siècle, toutes les disciplines susceptibles d'avoir recours à de tels types de données : l'actuariat, la physique (avec l'apparition de la théorie de la fiabilité), l'industrie (pharmaceutique, biomédicale),...etc [**?**].

Jusqu'en 1950, la communauté des statisticiens s'intéresse peu à l'analyse des données de survie, la principale contribution étant celle de Greenwood (1926), qui propose une formule donnant la variance de survie. En 1951, Weibull conçoit un modèle paramétrique dans le domaine de la fiabilité ; à cet effet, il fournit une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de survie : la **loi de Weibull**.

En 1958, Kaplan et Meier présentent d'importants résultats concernant l'estimation non-paramétrique de la fonction de survie. Ils étudient l'espérance, la

variance et les propriétés asymptotiques de l'estimateur.

L'année 1972 se révèle être une date fondamentale : en effet, un modèle statistique semi-paramétrique voit le jour, grâce aux travaux de Cox. Ce modèle comporte des variables exogènes qui sont introduites, dans la fonction de risque, par le moyen d'une régression paramétrique, le reste de la fonction de risque, non-paramétrique, demeurant indéterminé.

De ce modèle, sans doute le plus utilisé en analyse des données de survie, seront tirées une quantité de variantes, et notamment des formulations permettant de stratifier l'effet des covariables, d'introduire une dépendance vis-à-vis du temps, ou encore de prendre en compte une possible interdépendance des durées de vie observées.

L'estimation des fonctions de survie et de risque est une partie très importante en analyse de survie pour déterminer la durée de vie d'une certaine population, mais cette dernière est exposée à certains facteurs qui peuvent l'améliorer ou l'agraver, et qui sont à prendre en compte afin de faire une analyse plus précise.

L'analyse des durées de vie est un domaine de la statistique qui étudie l'apparition d'un événement au cours du temps. De ce fait, il est nécessaire de disposer du temps de suivi de tous les individus ainsi que le moment auquel l'événement ce produit. Ce qui est particulier avec ce type d'étude, c'est la présence des données censurées (données incomplètes, phénomène courant dans les applications médicales) pour les sujets qui n'ont pas subi l'événement d'intérêt, la présence des variables explicatives et l'assymétrie de la distribution des données. L'analyse des durées de vie est donc une partie particulièrement utile pour étudier plusieurs types d'événements notamment des pannes d'équipement, de tremblement de terre, de divorce, et évidemment des décès.

L'analyse de survie réalise des applications en science actuarialle, démographie, épidimiologie, recherche médicale, analyse de fiabilité et beaucoup d'autres champs. Les exemples des durées de dérrangement inclus les vies des composants de machines en fiabilité industrielle, les durées des grèves ou périodes de chômage dans les sciences économiques. Dans la recherche médicale, si le point final est la mort d'un patient, les données résultant sont littéralement des vies. Cependant, des

données d'une forme semblable peuvent être obtenues quand le point final n'est pas mortel. Les exemples de vies dans la recherche clinique incluent le temps de début du traitement ou soulagement d'une douleur, et le temps de début de l'infection au début de la maladie.

Concernant les modèles statistiques proprement dits, trois approches d'estimation sont possibles : paramétrique, non-paramétrique et semi-paramétrique.

L'approche paramétrique stipule l'appartenance de la loi de probabilité réelle des observations à une classe particulière de lois, qui dépendent d'un certain nombre (fini) de paramètres. L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres, ainsi que l'obtention d'intervalles de confiance et la construction de tests. L'inconvénient de la méthode paramétrique est l'inadéquation pouvant exister entre le phénomène étudié et le modèle retenu.

L'approche non-paramétrique ne nécessite aucune hypothèse quant à la loi de probabilité réelle des observations, et c'est là son principal avantage. Il s'agit dès lors d'un problème d'estimation fonctionnelle, avec les ambiguïtés que cela implique par exemple, la fonction de survie, qui est continue, sera estimée par une fonction discontinue. L'inconvénient d'une telle approche est la nécessité de disposer d'un nombre important d'observations, le problème de l'estimation d'un paramètre fonctionnel étant délicat puisqu'il appartient à un espace de dimension infinie.

L'approche semi-paramétrique est une sorte de compromis entre les deux approches précédentes. La loi de probabilité réelle des observations est supposée appartenir à une classe de lois composée de deux partie : une partie paramétrique et une partie non paramétrique. Elle est apparue au cours des années soixante-dix, cette approche est très répandue en analyse de survie, notamment au travers du modèle de régression de Cox (1972).

L'objectif de ce travail est de présenter les données de survie ainsi que les différentes méthodes d'estimation de la fonction de survie, en les illustrant par des exemples d'applications. Nous avons considéré deux applications : une sur les données d'assurance retraite [36] et une autre sur les données du stage que nous avons effectué au sein de l'hôpitale d'Amizour, dans le service d'oncologie.

Ce mémoire est composé d'une introduction générale, de trois chapitres, d'une conclusion générale, et d'une liste de références bibliographiques.

Dans le premier chapitre, nous présentons les principaux concepts de survie. Dans le second chapitre, nous présentons les trois modèles d'estimation en particulier le modèle de Cox proportionnel. Le troisième chapitre, est consacré aux deux applications.

Nous terminons ce mémoire par une conclusion générale.

CHAPITRE 1

NOTIONS PRÉLIMINAIRES D'ANALYSE DE SURVIE

1.1 Introduction

Le terme de la durée de survie désigne le temps écoulé depuis un instant initial jusqu'à la survenue d'un événement précis, cette durée peut représenter (par exemple) [15] :

* Le temps de suivie, après le diagnostique, d'un cancer de sein ;
* L'âge d'entrée dans une schizophrénie (une maladie mentale dans laquelle le sujet perd le contacte avec la réalité) ;
* Durée de fonctionnement d'une machine ;
* Durée avant une récidive d'un ancien détenu en prison,....

On distingue pour chaqu'une de ces durées un événement qui lui est propre :

* Décès.
* L'entrée dans une schizophrénie.
* La panne de la machine.
* L'arréstation d'un ancien détenu de prison.

On s'intéresse à l'étude des durées de survie jusqu'à la survenence de l'événement d'intérêt. L'orsque l'événement d'intérêt est le décés, on parle d'une analyse de survie relative à l'être humain, contrairement à l'analyse de fiabilité qui s'intéresse au matériel.

Les données de survie ont une caractéristique principale donnée par la difficulté d'observer complètement tous les temps d'événements. Par exemple, quand l'événement étudié est le décés, la date de cet événement n'est pas observée pour les sujets toujours vivants à la fin de l'étude. Ces obsérvations sont incomplètes et on parle de "censure", car pour certaines de ces données, on ne connaît qu'une borne inférieure ou supérieure et non une valeur précise [1]. Il existe aussi un autre type de problème dit "troncature", nous reviendrons à toute ces notions dans le paragraphe 1.3.2.

Dans ce chapitre, nous donnons quelques notions de base sur l'analyse de survie ainsi que les expressions mathématiques des fonctions d'intérêts.

1.2 Domaines d'application

L'analyse de survie est appliquée dans plusieurs domaines, nous citons quelques uns [19] :

En médecine, l'analyse de survie permet d'évaluer l'efficacité d'un traitement.

En démographie, l'analyse de survie sert à construire des tables de mortalité (tables de survie). Celles-ci sont utilisées par les actuaires pour déterminer le montant des assurances-vie, entre autres ; on parle de tables actuarielles quand les données sont regroupées dans des intervalles.

En ingénierie, l'analyse de survie permet d'estimer la fiabilité de machines ou de composants électroniques, ...

On peut aussi appliquer l'analyse de survie, dans d'autres domaines tels que : marketing [11], finance [10], assurance [5],....

1.3 La nature des données de survie

Les temps de survie mesurés à partir d'une origine appropriée ont deux caractéristiques. La première est qu'ils sont non négatifs et tels qu'une hypothèse de normalité n'est généralement pas raisonnable en raison d'une asymétrie prononcée. La seconde est structurelle et tient au fait que pour certains individus l'événement étudié ne se produit pas pendant la période d'observation, et en conséquence certaines données sont censurées à droite et d'autres sont incomplètes, par exemple, dans les enquêtes épidémiologiques, les données sont souvent reccueillies de façon incomplète. Cette censure à droite est la plus courante mais n'est pas la seule censure que l'on peut rencontrer avec des données de survie [3].

1.3.1 Quelques définitions

- **Cohorte :** Est l'ensemble de sujets concernés par une étude et suivis au même moment, dans des conditions standarisées pendant une durée prédéfinie [20].
- **Evénement d'interêt :** Est l'événement auquel on s'intéresse au cours de l'étude. Par exemple, le décés qui peut être lié à un AVC (Accident Vasculaire Cérébral), complication, rechute, disparition de symptômes,...

- **Date d'origine :** Elle correspond à 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 (elle n'est pas importante car c'est la durée qui nous intéresse) [4].
- **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.
- **Date des dernières nouvelles :** c'est la date la plus récente où des informations sur un sujet ont été recueillies.
- **Durée de survie :** Est le délai entre la date d'origine et la date de survenue de l'événement d'interêt ou la date des dernières nouvelles.
- **Perdu de vue :** Un sujet est perdu de vue lorsque sa surveillance est interrompue avant la date de point et que l'événement d'intérêt ne s'est pas produit.
- **Exclu vivant :** Est un sujet suivi régulièrement et vivant à la date de point ou à la date des dernières nouvelles.
- **Temps de recul :** Délai entre la date d'origine et la date de point, c'est-à-dire le délai maximum potentiel de suivi pour un sujet. Les reculs minimum et maximum d'une série de sujets définissent donc l'ancienneté de cette série.
- **Temps de participation :** Durée de surveillance pour chaque sujet. Il est calculé suivant les trois cas possibles :
  - **Premier cas :** l'événement a lieu au cours de la surveillance, alors
    Temps de participation= Date de survenue de l'événement-Date d'origine.
  - **Deuxième cas :** le sujet est vivant à la date de point, alors
    Temps de participation= Date de point-Date d'origine.
  - **Troisième cas :** le sujet est perdu de vue, alors
    Temps de participation= Date de dernière nouvelle-Date d'origine.

1.3.2 Censure et troncature

Une des caractéristiques des données de survie est l'existence d'obsevations incomplètes. En effet, les données sont souvent receuillies partiellement, notamment, à cause des processus

de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accés à toute l'information : au lieu d'observer des réalisations indépendentes et identiquement distribuées (*i.i.d.*) de durée X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié [4].

1.3.2.1 Censure

Observons une étude portant sur la durée de rémission X de patients atteints de leucimie aigue (cancer des cellules de la moelle osseuse). Le protocole prévoit une analyse des résultats deux ans après la première inclusion (à la date du point). Au bout de deux ans la plus part des patients ont subi l'événement d'intérêt (fin de la rémission), l'information est complète. Un nombre négligeable de patients ne présenterons pas de rechute , pour de tels patients, X n'est pas connue, mais pas totalement inconnue puisqu'elles sont au moins supérieure à la date d'observations du malade, ainsi l'information portant sur X est dite censurée [2].
A partir de ce qui précède on dira qu'une durée de survie d'un individu est censurée lorsque l'événement d'intérêt n'a pas été observé. Elle concerne les sujets perdus de vues et ceux vivant à la date du point. La censure est le phénomène le plus couramment rencontré lors du recueil des données de survie.

Exemple 1.3.2.1.1.

Nous prenons cet exemple pour illustrer la censure et la différence entre les types de données (complète et censurées) [21]. On considère la durée de vie de patients atteints de cancers observées sur une durée de temps déterminée allant de la date d'origine 01/01/2000 *; à la date de point* 01/4/2000*, tels qu'elle est illustrée dans la figure suivante :*

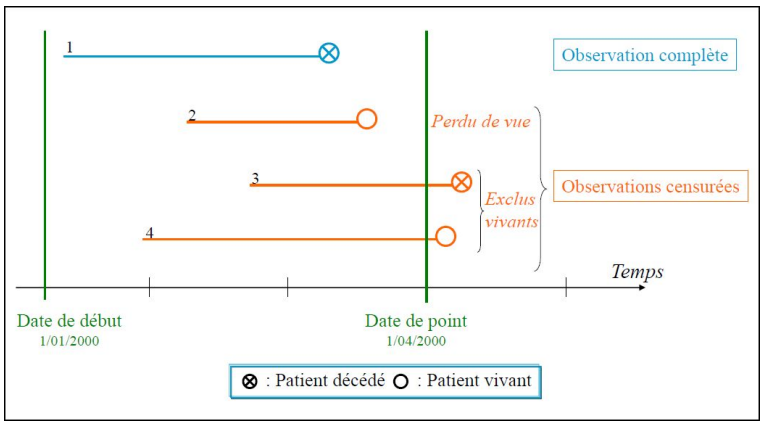

FIGURE 1.1 – l'illustration de la censure par 4 sujets

- *A la fin de l'étude, l'individu qui n'a pas subi l'événement d'intérêt (le décés), sa durée de vie, n'est pas exactement connue (observations censurée, patients 2, 3, et 4).*
- *Au cours de l'étude, le suivi de cetains patients peut-être intérompu pour plusieurs raisons indépendantes à l'étude. Le patient est perdu de vue et l'observation est censurée (patient 2).*
- *A la date de point, certains patients sont éxclus vivants, après un moment de l'exclusion, il est probable qu'ils subiront l'événement (patient 3) comme ils peuvent être toujours vivants (patient 4).*
- *Au cours de l'étude, certains patients subissent l'événement, cela est normale la donnée n'est pas censurée (patient 1).*

Remarque 1.3.1. *[3]*

Lorsque la censure est dûe à des conditions de l'expérimentation, elle est dite déterministe.

Considérons, pour l'individu i, i=1,...n.

— son temps de survie X_i,

— son temps de censure C_i,

— la durée réellement observée T_i.

1. **Censure à droite :**

   La durée de survie est dite censurée à droite si l'individu n'à pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas

toutes obsérvées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue. On considère trois types de censure à droite [4] :

La censure de type I :

Soit C une valeur fixée, au lieu d'observer toutes les variables $X_1, ..., X_n$, on se limite d'observer X_i, uniquement lorsque $X_i \leq C$, pour les autres, on aura $X_i > C$. On utilise la notation suivante :

$$T_i = X_i \wedge C = min(X_i, C) = \begin{cases} X_i, & \text{si la variable est observée} \\ C, & \text{sinon.} \end{cases}$$

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles et même souvent utilisé dans les études épidémiologie. Par exemple, tester la durée de vie de n objets identiques (ampoules) sur un intervalle d'observation fixé $[0, u]$.

La censure de type II :

Dans ce type de censure on décide d'observer les durées de survie des n patients jusqu'à ce que K d'entre eux soient décédés (le nombre d'événements est fixé) et d'arrêter l'étude à ce moment là, cela entraîne que la date de fin d'étude est aléatoire.
Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables X_i et T_i, i=1,...,n. La date de censure est donc X_K et on observe les variables suivantes :

$$\begin{cases} T_{(1)} = X_{(1)} \\ \vdots \\ T_{(k)} = X_{(k)} \\ T_{(k+1)} = X_{(k)} \\ \vdots \\ T_{(n)} = X_{(k)} \end{cases}$$

A partir du $K^{ème}$ événement la durée est dite censurée à droite de type II, alors

$$\begin{cases} T_{(k+1)} = X_{(k)} \\ \vdots \\ T_{(n)} = X_{(k)} \end{cases} \Leftrightarrow \begin{cases} X_i \geq X_k \\ \forall i = k, ..., n. \end{cases}$$

Ce modèle est souvent utilisé dans les études de fiabilité : l'observation de la durée de fonctionnement de n machines tant que K d'entre elles ne tombent pas en panne [26].

La censure de type III : (ou censure aléatoire de type I)

Soient $C_1, ..., C_n$, $n, (v.a.i.i.d)$ de temps de censure, et $X_1, ..., X_n$, $n, v.a.i.i.d$ de temps de survie. On observe les variables T_i tel que $X_i \wedge C_i$.

Dans ce cas, l'information disponible peut être résumée par :

— la durée réellement observée T_i,

— un indicateur $\delta_i = \mathbb{1}_{X_i \leq C_i}$.

Où, $\mathbb{1}_A$ est la fonction indicatrice sur l'ensemble A. Alors,

$\delta_i = 1$, si l'événement est obsérvé, (dans ce cas, $T_i = X_i$. On observe les durées complètes).

$\delta_i = 0$, si l'individu est censuré, (dans ce cas, $T_i = C_i$. On observe des durées incomplètes).

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

(a) La perte de vue : le patient quitte l'étude en cours et on ne le revoit plus (suite à un déménagement ou le patient décide de se faire soigner ailleurs,...). Ce sont des patients "perdus de vue".

(b) L'arrêt ou le changement de traitement à cause des effets secondaires ou l'inefficacité du traitement. Ces patients sont exclus de l'étude.

(c) La fin de l'étude : l'étude se termine, alors que, certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients "exclus-vivants". Les "perdu de vue" (et les exclusions) et les "exclus-vivants" correspondent à des observations censurées mais les deux mécanismes sont de natures différentes.

Exemple 1.3.2.1.2. *[3]*

Considérons une étude relative à la durée de survie de patients soumis à un traitement particulier. L'événement d'intérêt est la mort de la personne. Tous les individus sont suivis pendant les 52 *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.*

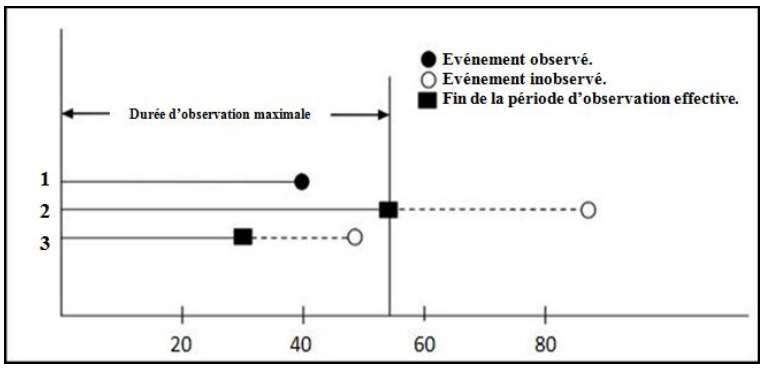

FIGURE 1.2 – Exemple de deux types de censures.

— *L'individu* 1 *est décédé* 40 *semaines après le début de traitement. Il s'agit d'une observation non censurée.*

— *La deuxième personne est toujours vivante au terme des* 52 *semaines d'observation. Elle décèdera après* 90 *semaines mais cette information n'est pas connue lorsque la constitution de la base de données est arrêtée. Même si l'information est incomplète, elle est utile puisque l'on sait que le temps de survie réel est supérieur à* 52 *semaines. Il ne faut donc pas l'éliminer de la base sous peine de biaiser vers le bas l'estimation de la durée moyenne de survie. Il s'agit d'une censure déterministe car elle ne dépend pas de l'individu considéré mais des conditions de l'expérimentation.*

— *La troisième personne décède après* 50 *semaines mais cet évènement n'est pas enregistré dans la base de données car le patient concerné n'a pu être effectivement suivi que pendant* 30 *semaines. C'est un exemple de censure aléatoire car elle échappe au contrôle de l'expérimentateur. Là encore l'information est incomplète mais non nulle. Par exemple, savoir que cet individu a survécu au moins* 30 *semaines est pertinent pour l'estimation du taux de survie à* 20 *semaines.*

Remarque 1.3.2. *[4]*

La censure causée par un déménagement ou par la fin de l'étude, est indépendante du temps de survie car elle n'apporte pas d'informations sur le temps de survie (censure non informative). Par contre, si la censure est dûe à un arrêt de traitement ou l'arrêt du suivi des patients les plus malades. Cette censure apporte de l'information sur le temps de survie (censure informative), d'où la dépendance.

Toute au long de ce document on considèrera que le délai de censure C est une v.a.réelle indépendante du temps de survie X .

2. **Censure à gauche :**

   Considérons, X la durée de survie des patients depuit l'infection par le VIH virus responsable du sida, (événement initial) jusqu'à la survenue du décés (événement terminal). Cette durée est calculée entre deux dates connues, mais ce n'est pas le cas pour tous les patients. Pour certains d'entre eux on ignore la date de l'événement initial. Cela nous amène à dire que les données sont incomplètes [27]. La censure à gauche correspond à l'individu qui a déjà subi l'événement avant qu'il soit obsérvé. On sait dans ce cas, que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu i, $i = 1, ..., n$ on peut associer un couple de variables aléatoires (T_i, δ_i) telque :

$$T_i = X_i \vee C_i = max(X_i, C_i) = \begin{cases} C_i, & si \quad X_i > C_i \\ X_i, & sinon \end{cases}$$

   et

$$\delta_i = \mathbb{1}_{X_i \geq C_i} = \begin{cases} 1 & si \quad X_i \leq C_i \\ 0 & si \quad X_i > C_i \end{cases}$$

3. **Censure par intervalle :**

   Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est que l'événement a eu lieu entre deux dates connues. C'est-à-dire au lieu d'observer X, on observe $C_i^L < C_i^R$ tels que $C_i^L < X < C_i^R$ (X est non observé). On dit dans ce cas, qu'il y a censure par intervalle[7]. En particulier, la censure à gauche peut être considérée comme une censure par un intervalle avec $C_i^L = -\infty$, et la censure à droite est une censure par un intervalle avec $C_i^R = +\infty$. On observe le couple (T_i, δ_i), $i = 1, ..., n$ où, $T_i = max[min(X_i, C_i^R), C_i^L]$, tel que C_i^R est une censure à droite, et C_i^L est une censure à gauche.

avec,

$$\delta_i = \begin{cases} 1, & si\ X_i = T_i \quad \text{(T observé)} \\ 0, & si\ X \leq C_i^L \quad \text{(Censure à gauche)} \\ -1, & si\ C_i^R \leq X \quad \text{(Censure à droite)} \end{cases}$$

Remarque 1.3.3.

(a) Ce type de censure et dit aussi censure mixte.

(b) Dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (non continu),on sait alors uniquement que l'événement s'est produit entre deux temps d'observations : la date de dernière visite et la date de la prochaine visite. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à la censure à droite [3].

(c) Cette censure se produit notamment si un patient se rend à l'hôpital à des dates régulières : s'il ne se présente pas à un rendez-vous, on sait seulement que son décès s'est produit dans l'intervalle entre la dernière visite et le rendez-vous [19].

(d) Dans le cas où l'incorporation d'un patient dans une étude est conditionnée par un évenement initial (par exemple, par la date associée à l'apparition de certains symptômes), il est alors fréquent de ne connaitre qu'un minorant de la date de ce premier événement. La durée de survie, prise comme la différence entre le temps où se produit l'événement d'intérêt et le temps où s'est produit l'événement initial, est ici supérieure ou égale à la durée effectivement observée par le statisticien. Les deux cas peuvent être combinés. on dispose ici de deux censures C_i^1 et C_i^2, où C_i^1 représente la durée censurée par rapport à l'événement initial (censure à droite) et C_i^2 représente la durée censurée par rapport à l'événement terminale (censure à gauche), avec $C_i^1 < C_i^2$. On observe alors le triplet (T, δ_1, δ_2) avec [28], où :

$\delta_i = \mathbb{1}_{C_i^2 \leq X \leq C_i^1}$

en ce cas :

$$T = \begin{cases} C_i^1, & si\ X \leq C_i^1 \quad \textit{(censure à gauche)} \\ X, & si\ C_i^1 < X < C_i^2 \quad \textit{(pas de censure)} \\ C_i^2, & si\ C_i^2 \leq X \quad \textit{(Censure à droite)} \end{cases}$$

La figure suivante illustre les trois types de censures :

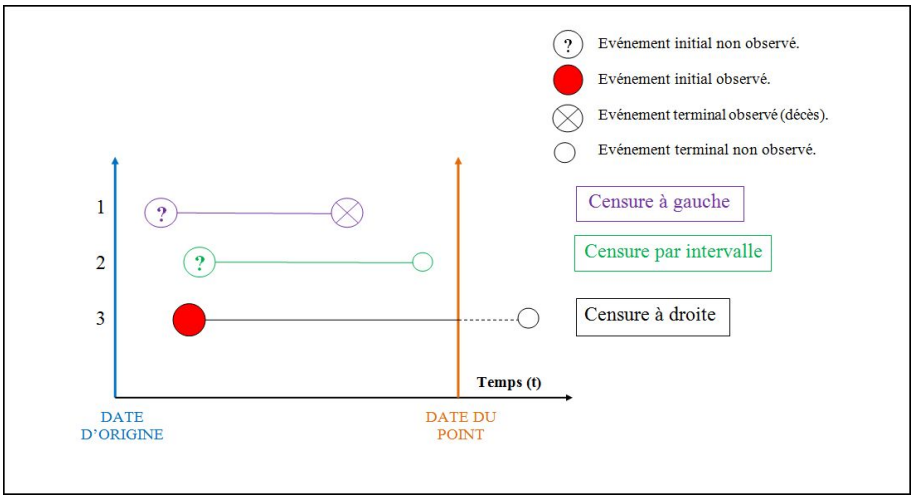

FIGURE 1.3 – Les trois types de censures.

1.3.2.2 Troncature

Les troncatures sont différentes des censures au sens où elles concernent l'échantillonnage lui même [4]. Ainsi, la variable X est tronquée par un sous ensemble éventuellement aléatoire A de $\mathbb{R}_+$, on observe X uniquement si $X \in A$. Les points de l'echantillion "tronqué" appartiennent tous à A, et suivent donc la loi de T conditionnée par l'appartenance à A. S'il y a troncature, une partie des individus (donc des X_i) ne sont pas obsevables et on n'étudie qu'un sous-échantillon (problème d'échantillonage).

1. **La troncature à gauche :**

   Soit Z une variable aléatoire de troncature à gauche indépendante de X, On dit qu'il y a troncature gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à Z. On observe le couple (X, Z), avec $X > Z$.

   Remarque 1.3.4.

   (a) Par exemple, si la durée de vie d'une population est étudiée à partir d'une cohorte tirée au sort dans une population, seule la survie des sujets vivants à l'inclusion pourra être étudiée (il y a troncature à gauche car seuls les sujets ayant survécus jusqu'à la date d'inclusion dans la cohorte sont obsevables).

   (b) Attention, il ne faut pas confondre la censure à gauche avec la troncature à gauche. S'il y a troncature, un certain nombre d'individus ne sont pas observables et

on n'étudie qu'un sous-échantillon ; tandis que s'il y a censure à gauche pour certains individus l'information est incomplète bien qu'ils soient présents dans l'échantillon.

2. **La troncature à droite :**

   De façon similaire à la troncature à gauche on peut définir la troncature à droite : X est tronquée à droite si elle n'est observable qu'à la condition $X < Z$.

   Exemple 1.3.2.2.1.

   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 $(Y + T)$ 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 1.5.

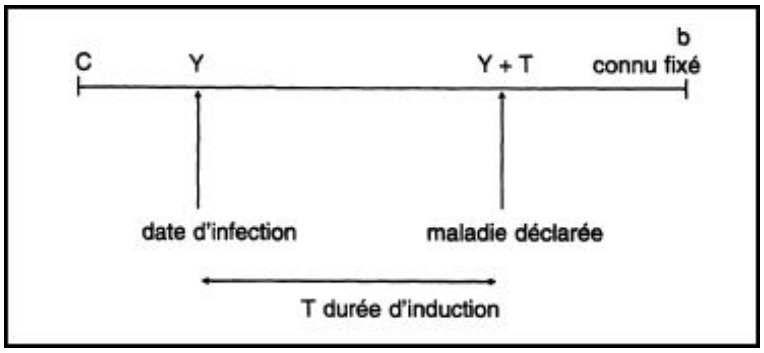

FIGURE 1.4 – Schéma correspondant à la trancature

3. **La troncature par intervalle :**

   Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.

La figure suivante schématise les différentes types de données dans l'analyse de survie :

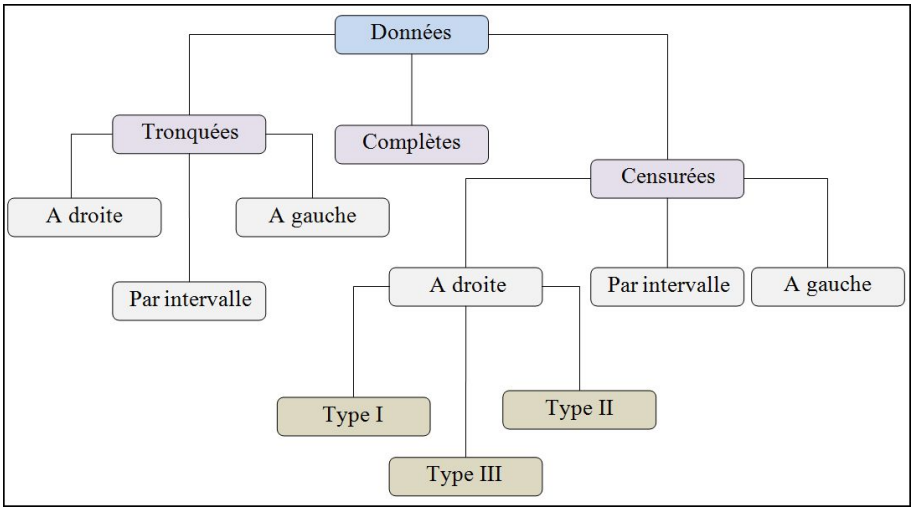

FIGURE 1.5 – Les différent types de données

1.4 Distribution de la durée de survie

Supposons que la durée de survie est représentée par une variable positive ou nulle notée X, et absolument continue. Sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions) [4].

Fonction de répartition F

La fonction de répartition (ou c.d.f. "cumulative distribution function") représente la probabilité de subir l'événement sur la durée $[0, t]$, c'est-à-dire :

$$F(t) = \mathbb{P}(X \leq t), \quad t > 0.$$

Pour t fixé, elle est aussi décrite comme une probabilité de mourir avant l'instant t.

Propriété 1.4.1. *[1]*

1. *F est une fonction croissante en t ;*
2. $F(0) = 0$;

3. $\lim_{t\to+\infty} F(t) = 1$.

Densité de probabilité f

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$

$$F(t) = \int_0^t f(u)du.$$

Si la fonction de répartition F admet une dérivée au point t alors,

$$f(t) = F'(t) = \lim_{h\to 0}\left[\frac{F(t+h) - F(t)}{h}\right] = \lim_{h\to 0}\frac{\mathbb{P}(t \leq X \leq t+h)}{h}.$$

Cela implique que f peut s'écrire aussi :

$$f(t)h \simeq \mathbb{P}(t \leq X \leq t+h), \quad h \text{ petit}$$

Pour t fixé, la densité de probabilité représente "la probabilité de mourir à l'instant t".

1.4.1 Fonction de survie S

La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire :

$$S(t) = \mathbb{P}(X > t), \quad t \geq 0.$$

La fonction de survie est la probabilité que l'individu survive entre l'instant de la découverte de la maladie (t=0) jusqu'à la survenu de l'événement d'intérêt.
C'est une fonction monotone décroissante continue du temps et tend vers 0.
En général la fonction de survie a la forme suivante :

Propriété 1.4.2. *[29]*

1. $S(t) = \mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^{+\infty} f(u)du \quad t > 0$;

2. $f(t) = F'(t) = -S'(t)$;

3. $S(o) = 1$;

4. $\lim_{t\to+\infty} S(t) = 0$;

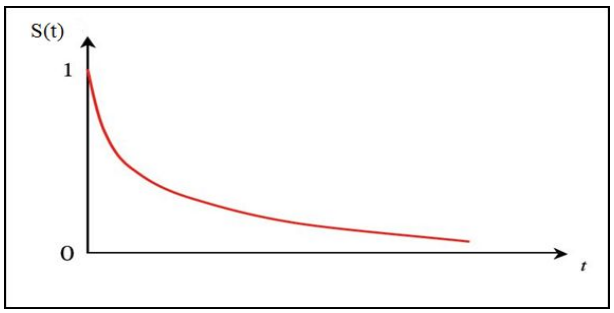

FIGURE 1.6 – Fonction de survie

5. $\mathbb{P}(t_1 < X < t_2) = F(t_2) - F(t_1) = S(t_1) - S(t_2), 0 < t_1 < t_2$

Remarque 1.4.1.

Il est arbitraire de décider que $S(t) = \mathbb{P}(X \geq t)$ *ou* $S(t) = \mathbb{P}(X > t)$*, cela n'a aucune importance quand la loi de* X *est continue car* $\mathbb{P}(X > t) = 0$*. Dans les cas où* F *a des sauts (le temps est discret, par exemple, compté en mois ou en semaine), on utilise les notations suivantes [4] :*

$$F^-(t) = \mathbb{P}(X < t) \text{ et } F^+(t) = \mathbb{P}(X \leq t)$$

où F^- *est la limite à gauche et* F^+ *la limite à droite de* F *(définitions et notations sont identiques pour la fonction* S*). Remarquons que* $F^- \leq F^+$ *et* $S^- \geq S^+$*.*

1.4.2 Risque instantané h (ou taux de hasard)

Le risque instantané (ou taux d'incidence), noté h(t) pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour les survivants). Il est alors :

$$h(t) = \lim_{dt \to 0} \left[\frac{\mathbb{P}(t \leq X \leq t + dt \mid X > t)}{dt} \right].$$

En utilisant le fait que : $f(t)dt \approx \mathbb{P}(t < X \leq t + dt)$.
On obtient : $h(t)dt \approx \mathbb{P}(t < X \leq t + dt | X > t)$.

— **Lien avec la survie**

$$\begin{aligned} h(t) &= \lim_{dt\to 0}\left[\frac{1}{dt}\mathbb{P}(t < X \leq t+dt | X > t)\right] \\ &= \lim_{dt\to 0}\left[\frac{1}{dt}\left(\frac{\mathbb{P}(t < X \leq t+dt)}{\mathbb{P}(X > t)}\right)\right] \\ &= \frac{1}{\mathbb{P}(X > t)}\lim_{dt\to 0}\left[\frac{1}{dt}\mathbb{P}(t < X \leq t+dt)\right]. \end{aligned}$$

D'où

$$h(t) = \frac{f(t)}{\mathbb{P}(X > t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{(1-F(t))} = \frac{-S'(t)}{S(t)}.$$

1.4.3 Les formes de taux de hasard

On analyse des durées de vie, on a cinq formes les plus usuelles de taux de hasard qui sont : Constant, croissant, décroissant, en cloche, en forme de baignoire [29].

★ **Taux de hasard constant :** indique que le taux est indépendant du temps c'est le cas où la mortalité est principalement dûe aux accidents (modèle exponentiel).

★ **Taux de hasard croissant :** un risque croissant augmente avec l'âge, est typiquement observé chez les adultes (loi gamma, loi weibull :à deux paramètres où le paramètre de forme > 1).

★ **Taux de hasard décroissant :** déminuation du taux de hasard chez les enfants de moins d'un an et augmente avec l'âge (la weibull :à deux paramètre avec le paramètre de forme $1 <$).

★ **Taux de hasard de forme cloche $\bigcap$:** un mélange de taux de hasard croissant puis décroissant (loi log-normale).

★ **Taux de hasard de forme en baignoire ou $\bigcup$:** un mélange de taux de hasard décroissant puis croissant (modèle Weibull à trois paramètre nomé weibull généralisé).

1.4.4 Les trois principales phases de survie

Referant à la fiabilité, la courbe en baignoire montre, l'évolution du taux de hasard d'un être humain né en bonne santé pendant toute sa durée de vie. Elle comprend trois phases,

chaqu'une avec un sens de variation différent (Voir figure1.8) .

— **Phase 1 :** Période d'enfance commence depuit la naissance où il se mis à s'adapter à la vie extra-utérine, durant laquelle il pourrait être particulièrement éxposé à des pathologies comme l'anoxie (manque d'oxygènation, du cerveau notamment), en cette période aussi il se mis à faire face aux multiples agressions infectieuses. Elle est caractérisée par un taux de hasard décroissant [Encyclopedie medicale].

— **Phase 2 :** Correspond à la vie active et vitale, pleine d'énergie. Elle est caractérisée par un taux de hasard constant.

— **Phase 3 :** Correspond à la période de vieillesse où l'état de santé se dégrade à cause habituellement des effets de vieillessement et des effets additifs de maladies passées actuelles chroniques ou aigues. Elle est caractérisée par un taux de hasard croissant.

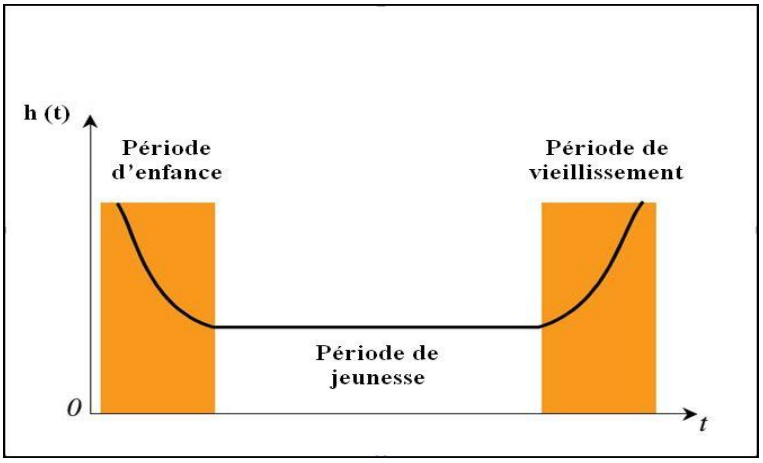

FIGURE 1.7 – Courbe en baignoire.

1.4.5 Fonction de hasard cumulé H

La fonction de hasard ou le taux de hasard cumulé est l'intégrale du risque instantané $h(t)$[4] :

$$H(t) = \int_0^t h(u)du = -ln(S(t)).$$

On peut déduire de cette formule une expression de la fonction de survie en fonction du taux de hasard cumulé (ou de risque instantané) :

$$S(t) = exp(-H(t)) = exp\left(-\int_0^t h(u)du\right), \quad t > 0.$$

D'où,

$$F(t) = 1 - exp\left(-\int_0^t h(u)du\right), \quad t > 0.$$

$$f(t) = h(t)exp\left(-\int_0^t h(u)du\right), \quad t > 0.$$

— **Fonction de survie résiduelle :**

La survie résiduelle correspond à la probabilité que l'individu ne soit pas décédé, sachant qu'à l'instant t_0, il était encore vivant[17].

$$\begin{aligned} S(t|t_0) &= \mathbb{P}(X > t + t_0 | X > t_0) \\ &= \frac{\mathbb{P}(X > t + t_0) \bigcap \mathbb{P}(X > t_0)}{\mathbb{P}(X > t_0)} \\ &= \frac{\mathbb{P}(X > t + t_0)}{\mathbb{P}(X > t_0)} \\ &= \frac{S(t + t_0)}{S(t_0)}. \\ &= exp\left(-\int_{t_0}^t h(u)du\right), \ \forall \ \ t > 0. \end{aligned}$$

— **Espérance de la durée de vie restante** :

Si, à la date t, l'individu est encore dans un état donné et si X désigne sa date de sortie de cet état, sa durée de vie résiduelle est : $X - t$. L'espérance de cette variable est obtenu e.i

$$\begin{aligned} r(t) &= \mathbb{E}(X - t \backslash X > t) \\ &= \frac{1}{S(t)} \left\{ \int_t^{+\infty} uf(u)du - tS(t) \right\} \\ &= \frac{1}{S(t)} \left\{ [-uS(u)]_t^{+\infty} + \int_t^{+\infty} uS(u)du - tS(t) \right\}. \end{aligned}$$

Si $\lim_{u \to +\infty} uS(u) = 0$, alors : $r(t) = \frac{1}{S(t)} \int_t^{+\infty} uS(u)$.

1.4.6 Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbb{E}(X)$ est :

$$\mathbb{E}(X) = \int_0^\infty tf(t)dt.$$

avec, f(t)= -dS/dt. En intégrant par parties, on obtient

$$\mathbb{E}(X) = [-tS(t)]_0^\infty + \int_0^\infty S(t)dt = \int_0^\infty S(t)dt.$$

de la même manière on trouve la variance,

$$\mathbb{V}(X) = 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2.$$

Ainsi on peut écrire l'espérance et la variance en fonction de n'importe laquelle des fonctions F, S, f, h, et H (mais pas l'inverse).

1.5 Fonction de vraisemblance

La vraisemblance est un outil fondamental pour l'inférence statistique. Rappelons qu'un modèle statisqtique $(\Omega, A, \mathbb{P})$ est tel que (Ω, A) est un espace mesurable, $\mathbb{P} = \{\mathbb{P}_\theta, \theta \subset \Theta\}$ est une famille de lois de probabiltés dépendant d'un paramètre θ à valeurs dans Θ(l'espace des paramètres). Pour écrire la vraisemblance, on doit prendre en compte les différents types de données [1] :

1.5.1 Fonction de vraisemblance en cas des données complètes

Soit $(T_1, ..., T_n)$ un échantillon de durées de survie réellement observées indépendantes identiquement distribuées (i.i.d) de densité f_θ. On appelle vraisemblance du modèle, l'application $L : \Theta \to \mathbb{R}^+$ définie par :

$$L(t, \theta) = \prod_{i=1}^{n} f_\theta(t_i).$$

1.5.2 Fonction de vraisemblance en cas des donnée censurées

La vraisemblance en cas des données censurées prend des expressions différentes selon les trois types de données [30] :

A) Dans la cas de la censure à droite :

1. **La censure de type I :**

   Soit un échantillion de durées de survie $(X_1, ..., X_n)$ et $C > 0$ fixé, la variable censurée ; la vraisemblance du modèle associé aux observations $(T_1, \delta_1), ..., (T_n, \delta_n)$ tels que,

$$T_i = X_i \wedge C \text{ et } \delta_i = \begin{cases} 1 & si \quad X_i \leq C \\ 0 & si \quad X_i > C. \end{cases}$$

   s'écrit comme produit d'une composante continue et une autre discrète [30] :

$$L(t, \theta) = \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(C)^{1-\delta_i}.$$

   lorsqu'on observe l'événement d'intérêt avant la censure, c'est le terme de densité qui intervient dans la vraisemblance, et dans le cas contraire on retrouve le terme de la fonction de survie à la date de censure. Pour obtenir cette formule, il suffit de calculer $\mathbb{P}_\theta(T_i \in [t, t + dt], d_i = \delta_i)$. On calcule sur $[0, C]$:

   a) La probabilité d'une durée non censuré :

$$\begin{aligned} \mathbb{P}_\theta(T_i \in [t, t + dt], \delta_i = 1) &= \mathbb{P}_\theta(X_i \wedge C \in [t, t + dt], X_i \leq C) \\ &= \mathbb{P}_\theta(X_i \in [t, t + dt]) \\ &\simeq f_\theta(t)dt. \end{aligned}$$

   (On peut toujours supposer dt suffisamment petit pour que $t + dt \leq C$)

b) La probabilité d'une durée censuré :

$$\begin{aligned}\mathbb{P}_\theta(T_i \in [t_i, t_i + dt], \delta_i = 0) &= \mathbb{P}_\theta(X_i \wedge C \in [t_i, t_i + dt], X_i \geq C) \\ &= \mathbb{P}_\theta(X_i \geq C) \\ &= S_\theta(C), \quad C \in [t, t + dt].\end{aligned}$$

De a) et b) on trouve :

$$\mathbb{P}_\theta(T_i \in [t, t + dt], d_i = \delta_i) = f_\theta(t)^{\delta_i} S_\theta(C)^{1-\delta_i}.$$

Pour une observation censurée, par définition $T_i = C$, et pour une observation non censurée $T_i = X_i$, l'expression ci dessus peut se réecrire :

$$L(t, \theta) = \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(t_i)^{1-\delta_i}.$$

On peut écrire la densité en fonction de la fonction de hasard et de la fonction de survie $f_\theta(t) = h_\theta(t)S_\theta(t)$. Dans ce cas, la vraisemblance s'ecrit :

$$L(t, \theta) = \prod_{i=1}^{n} S_\theta(t_i) h_\theta(t_i)^{\delta_i}.$$

2. **La censure à droite type II** :

   Soit un échantillon de durées de survie $(X_1, ..., X_n)$ et $r > 0$ fixe. On dit qu'il y a censure de type II pour cet echantillon si au lieu d'observer directement $(X_1, ..., X_n)$ on observe $(T_1, \delta_1), ..., (T_n, \delta_n)$.

   La censure ici est fixé avec $C = X_{(r)}$, r fixé. Alors, la vraisemblance a une forme proche du cas de la censure de type I ; on remarque pour l'écrire que, dans la partie discrète de la distribution, il convient de choisir les instants des r sorties parmi les n observations. Alors,

$$L(t, \theta) = \frac{n!}{r!(n-r)!} \prod_{i=1}^{r} f_\theta(t_i) \times S_\theta(t_r)^{n-r}.$$

On peut l'écrire de cette manière :

$$L(t,\theta) = \frac{n!}{r!(n-r)!}\prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(t_i)^{1-\delta_i}.$$

on peut aussi l'écrire en fonction de la fonction de risque :

$$L(t,\theta) = \frac{n!}{r!(n-r)!}\prod_{i=1}^{n} h(t_i)^{\delta_i} S_\theta(t_i), \quad \text{où,} \quad \delta_i = \begin{cases} 1 & si \quad X_i \leq C \\ 0 & si \quad X_i > C. \end{cases}$$

3. **La censure à droite type III :**

Soient un échantillon de durées de survie $(X_1, ..., X_n)$ et un second échantillon de données censurées $(C_1, ..., C_n)$ composé de variables positives indépendantes identiquements distribuées. On dit qu'il y a censure de type III pour cet échantillon si au lieu d'observer directement $(X_1, ..., X_n)$ on observe $(T_i, \delta_i), ..., (T_n, \delta_n)$.

Supposons que les variables X et C ont pour densités respectives f et g et pour fonctions de survies S et G. La vraisemblance de l'échantillon peut être écrite :

$L(t,\theta) = \prod_{i=1}^{n} [f_\theta(t_i)G_\theta(t_i)]^{\delta_i} [g_\theta(t_i)S_\theta(t_i)]^{1-\delta_i}$, où, $\delta_i = \begin{cases} 1 & si \quad X_i \leq C_i \\ 0 & si \quad X_i > C_i. \end{cases}$

a) probabilité d'une durée non censurée est :

$$\begin{aligned} \mathbb{P}_\theta(T_i \in [t, t+dt], \delta_i = 1) &= \mathbb{P}_\theta(X_i \wedge C_i \in [t, t+dt], X_i \leq C_i) \\ &= \mathbb{P}(X_i \in [t, t+dt], t_i \leq C_i) \simeq f_\theta(t)G_\theta(t)dt. \end{aligned}$$

b) probabilité d'une durée censurée est :

$$\begin{aligned} \mathbb{P}_\theta(T_i \in [t, t+dt], \delta_i = 0) &= \mathbb{P}_\theta(X_i \wedge C_i \in [t, t+dt], X_i > C_i) \\ &= \mathbb{P}_\theta(C_i \in [t, t+dt], X_i > t) \simeq S_\theta(t)g_\theta(t)dt. \end{aligned}$$

Alors, de a) et b) on écrire :

$$L(t,\theta) = \prod_{i=1}^{n} [f_\theta(t_i)G_\theta(t_i)]^{\delta_i} [S_\theta(t_i)g_\theta(t_i)]^{1-\delta_i} \quad \delta_i = \begin{cases} 1 & si \quad X_i \leq C_i \\ 0 & si \quad X_i > C_i. \end{cases}.$$

On fait alors l'hypothèse que la censure est non informative, c'est à dire que la loi

de censure est indépendante du paramètre θ. La vraisemblance se met dans ce cas sous la forme :

$$\begin{aligned} L(t,\theta) &= \prod_{i=1}^{n}[f_\theta(t_i)G_\theta(t_i)]^{\delta_i}[g_\theta(t_i)S_\theta(t_i)]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(t_i)^{1-\delta_i}. \end{aligned}$$

où, $\propto$ est un indice de proportionnalité.

B) Dans la cas de la censure à gauche :

Soit un échantillon de durée de survie $X_1, ..., X_n$, de densité f, et de fonction de répartition F et soit $C_1, ..., C_n$ un échantillon de durée censurées, de densité g et de fonction de répartition G. On observe $(T_1, \delta_1), ..., (T_n, \delta_n)$. De la même manière on calcule les probabilité suivantes :

a) Probablité d'une durée non censurée :

$$\begin{aligned} \mathbb{P}_\theta(T_i \in [t, t+dt], \delta_i = 1) &= \mathbb{P}_\theta(X_i \in [t, t+dt], X_i > C_i) \\ &\simeq f_\theta(t)G_\theta(t)dt \end{aligned}$$

b) Probabilité d'une durée censurée :

$$\begin{aligned} \mathbb{P}_\theta(T_i \in [t, t+dt], \delta_i = 0) &= \mathbb{P}_\theta(C_i \in [t, t+dt], X_i < C_i) \\ &\simeq g_\theta(t)F_\theta(t)dt. \end{aligned}$$

D'où la vraisemblance :

$$\begin{aligned} L(t,\theta) &= \prod_{i=1}^{n}[f_\theta(t_i)G_\theta(t_i)]^{\delta_i}[g_\theta(t_i)F_\theta(t_i)]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} F_\theta(t_i)^{1-\delta_i} \end{aligned}$$

C) Dans le cas de censures par intérvalle :

La contribution à la vraisemblance d'un sujet i dans le délai est compris entre t_{1i} et t_{2i}

la vraisemblance est donnée par[1] :

$$L(\theta) = S(t_{1i}) - S(t_{2i}).$$

1.5.3 Fonction de vraisemblance en cas des données tranquées :

le cas de troncature à gauche

Considérons $(X_1, ..., X_n)$ un échantillon (i.i.d) de durées de survie de densité f et de fonction de survie S, et $(Z_1, ..., Z_n)$ un échantillon de donnée tronquée de densité g. Les variables Z_i sont supposées indépendantes des X_i. La condition s'exprime comme une probabilité de subir l'événement en t_i sachant que t_i est supérieure à Z_i. La vraisemblance est [1] :

$$L(t, \theta) = \prod_{i=1}^{n} \frac{f_\theta(t_i)}{S_\theta(t_i)}$$

1.6 Conclusion

Dans ce chapitre, nous avons d'abord présenté les notions élémentaires de la théorie de survie, et nous avons donné des exemples de censures et de troncatures, et enfin nous avons donné l'écriture de la vraisemblance pour les différents types de censures et troncatures.

CHAPITRE 2

MÉTHODES D'ESTIMATION EN ANALYSE DE SURVIE

2.1 Introduction

Afin d'éstimer les durées de survies, trois types de modèles sont possibles : paramétrique, non paramétrique, ou semi-paramétrique.

Dans le modèle paramétrique, les temps de survie sont supposés être distribués selon une loi de forme parfaitement connue. Les distributions les plus couremment utilisées sont : Exponentielle, Weibull, Gompertz, Log-normale, Gamma, etc.

Le modèle non paramétrique ne necessite aucune hypothèse quant à la forme de loi de probabilité réelle des observations et c'est là son principal avantage. On distingue deux modèles d'estimation non-paramètrique (Méthode de Kaplan-Meier et Méthode Actuarialle).

La méthode de Kaplan-Meier repose sur des intervalles determinés par les dates d'événements, et la méthode actuarialle utilise des intervalles de temps fixés à priori.

Le modèle semi-paramétrique se situe entre les deux modèles précédents. On l'utilise lorsque la famille de lois à laquelle appartient la loi de la variable de durée T n'est pas totalement spécifiée, on cherche plus à évaluer l'effet des variables exogènes sur la variable T. Ce modèle est très répondu en analyse de survie, notament au travers le modèle de régression de Cox (1972) où la distribution exacte du risque ou de la survie n'est jamais connue même si les coefficients du modèle ont pu être estimés à partir d'un ensemble de données [17]. De Même, le risque de base et les fonctions de survie de base ne sont pas spécifiés. L'objectif est d'évaluer l'effet des covariables sur la durée de vie.

Parmi les problèmes rencontrés dans les situations non paramétrique et semi-paramétrique, le résultat de la courbe estimée peut être différent. En effet, [31] :

- S'il y a des données à disposition, la "courbe" aura plutôt une forme en escalier.
- S'il reste des sujets à risques de subir l'événement en fin d'études, alors la courbe de survie n'atteindra pas son minimum de zéro.

  Ces problèmes ne sont pas rencontrés dans les modèles paramétriques.

Dans ce chapitre nous énoncerons les trois modèles utilisés en analyse de survie, nous donnont le principe de chaque méthode d'estimation.

2.2 Estimation paramétrique

Soit T la $v.a$ représentant la durée de survie et supposons que la distribution de T appartient à une famille de lois paramétriques donnée. Ainsi, le modèle paramétrique peut être formulé en précisant la forme de l'une ou l'autre des cinq fonctions équivalentes qui définissent la loi de la durée : H, h, f, S ou F. Les estimateurs des paramètres du modèle sont ensuite obtenus en maximisant la vraisemblance des observations.

2.2.1 Loi exponentielle

La loi exponentielle est utilisée pour modéliser la survie au sens large (être humain, appareils, pièces mécaniques,...) [18]. Elle modélise le temps d'attente entre deux événements qui surviennent indépendamment de façon purement aléatoire. Elle dépend d'un paramètre $\lambda > 0$, et admet un risque instantané constant égale à λ [31].

- Fonction de densité :

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0.$$

- Fonction de survie :

$$S(t) = e^{-\lambda t}, \quad t \geq 0.$$

- Taux de hasard :

$$h(t) = \lambda = cste, \quad \forall\, t > 0.$$

- Espérance :

$$\mathbb{E}(T) = \frac{1}{\lambda}.$$

- Variance :

$$\mathbb{V}(T) = \frac{1}{\lambda^2}.$$

Remarque 2.2.1.1.

1. *Plus le taux de hasard h est grand, plus l'espérance de survie est faible.*
2. *Soit deux sous populations avec des taux de hasard constants* $h_1 = \lambda$, *et* $h_2 = u$, $\forall t$ *et d'une probabilité* w, *et* $1 - w$ *respectivement,* $w \in [0, 1]$,
   alors $\lambda(t) = \frac{u+w(\lambda e^{-(\lambda+u)}-u)}{1-w(1+e^{(\lambda+u)}-u)} \neq cst$ *le taux de hasard du mélange est non constant.*

	Population 1	Population 2	Mélange
Proportion	w	$1-w$	1
Densité	$\lambda e^{-\lambda t}$	ue^{-ut}	$w\lambda e^{-\lambda t} + u(1-w)e^{-ut}$
Survie	$e^{-\lambda t}$	e^{-ut}	$(1-w)e^{-ut} + we^{-\lambda t}$
Risque instantanée	λ	0	$\lambda(t) = \frac{u+w(\lambda e^{-(\lambda-u)}-u)}{1-w(1+e^{(\lambda-u)}-u)}$
Espérence	$\frac{1}{\lambda}$	∞	$\frac{w}{\lambda} + \frac{1-w}{u}$

TABLE 2.1 – Tableau de construction de risque à partir de deux sous populations

3. *Cette loi exponentielle est dite "sans mémoire" car la probabilité de décès pour un individu dans un certain laps de temps est la même quelque soit sa durée de vie [4]* $i.e, P(X > s + t_j | X > t_j) = P(X > s), \quad s, t > 0.$

2.2.2 Loi Weibull $W(\alpha, \lambda)$

La distribution de Weibull est souvent utilisée dans le domaine de l'analyse de la durée de vie. Grâce à sa flexibilité elle permet de représenter au moins approximativement une infinité de lois de probabilité [22]. C'est une généralisation de la loi exponentielle,

$$T \rightsquigarrow W(\alpha, \lambda) \quad si \quad T^{\alpha} \rightsquigarrow exp(\lambda)$$

Cette loi est caractérisée par un taux de hasard h croissant ou décroissant de façon monotone, et qui dépend de deux paramètres $\lambda > 0$ (paramètre d'echelle), $\alpha > 0$ (paramètre de forme) tels que pour,

$$\begin{cases} \alpha > 1, \text{le taux de hasard est croissant.} \\ \alpha < 1, \text{le taux de hasard est décroissant.} \\ \alpha = 1, \text{le taux de hasard est constant.} \end{cases}$$

- Fonction de densité :

$$f(t) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\alpha-1} exp\left(-(\frac{t}{\lambda})^{\alpha}\right), \qquad t > 0.$$

- Fonction de survie :

$$S(t) = exp\left(-(\frac{t}{\lambda})^{\alpha}\right), \qquad t > 0.$$

- Fonction de hasard :

$$h(t) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\alpha-1}, \qquad t > 0.$$

- Espérance :

$$\mathbb{E}(T) = \lambda\Gamma\left(1 + \frac{1}{\alpha}\right).$$

Où

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt.$$

2.2.3 Loi de Gompertz (Makeham)

Le modèle de Gompertz est largement utilisé dans de nombreux aspects de la biologie. Il a été fréquemment utilisé pour décrire la croissance des animaux et des plantes, pour décrire le nombre ou le volume de bactéries et de cellules cancéreuses [25].

Cette distribution s'obtient lorsque le taux de hasard varie de façon proportionnelle à sa valeur i.e., $h(t) = \lambda e^{\gamma t}$, $t > 0$ avec $\gamma > 0$. Dans le cas de taux de mortalité, λ est la mortalité de base et γ l'influence de l'âge.

Soit T une v.a de Gompertz à trois paramètres λ, γ et α, où $\lambda > 0$, $\gamma > 0$, le paramètre $\alpha > 0$ tient compte de la mortalité accidentelle. Dans ce cas,

- Fonction de densité :

$$f(t) = \left(\alpha + \lambda e^{\gamma t}\right) e^{\left(-\alpha t - \frac{\lambda}{\gamma}\left(e^{(\gamma t)} - 1\right)\right)}, \qquad t > 0.$$

- Fonction de survie :

$$S(t) = e^{\left(-\alpha t - \frac{\lambda}{\gamma}\left(e^{\gamma t} - 1\right)\right)}, \qquad t > 0.$$

- Fonction d'hasard :

$$h(t) = \alpha + \lambda e^{\gamma t}, \qquad t > 0.$$

2.2.4 La loi log-normale LN(μ, σ)

Une variable aléatoire continue et positive est distribué selon une loi log-normale de paramètres μ et σ^2 si son logarithme est distribuée suivant une loi normale de paramètres μ et σ^2 [24].

- Fonction de densité :

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left(\frac{-\left[\ln(t) - \mu\right]^2}{2\sigma^2}\right), t > 0.$$

- Fonction de survie :

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right), t > 0.$$

Où Φ est la fonction de répartition de la loi normale centrée réduite.

- Fonction de hasard :

$$h(t) = \frac{e^{\frac{-(\ln(t)-\mu)^2}{2\sigma^2}}}{t\int_0^\infty \sigma\sqrt{2\pi} f(x)dx}, t > 0.$$

Le domaine de définition n'étant jamais négatif, il n'y a aucune limitation à l'emploi de la distribution log-normale en fiabilité. Le taux de défaillance est croissant dans le début de vie puis décroissant en tendant vers zéro et la distribution est très dissymétrique.

2.2.5 La loi gamma $\mathbf{G}(\alpha, \beta)$

La loi Gamma est définie par un paramètre de forme $\alpha > 0$ et un paramètre d'échelle $\beta > 0$. Elle est souvent utilisée en analyse de survie.

- Fonction de densité :

$$f(t) = \frac{e^{-\frac{t}{\beta}} t^{\alpha-1}}{\beta^{\alpha} \int_t^{\infty} \Gamma(\alpha) f(x) dx}, t > 0.$$

- Fonction d'hasard :

$$h(t) = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}, t > 0.$$

où, $\Gamma(t)$ est une fonction d'Euler.

Remarque 2.2.5.1.
Si X suit la loi Gamma pour ($\alpha = 1$ et $\beta = \frac{1}{\lambda}$) on obtient la loi exponentielle ($exp(\lambda)$).

2.3 Estimation non-paramétrique

Le modèle non-paramétrique est appliqué lorsque aucune hypothèse n'est faite sur la distribution des temps de survie. Deux méthodes on été proposées pour estimer la durée de survie : la méthode de Kaplan-Meier et la méthode actuarielle.

2.3.1 Méthode de Kaplan-Meier

L'étimateur de Kaplan-Meier (Kaplan-Maier, 1958) aussi dénommé estimateur "produit limite", permet d'estimer la fonction de survie $S(.)$ à partir d'un échantillon de n sujets avec des durées de survies qui peuvent être censurées à droite. La méthode consiste à construire des intervalles qui sont déterminés par les dates d'événements, ils sont donc inégaux [1].
Le principe de la méthode repose sur l'idée de survivre aprés un temps t_2, c'est être en vie avant t_2 et ne pas être décédé au temps t donc survivre avant t et ne pas être décédé au temps t. On écrit cette propriété de la manière suivante : Pour tout temps t_1 et t_2 tels que $t_2 > t_1$,

$$S(t_2) = \mathbb{P}(X > t_2) = \mathbb{P}(X > t_2, X > t_1)$$

En utilisant le théorème des probabilité conditionnelle, on a

$$S(t_2) = \mathbb{P}(X > t_2 | X > t_1) Pr(X > t_1).$$

Tenant compte, de la censure, on observe tous les couples (T_i, δ_i), tels que T_i est la variable réellement observée, et δ_i l'indicateur de censure avec $i = 1, ..., n$. On peut ainsi généraliser la propriété précédente en ordonnant toutes les durées où $t_1 < t_2 < ... < t_n$.

$$S(t_n) = \mathbb{P}(X > t_n | X > t_{n-1})...(X > t_1 | X > t_0)\mathbb{P}(X > 0),$$

avec $t_0 = 0$.

D'où,

$$S(t_n) = \prod_{i=1}^{n} \mathbb{P}(X > t_i | X > t_{i-1})\mathbb{P}(X > t_0).$$

Si on souhaite la survie à n'importe quel temps t tel que $t < t_n$, il suffit d'arrêter le produit précédent au dernier individu suivi juste avant t ;

$$S(t) = \prod_{i=1, t_i<t}^{n} \mathbb{P}(X > t_i | X > t_{i-1}).$$

La probabilité conditionnelle est la probabilité de connaitre l'événement au temps t_i sachant qu'on ne l'a pas subi avant t_{i-1}. Elle est estimée par le nombre d'événements entre t_{i-1} et t_i parmi tous les sujets à risque au temps t_i. Nous considérons les temps d'événements (décés et censure) sont distincts et les notations suivantes :

- n_i, le nombre d'individus à risque de subir l'événement juste avant le temps t_i.
- d_i, le nombre de décés en t_i.
- c_i, le nombre d'observations censurées entre t_i et t_{i+1}.
- n_{i+1}, le nombre d'individus à risque à l'instant t_{i+1} égal à $n_i - d_i - c_i$.

Donc la probabité de mourir dans l'intervalle $[t_{i-1}, t_i]$ sachant que l'on était vivant en t_{i-1}, est estimée par :

$$\mathbb{P}(X \leq t_i | X > t_{i-1}) = \frac{\sum_j \mathbb{P}(\delta_j = 1, t_{i-1} < X_j \leq t_i)}{\mathbb{P}(X_j > t_{i-1})} = \frac{d_i}{n_i}.$$

On en déduit les probabilités de survie conditionnelles qui est l'événement contraire de l'événement précédent :

$$\mathbb{P}(X > t_i | X > t_{i-1}) = 1 - \frac{d_i}{n_i}.$$

Si le temps t_i correspond à une censure ($\delta_i = 0$), il n'y a aucun événement dans l'intervalle de t_{i-1} à t_i. D'où, cette probabilité de survie $S(t)$ ne change que si t_i correspond à l'observation d'un événement dans l'échantillon ($\delta_i = 1$). Ainsi, que l'estimateur de Kaplan-Meier de la survie est donné par :

$$\widehat{S}(t) = \prod_{i=1, t_i<t} (1 - \frac{d_i}{n_i}).$$

Remarque 2.3.1.1.

Supposons qu'un échantillon simple de temps de survie non censurées $(X_i)_{i=1,\dots,n}$. La fonction de survie peut être estimée par la fonction de survie empirique, donnée par [14]

$$\widehat{S}(t) = \left[\frac{\textit{Nombre d'individus avec des temps de survie} \geq t}{\textit{Nombre d'individus à l'étude}} \right].$$

Où,

$$\widehat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T_i > t}$$

Cette méthode d'estimation de la fonction de survie ne peut malheureusement pas être employée lorsque nous sommes en présence de temps de survie censurés.
L'estimateur de Kaplan Meier (KM) ne peut pas être utilisé avec des données censurées par intervalles puisque dans ce cas, les temps exacte d'événements ne sont pas connus [1].

Exemple 2.3.1.1. *(Cas des données non censurées) [3]*
On considère dans cette exemple des données complètes (non censurées), relatives à la réalisation d'un événement mesuré en jours et observé sur 10 individus : 6, 19, 32, 42, 42, 43, 94, 105, 105, et 120
Le tableau suivant donne la valeur de l'estimateur de KM de la durée de survie à ces instants.

t_i	d_i	n_i	$\frac{d_i}{n_i}$	$1 - \frac{d_i}{n_i}$	$\widehat{S}(t_i)$
0	0	10	0	1	1
6	1	10	0,1	0,9	0,9
19	1	9	0,111	0,889	0,8001
32	1	8	0,125	0,875	0,7
42	2	7	0,2857	0,7143	0,5
43	1	5	0,2	0,8	0,4
94	1	4	0,25	0,75	0,3
105	2	3	0,67	0,330	0,1
120	1	1	0	0	0

TABLE 2.2 – Exemple de calcul de survie pour des données complètes.

La figure suivante donne la représentation graphique de l'estimateur de KM pour la fonction de survie.

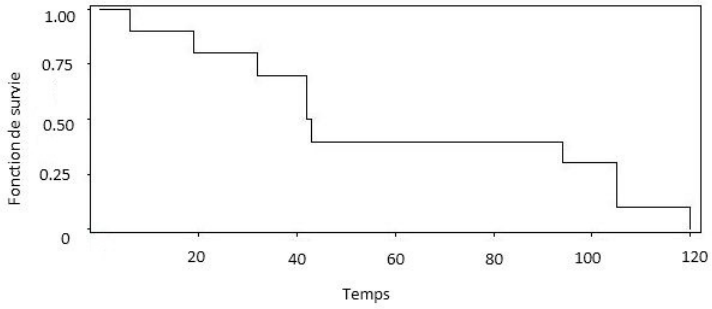

FIGURE 2.1 – Fonction de survie.

Interprétation

$\widehat{S}(t)$ est une fonction en escalier dont les valeurs changent uniquement aux temps correspondants à des événements observés.

Exemple 2.3.1.2. *(Cas des données avec censure)*

On introduit des données censurées à l'exemple précédent. Dans ce cas, la fonction de survie n'est estimée que pour les temps observés mais il faut naturellement ajuster le nombre d'individus à risque. La règle est que pour une durée donnée t_i on ne compabilise dans les individus risqués que ceux qui ont une date d'événement égale ou supérieure à t_i. Dans la liste ci dessous relatives à 19 durées censurées en jours, les données censurées sont signalées

par l'exposant :*

$6, 19, 32, 42, 42, 43^*, 94, 126^*, 207, 211^*, 227^*, 253, 255^*, 270^*, 310^*,$ *et* 316^*

Le tableau suivant donne la valeur de l'estimateur de KM de la durée de survie à ces instants.

t_i	d_i	n_i	$\frac{d_i}{n_i}$	$1 - \frac{d_i}{n_i}$	$\hat{S}(t_i)$
0	0	19	0	1	1
6	1	19	0,053	0,947	0,947
19	1	18	0,056	0,944	0,895
32	1	17	0,059	0,941	0,842
42	2	16	0,125	0,875	0,737
94	1	13	0,077	0,923	0,680
207	1	10	0,1	0,90	0,612
253	1	7	0,143	0,957	0,525

TABLE 2.3 – Exemple de calcul de survie pour des données censurées.

La représentation graphique de l'estimateur de KM pour la fonction de survie est donnée dans la figure suivante,

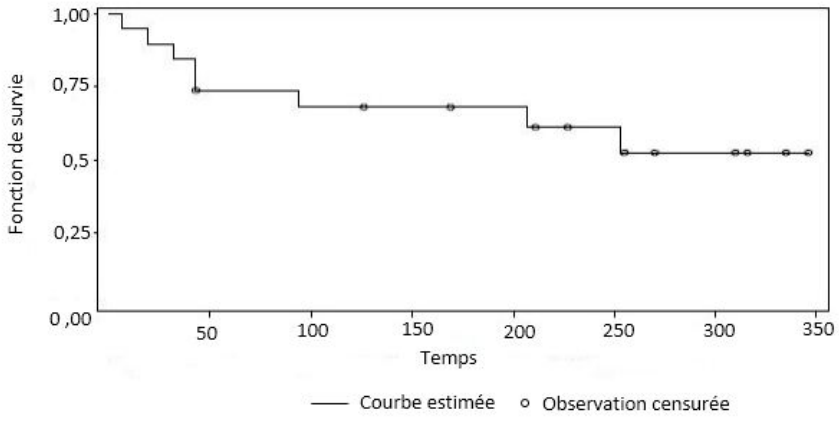

FIGURE 2.2 – Fonction de survie.

Interprétation :

Nous avons des durées supérieures à 253 jours mais elles sont censurées. En conséquence, dans cet exemple, la valeur estimée de la fonction de survie correspondant au temps d'événement maximal observé (soit 253 jours) ne s'annule pas.

2.3.2 Méthode actuarielle

Le terme actuarielle vient du latin actuarius qui signifie littéralement secrétaire aux comptes c'est la première méthode d'analyse de survie à voir le jour en 1912, en tant que théorie statistique [8]. Elle a été introduite pour la première fois dans le champs des applications médicales en 1950, c'était alors la seule méthode disponible pour estimer la survie. Elle suppose à priori une analyse univariée, c'est-à-dire une situation où seul un unique facteur influence la survie. Cette méthode utilise des durées pour lesquelles le moment exacte où se produit l'événement n'est pas connu, seul l'intervalle durant lequel il a eu lieu est utilisé dans les calculs Toutes les données observées, censurées ou non, doivent être prisent en compte dans les calculs.

L'idée de base est de survivre au bout de j intervalles, c'est à dire avoir survécu au 1^{er} intervalle, puis au $2^{ème}$,..., puis au $j^{ième}$ intervalle.

On considère, une échelle des temps dévisée en r intervalles de temps arbitrairement choisis, $[a_0, a_1[, [a_1, a_2[, [a_2, a_3[, ..., [a_{j-1}, a_j[, ..., [a_{r-1}, \infty[$ ces derniers sont généralement de même étendu c'est-à-dire $[a_{j-1}, a_j[$= 12 mois, par exemple. Pour tout t dans l'intervalle $[a_{j-1}, a_j[$ la fonction de survie est :

$$\begin{aligned} S(t) &= \mathbb{P}(T \geq t) \\ &= \mathbb{P}(T \geq t | T \geq a_{j-1})\mathbb{P}(T \geq a_{j-1} | T \geq a_{j-2})...\mathbb{P}(T \geq a_1 | T \geq a_0)\mathbb{P}(T \geq a_0) \\ &= q_j q_{j-1}...q_1 \mathbb{P}(T > a_0), t \in [a_{j-1}, a_j]. \end{aligned}$$

Où,

q_j est la probabilité de survivre en a_j sachant qu'on a survécu en a_{j-1}.

Pour $a_0 = 0$, $\mathbb{P}(T > a_0) = 1$.

Pour estimer q_j, il faut déterminer dans l'intervalle $[a_{j-1}, a_j[$:

n_j : le nombre de sujets exposés au risque au début de l'intervalle

d_j : le nombre d'événements survenant dans l'intervalle.

c_j : le nombre de données censurées à droite de l'intervalle.

e_j : le nombre de sujet exposés au risque dans un intervalle tel que : $n_{j+1} = n_j - d_j - \frac{c_j}{2}$.

Procédant par le calcul de hasard pour chaque intervalle $[a_{j-1}, a_j[$, la probabilité

$\mathbb{P}(T \leq a_j | T > a_{j-1})$ est estimée par, $\widehat{P_j} = \frac{d_j}{n_j}$

D'où, $\widehat{S}(t) = 1 - \frac{d_j}{n_j}$, $t \in [a_{j-1}, a_j]$.

D'où, l'estimateur de la fonction de survie est, $\widehat{S}(t) = \prod_{j=1, t_j \leq t}(1 - \frac{d_j}{n_j}), \quad t > 0,$

C'est donc la même forme que KM.

Exemple 2.3.2.1. *[3]*

Considérons des données regroupées, on ne connaît pas pour chaque individu la date exacte de l'événement ou la censure mais seulement son appartenance à un intervalle de temps correspondant ici à un découpage en trimèstres d'informations mensuelles. Les trois premièrs colonnes du tableau suivant correspondent aux informations de chaque intervalles. Le tableau suivant, donne la valeur de la durée de survie par la méthode actuarielle, à ces instants.

	n_i	d_i	c_i	e_i	$\frac{d_i}{n_i}$	$1 - \frac{d_i}{n_i}$	$\widehat{S}(t)$
$[0, 3[$	20	2	0	20	0,1	0,9	1
$[3, 6[$	18	5	0	18	0,28	0,72	O,9
$[6, 9[$	13	0	0	13	0	1	0,65
$[9, 12[$	13	3	0	13	0,23	0,77	0,65
$[12, 15[$	10	2	2	9	0,22	0,78	0,5
$[15, 18[$	6	0	2	5	0	1	0,39
$[18, 21[$	4	2	0	4	0,5	0,5	0,39
$[21, \infty[$	2	0	2	1	0	1	0,19

TABLE 2.4 – Exemple de calcul pour une table de survie : méthode actuarielle.

La figure suivante donne la représentation graphique de l'estimateur de KM pour la fonction de survie.

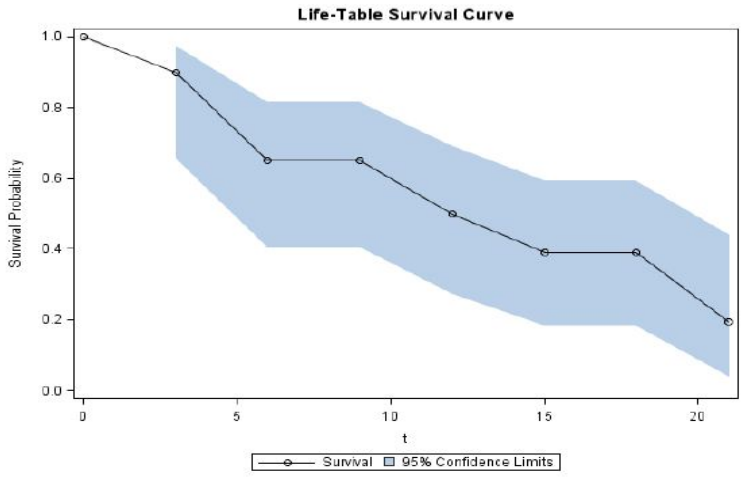

FIGURE 2.3 – Survie estimée par la méthode actuarielle.

Interprétation :

La fonction de survie estimées par la méthode actuarielle est décroissante et différente de celle de Kaplan Meier, elle n'est pas en escalier mais change d'un intervalle à un autres.

2.3.3 Comparaison des deux méthodes

Le tableau suivant nous donne la différence des deux méthodes [33].

La méthode Kaplan Meier	La méthode actuarielle
- Condition d'application : échantillon < 200 individus. - Intervalles construits non forcément égaux. - Les dates de survenu d'événements sont connues. - Courbe en escalier dont l'étendue de chaque marche d'escalier varie d'un palier à l'autre. - L'estimateur donne une estimation de la survie pour tous les temps d'événements observés, et l'estimateur reste constant entre deux temps d'événement observés.	- Condition d'application : échantillon > 200 individus. - Intervalles fixés (mois, trimestre, année) a priori pas en fonction de l'évènement. - Les dates de survenu d'événements ne sont pas connues. - Calcul de la probabilité de survie dans chaque intervalle. - Donne des estimations pour les durées correspondant aux bornes supérieures des intervalles.

TABLE 2.5 – Tableau comparatif des deux méthodes

2.3.4 Comparaison de deux ou plusieurs fonctions de survie

Comparer les fonctions de survies revient à comparer les durées de vies entre deux ou plusieurs groupes en fonction du sexe, âge, etc. Pour ce faire, on utilise des tests statistiques. En absence de données censurées, on peut utiliser le test de Kolmogorov Smirnow, et le test de Mann-Withney, par contre la présence des données censurées nécessite l'utilisation d'autres tests tels que le test Wilcoxon généralisé (test de Gehan), et le test de Log-Rank.
Le test du logrank est le test le plus populaire pour comparer plusieurs courbes de survie. C'est un test dit non-paramétrique. En effet, il permet de prendre en compte toute l'information sur l'ensemble du suivi sans la nécessité de faire des hypothèses sur la distribution des temps de survie. Par souci de simplicité, le test est présenté pour la comparaison de deux groupes, mais il est généralisable à un nombre quelconque de groupes de comparaison [13].
Considérons deux groupes G_A et G_B et Posons :

- $S_A(t)$ la fonction de survie du groupe G_A.
- $S_B(t)$ la fonction de survie du groupe G_B.

Les hypothèses du test de comparaison de Lang-Rank sont les suivantes[32] :

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

Principe du test Log-Rank :

Considérons $t_1, t_2, ..., t_k$ les temps de décés observés dans les deux groupes G_A et G_B,tels que :

- m_{Ai} : Nombre de décés observés dans G_A en t_i.
- m_{Bi} : Nombre de décés observés dans G_B en t_i.
- m_i : Nombre de décés observés en t_i.
- n_{Ai} : Nombre de sujets exposés au risque dans G_A en t_i.
- n_{Bi} : Nombre de sujets exposés au risque dans G_B en t_i.
- n_i : Nombre de sujets exposés en t_i

calculons en chaque temps de décés observés t_i les quantités suivantes :

- e_{Ai} : le nombre de décés attendus en t_i dans le groupe G_A sous H_0

$$e_{Ai} = \frac{m_i n_{Ai}}{n_i}$$

- e_{Bi} : le nombre de décés attendus en t_i dans le groupe G_B sous H_0

$$e_{Bi} = \frac{m_i n_{Bi}}{n_i}$$

Puisque ce n'est pas evident de faire le test pour chaque t_i alors nous calculons les quantités précédentes pour tous les temps t_i on considère alors,

- $E_A = \sum_{i=1}^{k} e_{Ai}$: le nombre total de décés attendus dans G_A sous H_0.
- $E_B = \sum_{i=1}^{k} e_{Bi}$: le nombre total de décés attendus dans G_B sous H_0.
- $O_A = \sum_{i=1}^{k} m_{Ai}$: le nombre total de décés observés dans G_A.
- $O_B = \sum_{i=1}^{k} m_{Bi}$: le nombre total de décés observés dans G_B.

Sous H_0,

$$\frac{O_A - E_A}{E_A^{1/2}} \to N(0,1) \quad \Leftrightarrow \quad \frac{(O_A - E_A)^2}{E_A} \to \chi_1^2$$

De même,

$$\frac{O_B - E_B}{E_B^{1/2}} \to N(0,1) \quad \Leftrightarrow \quad \frac{(O_B - E_B)^2}{E_B} \to \chi_1^2$$

Sous H_0,

$$\begin{aligned} \chi^2 &= \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \\ &= (O_A - E_A)^2(\frac{1}{E_A} + \frac{1}{E_B}) \rightsquigarrow \chi_1^2, \quad car \quad O_A + O_B = E_A + E_B \end{aligned}$$

Sous H_0 la décision est :

— Si la khi-deux calculée est supérieure au khi-deux tabulée on rejette H_0.

— Si la khi-deux calculée est inférieure au khi-deux tabulée on rejette H_1.

Exemple 2.3.4.1. *[3]*

Donnons un exemple illustratif de la conparaison, les données suivantes décrivent les temps de survie de patients leucemique avec traitement 6-MP (groupe A, 21patients) et sans traitement (groupe B, 21patients). le signe signale une donnée censurée :*

— *Groupe A : 6, 6, 6,* 6^**, 7,* 9^**, 10,* 10^**,* 11^**, 13, 16,* 17^**,* 19^**,* 20^**, 22, 23,* 25^**,* 32^**,* 32^**,* 34^**,* 35^*

— *Groupe B : 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.*

	d_{1i}	n_{1i}	d_{2i}	n_{2i}	d_i	n_i	e_{1i}	e_{2i}
1	0	21	2	21	42	2	1	1
2	0	21	2	19	40	2	1,05	0,95
3	0	21	1	17	38	1	0,553	0,447
4	0	21	2	16	38	2	1,135	0,865
5	0	21	2	14	37	2	1,2	0,8
6	3	21	0	12	35	3	1,909	1,091
7	1	17	0	12	39	1	0,586	1,714
8	0	16	4	12	29	4	2,286	1,714
10	1	15	0	8	28	1	0,652	0,348
11	0	13	2	8	23	2	1,238	0,762
12	0	12	2	6	21	2	1,333	0,667
13	1	12	0	4	18	1	0,75	0,25
15	0	11	1	4	16	1	0,733	0,267
16	1	11	0	3	15	1	0,786	0,214
17	0	10	1	3	14	1	0,769	0,231
22	1	7	1	2	9	2	1,556	0,444
23	1	6	1	1	7	2	1,714	0,286
	$O_1 = 9$		$O_2 = 21$				$E_1 = 19,25$	$E_2 = 10,75$

TABLE 2.6 – Tableaux de données de deux groupes

Interprétation :

On $O_A - E_A = -10,25$. la valeur négative de la statistique de Log-Rank signale que le traitement 6-MP affecte favorablement le temps de survie des patients traités.

On as aussi $O_B - E_B = 19,25$. La valeurs positive de la statistique de Log-Ran relative au groupe B signale que la maladie affecte négativement sur le temps de survie des patients sans traitement. Par ailleurs, la khi-deux associée égale à :

$$\frac{(9-19,25)^2}{19,25} + \frac{(21-10,75)^2}{10,75} = 15,23 > \chi_1^2 = 3,8415.$$

Si on compare ce chiffre à la valeur tabulée d'un seuil de risque usuel (0,05) de la distribution de khi-deux à 1 degré de liberté, on conclut que l'avantage du traitement est significatif d'où, on rejette l'hypothèse H_0.

Remarque 2.3.1.

Lorsqu'on désire prendre en compte un ou plusieurs facteurs pouvant influencer sur la survie (analyse multivariée), on utilise le modèle de Cox.

2.4 Estimation semi-paramétrique

Dans cette étude, des facteurs explicatifs (ou les covariables) sont fréquemment disponibles. Le désire de connaître les effets de ces facteurs sur le changement du taux de risque à travers le temps permet de chercher le modèle adéquat pour ce phénomène. La nature semi-paramétrique du modèle a permis de quantifier l'effet des covariables comme l'âge, la pression sanguine, etc, sur le risque de la survie d'un individu. Pour cette raison, le modèle des risques proportionnels (Cox, 1972) a connu ces dernières années un remarquable intérêt pour analyser les données de temps de survie [33].

2.4.1 Les modèles à hasards proportionnels

Le modèle à hasard proportionnelle est un modèle qui est modélisé par le produit d'une fonction qui ne dépond que du temps (fonction de risque de base $h_0(t)$ partie non-paramétrique) et d'une fonction positive qui n'en dépend pas r (partie paramétrique). Alors il s'écrit comme suit [4] : pour tout $t > 0$,

$$h(t|Z, \beta) = h_0(t)\ r(\beta', Z).$$

Où,

Z est un vecteur de covariables.

β est le vecteur de paramètres d'intérêt.

Ce modèle est dit à risques proportionnels car, quels que soient deux individus i et j qui ont pour covariables Z_i et Z_j, Le rapport des fonctions de hasard ne varie pas au cours du temps. C'est à dire,

$$\frac{h(t|Z_i, \beta)}{h(t|Z_j, \beta)} = \frac{r(\beta', Z_i)}{r(\beta', Z_j)}, \qquad \text{indépendant de temps t.}$$

Le rapport des fonctions de hasard est par définition un risque relatif à l'instant t des sujets de caractéristiques Z_i par rapport aux sujets de caractéristiqueq Z_j.

Un des cas particulier, trés important est le modèle de Cox, qui suppose que la fonction r est la fonction exponentielle, c'est à dire :

$$h(t|Z, \beta) = h_0(t)exp(\beta' Z).$$

Remarque 2.4.1.

1. *D'autre choix de la fonction r sont possibles. Néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et* $exp(0) = 1$.
2. *Si* h_0 *a une forme inconnue, le modèle est dit semi-paramétrique.*

2.4.2 Modèle de Cox

Ce modèle est devenu un modèle de référence pour l'analyse statistique des enquêtes de cohorte en épidémiologie. L'hypothèse principale est que ce modèle est à risque proportionnels (ou ne dépond pas du temps), avec une fonction de risque de base $h_0(t)$ (partie non paramétrique du modèle), et une fonction exponentielle qui dépend des caracteristiques Z du sujet (partie paramétrique du modèle).

Construction du modèle :

Le rapport des fonctions de risque de deux sujets i et j avec des vecteurs de variables explicatives Z_i et Z_j est :

$$\frac{h_i(t, Z_i, \beta)}{h_j(t, Z_j, \beta)} = \frac{h_0(t)exp(\beta' Z_i)}{h_0(t)exp(\beta' Z_j)} = \frac{exp(\beta' Z_i)}{exp(\beta' Z_j)}.$$

Le rapport est constant au cours du temps, ce qui entraine que les fonctions de risque sont proportionnelles, d'où le nom du modèle.

Exemple d'interprétation :

Soit la relation suivante [13] :

$$h(t|Z = z_1, Z_2 = z_2, ...) = h_0(t)exp(\beta_1 z_1 + \beta_2 z_2 + ...).$$

On considère que $X_1 = \begin{cases} 1 & \text{chez les hommes} \\ 0 & \text{chez les femmes} \end{cases}$ et que toutes les autres covariables sont fixées alors,

$$h(t|Z_1 = z_1, Z_2 = z_2, ...) = \begin{cases} h_0(t)exp(\beta_1 + \beta_2 x_2 + ...), & \text{chez les hommes} \\ h_0(t)exp(\beta_2 x_2 + ...), & \text{chez les femmes} \end{cases}$$

D'où, le rapport des risques (hasard ratio :HR) entre femme et homme est :

$$HR_{H/F} = \frac{h(t|Z_1 = 1, Z_2 = z_2, ...)}{h(t|Z_1 = 0, Z_2 = z_2, ...)} = exp(\beta_1).$$

On conclut l'influence de facteur Z_1 :

Si $\beta_1 > 0$, alors $exp(\beta_1) > 1$, et Z_1 est un facteur de risque.

Si $\beta_1 < 0$, alors $exp(\beta_1) < 1$, et Z_1 est un facteur protecteur.

Si $\beta_1 = 0$, alors $exp(\beta_1) = 1$, et Z_1 n'est pas un facteur influençant le temps d'événement.

2.4.2.1 Vraisemblance partielle

Pour estimer les coefficients de régression β, on maximise la vraisemblance d'un n-échantillon de durées $T_1, T_2, ..., T_n$ [35] :

$$L = \prod_{i,Décés} f(t_i|z_i) \prod_{j,Censurée} S(t_j|z_j)$$

où,

$$\begin{aligned} S(t_j|z,\beta) &= exp(-\int_0^{t_j} h(u|z)du) \\ &= exp(-\int_0^{t_j} h_0(u)exp(\beta' z)du) = exp\left(-exp(\beta' z)\int_0^{t_j} h_0(u)du\right) \\ &= \left[exp\left(-\int_0^{t_j} h_0(u)du\right)\right]^{exp(\beta' z)}. \end{aligned}$$

or, $exp\left(-\int_0^{t_j} h_0(u)du\right) = S_0(t_j)$ (fonction de survie initiale associée au taux de hasard $h_0(t)$),
On pose $\alpha_j = [S_0(t_j)]^{exp(\beta' z)}$.

et,

$$\begin{aligned} \gamma_i &= f(t_i|z,\beta) = h(t_i|z)S(t_i|z) \\ &= S(t_i/z)h_0(t_i)exp(\beta' z) \\ &= [S_0(t_i)]^{exp(\beta' z)} h_0(t_i)exp(\beta' z). \end{aligned}$$

En remplaçant dans la fonction de vraisemblance, on obtient :

$$L(\beta) = \prod_{i,Décés} [\gamma_i] \prod_{j,Censurée} [\alpha_j]$$

D'où,

$$L(\beta) = \prod_{i,Décés} \left[[S_0(t_i)]^{exp(\beta' z)} h_0(t_i) exp(\beta' z) \right] \prod_{j,Censurée} \left[[S_0(t_j)]^{exp(\beta' z)} \right].$$

La vraisemblance contient $h_0(t)$, que nous ne cherchons pas à estimer, de ce fait Cox à proposé d'estimer les coefficients β sans prendre en compte h_0, et le considérer comme un paramètre de nuisance. Cette vraisemblance est dite partielle (vraisemblance de Cox). L'information ne peut pas être donnée sur β par les intervalles dans lesquels aucun événement n'a eu lieu, car on peut concevoir que h_0 soit nulle dans ces intervalles. Cela veut dire que les moments où se produisent les censures n'apportent que peu ou pas d'informations.

Cette vraisemblance est obtenue en factorisant la probabilité conditionnelle (elle ressemble à une probabilité conditionnelle mais n'est pas exactement égale), que le sujet i subisse l'événement en t_i s'achant qu'il est à risque au temps t_i et qu'il n'y a qu'un seul événement en t_i. En effet $t_1 < t_2 <, ..., < t_N$, sont les différents temps d'événements observés, et $Z_1, Z_2, ..., Z_n$ sont les variables explicatives des sujets ayants subi l'événement avec un taux de hasard identique pour tous les individus.

La probabilité conditionnelle s'écrit comme suit :

$$\mathbb{P}_i = \frac{h(t_i|z_i, \beta)}{\sum_{j \in R(t_i)} h(t_i|z_j, \beta)}.$$

où, $R(t_i)$ l'ensemble des invidus encore à risque juste avant t_i. On utilisant le modèle à risque proportionnels, la probabilité devient :

$$\mathbb{P}_i = \frac{exp(\beta' z_i)}{\sum_{j \in R(t_i)} exp(\beta' z_j)}.$$

Cette dernière ne dépond pas de la fonction à risque de base $h_0(t)$, alors la vraisemblace s'ecrit :

$$L(z, \beta) = \prod_{i=1}^{N} \mathbb{P}_i = \prod_{i=1}^{k} \frac{exp(\beta' z_i)}{\sum_{j \in R(t_i)} exp(\beta' z_j)}.$$

2.4.2.2 Événements simultanés

La vraisemblance partielle du modèle de Cox nécessite l'hypothèse de données continues, c'est-à-dire qu'il n'y a pas plusieurs événements (ex-æquo), à la même date [1]. En pratique, cette hypothèse n'est pas toujours vérifiée à cause de la discrétisation du temps inférente au receuil des données. Plusieurs corrections de la vraisemblance partielle ont été proposées,(méthode d'Efron, méthode exacte, méthode discrète,...) mais les résultats sont quasiment identiques [**?**].
Cette méthode est la plus utilisée, elle est dénomée "approximation de breslow". La vraisemblance est approchée par :

$$L(z,\beta) = \prod_{i=1}^{N} \frac{exp\left(\beta'\left(\sum_k z_k\right)\right)}{\left(\sum_{j\in R(t_i)} exp(\beta' z_j)\right)^{r_i}}$$

Où, k unités sont décédées en t_i, et r_i sujets ayant subis l'événement au temps t_i. Cette corrections est assez précise lorsque le nombre d'ex-æquo n'est pas trop élevé [1].

2.4.2.3 Estimation des paramètres

L'estimation des paramètres β_j du vecteur β est simple. Il suffit de maximiser la vraisemblance partielle [4]. La log vraisemblance est donnée par :

$$L_{log}(z,\beta) = \log(L(z,\beta)) = \sum_{i=1}^{N}[\beta' z_i - \log \sum_{j\in R(T_i)} exp(\beta' z_i)]$$

$U(\beta)$ est la fonction de score, c'est à dire le vecteur p des dérivées-premières de $L_{log}(\beta)$,tel que

$$\begin{aligned} U(\beta) &= \frac{\partial L_{log}(\beta)}{\partial \beta} \\ &= \left(\frac{\partial L_{log}(\beta)}{\partial \beta_1}, ..., \frac{\partial L_{log}(\beta)}{\partial \beta_p}\right) \\ &= \sum_{i=1}^{N}\left[z_i - \frac{\sum_{j=1}^{n} z_j exp(\beta' z_j)}{\sum_{j=1}^{n} exp(\beta' z_j)}\right] \end{aligned}$$

- La matrice de Fisher est obtenue comme suit :

$$I(\beta) = \left(-\mathbb{E}\left[\frac{\partial^2 L_{log}(\beta)}{\partial \beta_i \partial \beta_j} \right] \right)_{i,j=1,\dots,n.}$$

- La matrice de covariance de β peut être calculée à partir de la matrice de Fisher :

$$\widehat{Var} = [I(\widehat{\beta})]^{-1}$$

2.4.3 Test de wald

Le test de wald permet de verifier l'hypothèses de significativité du modèle ou des variables. Le test déduit les propriétés asymptotique de $\widehat{\beta}$ qui suit une loi normale asymptotiquements. Il porte sur les hypothèses suivantes [30] :

$$\begin{cases} H_0 : \beta_k = 0, & \text{aucune variable explicative n'apporte de l'information} \\ H_1 : \beta_k \neq 0, & \text{au moins une variable explicative apporte de l'information.} \end{cases}$$

L'idée est de mesuré l'écart entre $\widehat{\beta}$ et β_0, la statistique du test s'écrit de cette manière :

$$\chi^2_w = (\widehat{\beta} - \beta_0)' I(\widehat{\beta})(\widehat{\beta} - \beta_0).$$

Cette statistique suit asymptotiquement une loi du χ^2_1, d'où la région critique du test est :

$$\chi^2_w > \chi^2_1.$$

2.4.4 L'adéquation du modèle

Pour valider un modèle de Cox, il faut que l'hypothèse des risques proportionnels soit vérifiée. Autrement dit, le risque doit être constant au cours du temps, pour chaque variable explicative [36].

Contrôler l'hypothèse de risques proportionnels revient à vérifier que les courbes $Log(H(t))$ tracées par rapport aux $Log(t)$, tracées pour chaque modalité, sont parallèles entre elles. Il aurait été possible de contrôler cette hypothèse à partir de la fonction de survie, mais comme les fonctions de survie des différentes modalités se ressemblent, il est plus facile de passer

par la fonction de risques cumulés $H(t)$. Il faut effectuer ce test graphique pour toutes les variables. Pour les variables quantitatives, il est difficile à mettre en oeuvre le tracé des courbe puisqu'une courbe est tracée pour chacune des valeurs de la variable.

Remarque 2.4.4.1.

Il existe d'autres moyens pour verifier la proportionnalité comme les résidus de schoenfeld [1].

2.4.5 Cas où l'hypothèse de proportionnalité n'est pas verifiée

En cas ou une variable explicative ne respecte pas l'hypotèse de proportionnalité des risques nous proposons quelques solutions :

- La stratification :

  Si l'hypothèse de proportionnalité des risques n'est pas verifiée pour une variable explicative qualitative ou quantitative, et que l'on veuille tout de même prendre en compte l'effet de cette variable, la solution est de stratifier le modèle sur cette variable. Mais, dans ce cas, il n'y a pas d'estimation de l'effet de cette variable.

- Eliminer des variables :

  Si la présence d'une variable n'est pas obligatoire dans un modèle, pour diverses raisons, et si cette variable ne verifie pas l'hypothèse de proportionnalité des risques, il ya encore moins de raison de la garder. Evidemment, avant de l'enlever définitivement du modèle on doit d'abord vérifier qu'elle n'est pas un facteur de confusion.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les trois modèles d'estimation de la fonction de survie à savoir le modèle paramétrique, le modèle non-paramétrique et le modèle semi-paramétrique (de Cox). Dans le chapitre suivant nous appliquons ces modéles sur des données réelles afin de les comparer.

CHAPITRE 3

APPLICATIONS

3.1 Introduction

Le cancer, d'après l'Organisation Mondiale de la Santé [**?**] est la première cause de décès dans le monde. En effet, $8,8$ millions de décès dus au cancer ont été enregistrés en 2012 [**?**] dont les principales localisations étaient, dans l'ordre du nombre de décès, le poumon, le foie, l'estomac, le cancer colorectal, le sein et l'oesophage.

Les patients en rémission à la suite d'un traitement curatif du cancer entrent dans la phase de surveillance post-thérapeutique. La surveillance repose sur des examens cliniques, de l'imagerie, éventuellement des endoscopies et des études de marqueurs biologiques. Idéalement, la phase de surveillance devrait durer jusqu'à ce que le patient soit considéré comme guéri, c'est-à-dire ne présentant plus de risque de rechute. Quant aux visites, elles devraient être les plus nombreuses possibles. Ceci n'est bien évidemment pas possible pour des raisons pratiques et de coût. Il est alors nécessaire de déterminer un calendrier de surveillance. Ce calendrier devra proposer des visites suffisamment fréquentes mais aussi en nombre suffisamment réduit afin de limiter la charge pour le patient, le praticien et l'établissement de soins. Il est important de faire une étude sur la maladie car elle nous permettera de définir l'effet des facteurs de cette dernière sur les patients.

Plusieurs études ont été faitent en analyse de survie basées sur l'analyse de durée. L'application de cette dernière ne se limite pas que dans le domaine médicale, mais aussi dans d'autres domaines comme le domaine économic et le scocial [**?**]. L'analyse des durée est importante dans le domaine d'assurance car elle permet d'apporter un éclairage sur l'événement étudier en mettant en exergue les facteurs influençant la durée étudié, tel que l'étude de durée du cumul emploi-retraite.

Dans ce chapitre, deux applications sont proposées afin de mettre en pratique la théorie developpées dans les chapitres précédents.

- La première application concerne les données d'assurance retraite de france.
- La deuxième application concerne les données du stage effectué au sein de

l'hopital d'Amizour (service Oncologie) pour les données relatives aux patientes atteintes du cancer du sein, l'une des maladies les plus fréquentes chez les femmes, plus de 300 patientes reçus chaque années.

Nous avons utilisé le logitiel SAS, (Statistical Analysis System) version 9.3 qui est un logiciel de statistique polyvalent. Il est assez ancien (ses débuts remontent aux années 1960) mais il est constamment enrichi de nouvelles méthodes d'analyse statistiques des données. Par conséquent, on trouve souvent des présentations différentes pour le même problème traité [**?**].

3.2 Traitement des données d'assurances retraite

On considère les données d'assurance retraite extraites de la caisse nationale de l'assurance vieillesse (CNAVF) qui gère le régime général (RG) et l'organisme de retraite "régime social indépendant" (RSI). On sintéresse de cette application, à l'étude de la durée du cumul emploi-retraite RSI-RG. La durée du cumul emploi-retraite (exprimée en année) est la durée qui sépare la date du début d'activité d'un retraité au régime général et la date de fin d'activitées concernent la période allant du 1/01/2008 au 31/12/2012. 49% des données sont censurées à droite.

3.2.1 Présentation des données

Le tableau suivant (3.1) présente la répartition des cotisants selon leur groupe professionnel et le statut d'auto-entrepreneur. La fonction exercée, pour le cotisant, en tant qu'indépendant peut être : Artisant, commerçant ou fonction libérale. Il peut avoir le statu d'auto-entrepreneur ou non.

	Cotisants au RSI	**Cotisants RSI et retraités au régime général**	**Répartition des cotisants au RSI**	**Répartition des cotisants au RSI et retraités au régime général**	**Taux de cumulants**
Artisans	163 253	33 889	28,5%	23,2%	20,8%
Commerçants	240 704	60 391	42,0%	41,3%	25,1%
Professions libérales	169 514	51 972	29,6%	35,5%	30,7%
Auto-entrepreneurs	135 676	54 852	23,7%	37,5%	40,4%
Non autoentrepreneurs	135 676	91 400	76,3%	62,5%	20,9%
Ensemble	573 471	146 252	100%	100%	25,5%

TABLE 3.1 – Cotisants au RSI de 55 ans et plus au 31 décembre 2012 selon leur groupe professionnel et le statut d'auto-entrepreneur.

Les données du RSI et de la CNAV, sur la population agée de 55ans et plus, ont permi d'identifier les indépendants ayant demandé leurs retraite du régime général : Parmi 573 471 cotisants de 55ans et plus du RSI, 26% (146 252) sont également retraité au régime général (RG).

Nous présentons une analyse de la durée passé en cumul, en appliquant des modèles de durée (sous SAS) [36]. L'événement étudié correspond à la fin du cumul emploi-retraite, qui se traduit par la fin de l'activité d'indépendant.

3.2.2 Estimation non-paramétrique

Deux méthodes non-paramétrique sont appliquées, à savoir : la méthode de Kaplan-Meier (KM) et la méthode actuarielle. Dans la méthode (KM), ou on considère les durées exactes et que les observations censurées sont exposées au risque jusqu'à la durée t (la censure arrive trés rapidement aprés la durée t). Dans la méthode actuarielle, les durées sont regroupées sur des intervalles de temps de sorte que la censure survient de manière uniforme sous un intervalle. Ainsi les individus censurées sont exposées au risque uniquement pendant la moitié de l'intervalle de temps, alors que dans KM ils font pertie de la population soumise au risque.

Mise en oeuvre des deux méthodes : KM et Actuarielle sous (SAS), respectivement.

```
Proc lifetest data=cohorte2008 method=KM;
Time duree*censure (0);
Run;
```

```
Proc lifetest data=cohorte2008 method=ACT;
Time duree*censure (0);
Run;
```

Le Système SAS

Procédure LIFETEST

Product-Limit Survival Estimates

duree		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000		1.0000	0	0	0	15017
0.08219		.	.	.	1	15016
0.08219		.	.	.	2	15015
3.04932		0.5937	0.4063	0.00402	6073	8739
3.04932	*	.	.	.	6073	8738
3.05205		0.5936	0.4064	0.00402	6074	8737
3.05479		0.5935	0.4065	0.00402	6075	8736
3.05479	*	.	.	.	6075	8735
4.99726	*	.	.	.	7626	1
4.99726	*	.	.	.	7626	0

TABLE 3.2 – L'estimateur de la fonction de survie par la méthode de Kaplan-Meier.

Procédure LIFETEST

Life Table Survival Estimates

Interval		3	4	5	6	7	8	9	10	11	12	14 Evaluated at the Midpoint of the Interval			16
1 [Lower,	2 Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	13 PDF	PDF Standard Error	15 Hazard	Hazard Standard Error
0	0.5	1057	4	15015.0	0.0704	0.00209	1.0000	0	0	4.2868	0.0827	0.1408	0.00418	0.145929	0.004486
0.5	1	1354	10	13951.0	0.0971	0.00251	0.9296	0.0704	0.00209	.	.	0.1804	0.00468	0.204008	0.005537
1	1.5	1042	37	12573.5	0.0829	0.00246	0.8394	0.1606	0.00300	.	.	0.1391	0.00416	0.17291	0.005352
1.5	2	1033	28	11499.0	0.0898	0.00267	0.7698	0.2302	0.00344	.	.	0.1383	0.00415	0.188117	0.005847
2	2.5	785	45	10429.5	0.0753	0.00258	0.7007	0.2993	0.00374	.	.	0.1055	0.00366	0.156421	0.005579
2.5	3	775	76	9584.0	0.0809	0.00278	0.6479	0.3521	0.00390	.	.	0.1048	0.00368	0.168542	0.006049
3	3.5	605	98	8722.0	0.0694	0.00272	0.5955	0.4045	0.00402	.	.	0.0826	0.00329	0.143714	0.005839
3.5	4	579	119	8008.5	0.0723	0.00289	0.5542	0.4458	0.00407	.	.	0.0801	0.00326	0.150019	0.00623
4	4.5	286	2822	5959.0	0.0480	0.00277	0.5142	0.4858	0.00411	.	.	0.0494	0.00287	0.098349	0.005814
4.5	5	110	4152	2186.0	0.0503	0.00468	0.4895	0.5105	0.00416	.	.	0.0493	0.00460	0.103238	0.00984
5	.	0	0	0.0	0	0	0.4648	0.5352	0.00457	.	.	.	.	.	.

TABLE 3.3 – Méthode durée actuarielle, fonctions de survie et de risque.

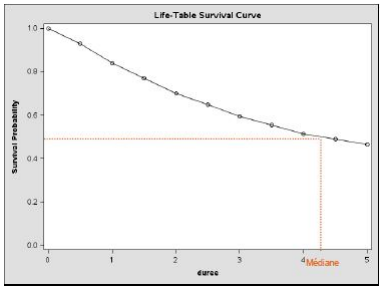

FIGURE 3.1 – Fonction de survie avec la méthode actuarielle.

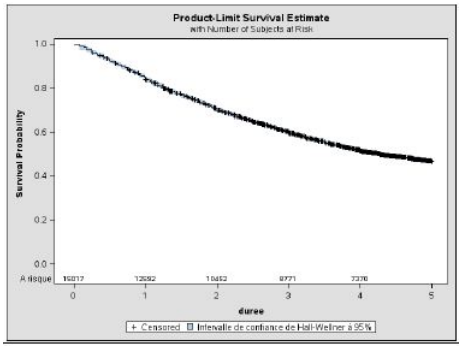

FIGURE 3.2 – Fonction de survie avec la méthode de Kaplan-Meier.

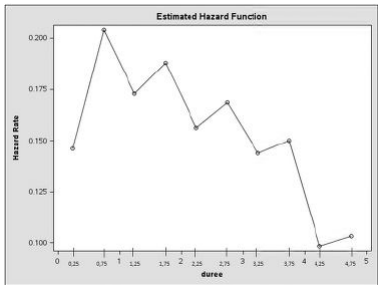

FIGURE 3.3 – Fonction de risque instantané avec la méthode actuarielle.

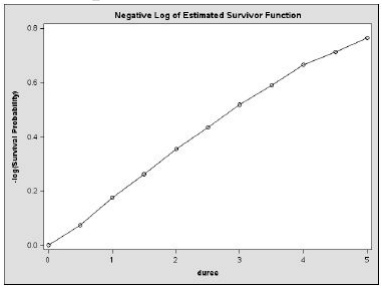

FIGURE 3.4 – Fonction de risques cumulés avec la méthode actuarielle.

Interprétation :

Le tableau (3.2) représente l'estimateur de survie par la méthode de Kaplan Meier. La survie à la fin de l'étude est de 0,5935, l'écart-type de la fonction de survie est $0,00402$, ainsi l'effectife ayant connu l'événement juste avant 3 ans et 2 semaines (3,049ans) est de 6073 personnes ont arrêté le cumul, et la population encore soumise au risque aprés 3 ans et 2 semaines est de 8739 personnes encore en cumul. 7626 personnes ont arrêté le cumul, 7391 personnes censurés.

Le tableau (3.3) représente l'estimateur de survie par la méthode actuarielle, cette méthode donne plus de précision sur les informations, tel que la médian résiduelle qui est la durée de t_i et l'instant où $S(t) = S(t_i)/2$. La survie à la fin de l'étude est de $0,4648$, l'écart-type de la fonction de survie est $0,00457$.

D'aprés la figure (3.1) plus de la moitié de la population est en cumul emploi-retraite, pendant au moins 4 ans. Aprés 5 ans de cumul emploi-retraite, 47% de la cohorte est toujours en situation cumul (tableau 3.3, encadré vert). En parallèle 23% de la population arrête le cumul emploi-retraite avant un an et demi (tableau 3.3, encadré rouge).

D'aprés la figure (3.2) le risque de sortir du cumul emploi-retraite est assez élevé au

cours de la première année. Aprés 6 mois de cumul, on as presque une chance sur 10 de sortir du dispositif (tableau 3.3, encadré bleu). Au delà d'un an, le risque décroit progressivement.

Comparaison des deux méthodes :

La fonction obtenu par la méthode de (KM) est une fonction décroissante en escalier dont les valeurs changent uniquement aux temps correspondants aux événements observés, cela n'apparaisse pas bien vue qu'on a un nombre important de valeurs censurées. La fonction de survie estimées par la méthode actuarielle est décroissante et différente de celle de (KM), elle n'est pas en escalier mais change d'un intervalle à un autres.

3.2.3 Estimation semi-paramétrique

Nous mettons en oeuvre un modèle de Cox, et verifiant par la suite la proportionnalité des variables.

Voici la procédure SAS du modèle de Cox :

```
Proc phreg data= cohorte2008 ;
Class activité_der(rf="C") statut(ref="2") sect_mod(ref="serv") ;
Model duree*censure(0)=activité_der statut sect_mod generation ;
Run ;
```

Les résultats du modèle sont les suivants :

Le tableau (3.4) suivant donne les caractéristiques principales du modèle : nom de la table SAS utilisée, nom des variables de durée et de censure, méthode de gestion des événements simultanés et nombre d'observations.

Model Information	
Data Set	WORK.COHORTE2008
Dependent Variable	duree
Censoring Variable	censure
Censoring Value(s)	0
Ties Handling	EFRON

Number of Observations Read	15017
Number of Observations Used	15017

TABLE 3.4 – Information générale sur le modèle.

On considère les codes suivant :

- Activité_der indique le groupe professionnel peut être,
  - Commerçant → le code C = modalité de référence
  - Artisan → le code A
- Statut indique la situation avant la liquidation peut être,
  - Salarié → le code1
  - Indépendant → le code2 = modalité de référence
  - Ni indépendant, ni salarié → le code3
- Sect_mod indique le secteur d'activité peut être,
  - Secteur non renseigné → le code NR
  - Commerce → le code Com
  - Construction → le code Const
  - Industrie → le code Indus
  - services → le code Serv

Le tableau (3.5) suivant indique quelles sont les variables qualitatives, leurs modalités et les modalités de référence du modèle (vecteur nul).

Class Level Information					
Class	Value	Design Variables			
activite_der	A	1			
	C	0			
statut	1	1	0		
	2	0	0		
	3	0	1		
sect_mod	NR	1	0	0	0
	com	0	1	0	0
	const	0	0	1	0
	indus	0	0	0	1
	serv	0	0	0	0

TABLE 3.5 – Information sur les variables qualitatives du modèle.

Le tableau (3.6) suivant rappelle le nombre d'événements observés (les sorties du cumul), le nombre d'observations censurées et leur proportion.

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
15017	7626	7391	49.22

TABLE 3.6 – Tableau du nombre d'évènements observés et du nombre d'observations censurées.

Les tableaux (3.7), (3.8)suivants sont émis en raison de l'option simple. Ils éditent les statistiques descriptives des variables qualitatives et quantitatives.

Descriptive Statistics for Continuous Explanatory Variables					
Total Sample					
Variable	N	Mean	Standard Deviation	Minimum	Maximum
generation	15017	1946	3.99761	1912	1953

TABLE 3.7 – Tableau du nombre d'évènements observés et du nombre d'observations censurées.

Frequency Distribution of CLASS Variables		
Total Sample		
Class	Value	Frequency
activite_der	A	5680.0
	C	9337.0
statut	1	3368.0
	2	10734.0
	3	915.0
sect_mod	NR	1286.0
	com	4131.0
	const	2007.0
	indus	1249.0
	serv	6344.0

TABLE 3.8 – Statistiques descriptives des variables qualitatives du modèle.

Le tableau (3.9) suivant indique que le modèle est convergent, le critère utilisé est GCNV=1E-8.

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

TABLE 3.9 – Statut du modèle vis-à-vis de la convergence.

Le tableau (3.10) suivant donne les critères utilisés (-2logl, AIC, SBC) pour le choix du modèle parmi les résultats des critères (ayant le moins de variables) afin de satisfaire le critère de parcimonie. Le meilleur modèle est celui pour lequel le critère d'information d'Akaike (AIC) ou le critère bayésien de Schwartz (SBC) est le plus faible. Donc, nous choisissons le modèle

avec covariable car le critère AIC est inférieur .

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	141373.86	137305.33
AIC	141373.86	137321.33
SBC	141373.86	137376.84

TABLE 3.10 – Résultats des critères de qualité du modèle.

Les résultats des tests donnés par le rapport de vraisemblance, Wald et score de l'hypothèse : H_0 : "aucune variable explicative n'apporte de l'information", contre H_1 : "au moins une variable explicative apporte de l'information" sont résumés dans le tableau (3.11) suivant. D'aprés le tableau, l'hypothèse nulle est rejetée, il ya au moins une variable utile dans le modèle car les valeurs des statistiques des tests sont supérieure à celles des p-values.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > Khi-2
Likelihood Ratio	4068.5344	8	<.0001
Score	9277.5886	8	<.0001
Wald	5917.3090	8	<.0001

TABLE 3.11 – Résultats des tests globaux d'hypothèse nulle.

Pour la signification des variables explicatives, on utilise le test de Wald pour lequel l'hypothèse nulle est H_0 :"la variable n'est pas significative", et l'hypothèse altérnative H_1 : "la variable est significative".
D'apres le tableau (3.12) suivant toutes les variables sont significatives, car la valeur de la statistique du test pour chaque variable est supérieure à celle de p-value.

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > Khi-2
activite_der	1	14.0901	0.0002
statut	2	283.8912	<.0001
sect_mod	4	5538.0908	<.0001
generation	1	47.4792	<.0001

TABLE 3.12 – Résultats des tests de significativité des variables du modèle.

Le tableau (3.13) suivant donne les résultats du modèle de Cox :

Lorsque le coefficient est positif, les personnes ayant la caractéristique étudiée ont un risque h plus élevé que les personnes ayant la caractéristique de référence de mettre fin au cumul emploi-retraite. Ainsi, les artisans ont comparativement aux commerçants une probabilité plus grande de sortir du dispositif. Inversement, lorsque le coefficient est négatif, les personnes ayant la caractéristique étudiée ont une plus faible probabilité de mettre fin au cumul emploi-retraite que les personnes ayant la caractéristique de référence.

Le coefficient estimé est négatif pour les personnes salariées avant la liquidation, elles ont une probabilité plus faible de sortir du cumul que les personnes qui étaient indépendantes avant la liquidation de leur retraite du régime général.

Pour une variable quantitative, l'interprétation n'est pas la même : une augmentation d'une unité de la variable quantitative conduit à une variation de $(exp(\beta) - 1)\%$ de risque de survenue de l'évènement. Par exemple, si la génération augmente d'une année, le risque de mettre fin à un cumul emploi-retraite augmente de $2,3\%(exp(0.02270) - 1)$. Ce résultat, $2,3\%$, peut être lu directement dans la colonne hazard ratio.

Le risque de première espèce, correspont au risque de considérer que la variable a un effet sur l'événement étudié alors qu'elle n'en a pas (=rejetter à tort l'hypothèse nulle). Ainsi, si l'on considère que travailler dans l'industrie a un effet sur la durée passée en cumul emploi-retraite, nous aurons 44% de risques de nous tromper. En revanche, nous pouvons affirmer que toutes les autres variables de l'étude ont un effet sur le cumul emploi-retraite avec moins 1% de risques de nous tromper.

Le risque ratio mesure le gain d'occurrence du risque chez les personnes ayant la caractéristique étudiée par rapport à ceux possédant la modalité de référence toutes choses égales par ailleurs. Il correspond aux odds-ratio des modèles de régression. Dans notre étude, le risque de sortir du cumul emploi-retraite est 1,148 fois plus grand pour les personnes travaillant dans le secteur de la construction que pour celles travaillant dans le secteur des services, toutes choses égales par ailleurs. Le risque ratio peut être obtenu à partir des paramètres estimés : $exp(0.13783) = 1.148$.

Les colonnes 9 et 10 sont obtenues grâce à l'option risklimits. Elles donnent l'intervalle de confiance du risque ratio, toutes choses égales par ailleurs, il y a 95% de chances que le risque de mettre fin au cumul emploi-retraite pour les cumulants du secteur de la construction soit entre $1,06$ et $1,24$ fois supérieure à celui des personnes travaillant dans le domaine des ser-

vices.

Analysis of Maximum Likelihood Estimates										
Parameter		DDL	Parameter Estimate	Standard Error	Chi-Square	Pr > Khi-2	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
activite_der	A	1	0.10854	0.02892	14.0901	0.0002	1.115	1.053	1.180	activite_der A
statut	1	1	-0.48478	0.03121	241.2227	<.0001	0.616	0.579	0.655	statut 1
statut	3	1	-0.46797	0.05419	74.5629	<.0001	0.626	0.563	0.696	statut 3
sect_mod	NR	1	2.57490	0.03702	4837.2513	<.0001	13.130	12.211	14.118	sect_mod NR
sect_mod	com	1	0.08962	0.03022	8.7927	0.0030	1.094	1.031	1.161	sect_mod com
sect_mod	const	1	0.13783	0.04065	11.4972	0.0007	1.148	1.060	1.243	sect_mod const
sect_mod	indus	1	0.03585	0.04673	0.5887	0.4429	1.037	0.946	1.136	sect_mod indus
generation		1	0.02270	0.00329	47.4792	<.0001	1.023	1.016	1.030	

TABLE 3.13 – Estimateurs du modèle de Cox et risque-ratio.

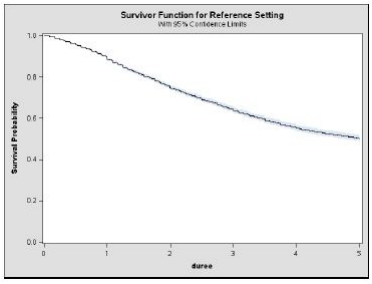

FIGURE 3.5 – Fonction de survie obtenue avec le modèle de Cox.

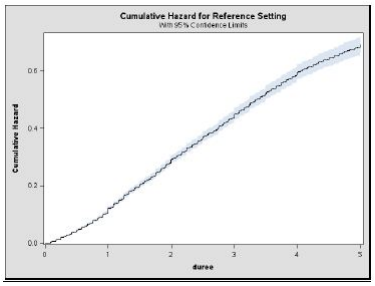

FIGURE 3.6 – fonction de risques cumulés obtenue à partir du modèle de Cox.

Vérification de la forme fonctionnelle des variables continues :

Pour réaliser un modèle de Cox, les variables continues doivent respecter la règle suivante un changement d'une unité dans la variable continue doit avoir le même effet sur l'évènement considéré, et ce qu'elle que soit la valeur. Pour contrôler cette hypothèse, les résidus de Martingale sont utilisés. Ils peuvent être interprétés comme la différence au cours du temps entre le nombre d'évènements observés et le nombre d'évènements prédit par le modèle de Cox. Avec la procédure PHREG, il est plus simple d'utiliser les résidus de Martingale cumulatifs pour vérifier cette hypothèse. Une option permet de représenter un graphique des résidus de Martingale observés en fonction de la variable continue. Sur ce graphique, des simulations de résidus sont réalisées sous l'hypothèse que la variable a une forme adéquate. Il suffit alors d'observer si les résidus cumulatifs observés diffèrent des simulations. Si c'est

le cas, alors la variable continue n'a pas la bonne forme.

Sous SAS, pour effectuer cette vérification, il suffit d'intégrer au modèle de Cox, créé à partir de l'instruction phreg, une instruction assess :

```
prog freg data=cohorte2008 ;
class activité_der(rf="C") statut(ref="2") sect_mod(ref="serv") ;
model durée*censure(0)= activité_der statut sect_mod generation ;
assess var=(generation) resample=50 seed=27513 ;
```

La figure (3.7) suivante indique que la distribution des résidus de Martingale des sorties de cumul emploi-retraite en fonction de la variable generation fait partie des distributions de résidus simulés. Par ailleurs, le test de Kolmogorov conduit à ne pas rejeter l'hypothèse nulle : la variable generation peut donc être considérée comme de la forme adéquate.

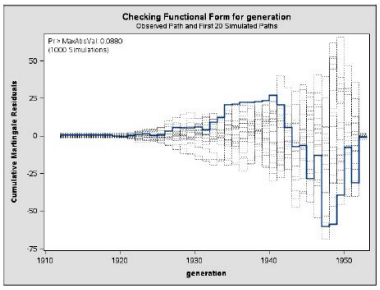

FIGURE 3.7 – Vérification de la forme de la variable generation à partir des résidus de Martingale.

Vérification de l'hypothèse des risques proportionnels :

Pour valider le modèle de Cox, il faut que l'hypothèse des risques proportionnels soit vérifiée : autrement dit, le risque doit être constant au cours du temps, pour chaque variable explicative. Dans notre étude, il s'agit de contrôler cette hypothèse pour les variables sur le secteur d'activité, le groupe professionnel, la situation au moment de la liquidation et la génération.

Il suffit de tracer les courbes log(H(t)). Voici le programme SAS correspondant :

```
Proc lifetest data=cohorte2008 conftype=loglog
Plots=(loglogs) graphics ;
Time duree*censure(0) ;
strata activite_der ;
Run ;
```

Les figure (3.8), (3.9) et (3.10) suivantes indiquent que la variable "**activite_der**", concernant les groupes professionnels, semble plutôt respecter l'hypothèse de risque proportionnels. La courbe représentant le logarithme de la fonction des risques cumulés des artisans est plutôt parallèle à celle des commerçants.

En revanche, les variables concernant la situation avant la liquidation (statut) et le secteur d'activité ne semblent pas respecter l'hypothèse des risques proportionnels, les courbes semblent se rejoindre.

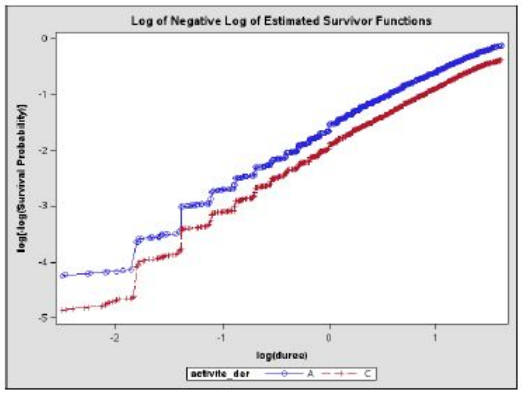

FIGURE 3.8 – Vérification de l'hypothèse des risques proportionnels pour la variable activite_der.

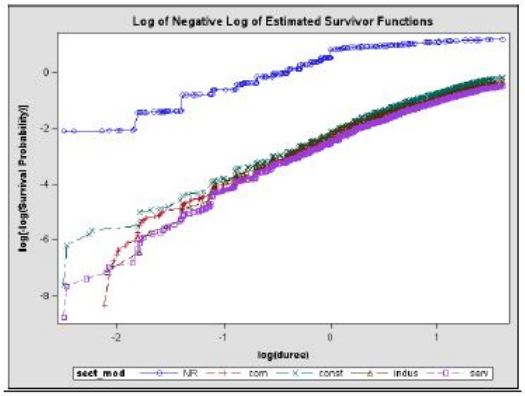

FIGURE 3.9 – Vérification de l'hypothèse des risques proportionnels pour la variable sect_mod.

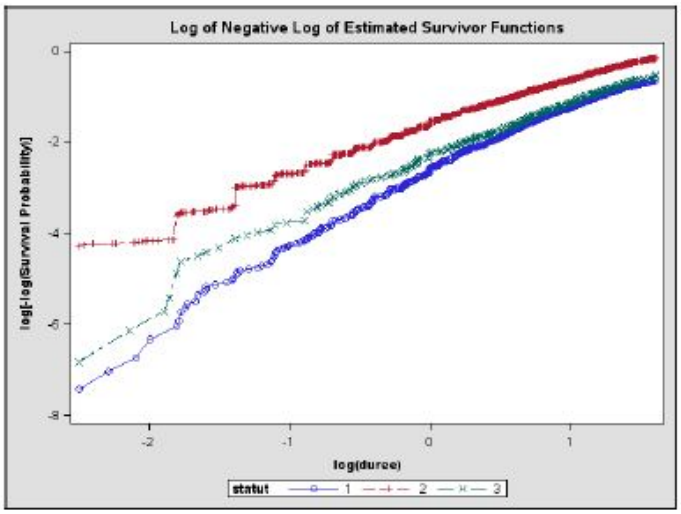

FIGURE 3.10 – Vérification de l'hypothèse des risques proportionnels pour la variable statut.

La proportionnalité n'est pas verifiée d'où nous ne validons pas le modèle. C'est le modèle semi-paramétrique non proportionnel qui conviendra (voir [36]). Nous ne le présentons pas dans ce document.

Remarque 3.2.3.1.

L'inconvénient de cette méthode de vérification est de reposer sur un graphique. En général, les courbes ne sont pas strictement parallèles et il y a donc une part de subjectivité pour considérer qu'elles le sont. Il existe d'autres moyens pour le confirmer comme l'utilisation des résidus du score [1].

Conclusion :

D'aprés les modèles non-paramétriques et le modèle de Cox, plus de la moitié des indépendants reste en cumul pendant au moins 4 ans. La durée du cumul emploi-retraite s'explique principalement par la carrière des individus, plus que par l'activité exercée en parallèle de la retraite.

3.3 Traitement des données du stage au sein de l'hôpital d'Amizour

La deuxième application concerne les données réelles de la durée de survie des patientes souffrant du cancer de sein de type carcinome canalaire infiltrant, l'un des plus dangereux type de cancer. Les données utilisées proviennent de l'hopital d'Amizour plus précisemment du service d'oncologie (traitement des tumeurs). La durée de surveillance est en général de 10 ans mais les médecins adaptent l'intensité et les méthodes de suivi en fonction des facteurs de risque. Les patientes traitées sont à risque par rapport à trois types d'événements : une récidive, des métastases à distances, décès. On considère dans notre étude l'événement "Décès".

On a éffectué un stage de durée d'un mois et les données récoltées concernent la période allant du 01/01/2007 jusqu'à 31/12/2018.

3.3.1 Description des données

L'ensemble des données collectées concernent 228 patientes agées entre 23 et 80 ans à la date d'inclusion à la cohorte. Plus de 300 patientes sont observées chaque année. La date d'origine est 01/01/2007 et la date du point 31/12/2018. La variable d'intérêt est la durée de survie des patientes. Si le décés est réalisé avant la date du point, la survie est la différence entre la date du décés et la date d'origine, sinon on considère l'écart entre la date de point et la date d'origine (les dates d'origines ne sont pas les mêmes pour toutes les patientes). Dans ce cas, les données sont censurées à droite de type I, et sont quantifiées par mois. les variables utilisées dans l'étude sont :

- Var1 : allaitement, elle peut être,
  $\begin{cases} \text{allaité} \rightarrow \text{le code=OA} \\ \text{non allaité} \rightarrow \text{le code=NA} \end{cases}$
- Var2 : la situation, elle peut être,
  $\begin{cases} \text{mariée} \rightarrow \text{le code=Mariée} \\ \text{célibataire} \rightarrow \text{le code=Célibataire} \end{cases}$
- Var3 : la contraception, elle peut être,

$\begin{cases} \text{Prend la pilule, } \rightarrow \text{ le code} = OP, \\ \text{Pas d'utilisation de la pilule } \rightarrow \text{ le code} = NP; \end{cases}$

— var4 : la profession, elle peut être,

$\begin{cases} \text{Travaille } \rightarrow \text{ le code =OT,} \\ \text{Femme au foyer } \rightarrow \text{ le code=NT;} \end{cases}$

— Var5 : le grade du cancer à trois modalités :

$\begin{cases} \text{gradeI } \rightarrow \text{ le code} = G1, \\ \text{gradeII } \rightarrow \text{ le code} = G2, \\ \text{gradeIII } \rightarrow \text{ le code} = G3; \end{cases}$

— Var6 : l'âge

3.3.2 Résumé statistique des données

Le tableau (3.14) suivant donne le nombre d'événements observés (Décès) et le nombre de données censurées, selon le fait que la femme a procédé à l'allaitement ou non, sa situation (mariée ou non), et l'utilisation d'un moyen contraception ou non, situation professionelle, grade de la maladie et sa tranche d'âge.

Variables explicatives	Situation	Décès observés	valeurs Censurées	Total
Var1	OA	89	64	153
	NA	49	26	75
Var2	Mariée	118	75	193
	Célibataire	20	15	35
Var3	OP	65	50	115
	NP	73	40	113
Var4	OT	10	6	16
	NT	128	84	212
Var5	G1	15	12	27
	G2	84	53	137
	G3	39	25	64
var6	< 40	39	23	62
	$40 < 50 < 60$	76	50	126
	≥ 60	23	17	40
Total		138	90	228

TABLE 3.14 – Nombre d'événement obsevées et valeurs censurées.

Interprétation :

D'aprés le tableau (3.14), l'événement de décès est le plus observé pour toutes les variables. On a 183 données observées et 90 données censurées. Le pourcentage des valeurs

observées $60,53\%$ est plus élevé que celui des valeurs censurées $39,47\%$.

Le tableau (3.15) suivant donne les valeurs moyennes, minimales et maximales des données.

Variables explicatives	Situation	Moyenne		Maximum		Minimum	
		Décédé	Censurée	Décédé	Censurée	Décédé	Censurée
Var1	OA	43,97	77,45	127	136	10	19
	NA	31,35	72,12	86	142	1	9
Var2	Célébataire	23,9	65,4	55	101	2	14
	Marier	42,13	78,01	127	142	1	9
Var3	OP	41,86	76	127	136	1	9
	NP	36,82	75,8	86	142	2	14
Var4	OT	49	61,67	74	101	11	19
	NT	38,74	76,93	127	142	1	9
Var5	G1	43,2	75,42	76	139	8	9
	G2	40,12	76,30	127	142	2	14
	G3	36,69	75,32	76	132	1	19

TABLE 3.15 – Valeurs moyens, mininimales et maximales des durées de survie.

Iterprétation :

D'aprés le tableau(3.15) la moyenne des durées censurées est plus élevé par rapport à la moyenne des durées observées des femmes décédées. On remarque aussi que les durées maximales sont plus élevées que les durées minimales.

L'étude de la durée de survie peut être effectuée par trois méthodes différentes : méthode paramétrique, non paramétrique ou semi-paramétrique.

3.3.3 Estimation paramétrique

La figure (3.11) suivante nous donne l'histogramme de la densité de la durée de survie, ajustée par la loi de weibull de paramètres α=1,71, β=60,3, notée W(1,71 ;60,3). On as utilisé le logiciel SAS, pour le tracé de la courbe.

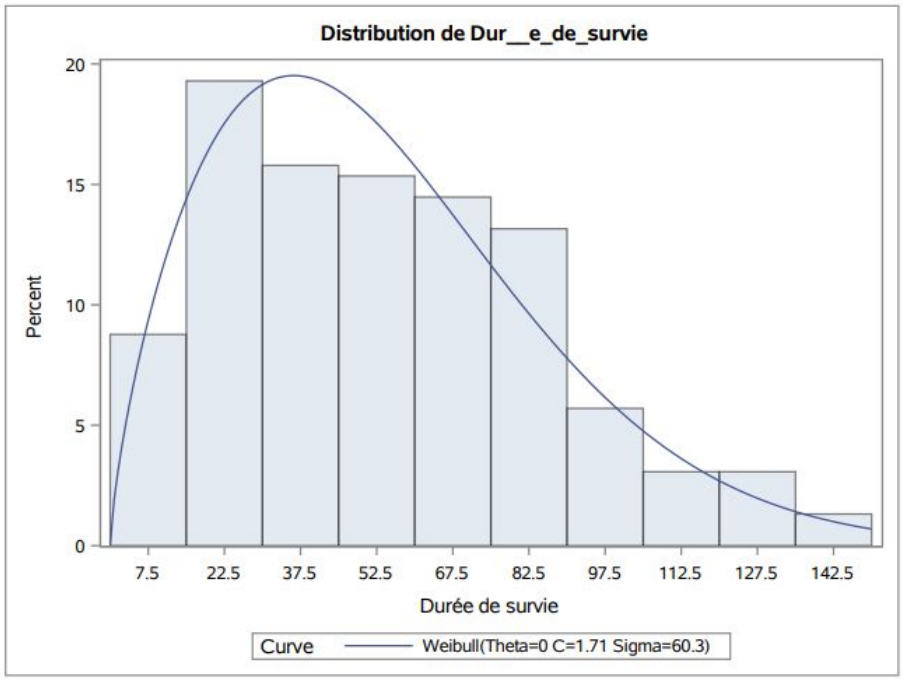

FIGURE 3.11 – Histogramme de densité de la durée de survie.

Le test de Kolmogorove Smirnov permet de valider l'approximation de la durée de survie par la loi de weibull. Les résultats du test sont données dans le tableau 1.11 avec :

KS cal : c'est la valeur empirique de la statistique de khi-deux.

KS th : c'est la valeur tabuleé de la statistique au niveau de signification $\alpha = 5\%$.

N : la taille de de l'échatillon.

Loi	Nombre de classe	KS cal	KS th	N	Paramètres
Weibull	10	7,15	14,07	228	$\alpha = 1,71,\ \beta = 60,3$

TABLE 3.16 – Résultats du test de KS pour l'ajustement des durées de survie.

D'aprés le tableau ci-dessus on remarque que la valeur calculée est inférieure à la valeur tabulée, d'où on accèpte l'hypothèse que l'échantillon suit la loi weibull avec les paramètres $\alpha = 1,71$, et $\beta = 60,3$. Dans ce cas,

Fonction de survie : $S(t) = exp(-(t/60.3)^{1.71}), \quad t > 0$

Fonction de risque : $h(t) = (0.0283)(t/60.3)^{(0.71)}, \quad t > 0$

Espérance=53,8479 mois.

Interprétation :

La figure de risque de weibull(1,71 ;60,3) est une fonction croissante du temps (voir

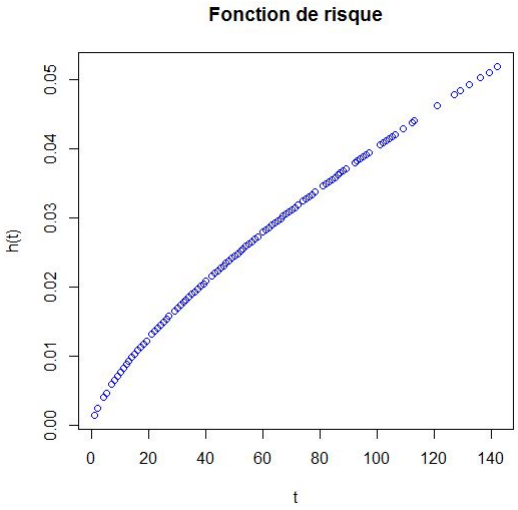

FIGURE 3.12 – Graphe de la fonction de risque pour la weibull.

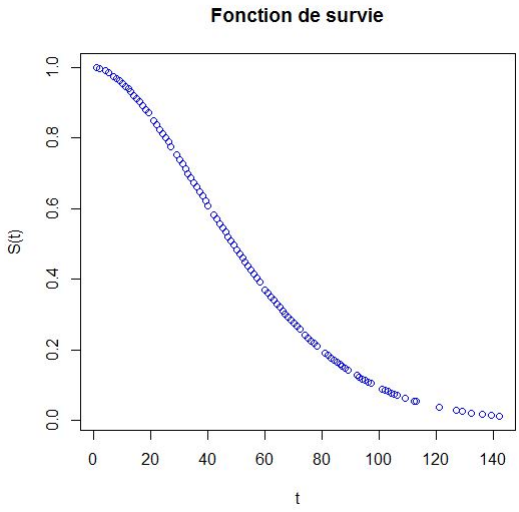

FIGURE 3.13 – Courbe de la fonction de survie de la weibull.

figure 3.12). D'où, le traitement effectué à l'hôpital d'Amizour pour les malades atteintes du cancer de sein semble être non efficace dans le sens où il empêche de diminuer le risque de décés chez ces femmes.

D'aprés le traitement administré au service d'oncologie de l'hôpital d'Amizour, l'espérance de vie chez ces femmes est de 53,8479 mois.

3.3.4 Estimation non-paramétrique

Etant donné que les durées ne sont pas connues avec précision, on applique la méthode d'estimation non-paramétrique Actuarielle, en supposant que la distribution exacte n'est pas connues. Le tracé de la courbe de la fonction de survie est effectué selon les commandes écrites sous SAS.

```
Proc lifetest data=fin outsurv=actu method=act
conftype=loglog
plots=(survival(atrisk cb=hw), hazard, logsurv) graphics;
Time Dur_e_de_survie*D_V(0);
Run ;
```

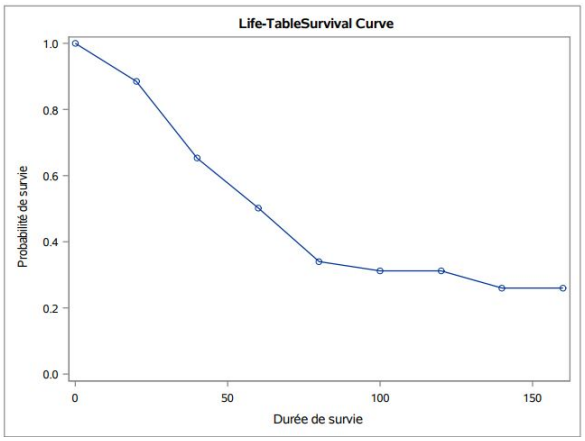

FIGURE 3.14 – Fonction de survie par la méthode actuarielle.

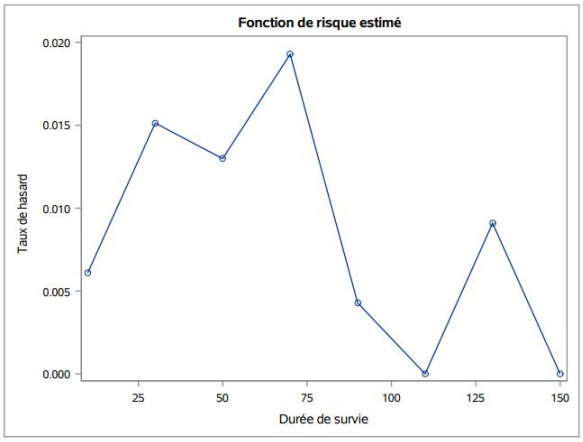

FIGURE 3.15 – Fonction de risque instantané par la méthode actuarielle.

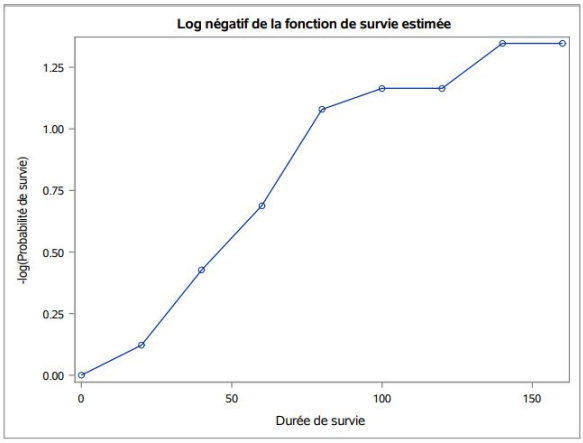

TABLE 3.17 – Fonction de risque cumulés par la méthode actuarielle.

Interprétation :

Les résultats du modèle non-paramétrique montrent que le décés des femmes atteintes du cancer de sein s'étale sur plusieurs mois. Plus de la moitié de la cohorte est décèdée aprés au moins 60mois (≈5ans). Aprés 11 ans (la date de fin de la période d'observation), 30% de la cohorte est toujours en vie. En parallèle, plus d'un tiers de la cohorte décède pendant 40 mois (≈3ans) (voir figure 3.14).

Le risque (voir figure 3.15) de décès est un peu élevé au cours des premiers mois, et au bout de 70 mois (≈6 ans) le risque de décédé est devenu plus élevé et décroit progréssivement jusqu'à ce qu'il devienne nul.

Analyse non paramétrique des sous-populations :

Afin de comparer les survies de différents groupes, nous avons regroupé les durées de

chaque groupe pour déterminer l'effet de chaque variable sur la survie des patientes.

A]-Variable1 : Allaitement :(OA, NA).

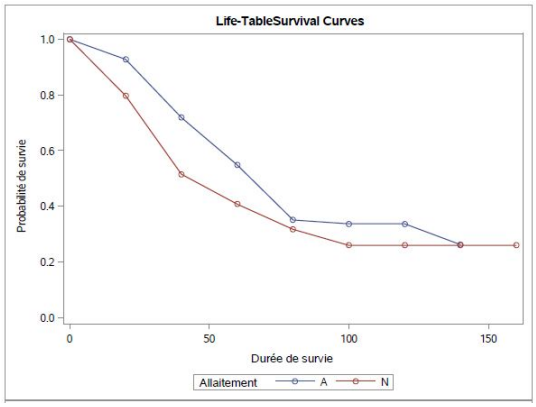

FIGURE 3.16 – Fonction de survie avec la méthode actuarielle stratifiée par l'Allaitement.

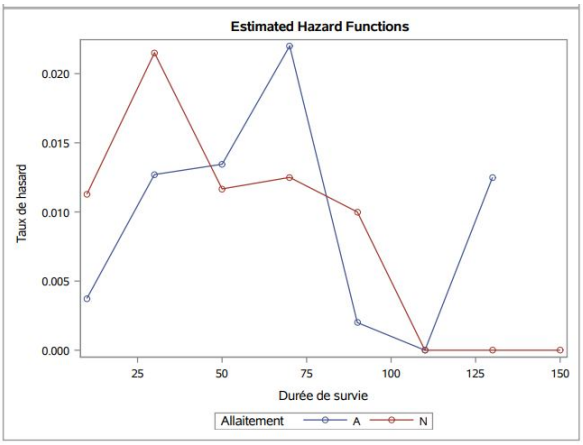

FIGURE 3.17 – Fonction de risque instantané avec la méthode actuarielle stratifiée par l'Allaitement.

Le rouge indique la courbe des femmes NA, le bleu disigne la courbe des femmes OA.

Interprétation :

D'après la figure (3.16), les femmes qui ont allaité ont une durée de survie supérieur à celle des femmes qui n'ont pas allaité. Au bout de 30 mois ($\approx$ 2ans et demi) 25% de ces femmes ont décédées, alors que ce pourcentage est atteint pour les femmes "allaité" aprés 40mois ($\approx$ 3ans).

Avant 50 mois de traitement, le risque des femmes allaité est un peu ; élevé (voir figure 3.17) mais aprés cette période de risque s'élève pour ensuite varier.

Chez les femmes qui n'ont pas allaité le risque diminu aprés 30 mois de traitement.

La statistique de Log-Rank prend la valeur 4,1787, confirme une différence significative entre les deux courbes avec une p-value de 0,0409.

B]-Variable2 : Situation :(Mariée, Célibataire).

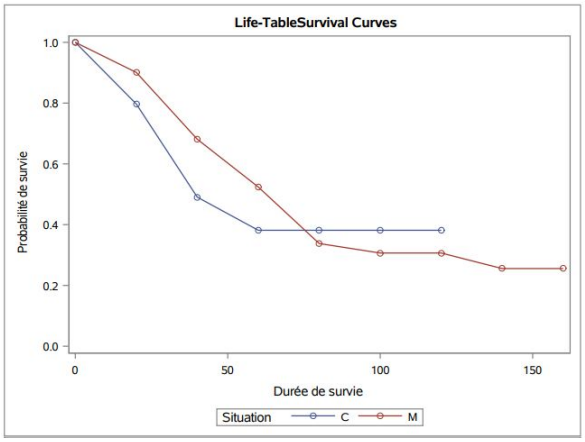

FIGURE 3.18 – Fonction de survie avec la méthode actuarielle stratifiée par la situation.

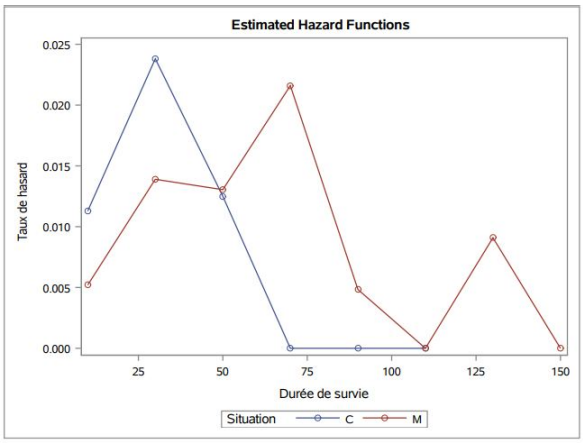

FIGURE 3.19 – Fonction de risque instantané avec la méthode actuarielle tratifiée par la situation.

Le rouge indique la courbe des femmes mariées, le bleu disigne la courbe des femmes célibataires.

Interprétation :

D'après la courbe (3.18), les femmes mariées ont une durée de survie plus grande que celle des femmes célibataires avant les 75 mois ($\approx$ 6ans), et dès qu'on dépasse cette période la situation est inversée. Environs 20% des célibataires ont décédé au bout de 25 mois (2ans) et seulement la moitié de ce nombre a subi cet événement pour les mariées.

D'aprés la courbe (3.19), le risque est trés élevé chez les femmes célibataires mais par la suite il déminu sensiblement à partir de 30 mois pour devenir canstant ($\approx$ 6 ans et demi).

Par contre, chez les femmes mariées le risque est variable. La statistique de Log-Rank prend la valeur 0,8571, confirme une différence significative des deux courbes avec une p-value de 0,3545.

C]-Variable3 : Situation professionnelle :(OT et NT).

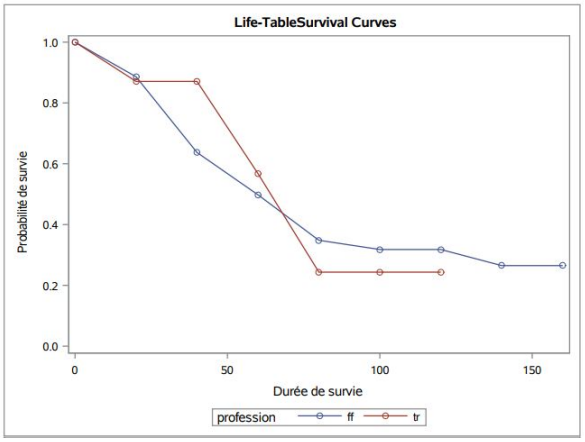

FIGURE 3.20 – Fonction de survie avec la méthode actuarielle stratifiée par la profession.

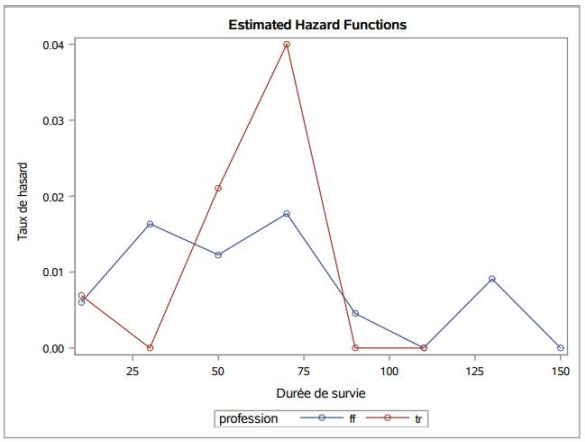

FIGURE 3.21 – Fonction de risque instantané avec la méthode actuarielle tratifiée par la profession.

Le rouge indique la courbe des femmes qui travaille (OT), le bleu disigne la courbe des femmes au foyer (NP).

Interprétation :

D'aprés la courbe (3.20), en moyenne il n'y a pas une grande différence entre les deux courbes de la fonction de survie.

Consernant le risque, il est moins élevé pour les femmes qui travaillent et en croissance avant 75 mois ($\approx$ 6 ans) de traitement, mais à partir de cette date il déminu sensiblement. Pour les femmes au foyer, le risque est trés élevé et évolu de façon variable (voir figure 3.21).

La statistique de Log-Rank prend la valeur 0,3884, ne confirme aucune différence significative entre les deux courbes avec une p-value de 0,8235.

D]-Variable4 : Contraception :(OP et NP).

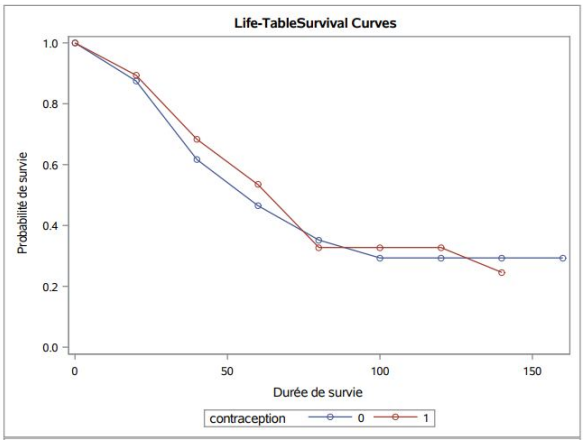

FIGURE 3.22 – Fonction de survie avec la méthode actuarielle stratifiée par la contraception.

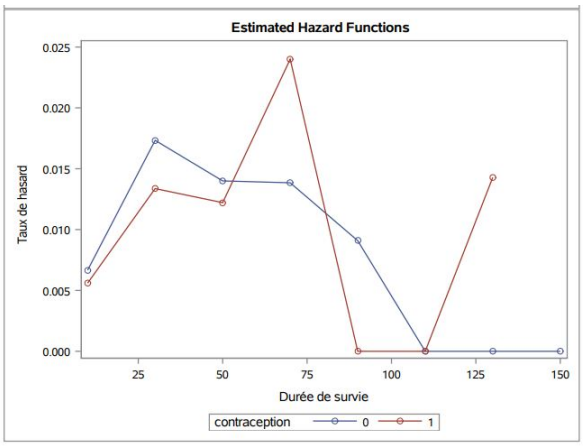

FIGURE 3.23 – Fonction de risque instantané avec la méthode actuarielle tratifiée par la contraception.

Le rouge indique la courbe des femmes ayant utilisée des moyens contraceptifs (OP), le bleu disigne la courbe des femmes n'ayant pas utilisée des moyens contraceptifs (NP).

Interprétation :

D'aprés la courbe (3.22), en moyenne il n'y a pas une grande différence entre les deux courbes de la fonction de survie.

Concernant le risque, il est moins élevé pour les femmes ayant pris des moyens contraceptives et en croissance avant 74 mois ($\approx$ 6ans) de traitement, mais appartir de cette date il déminu sensiblement. pour les femmes n'ayant pas pris des moyens contraceptifs le risque trés élevé et diminu de façon progressive (voir figure 3.23).

La statistique de Log-Rank prend la valeur 0,3884, ne confirme pas une différence significative entre les deux courbes avec une p-value de 0,9735.

E]-Variable5 : Grade :(G1, G2, G3).

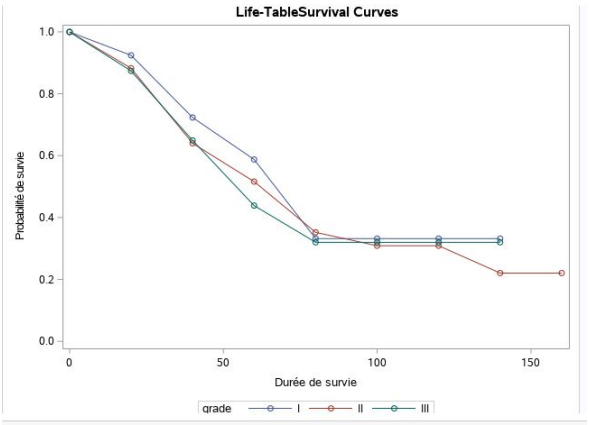

FIGURE 3.24 – Fonction de survie avec la méthode actuarielle stratifiée par grade.

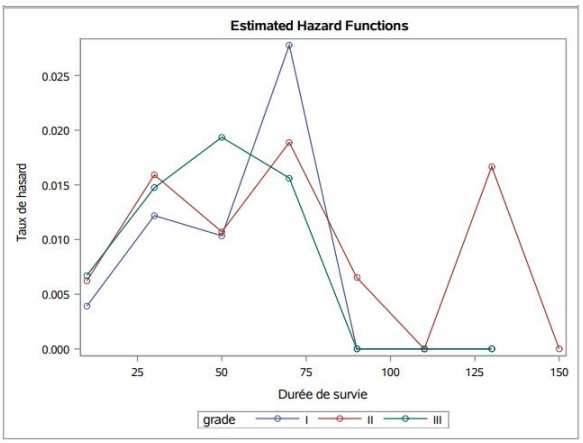

FIGURE 3.25 – Fonction de risque instantané avec la méthode actuarielle tratifiée par grade.

La courbse rouge indique les femmes de grade 2, en vert indique les femmes de grade 3 et en bleu les femmes de grade 1.

Interprétation :

D'aprés la courbe (3.24), en moyenne il n'y a pas une grande différence entre les trois courbes de la fonction de survie.

Concernant le risque, il est moins élevé pour les femmes de grade 1 et en croissance avant 75 mois (≈6ans) de traitement, mais à partir de cette date il diminu sensiblement. Ainsi pour les femmes ayant le grade 3 le risque est très élevé et en croissance avant 50 mois (≈4ans) de traitement, mais à partir de cette date il diminu sensiblement. Pour les femmes ayant le grade 2 le risque est moins élevé et évolu de façon variable (voir figure 3.25).

La statistique de Log-Rank prend la valeur de 0,3884, ne confirme aucune différence significative entre les trois courbes avec une p-value de 0,8235.

F]-Variable6 : âge :(< 40, $[40, 60]$ **ou** $>= 60$).

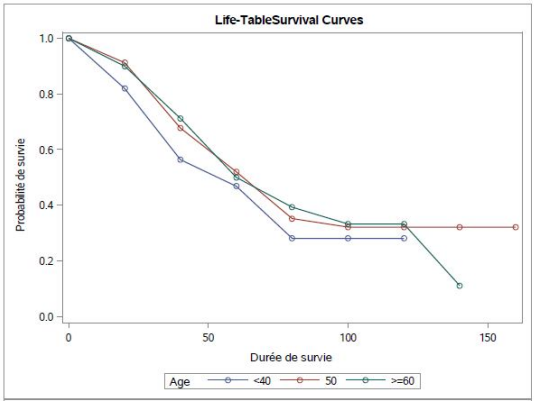

FIGURE 3.26 – Fonction de survie avec la méthode actuarielle stratifiée par l'âge.

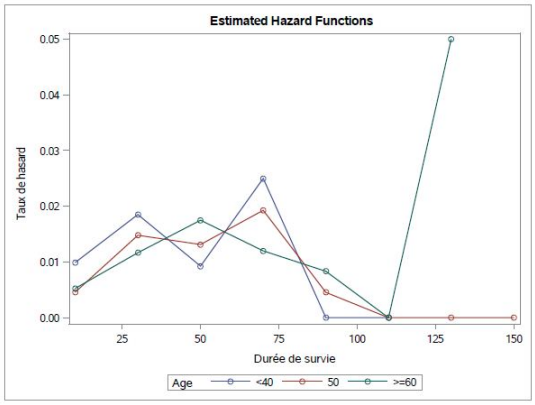

FIGURE 3.27 – Fonction de risque instantané avec la méthode actuarielle stratifiée par l'âge.

La courbe bleu indique les femmes d'une tranche d'âge < 40, en rouge celles de tranche d'âge $[40, 60]$ et en vert les femmes plus de 60 ans.

Interprétation :

D'aprés la courbe (3.26), les femmes moins de 40 ans ont une durée de survie plus petite que les aures. Pour les autres, en moyenne il n'y a pas de grande différence entre les deux courbes de fonction de survie.

Concernant le risque (voir figure 3.27), en moyenne il n'y a pas une grande différence entre les courbes de risque sauf la courbe des femmes plus 60 ans. Elle est en croissance aprés 130 mois (≈10ans) de traitement. La statistique de Log-Rank prend la valeur $2,2341$, confirme une différence significative entre les trois courbes avec une p-value de $0,3273$.

En conclusion :

D'après le test de Log-Rank, on constate que seulement les variables : allaitement, situation et âge qui agissent par rapport au traitement.

3.3.5 Estimation semi-paramétrique

Nous mettons en oeuvre le modèle de Cox, les instructions (sous SAS) utilisées pour le modèle de Cox sont les suivantes :

```
PROC PHREG DATA=fin simple outest=rslCOX plots=(survival, cumhaz) ;
class Allaitement(ref="A") grade(ref="I") Situation(ref="M") profession(ref="tr")
contraception(ref="0") ;
MODEL Dur_e_de_survie*D_V(0)= Age Allaitement grade situation profession
risklimits ;
RUN ;
```

Nous avons obtenu les résultats suivants :
Le modèle est convergent, le critère utilisée est GCONV=1E-8 (voir tableau 3.18).

Etat de convergence
Critère de convergence (GCONV=1E-8) respecté.

TABLE 3.18 – Statut du modèle vis-à-vis de la convergence.

Le meilleur modèle est celui pour lequel le critère d'information d'Akaike (AIC) est le plus faible. Dans notre cas le meilleur modèle est le modèle sans covariables (voir tableau 3.19).

Statistique d'ajustement du modèle		
Critère	Sans covariables	Avec covariables
-2 LOG L	1346.695	1341.944
AIC	1346.695	1349.944
SBC	1346.695	1361.653

TABLE 3.19 – Résultats des critères de qualité du modèle.

Les résultats des tests donnés par le rapport de vraisemblance, Wald et score, de l'hypothèse : H_0 (aucune variable explicative n'apporte de l'information) contre H_1(au moins une variable explicative apporte de l'information), sont résumés dans le tableau (3.20).
D'aprés le tableau, l'hypothèse nulle est rejetée, il ya une variable utile dans le modèle car les valeurs des statistiques des tests sont supérieures à celles des p-values.

Test de l'hypothèse nulle globale : BETA=0			
Test	**khi-2**	**DDL**	**Pr > khi-2**
Rapport de vrais	5.0279	6	0.5402
Score	5.2277	6	0.5150
Wald	5.1838	6	0.5205

TABLE 3.20 – Résultats des tests globaux d'hypothèse nulle.

Pour la signification des variables explicatives, on utilise le test de wald pour lequel : H_0 :"la variable n'est pas significative" et H_1 :"la variable est significative".
D'après le tableau (3.21), la seule variable accepté dans le modèle est la variable Allaitement, parmi les variables testées : Age, Situation contraception, grade. Alors, nous allons procéder à éliminer toutes les variables non significative et nous ne gardons que la variable allaitement dans le modèle.

Tests Type 3			
Effet	**DDL**	**Khi-2 de Wald**	**Pr > khi-2**
Age	1	0.4878	0.4849
Allaitement	1	3.0714	0.0797
Situation	1	0.1731	0.6773
contraception	1	0.0409	0.8398
grade	2	0.4604	0.7944

TABLE 3.21 – Résultats des tests de significativité des variables du modèle.

On réapplique alors le modèle de Cox (voir le tableau (22),(23)). Le meilleur modèle est celui pour lequel le critère d'information d'Akaike (AIC) est le plus faible. Dans notre cas le meilleur modèle est le modèle avec covariables (voir tableau 3.21).

Statistique d'ajustement du modèle		
Critère	**Sans covariables**	**Avec covariables**
-2 LOG L	1346.695	1342.773
AIC	1346.695	1344.773
SBC	1346.695	1347.700

TABLE 3.22 – Résultats des critères de qualité du modèle.

Analyse des valeurs estimées du maximum de vraisemblance										
Paramètre		DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
Allaitement	N	1	0.35982	0.17814	4.0801	0.0434	1.433	1.011	2.032	Allaitement N

TABLE 3.23 – Estimateurs du modèle de Cox et rapport de risque.

Intèrprétation :

D'aprés le tableau (3.23), le coefficient de la variable Allaitement est $0,35982 > 0$ donc le risque h des femmes qui n'ont pas allaité (OA) est plus élévé. Ainsi, elles ont une probabilité plus grande de décés. Le risque de première espèce, correspond au risque d'erreur lorsqu'on considère que les variables ont un effet sur l'événement étudié. Ainsi, si l'on considère qu'une femme non allaité a un effet sur la durée de vie, nous aurons 4% de risques de nous tromper. Le rapport de risque mesure le gain d'occurrence du risque de décés chez les femmes NA par rapport OA, toutes choses égales par ailleurs. Alors, le risque de décés est 1,433 fois plus grand pour les patientes NA que pour OA. Le rapport de risque peut être obtenu à partir du coefficient estimé : $exp(0.35982) = 1.433$. Les colonnes 9 et 10 dans le tableau (3.23) sont obtenues grâce à l'option risklimits. Elles donnent l'intervalle de confiance du rapport de risque, toutes choses égales par ailleurs. Il y a 95% de chance que le risque de décés pour les femmes NA soit entre 1,011 et 2,032 fois supérieure à celui des femmes OA. Les graphiques des fonctions de survie et de risque pour la modalité de référence OA sont les suivants :

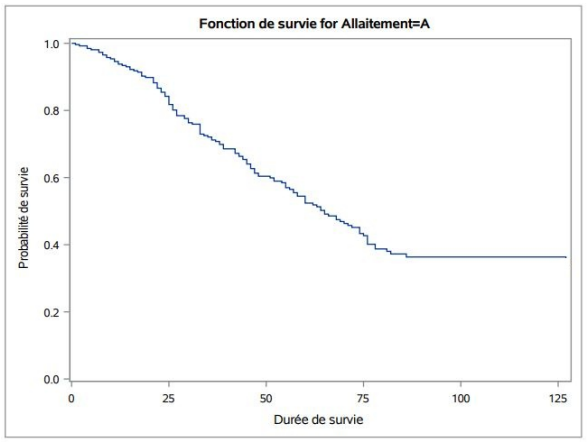

FIGURE 3.28 – Fonction de survie obtenue avec le modèle de Cox.

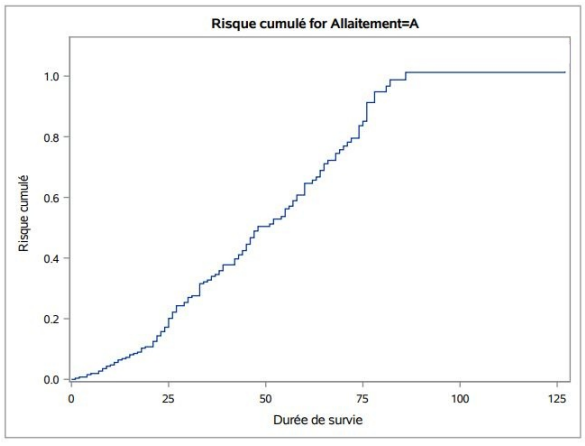

FIGURE 3.29 – Fonction de risques cumulés obtenue à partir du modèle de Cox.

Interprétation :

La courbe de la fonction de survie pour les femmes OA décroit avec le temps et au bout de 85mois ($\approx$ 7 ans) elle devient constante (voir figure 3.28) . Par contre, la courbe de la fonction du risque cumulé augmente avec le temps ensuite constante à partir de 85 mois (voir figure 3.29).

verification de l'hypothèse de proportionnalité :

Afin de valider le modèle il faut vérifier l'hypothèse de proportionnalité des variables. Dans notre cas, il s'agit de contrôler cette hypothèse sur la variable "allaitement".
Sous SAS pour obtenir ce graphique, il suffit de demander le tracé de la courbe $Log(H(t))$ avec la procédure lifetest. Voici le programme SAS correspondant :

```
Proc lifetest data=fin conftype=loglog
plots=(loglogs) graphics ;
Time Dur_e_de_survie*D_V(0) ;
strata Allaitement ;
Run ;
```

Le graphe correspondant est le suivant.

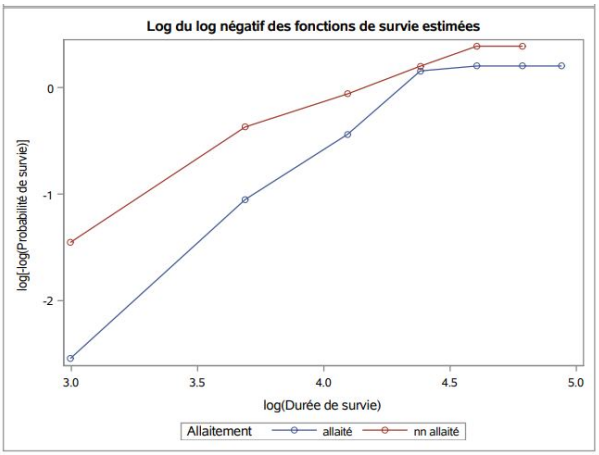

FIGURE 3.30 – Proportionnalité des risques.

Interprétation :

D'aprés la figure (3.30) les courbes sont globalement parallèles d'où la proportionnalité des risque. L'allaitement a un effet sur les femmes atteintes du cancer de sein. Le modèle de Cox confirme que les femmes NA ont une grande probabilité de décés par rapport aux femmes OA, cela est dû a l'activité des hormones dans le sein chez ces femmes. Le risque de décés est alors inférieur chez ces femmes. C'est pourquoi les medecins incitent les femmes à allaiter.

3.4 Conclusion

Les résultats dans les trois méthodes d'estimation appliquées aux données de stage indiquent que le taux de mortalité chez les femmes atteintes du cancer du sein est en augmentation ce qui permet de dire que les traitements effectués au sein du service d'oncologie empêche de diminuer le risque de mortalité. Il n'y a que $39,47\%$ qui répond au traitement.

CONCLUSION GÉNÉRALE

Dans ce mémoire, nous nous sommes intéréssé à l'étude des données de survie. En particulier, à l'estimation de la fonction de survie et la fonction de hasard. Pour cela, nous avons présenté les trois approche d'estimations appliquées à ce types de données, à savoir : la méthode paramétrique, non-paramétrique et semi-paramétrique. Afin d'appliquer ces méthodes, nous avons considéré deux applications différentes.

Dans la première application, nous avons présenté la fonction de survie ainsi que la fonction de hasard correspondants aux données d'assurance retraite [36], en appliquant la méthode non-paramétrique de Kaplan Meier ainsi que le modèle semi-paramétrique de Cox proportionnel.

Dans la deuxième application, nous avons travaillé avec des données du stage que nous avons effectué au niveau du service d'oncologie de l'hôpitale d'Amizour, et nous avons ajusté ces données par le modèle paramétrique de Weibull, ensuite nous avons appliqué la méthode non-paramétrique actuarielle et enfin le modèle semi-paramétrique de Cox proportionnel. Et ce, pour calculer la fonction de survie ainsi que la fonction de hasard.

Bibliographie

[1] D. Aissani and A. Aissani *Méthodes statistiques en fiabilité*. 2005.

[2] C.Alberti, J.-F. Timsit, and S.Checverolet.*Analyse de survie*. Revue des maladies respiratoires, 2005.

[3] L.N.Allen and L.C.Rose.*financial survival analysis of defaulted debors*. journal of the Operationnal research sociéty, 57(6) :6306636,2006.

[4] A.Berchtold,*les données longitudinales et modèles de survie logrank*, Universite données Université de Genève,2017.

[5] J,l Bon. *Fiabilité des susteme* Fialibité des systémes : méthodes mathématiques. Masson, 1995.

[6] K.Doudane. *estimation de la fonction de risque conditionnelle pour des données Markovienne*, PHD thesis, UMMTO, 2013.

[7] V-M, Castonguay,*Modèlisation de la survie relative*, 2004.

[8] G. Colletaz. *Modèles de survie*,notes de cours, document de travail, 2012.

[9] D. Commenges and H. Jacqmin-Gadda. *Modèles biostatistiques l'épidimiologie*, de boeck supervieurn 2015.

[10] F.Corbière, *Modèles de mélange en analyse de survie en présence de données groupées*. PhD thesis, Université Victor Segalan-Bordeaux II, 2007.

[11] A. Dardier. *Durée du cumul RG/RSI* : une application des modèles de durée, 2016.

[12] B. Falissard .*Comprendre et utiliser les statistiques dans les sciences de la vie*. (DEPRECIATED), 2005.

[13] M. Fioc. *Analyse de survie*. Technical report, Ecole doctorale d'astronomie et d'astrophysique d'Ile de France, 2013.

[14] M. Genin. *Introduction à l'analyse de survie*. Technical report, Université de Lille 2,2015.

[15] R. Giorgi. *Estimation dun taux de survie*. Technical report, Aix-Marseille Université, Marseille, France 2016.

[16] K. E. GNEYOU. *Cours d'analyse de survie*. Technical report, Université de Lomé, 2012.

[17] H. Hamisultane. *Initiation au logiciel sas (version 9.1. 3 sous windows)*, 2002.

[18] C. Huber-Carol. *Durées de survie tronquées et censurées*. Journal de la société francaise de statistique, 1994.

[19] E. Leconte. *Contributions à l'analyse statistique des données censurées à droite*. PhD thesis, Ecole Doctorale Mathématiques Informatique et Télécommunications de Toulouse, février 2018.

[20] S. Lemler. *Modèles de durée, analyse de survie*. Technical report, Laboratoire Statistique et Génome, 2012-2013.

[21] J. Lenoir. *analyse de survie*. Technical report, université jules verne 2013.

[22] T. les membres de la Plateforme CEPS. *Méthodes quantitatives d'évaluation des interventions non médicamenteuses*. Technical report, Université de de Montpellier, 2017.

[23] T. Lorino. *Modèles statistiques pour des données de survie corrélées*, PhD thesis, Institut national agronomique paris-grignon-INA PG, 2002.

[24] T. Lorino. *Modèles statistiques pour des données de survie corrélées*, PhD thesis, Institut national agronomique paris-grignon-INA PG, 2002.

[25] A. Mahdia. *L'échantillonnage descriptif amélioré dans les modéles de durée*. PhD thesis, Université de Béjaia, 2015.

[26] J.M. Marion. *Methodes déstimation de durees de vie de contrats d'assurances automobiles*, In DMAS, 2005.

[27] S. Mcguire. *World cancer report 2014*. geneva, switzerland : World health organization, international agency for research on cancer, who press, 2016.

[28] A. Pailhé. *Durée et conditions de retour à l'emploi des mères après une naissance*, Retraite et société, 2012.

[29] R. Perrigot. *La pérennité des réseaux de points de vente* : une approche par l'écologie des populations et les analyses de survie. Recherche et Applications en Marketing (French Edition), 2008.

[30] F. PLANCHET. *Modèles de durée support de cours statistique des modèles paramétriques et semi-paramétriques*. Technical report, institut de science financiere et d'assurance, 2018-2019.

[31] P. Saint-Pierre. *Introduction à l'analyse des dures de survie*, Cours Université Pierre et Marie Curie, 2015.

[32] K. SALLAH. *Méthode actuarielle d'estimation des courbes de survie*, principe, différences avec la méthode de kaplan-meier, 2016.

[33] Staccini. *Analyse de la survie*. Fiche biostatistique, 2011-2012.

[34] B. Stewart and C. Wild. *World cancer report*, Lyon Cedex : IARC Nonserial publication, 2014.

[35] K. M. Tjørve and E. Tjørve. *The use of gompertz models in growth analyses, and new gompertz-model approach* : An addition to the unified-richards family, Plos one, 2017.

[36] F.Yohann. *Introduction aux analyses de survie*. Technical report, Université Nantes 2018.

Résumé

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant la mort pour des organismes biologiques ou le temps restant avant l'échec ou la panne dans les systèmes artificiels.

Nous nous sommes intéréssé à l'étude des données de survie. En particulier, à l'estimation de la fonction de survie et la fonction de hasard. Pour cela, nous avons présenté les trois méthodes d'estimations appliquées à ce types de données, à savoir : la méthode paramétrique, non-paramétrique et semi-paramétrique.

Dans ce travail nous avons présenter les données de survie ainsi que les différentes méthodes d'estimation de la fonction de survie, en les illustrant par des exemples d'applications. Nous avons considéré deux applications : une sur les données d'assurance retraite et une autre sur les données du stage que nous avons effectué au sein de l'hôpitale d'Amizour, dans le service d'oncologie.

Mots clés : analyse de survie, vraisemblance partielle, Kaplan Meier, Cox, taux de hasard, survie.

Abstract

Survival analysis is a branch of statistics that search to model the time remaining before death for biological organisms or the time remaining before failure or failure in artificial systems.

We were interested in the study of survival data. In particular, to the estimation of the function of survival and the hazard function. For this, we presented the three estimation methods applied to this type of data, namely: parametric, non-parametric and semi-parametric methods.

In this work we present the survival data as well as the various methods of estimating the survival function, illustrating them by examples of applications. We considered two applications: one on the retirement insurance data and another on the data of the internship that we carried out in the hospital of Amizour, in the oncology department.

Keywords: analyse de survie, partial likelihood, Kaplan Meier, Cox, hazard rate, survival.