REPUBLIQUE ALGERIENNE DEMOCRATIQUE et POPULAIRE.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.
UNIVERSITE MOULOUD MAMMERI de TIZI-OUZOU

Faculté des Sciences
Département de Mathématiques

Mémoire de Master
en
Mathématiques
Option
Statistique et Probabilité

Thème

Analyse statistique des modèles de survie

Présenté par

Lyasmine HARROUCHE

Devant le jury

Mme.ATIL Lynda	MCA	UMMTO	Présidente
Mr.FELLAG Hocine	Professeur	UMMTO	Rapporteur
Mr.MAMOU Mohammed	MAA	UMMTO	Examinateur

Soutenu le 17/09/2018

Analyse statistique des modèles de survie

Remerciement

Louange à Dieu, le tout-puissant et miséricordieux qui m'a prodiguée le courage et la force à fin de mener à terme mon travail.

Je remercie particulièrement Pr FELLAG Hocine de m'avoir honoré avec son encadrement ,un bon suivi et pour sa disponibilité, non seulement tout au long de mon travail mais aussi durant mon cursus de master.

Mes sincères remerciements vont aux membres de jury pour l'intérêt qu'ils ont porté à mon travail en acceptant de l'examiner et de l'enrichir par leurs propositions et remarques.

Je tiens à remercier tous les membres de l'équipe de formation du master Probabilités et Statistique.

Je remercie également tous les membres de service d'oncologie de l'hopital de Tizi-Ozou pour leur accueil et leur aide durant mon enquête au sein de leur service et de m'avoir donnée tous les moyens pour accomplir mon étude pratique .

Mes vifs remerciements vont également à ma famille ainsi qu'à mes amis qui m'ont toujours soutenue et encouragée et aidée durant ma formation.

Je n'oublie pas de remercier les membres de l'association El Fedjr et les membres du service d'état civil de l'APC d'Ait Aissa Mimoun de Ouaguenoun.

Dédicaces

Je dédie mon travail à mes chers et honorables parents, mes sœurs et mon frère qui m'ont aidée,soutenue et qui ont eu cru à mes capacités de réussir et de les honorer et qui m'ont tenu la main aux moments de mes faiblesses et m'ont poussé à aller en avant.

Table des matières

Introduction générale

Une variable aléatoire est une fonction définie par l'ensemble des résultats d'une expérience aléatoire donnée.
Dans le cas où ces résultats sont définis comme un temps qui s'écoule jusqu'à la survenue d'un éventement particulier, nous parlons de durée de survie.

L'analyse des modèles de survie est une branche de la statistique qui a pris son développement depuis la deuxième guerre mondiale.
La première méthode d'analyse de survie c'est la méthode actuarielle est apparue en 1912
.
Sa première utilisation fut dans le domaine médical .Ensuite elle est devenue une branche indispensable dans de divers domaines comme l'actuariat, les séismes etc $\cdots$.
Dans la fiabilité , la durée de survie est, par exemple, définie comme le temps qui sépare la mise en marche d'une machine de la panne de celle-ci.
Elle est aussi utilisée dans d'autres domaines comme l'économie, les assurance etc$\cdots$.

Le but de ce travail est de présenter un état des lieux sur les modèles de survie qui ont une très grande importance dans la pratique et qui ont résolu différents problèmes.

Diverses approches de recherche fondamentale et empirique sont développées depuis des années sur la survie qui se caractérisent par des données censurées.

Ce travail s'articule sur 04 chapitres. Dans le premier nous allons introduire le concept de survie, les différentes distributions ainsi que les caractéristiques des données de survie. Dans le deuxième chapitre nous discuterons les différentes approches de modélisation paramétriques, semi-paramétriques, non paramétriques et Bayésiennes des modèles de survie. Dans l'avant-dernier chapitre, nous exposerons quelques orientations récentes et review dont nous allons citer quelques travaux de recherche réalisés par certains auteurs .

Nous terminons notre travail par le chapitre d'application qui est la présentation d'un cas pratique basé sur des données réelles et portant sur un problème d'analyse de survie des individus atteints d'un cancer du sein au centre hospitalo-univérsitaire de Tizi-ouzou au niveau du service d'oncologie.

Chapitre 1

Le concept de survie

1.1 Introduction

L'analyse des données de survie est l'étude qui s'intéresse à l'apparition d'un évènement au fil du temps de manière aléatoire sur un individu, dit évènement d'intérêt ou évènement attendu.

Nous définissons une variable d'intérêt par une variable aléatoire qui est la période allant de l'instant du début d'observation jusqu'à l'apparition de l'évènement d'intérêt, dont l'apparition est l'instant de passage irréversible d'un état à un autre de chaque individu. Ceci est illustré par le schéma suivant.

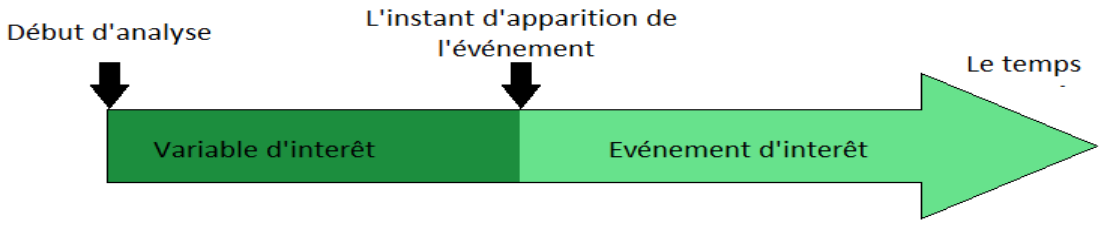


FIGURE 1.1 –

Lorsque cet instant est inconnu, la variable d'intérêt sera définie comme une variable aléatoire notée T dite durée de vie.
Pour prédire et comparer des évènements particuliers, améliorer et mesurer des phénomènes liés au temps, l'étude des modèles de survie est l'un des moyens les plus performants.

Les données de survie sont caractérisées par :
- La prise en considération des observations incomplètes que nous définirons dans la suite.
- Les données de durées sont générées par des variables aléatoires positives ou nulles qui sont interprétées par la fonction de survie et non par la fonction de répartition.

- Les données de survie utilisent des variables explicatives qui influent sur la durée de survie.

Dans plusieurs domaines, nous nous intéressons souvent aux durées de survie particulière. Par exemple :

Dans le domaine médical : dans le but d'améliorer la survie des malades atteints du cancer du sein, nous analysons la durée de survie après le diagnostic, l'évènement d'intérêt est l'observation du décès causé par ce type de tumeur, la variable aléatoire T est la différence entre la date du décès et la date de diagnostic.

Dans la fiabilité : dans le but de mesurer la qualité de certains matériaux comme des ampoules électiques, nous nous intéressons à la durée de fonctionnement celles-ci jusqu'à sa défaillance, T dans ce cas, est la différence entre la date de défaillance de l'ampoule et la date de la mise en marche de celle-ci.

Dans le domaine des assurances : dans le but de mettre en place des moyens nécessaires pour préserver les clients, nous nous intéressons à la durée du contrat des assurances auto, l'évènement d'intérêt est la date de résiliation du contrat donc T est définie comme la différence entre la date de résiliation et celle de création du contrat.

L'analyse des modèles de survie est applicable dans plusieurs domaines, et pour le définir nous gardons le terme de survie.

Dans ce chapitre nous commençons par la définition des distributions de la durée de survie ,par suite nous introduisons quelques notions liées à cette l'analyse , enfin nous allons introduire quelques distributions de probabilité utilisées dans la survie.

1.2 Les distributions de la durée de survie

Supposons qu'au départ nous ayons certains sujets dans un état initial donné. Le passage à l'évènement d'intérêt se produit d'une manière aléatoire, donc nous devons faire appel aux différentes méthodes de calcul de probabilités pour analyser et résoudre les problèmes de survie en tenant compte de la durée de surveillance et le moment où chacun de ces événements sE produit pour chaque individu.

Soit un espace probabilisé $(\mathbb{R}^+, \mathbb{F}, \mathbb{P})$, nous définissons une variable aléatoire T continue, à valeurs réelles positives et représentent la durée de vie qui est le délai entre la date initiale et la date ou un événement particulier se manifeste.

Pour un t fixé, nous définissons les distributions de la survie comme suit :

1.2.1 La densité de probabilité f

La densité de probabilité de T est notée par f définie sur $[0, +\infty[$ avec :

$$\begin{aligned} f(t) &= \lim_{dt \to 0} \frac{p(t \leq T < t + dt)}{dt} \\ &= \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt} \end{aligned} \tag{1.1}$$

1.2.2 Fonction de répartition F

C'est la probabilité de realisation d'un événement avant l'instant t , elle est définie comme suit :

$$F(t) = P(T < t) \tag{1.2}$$

1.2.3 Fonction de survie S

C'est la probabilité que l'événement arrive après l'instant t ,c'est-à-dire survivre après cet instant , elle est définie comme suit :

$$S(t) = P(T \geq t) \tag{1.3}$$
$$S(t) = 1 - F(t) \tag{1.4}$$

$S(t)$ est une fonction monotone décroissante et continue avec $S(0) = 1$ et $\lim_{t \to +\infty} S(t) = 0$

1.2.4 Fonction de risque instantané λ

Elle est aussi ppelée aussi le taux de défaillance ou taux instantané d'avarie en fiabilité. Dans le domaine biomédical elle est dite force de mortalité ou risque instantané de décès.

Si T définit la durée de séjour dans un état donné, $\lambda(t)$ est la probabilité de passage irréversible d'un état à un autre dans un intervalle de temps, $[t, t + dt]$ sachant que le sujet était dans l'état initial jusqu'à la date t,pour $(T > t)$, $\lambda(t)$ fournit la description la plus cohérente d'une distribution de survie.

Elle nous permet de mesurer la vitesse de changement d'état d'un individu à travers le temps. La fonction de risque instantané est définie comme suit :

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \backslash T \geq t)}{dt} \tag{1.5}$$

1.2.5 Fonction de risque cumulé Λ

Le taux de risque cumulé $\Lambda(t)$ sur l'intervalle [0,t]est défini comme suit :

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{1.6}$$

C'est la probabilité de passage à l'évènement d'intérêt dans $[t, t + dt]$ en prenant en considération les informations précédentes .

1.2.6 Relations entre les fonctions

Nous pouvons définir la distribution de probabilité de la variable aléatoire T à partir de l'une des fonctions suivantes S, F, f, λ, Λ , comme suit :

$$
\begin{aligned}
S(t) &= P(T \geq t) \\
&= 1 - F(t) \\
&= 1 - \int_0^t f(u)du \\
&= exp\{-\int_0^t \lambda(u)du\} \\
&= S(t) = -exp\Lambda(t)
\end{aligned}
$$

$$
\begin{aligned}
F(t) &= P(T < t) \\
&= 1 - S(t) \\
&= \int_0^t f(u)du \\
&= 1 + exp \int_0^t \lambda(u)du \\
&= 1 + exp\Lambda(t)
\end{aligned}
$$

$$
\begin{aligned}
f(t) &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt)}{dt} \\
&= \frac{dF(t)}{dt} \\
&= \frac{-dS(t)}{dt}
\end{aligned}
$$

$$
\begin{aligned}
\lambda(t) &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt \backslash T \geq t)}{dt} \\
&= \frac{-d}{dt} ln(S(t)) \\
&= \frac{f(t)}{1 - F(t)} \\
&= \frac{d}{dt}\Lambda(t)
\end{aligned}
$$

$$
\Lambda(t) = \int_0^t \lambda(u)du = -ln(S(t)
$$

1.2.7 Quantités associées à la distribution de survie

a) Les paramètres de position de la distribution de survie

- Le temps moyen de survie $E(T)$

$$E(T) = \int_0^{+\infty} tf(t)dt \tag{1.7}$$

Puisque $T \geq 0$, alors :

$$\begin{aligned} E(T) &= \int_0^{+\infty} p(T > t)dt & (1.8)\\ &= \int_0^{+\infty} 1 - F(t)dt & (1.9)\\ &= \int_0^{+\infty} S(t)dt & (1.10) \end{aligned}$$

- La variance de la durée de survie $V(T)$

$$V(T) = E(T^2) - (E(T))^2 \tag{1.11}$$

En utilisant l'intégration par partie nous obtenons :

$$V(T) = 2\int_0^{+\infty} tS(t)dt - (E(T))^2 \tag{1.12}$$

Rmarque : nous pouvons déduire l'espérance et la variance de la durée de survie à partir de n'importe laquelle des fonctions vues précédemment.

b) Les paramètres de dispersion de la distribution de survie

- Les quantiles de la durée de survie pour $0 \leq p \leq 1$ **sont définis par**

$$\begin{aligned} q(p) &= \inf(t \backslash F(t) \geq p)\\ &= \inf(t \backslash S(t) \leq 1 - p) \end{aligned}$$

Dans le cas où F est strictement croissante et continue, alors :

$$q(p) = F^{-1}(p) = S^{-1}(1 - p) \tag{1.13}$$

Le quantile d'ordre p est le temps où une proportion d'une population a subi l'événement.

- La médiane Dans le cas où la moitiée de la population a subi l'événement nous le définissons par un quantile particulier qui est la médiane définie par la valeur t_m satisfaisant $S(tm) = 0.5$

1.3 Les caractéristiques des données de survie

1.3.1 Les dates et délais principaux

Dans l'étude de survie, nous disposons d'un échantillon de n individus ou bien sujets qui possède des informations principales sur leurs dates et les différentes durées de survie.

Les principales dates

1- Date d'origine DO C'est la date de début de l'observation, c'est-à-dire l'origine du début de l'analyse de survie ; elle correspond au temps égal à zéro $t = 0$.

2- Date des dernières nouvelles DDN Pour faire l'analyse des résultats, chaque individu dispose d'une date des dernières nouvelles, qui est la date la plus récente où les informations ont été recueillies.

Par exemple, dans le cas où le sujet a subi l'événement, sa date de dernière nouvelle est la date de survenue de cet événement.

3- Date de point DP Nous ne pouvons pas attendre la survenue de l'évènement pour tous les sujets pour faire l'analyse des observations, donc il y a ce que nous appellons la date de point qui est une date au-delà de laquelle nous arrêtons l'observation et nous ne tenons plus compte de l'état du sujet après cet instant.

Les principaux délais

1- Durée de surveillance C'est la durée entre la date initiale et la date des dernières nouvelles, c'est la période de surveillance de l'instant de la survenue de l'événement d'un sujet i avec $T_i \in [DO, DDN]$.

2- Recul Noté par L_i la variable de la censure d'un individu i avec $L_i \in [DO, DP]$. c'est le délai écoulé entre la date d'origine et la date de point c'est-à-dire,le délai maximal d'observation du sujet.Il est dit aussi le délai de la censure ; il peut être connu ou inconnu, déterministe ou aléatoire, dépendant de l'individu ou non.

3- Le temps de participation Noté par t_i ,c'est le délai réellement observé pour chaque individu i. noté t_i c'est une variable aléatoire qui dépend de la date de point et la date des dernières nouvelles. Nous distinguons deux cas :

a- La date des dernières nouvelles est inférieure à la date du point

$$DDN < DP \Leftrightarrow t_i \in [DO, DDN]$$

Dans ce cas, le temps de participation est identique à la durée de surveillance ie :

$$t_i = T_i$$

Si le sujet n'a pas subi l'événement à la date des dernières nouvelles, il sera considéré comme perdu de vue, le temps de participation nous donnera une observation incomplète or que dans le cas où l'individu a subi l'événement à la date des dernières nouvelles, le temps de participation nous donnera une observation complète de son état.

b- La date des dernières nouvelles est supérieure à la date du point

$$DDN > DP \Leftrightarrow t_i \in [DO, DP]$$

Dans ce cas le temps de participation est identique au temps de recul :

$$t_i = L_i$$

Le sujet est considéré comme exclu vivant, car s'il n'a pas encore subi l'événement avant la date de point, c'est son état initial qui sera pris en considération même s'il y a un changement d'état après la date de point.

Remarque :

Les sujets perdus de vue et les sujets exclus-vivants correspondent aux observations censurés mais les deux mécanismes de censure sont de natures différentes que nous expliquerons en détail prochainement.

Nous résumons dans le schéma suivant, les différentes dates et les durées de survie vues dans cette section.

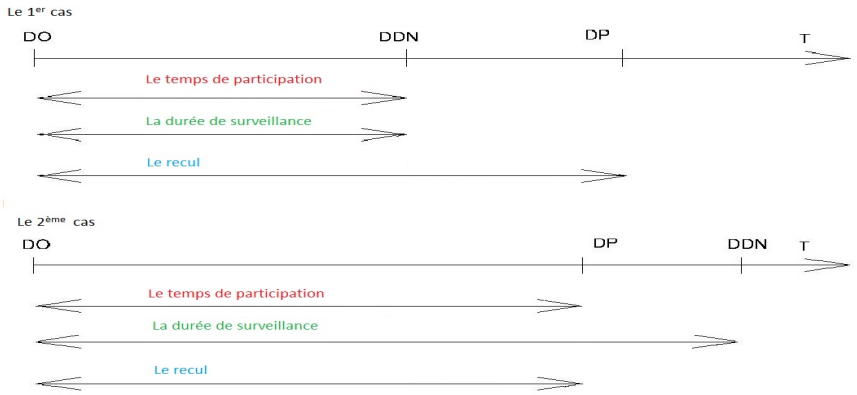


FIGURE 1.2 – les différentes dates et les durées de survies

1.3.2 Données censurées

Parmi les caractéristiques de l'étude de survie, nous considérons les données censurées. Soit un échantillon de n individus.Pour chaque individu i nous associons :

- L_i : Délai (temps) de la censure.
- T_i : Variable aléatoire qui associe à l'individu i son temps de survenue de l'événement $T_i \geq 0$.
- t_i : Durée réellement observé (Temps de participation) .
- d_i : La variable de la censure définie comme une variable indicatrice avec :

$$d_i = \begin{cases} 1 & \text{si l'individu a subi l'événement} \\ 0 & \text{si l'individu n'a pas subi l'événement.} \end{cases}$$

Puisque nous prenons en considération les données censurées, au lieu d'observer les T_i, nous observons les t_i qui dépendent du type d'observation, donc nous pouvons présenter les observations par n couples aléatoires (t_i, d_i).

Définitions des observations et la censure

L'échantillon à étudier est composé de deux types d'observations ou informations, à savoir :

1 - Observation incomplète Quand le sujet n'a pas subi l'événement à la date des dernières nouvelles (DDN), nous disposons d'une information incomplète, c'est le cas des données censurées où $d_i = 0$.

2 -Observation complète C'est l'information complète de l'individu, la durée de vie est totalement observée, le sujet a subi l'événement à la date des dernières nouvelles (DDN) il s'agit donc des données non censurées, correspondant au cas où $d_i = 1$.

Détermination des observations(t_i, d_i)

A fin de déterminer le couple aléatoire (t_i, d_i), nous devons prendre en considération les trois cas suivants :

1- Données censurées à droite La date des dernières nouvelles est ultérieure à la date de point, la durée de survie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière nouvelle. nous savons que la durée de vie du sujet est supérieure à une certaine valeur connue.
nous savons T_i si $T_i \leq L_i$ si non, nous savonst uniquement que $T_i > L_i$ donc nous observons les t_i avec

$$t_i = inf\{T_i, L_i\}$$

$$d_i = \begin{cases} 0 & \text{si } T_i > L_i \\ 1 & \text{si } T_i < L_i \end{cases}$$

Donc l'observation est de la forme : $(t_i, d_i) = \begin{cases} (L_i, 0), & \text{si } T_i > L_i \\ (T_i, 1) & \text{si } T_i < L_i \end{cases}$

2-Données censurées à gauche La durée de vie est dite censurée à gauche si la valeur exacte de l'observation n'est pas connue avant Li, c'est-à-dire que l'individu a subi l'événement avant qu'il soit observé ; dans ce cas nous savons uniquement que la date de l'événement est inferieure à une certaine date connue, donc pour chaque individu nous associons le couple aléatoire (t_i, d_i) avec :

$$t_i = \sup\{T_i, L_i\}$$

$$d_i = \begin{cases} 0 & \text{si } T_i < L_i \\ 1 & \text{si } T_i > L_i \end{cases}$$

Donc l'observation est de forme : $(t_i, d_i) = \begin{cases} (L_i, 0), & \text{si } T_i < L_i \\ (T_i, 1) & \text{si } T_i > L_i \end{cases}$

Exemple :

On considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger, sachant que les babouins passent la nuit sur les arbres.

Le temps d'événement, qui est textbf la descente de l'arbre, est observé si le babouin descend de l'arbre après l'arrivée des observateurs.

Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas nous savons uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs.

Nous observons donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

3-Données censurées par intervalle Soit l'intervalle $[l_i, L_i]$, donc,

$$t_i = sup\{inf\{T_i, L_i\}, l_i\}$$

Au lieu d'observer avec certitude le temps de surveillance de l'évènement, la seule information disponible est que l'évènement a eu lieu entre deux dates connues.

Modèles de censure

Plusieurs mécanismes peuvent être envisagés, en particulier les L_i peuvent être fixés à priori ou bien aléatoirement.

N ous avons dans le cas de la censure à droite les trois modèles suivants :

1- Censure non aléatoire (censure de typeI) Dans ce cas, les L_i sont fixés c'est-à-dire que $L_i = L$,à chaque individu a son temps de survie T_i, la censure n'est pas aléatoire, alors que le nombre de sujets ayant subi l'événement ainsi que leur durée de survie sont aléatoires . Au lieu d'observer les variables aléatoires $T_1, T_2, T_3, \ldots, T_n$, nous observons les T_i sauf si $T_i < L$,si non, on sais uniquement que $T_i > L$.

2- Censure aléatoire(censure de typeIII) Dans ce cas les L_i sont aléatoires, à chaque individu i nous associons son délai de survie T_i et son temps de censure L_i. Supposons que les T_i et L_i sont des variables aléatoires iid, donc la censure aléatoire nous donne les observations qui peuvent être résumées comme suit :

- Perte de vue.
- Arrêt de l'expérience à cause de faits indésirables inattendus.
- Fin d'études alors que certains individus n'ont pas encore subi l'événement ; nous parlons dans ce cas des exclus vivants.

3- Autres mécanismes (censure de typeII) Si le nombre de sujets qui doivent subir l'événement est limité à r individus, on observe les durées de survie de n individus jusqu'à ce que r d'entre eux subissent l'événement, nous arrêtons l'observation à ce moment. Soit la statistique d'ordre suivante : $T_{(1)}, T_{(2)}, T_{(3)}, \ldots, T_{(n)}$, la date de la censure est donc $T_{(r)}$ avec

$$\begin{aligned} T_{(1)} &= t_{(1)} \\ T_{(2)} &= t_{(2)} \\ &\vdots \\ T_{(r)} &= t_{(r)} \\ T_{(r)} &= t_{(r+1)} \\ T_{(r)} &= t_{(r+2)} \\ &\vdots \\ T_{(r)} &= t_{(n)} \end{aligned}$$

1.4 Quelques distributions de probabilités utilisées dans la survie

Supposons que la durée de survie appartienne à une famille de lois paramétriques donnée. En spécifiant la forme des courbes de risque instantané λ, il existe :

- Un risque instantané constant comme la loi exponentielle.
- Un risque instantané monotone comme la loi de weilbull, gamma, loi de weilbull exponenciée et le mélange de deux lois exponentielles.
- Un risque instantané sous forme d'une cloche ou baignoire comme la loi de weibull

généralisée, la loi log logistique, log normale et gaussienne inverse.
Dans ce qui suit, nous détaillerons les modèles exponentiel, gamma, loi de weilbull ainsi que la loi de weilbull généralisée.

1.4.1 Le modèle exponentiel

Définition

Si on considère des systèmes qui ne possèdent pas de période de jeunesse ni de vieillesse, c'est-à-dire dans le cas d'absence de mémoire, la fonction de risque instantané $\lambda(t)$, ne dépend pas de temps t, elle est constante.

$$\lambda(t) = \lambda$$

les distributions de survie correspondantes

En utilisant les relations entre les fonctions de survie vue dans le paragraphe 2, nous déduisons les distributions de survie comme suit :

$$\begin{aligned} \Lambda(t) &= \lambda t \\ S(t) &= \exp\{-\lambda\} \\ F(t) &= 1 - S(t) \\ &= 1 - \exp\{-\lambda\} \end{aligned}$$

Donc la variable aléatoire de durée de vie T a pour densité de probabilité :

$$f(t) = \begin{cases} \lambda \exp\{-\lambda t\} & \text{si } T \geq 0 \\ 0 & \text{si non} \end{cases}$$

D'où T suit une loi exponentielle de paramètreλ

$$T \sim \exp(\lambda)$$

Donc lorsque $\lambda(t)$ est une constante positive, la durée de vie suit une loi exponentielle.

Les quantités associés à cette distribution

La moyenne de vie est$\mathbf{E}(T) = \frac{1}{\lambda}$
La variance de vie est $\mathbf{V}(T) = \frac{1}{\lambda^2}$

1.4.2 La loi gamma $\Gamma(\theta, a)$

Définition

C'est une loi qui est beaucoup plus utilisée dans l'approche Bayésienne, elle représente la loi conjuguée naturelle de la loi exponentielle de paramètre λ.

On définie $\Gamma(a)$ par :$\Gamma(a) = \int_0^{+\infty} u^{a-1} \exp\{u\} du$

Les distributions de survie correspondantes

Soit $(\theta, a) > (0, 0)$

$$\begin{aligned} f(t\backslash\theta, a) &= \theta^a \Gamma(a)^{-1} t^{a-1} \exp\{\theta t\} \\ F(t\backslash\theta, a) &= \Gamma(a)^{-1} \int_0^{\theta t} u^{a-1} \exp\{-u\} du \\ \lambda(t\backslash\theta, a) &= \frac{f(t\backslash\theta, a)}{1 - F(t\backslash\theta, a)} \end{aligned}$$

Si $a = 1$ nous trouvons la loi exponentielle $\Gamma(\theta, 1) = \exp(\theta)$
Si $a > 1$ le risque instantané est croissant.
Si $0 < a < 1$ le risque instantané est décroissant.

Les quantités associées à cette distribution

La moyenne de vie est :$\mathbf{E}(T\backslash\theta, a) = \frac{a}{\theta}$
La variancede vie est :$\mathbf{V}(T\backslash\theta, a) = \frac{a}{\theta^2}$

1.4.3 Loi de weibull $W(\alpha, \lambda)$

Définition

C'est la loi la plus populaire, qui généralise la loi exponentielle et pour laquelle le risque instantané est une puissance du temps.
C'est une loi à deux paramètres notée $W(\alpha, \lambda)$

Les distributions de survie correspondantes

Pour $(\lambda, \alpha) > (0, 0), t \geq 0)$

$$\begin{aligned} S(t\backslash\alpha, \lambda) &= \exp\{-(\lambda t\lambda)^\alpha\} \\ \lambda(t\backslash\alpha, \lambda) &= \alpha(\lambda)^\alpha t^{\alpha-1} \end{aligned}$$

Lorsque $\alpha = 1$ on trouve la loi exponentielle $W(1, \lambda) = exp\{\lambda\}$
Si $0 < \alpha < 1$, le risque instantané est décroissant
Si$\alpha > 1$, le risque instantané est croissant

$$f(t\backslash\alpha, \lambda) = (\alpha\lambda)^\alpha t^{\alpha-1} \exp\{-(\lambda t)^\alpha\}$$

Les quantités associées à cette distribution

La moyenne de vie est :$\mathbf{E}(T\backslash\alpha, \lambda) = \frac{1}{\lambda}\Gamma(1 + \frac{1}{\alpha})$
La variance de vie est :$\mathbf{V}(T\backslash\alpha, \lambda) = (\frac{1}{\lambda})^2(\Gamma(1 + \frac{1}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha}))$

Chapitre 2

Modélisation de la survie

2.1 Introduction

Comme dans toute analyse statistique, pour décrire les observations, fournir les estimations des paramètres pertinents, comparer la survie de plusieurs groupes d'individus, expliquer et prédire la durée de survie en fonction de certain facteurs , les différentes approches de modélisation paramétrique, semi-paramétrique, non paramétrique et Bayésienne sont les moyens les plus efficaces pour répondre à ces objectifs.

Dans ce qui suit nous allons introduire le modèle de survie exponentiel pour l'estimation paramétrique, le modèle de Kaplan-Meier pour l'estimation non paramétrique de la fonction de survie, et le modèle de Cox, ainsi que l'estimation Bayésienne.

2.2 Méthodes paramétriques

2.2.1 Définition

Le modèle paramétrique de survie est un modèle dans lequel la fonction risque $\lambda(t)$ dépend d'un ou plusieurs paramètres, il permet d'évaluer les effets propres de chaque facteur sur la survie et estimer de façon explicite les différents paramètres de la distribution des données de survie.

Il existe plusieurs modèles paramétriques, comme le modèle exponentiel, le modèle de Weibull , le modèle de Pareto ,le modèle Log-normal etc ...

Le modèle exponentiel est beaucoup plus utilisé et développé, c'est le seul modèle qui a un risque instantané constant, son avantage est que l'estimatoin de son paramètre s'obtiennent facilement par la méthode du maximum de vraisemblance.

Cette méthode est aussi applicable sur les autres modèles paramétriques comme le modèle de Weibull mais maximiser la log-vraisemblance ne se fait pas d'une manière explicite ,il faut donc utiliser d'autres méthodes d'estimation .
Dans ce qui suit nous travaillerons sur le modèle exponentiel.

2.2.2 La fonction de vraisemblance des données de survie

Soit la densité de probabilité $f_T(t_i)$ associée à la variable aléatoire T_i iid et $S_T(t_i)$ la fonction de survie correspondante.

La distribution du couple d'observation (t_i, d_i) peut s'écrire

$$\Phi(t_i, d_i) = f_T(t_i)^{d_i} S_T(t_i)^{1-d_i} \tag{2.1}$$

La fonction de vraisemblance correspondante est :

$$\prod_{i=1}^{n} \Phi(t_i, d_i) = \prod_{i=1}^{n} f_T(t_i)^{d_i} S_T(t_i)^{1-d_i}$$

Soit un échantillon de n individus. Pour déterminer la fonction de vraisemblance, nous distinguons deux cas :

1- Cas d'une observation complète

Supposons qu'à partir du $t = 0$, nous avons n individus observés jusqu'à ce qu'ils subissent tous l'événement c'est-à-dire $d_i = 1$.
Leurs instants de changement d'état sont : $(t_1, t_2, \cdots, t_n)$,nous obtenons la densité jointe de s paramètres inconnus $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$
Les T_i sont iid, dans ce cas la vraisemblance est de forme :

$$L(\theta, t_1, t_2, \cdots, t_n) = L(t_i, \theta) = \prod_{i=1}^{n} f_T(t_i, \theta) \tag{2.2}$$

2- Cas des données censurées

Si nous avons un échantillon de n individus dont m parmi eux ont subi l'événement avant la fin de l'expérience $(d_i = 1)$,et les $(n - m)$individus restants non, $di = 0$.
Dans le cas des données censurées à droite, comme vu précidament.Nous avons tois types :

- Cas de la censure aléatoire et non aléatoire La vraisemblance est :

$$L(\theta, t_1, t_2, \cdots, t_n) = L(t_i, \theta) = \prod_{i=1}^{n} f_T(t_i, \theta)^{d_i} S_T(t_i, \theta)^{1-d_i} \tag{2.3}$$

Une notation alternative qui sera utile prochainement EST donnée par :

$$L(\theta, t_1, t_2, \cdots, t_n) = L(t_i, \theta) = \prod_{i \in D} f_T(t_i \backslash \theta) \prod_{i \in L} S_T(t_i \backslash \theta) \tag{2.4}$$

Avec,
D : représente les individus dont la durée de vie est observable.
L : représente les individus dont la durée de vie est censurée.

- Cas de la censure de type II Nous avons

$$L(\theta, t_1, t_2, \cdots, t_n) = L(t_i, \theta) = C_m^n \prod_{i=1}^{n} f_T(t_i) S(t_m)^{n-m} \tag{2.5}$$

avec : $C_m^n = \frac{n!}{m!(n-m)!}$

2.2.3 Estimation du paramètre λ dans le cas du modèle exponentiel

Rappelons la distribution de la variable de survie exponentielle

$$S(t) = exp\{-\lambda t\}$$

$$f(t) = \lambda exp\{-\lambda t\}$$

Soit un échantillon de n individus à étudier .Afin d'estimer le paramètre inconnu λ , nous devons calculer la fonction du maximum de vraisemblance.
Pour ceci, nous distinguons deux cas ; le cas où les observations sont complètes et le cas des données censurées.

1- Cas où toutes les données sont observées

La vraisemblance est définie comme suit :

$$\begin{aligned} L(\lambda, t_1, t_2, \cdots, t_n) &= \prod_{i=1}^{n} f_T(t_i, \lambda) \\ &= \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) \\ &= \lambda^n \exp\{-\lambda(\sum_{i=1}^{n} t_i)\} \end{aligned}$$

Par passage au logarithme nous obtenons

$$\begin{aligned} \ln(L(\lambda, t_1, t_2, \cdots, t_n)) &= l(\lambda, t_1, t_2, \cdots, t_n) \\ &= \ln(\lambda^n \exp\{-\lambda(\sum_{i=1}^{n} t_i)\}) \\ &= n \ln(\lambda) - \lambda \sum_{i=1}^{n} t_i \end{aligned}$$

Calcul de la première dérivée

$$\frac{\partial}{\partial\lambda}l(\lambda,t_1,t_2,\cdots,t_n)=\frac{n}{\lambda}-\sum_{i=1}^{n}t_i$$

Nous allons résoudre l'équation suivante afin de déterminer $\hat{\lambda}$

$$\frac{n}{\lambda}-\sum_{i=1}^{n}t_i=0$$

Donc

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}t_i}$$

La deuxième dérivée est :

$$-\frac{n}{\lambda^2}<0,\forall\lambda>0$$

D'où $\hat{\lambda}$ est un estimateur du maximum de vraisemblance qui est égale à l'inverse de la moyenne des observations.

2- Cas des données censurées

Si nous observons m individus qui ont subi l'événement parmi l'échantillon, la vraisemblance d'un décès observé à l'instant t_i est $f(t_i)$ et la vraisemblance d'une donnée censurée à droite est de $S(t_i)$:

$$\begin{aligned}
L(\lambda,t_1,t_2,\cdots,t_n) &= \prod_{i\in D}f_T(t_i,\lambda)\prod_{i\in C}S_T(t_i,\lambda)\\
&= \prod_{i=1}^{m}f_T(t_i,\lambda)\prod_{i=m+1}^{n}S_T(t_i,\lambda)\\
&= \prod_{i=1}^{m}\lambda\exp\{-\lambda t_i\}\prod_{i=m+1}^{n}\exp\{-\lambda t_i\}\\
&= \lambda^m\exp\{-\lambda\sum_{i=1}^{n}t_i\}
\end{aligned}$$

Par passage au logarithme nous obtenons

$$l(\lambda,t_1,t_2,\cdots,t_n)=m\ln\lambda-\lambda\sum_{i=1}^{n}t_i$$

La première dérivée est :

$$\frac{\partial}{\partial\lambda}l(\lambda, t_1, t_2, \cdots, t_n) = \frac{m}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

Nous obtenons l'estimateur

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}$$

La deuxième dérivée est :

$$\frac{-m}{\lambda^2} < 0, \forall\lambda > 0$$

Donc $\hat{\lambda}$ est un estimateur du maximum de vraisemblance.

Si nous comparons les estimateurs dans lesdeux cas ; nous remarquons que dans le premier cas, l'estimateur prend en considération tout l'échantillon mais dans le deuxième cas l'estimateur prend en considération que les m individus qui ont subi l'événement .

Le risque instantané de la distribution exponentielle est estimé par :

$$\frac{\text{Nombre d'individus qui ont subi l'évènement}}{\text{somme des temps de participation}}$$

2.3 Méthodes non paramétriques

2.3.1 Définition

L'approche non paramétrique est une estimation fonctionnelle, elle ne nécessite aucune hypothèse sur la loi de probabilité des observations ordonnées par ordre croissant des temps de participation.

Cette méthode est très efficace quand nous considèrons un échantillon important d'observations. Plusieurs méthodes sont mises en place à savoir

- L'estimateur de Kaplan-Meir de la fonction de survie.
- L'estimateur de Nelson-Aalen de risque cumulé.
- L'estimation de la survie par la méthode actuarielle.
- L'estimation par noyaux du risque instantané.

Dans cette section nous allons présenter la méthode de Kaplan-Meir de la fonction de survie. Une fois la fonction de survie estimée, nous pouvons déduire les autres fonctions à travers les relations vues en (1.2.6)

2.3.2 Estimation de la fonction de survie

Estimation dans le cas des observations complètes

Dans le cas des données non censurées, nous estimons la fonction de répartition

$$F(t) = P(T \leq t)$$

par la fonction de répartition empirique F_n.

Considérons un échantillon $T_1, T_2, \cdots, T_n$, iid de fonction de répartition F, et $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$, les statistiques d'ordre associée.

La fonction de répartition empirique F_n est donnée par :

$$\begin{aligned} F_n(t) &= \frac{\text{nombre d'observations} \leq t}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i \leq t\}} \end{aligned}$$

donc

$$F_n(t) = \begin{cases} 0 & \text{Si } t < T_{(1)} \\ \frac{k}{n} & \text{Si } T_{(k)} \leq t < T_{(k+1)},\ k = 1, 2, 3, \cdots, n-1 \\ 1 & \text{Si } t \geq T_{(n)}. \end{cases}$$

Puisque $S(t) = 1 - F(t)$, nous estimons la fonction de survie empirique par :

$$S_n(t) = \begin{cases} 1 & \text{Si } t < T_{(1)} \\ \frac{n-k}{n} & \text{Si } T_{(k)} \leq t < T_{(k+1)},\ k = 1, 2, 3, \cdots, n-1 \\ 0 & \text{Si } t \geq T_{(n)}. \end{cases}$$

La fonction de survie empirique est définie aussi par :

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

Estimation dans le cas des observations censurées

Dans le cas des données censurées, la méthode non paramétrique la plus utilisée est la méthode de Kaplan-Meir de la fonction de survie (1958) qui repose sur la définition suivante :

L'individu est à l' état initial après l'instant t_i , c'est être en état initial juste avant l'instant $t_i ie$ à l'instant $ti - 1$ et ne pas subir l'évènement en t_i.
avec t_i est l'instant où l'évènement a été observé.

Cette idée se traduit par la relation probabiliste suivante :

Soit les statistiques d'ordre

$$T_{(1)}, T_{(2)}, T_{(3)}, \cdots, T_{(i-1)}, T_{(i)}, \cdots, T_{(n)}$$

Être en vie après t jours, implique avoir survécu au jour 1, au jour 2, puis au jour t

Nous définissons la fonction de survie comme suit :

$$\begin{aligned}
S(t) &= P(T > t_{(n)}) \\
&= P(T > t_{(n)}, T > t_{(n-1)}) \\
&= P(T > t_{(n)} \backslash T > t_{(n-1)}) P(T > t_{(n-1)}) \\
&= P(T > t_{(n)} \backslash T > t_{(n-1)}) P(T > t_{(n-1)}, T > t_{(n-2)}) \\
&= P(T > t_{(n)} \backslash T > t_{(n-1)}) P(T > t_{(n-1)} \backslash T > t_{(n-2)}) P(T > t_{(n-2)}) \\
&\vdots \\
&= P(T > t_{(n)} \backslash T > t_{(n-1)}) \cdots P(T > t_{(i)} \backslash T > t_{(i-1)}) \cdots P(T > t_{(1)} \backslash T > t_{(0)}) P(T > t_{(0)})
\end{aligned}$$

Avec : $t_{(0)} = 0$ et $P(T > t_{(0)}) = 1$

Donc :

$$\begin{aligned}
S(t) &= P(T > t_{(n)}) \\
&= \prod_{i=1}^{n} P(T > t_{(i)} \backslash T > t_{(i-1)})
\end{aligned}$$

Soit P_i indique la probabilité de survivre jusqu'au jour T_i, sachant qu'au fait d'être en vie juste avant ce jour ie en T_{i-1}, c'est-à-dire la probabilité de ne pas subir l'événement jusqu'au jour T_i sachant que l'individu était toujours à l'état initial en T_{i-1}, avec :

$$P_i = P(T > t_{(i)} \backslash T > t_{(i-1)})$$

Donc :

$$\begin{aligned}
S(t) &= P_0 \times P_1 \times P_2 \times \cdots \times P_{n-1} \\
&= \prod_{i=1}^{n} P_{i-1}
\end{aligned}$$

Soit $\hat{P}_i$ estimateur de P_i donc $S(t)$ sera estimé par $\hat{S}(t)$ le produit des $\hat{P}_i$, c'est-à-dire

$$\hat{S}(t) = \prod_{i=1}^{n} \hat{P}_{i-1}$$

Soit :

$$\begin{aligned}
1 - P_i &= 1 - P(T > t_{(i)} \backslash T > t_{(i-1)} \\
&= P(T < t_{(i)} \backslash T > t_{(i-1)}
\end{aligned}$$

$1 - P_i$ est la probabilité que l'événemnt ait lieu dans l'intervalle $]T_{(i-1)}, T_{(i)}]$, sachant que le sujet n'a pas subi l'événemnt en $T_{(i-1)}$

L'idée c'est de couper l'intervalle de temps en petits intervalles indépendants.

Posons :

n_i : le nombre d'individus ayant un risque de subir l'événement juste avant l'instant $t_{(i)}$,(le nombre de cas possibles)

m_i : le nombre de sujets qui ont subi l'évènement qui a été observé en $t_{(i)}$ (c'est le nombre de cas favorable)

$1 - P_i$ peut être estimé par

$$1 - \hat{P}_i = \frac{n_i}{m_i}$$

D'où

$$\begin{aligned} \hat{P}_i &= 1 - \frac{n_i}{m_i} \\ &= \frac{n_i - m_i}{m_i} \end{aligned}$$

qui définit la proportion de survie à l'instant t_i

Estimation de la fonction de survie

1- Cas où les événements sont distincts Nous avons supposé que T est une variable aléatoire continue, donc dans ce cas deux sujets ne peuvent pas subir un événement au même instant, donc :

$m_i = 1$ pour tout $i(m_1 = 1, m_2 = 1, m_3 = 1, \cdots)$

Si $m_i = 0$, c'est le cas de censure en $t_{(i)}$ quand $d_i = 0$

Si $m_i = 1$ où l'événement a été subi $d_i = 1$

Par suit $\hat{P}_i = 1 - \frac{d_i}{n_i}$

$n_i = n - (i-1)$,avec $i - 1 =$ le nombre de de sujet qui ont subi l'événements, observés à l'instant t_{i-1}

Nous obtenons l'estimateur de Kaplan-Meier :

$$\begin{aligned} \hat{S(t)} &= \prod_{T_{(i)}<t} (1 - \frac{d_i}{n_i}) \\ &= \prod_{T_{(i)}<t} (1 - \frac{d_i}{n - (i-1)}) \\ &= \prod_{T_{(i)}<t} (\frac{n-i}{n-i+1})^{d_i} \end{aligned}$$

2- Cas où des événements surviennent simultanément Dans le cas pratique la précision de la mesure des délais est limitée du fait qu'il peut avoir des ex-æquos qui peuvent subir un événement au même instant , donc $m_i > 1$

Nous obtenons l'estimateur de Kaplan-Meier :

$$\hat{S(t)} = \prod_{T_{(i)}<t} (1 - \frac{m_i}{n_i})$$

Remarque : Nous pouvont démontrer que l'estimateur Kaplan-Meier maximise la fonction de vraisemblance.

2.3.3 La représentation et lecture d'une courbe de survie

Nous avons dit que la fonction de survie théorique est représentée par une courbe continue décroissante qui retrace la probabilité de survie en fonction du temps, avec

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

Nous savons qu'au début d'observation, tous les sujets sont encore dans l'état initial, c'est-à-dire 100% de la population n'ont pas encore subi l'événement.
La fonction de survie empirique estimée par la méthode de Kaplan-Meir est représentée par une courbe décroissante sous forme d'escaliers.
A chaque fois qu'un événement d'intérêt se produit, il se traduit par une marche, sa hauteur dépend du nombre d'individus qui ont subi l'événement à cet instant.

Comme nous travaillons dans le cas des données censurées, les perdus de vues sont représentés par des traits verticaux. Dans le cas ou le nombre de censures devient de plus en plus grand, les barres ne seront pas représentées sur le graphe.

Pour que le graphe soit complet, il faut y faire apparaitre le nombre total des sujets exposés à l'événement ainsi que le nombre total d'évènements observé et défini les axes de coordonnées
L'axe des abscisses représente le temps , le taux d'individus exposés aux risques est représenté sur l'axe des ordonnées .comme le montre la figure ci-dessus.

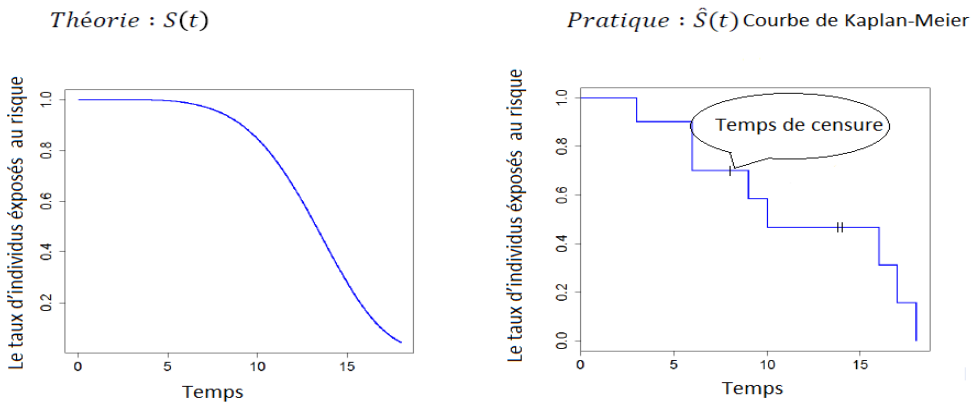


FIGURE 2.1 –

Dans le cas où la majorité des sujets ont subi l'évènement d'intérêt, nous prenons en considération dans la lecture, la variable expliquée, dans le cas contraire, nous interprètons l'évènement d'intérêt comme le montre l'exemple suivant

Exemple

Dans une étude de survie après le diagnostic d'une maladie, si nous observons :

- 80% de décès et 20% de vivants à la fin d'étude, nous nous intéressons à la survie (variable expliquée)

- 10% de décès et 90%de vivants à la fin d'étude, nous nous intéressons au nombre de décès (événement d'intérêt)

La précision de l'estimateur dépend du nombre de patients disponibles , ou d'écart-type de la fonction $S(t)$ ou par intervalle de confiance qui peut ne pas être figuré sur le graphe car il surcharge la représentation, ce qui rend la lecture difficile.

À la fin du graphe, il reste toujours un groupe de sujets qui n'ont pas encore subi l'événement, nous l'interprètons soit par l'apparence de stabilité de survie à long terme, ou bien par la disparition du risque à un certain temps.

2.4 Estimation semi paramétrique

Soit un échantillon de n individus , nous observons le couple (t_i, d_i).
Dans le cas où nous prenons en considération des variables explicatives de chaque individu i ,nous définissons le vecteur des covariables par :

$$Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{ip})$$

qui est indépendant du temps, dans ce cas nous observons le triplet :(t_i, d_i, Z_i)

2.4.1 Modèles à risque proportionnel $\lambda(.\backslash Z)$

Tous les modèles ou la fonction de risque est un produit de deux termes dont l'un dépend seulement du temps et l'autre n'en dépend pas, ce sont des modèles dits à risque proportionnels qui se caractérisent par la relation suivante :

Pour $t > 0$ nous avons

$$\begin{aligned} \lambda(t\backslash Z) &= \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \backslash T \geq t, Z)}{\delta t} \\ \lambda(t\backslash Z) &= \lambda_0(t) h(\beta, Z) \end{aligned}$$

Avec :

$\lambda_0(t)$:c'est le risque instantané de base,il dépend du temps et est le même pour chaque individu. Il peut être inconnu et doit l'estimée par une méthode non paramétrique.

Z : vecteur des variables explicatives spécifiques de chaque individu i, de dimension $(1 \times p)$.

β : vecteur des coefficients de régression, c'est le paramètre d'intérêt c'est-à-dire nous estimons ses coefficients pour évaluer l'impact de chacun des facteurs sur la durée étudiée, β est de dimension $(1 \times p)$.

$h(\beta, Z)$: est une fonction positive contenant la partie paramétrique indépendante du temps, elle caractérise l'individu.

Supposons que l'effet des covariables se ramène à une quantité réelle $\beta' Z$ c'est-à- dire :

$$\lambda(t \backslash Z) = \lambda_0(t) h(\beta' Z)$$

$\lambda(. \backslash Z)$ est appelé risque proportionnel, car soit deux individus i et j de l'échantillon de n sujets ont pour covariables $Z_i$et Z_j respectivement, le rapport des fonctions à risque proportionnel ne dépend pas du temps.

$$\frac{\lambda(t \backslash Z_i)}{\lambda(t \backslash Z_j)} = \frac{\lambda_0(t) h(\beta' Z_i)}{\lambda_0(t) h(\beta' Z_j)} = \frac{h(\beta' Z_i)}{h(\beta' Z_j)}$$

2.4.2 Modèles de Cox

Si$h(\beta' Z)$ est une fonction exponentielle, ie

$$h(\beta' Z) = \exp\{\beta' Z\} = \exp\{\beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_p Z_{ip}\}$$

nous nous retrouvons dans un modèle appelé modèle de Cox qui est un modèle miltivarié régressif semi-paramétrique. C'est l'approche la plus populaire dans l'analyse des modèles de survie, ce modèle est employé lorsque nous cherchons à évaluer l'effet de certains facteurs explicatifs sur la durée de survie.
Le modèle est défini comme suit :

$$\lambda(t \backslash Z) = \lambda_0(t) \exp\{\beta' Z\}$$

avec le rapport :

$$\frac{\lambda(t \backslash Z_i)}{\lambda(t \backslash Z_j)} = \frac{\exp\{\beta' Z_i\}}{\exp\{\beta' Z_j\}}$$

2.4.3 La fonction de vraisemblance partielle

Soit le modèle de Cox suivant :

$$\begin{aligned} \lambda(t \backslash Z) &= \lambda_0(t) h(\beta' Z) \\ &= \lambda_0(t) \exp\{\beta' Z\} \end{aligned}$$

A fin d'estimer la vecteur des paramètres β il faut calculer la vraisemblance L.
Dans le cas de la censure à droite, la vraisemblance d'un échantillon de n individus indépendants est définie comme suit :

$$L = \prod_{i=1}^{n} [f_i(t_i)]^{d_i} [S_i(t_i)]^{1-d_i}$$

Cas d'évènements distincts

Dans ce cas, en utilisant les relations entre les fonctions de survie définie en $(1, 2, 6)$, nous obtenons la vraisemblance suivante :

$$\begin{aligned} L &= \prod_{i=1}^{n} [\lambda_i(t_i \backslash Z_i) S_i(t_i)]^{d_i} [S_i(t_i)]^{1-d_i} \\ &= \prod_{i=1}^{n} [\lambda_i(t_i \backslash Z_i)]^{d_i} [S_i(t_i)] \end{aligned}$$

Avec :

$$\begin{aligned} \lambda(t_i \backslash Z_i) &= \lambda_0(t_i) h(\beta' Z_i) \\ &= \lambda_0(t_i) \exp\{\beta' Z_i\} \end{aligned}$$

La fonction de vraisemblance L prend en considération la fonction de risque de base $\lambda_0(t_i)$ qui ne nous intéresse pas et considérée comme un paramètre nuisible. Cox propose de l'éliminer c'est-à-dire construire la fonction de vraisemblance partielle.

Nous définissons d'abord les probabilités suivantes :

$$\lambda(t_i \backslash Z_i) = \lambda_0(t_i) \exp\{\beta' Z_i\}$$

C'est la probabilité qu'un individu i subisse l'évènement en t_i

$$\sum_{j \in \mathfrak{R}(T_i)} \lambda(t_i \backslash Z_i) = \sum_{j \in \mathfrak{R}(T_i)} \lambda_0(t_i) \exp\{\beta' Z_j\}$$

Définie la probabilité qu'il y ait un évènement en t_i

$\mathfrak{R}(T_i)$ est l'ensemble des individus à risque au temps d'événements t_i

La construction de la vraisemblance partielle de Cox notée PL Nous multiplions et nous divisons la vraisemblance totale par la même quantité. Nous obtenons :

$$L = \prod_{i=1}^{n} \frac{[\lambda_i(t_i \backslash Z_i)]^{d_i}}{[\sum_{j \in \mathfrak{R}(T_i)} \lambda(t_i \backslash Z_i)]^{d_i}} \left[\sum_{j \in \mathfrak{R}(T_i)} \lambda(t_i \backslash Z_i) \right]^{d_i} [S_i(t_i)]$$

L'idée de Cox est de considérer une partie de la vraisemblace totale , nous obtenons la fonction de vraisemblace partielle PL comme suit :

$$\begin{aligned} PL &= \prod_{i=1}^{n} [\frac{\lambda_i(t_i \backslash Z_i)}{\sum_{j \in \Re(T_i)} \lambda(t_i \backslash Z_j)}]^{d_i} \\ &= \prod_{i=1}^{n} [\frac{\lambda_0(t_i) \exp\{\beta' Z_i\}}{\sum_{j \in \Re(T_i)} \lambda_0(t_i) \exp\{\beta' Z_j\}}]^{d_i} \\ &= \prod_{i=1}^{n} [\frac{\exp\{\beta' Z_i\}}{\sum_{j \in \Re(T_i)} \exp\{\beta' Z_j\}}]^{d_i} \\ &= \prod_{i=1}^{D} \frac{\exp\{\beta' Z_i\}}{\sum_{j \in \Re(T_i)} \exp\{\beta' Z_j\}} \end{aligned}$$

Rappelons que

D : nombre de décès observés parmi n sujets soumis à l'étude.

$T_1 < T_2 < T_3 < \cdots < T_D$: Les statistiques d'ordre qui définit les temps d'évènements distinctes.

$(1), (2), (3), \cdots, (D)$, indices des individus décédé aux instants $T_1, T_2, T_3, \cdots, T_D$ respectivement.

Le terme $\frac{\exp\{\beta' Z_i\}}{\sum_{j \in \Re(T_i)} \exp\{\beta' Z_j\}}$ définit une probabilité conditionnelle que le sujet i subisse l'évènement en T_i sachant qu'un évènement a eu lieu en T_i

Donc la vraisemblance partielle ne dépend que de β, nous pouvons alors estimer ce dernier sans connaitre la fonction de hasard de base $\lambda_0(t)$ par la maximisation de la vraisemblance partielle de Cox.

LP n'est pas une vraisemblance au sens statistique de terme mais elle se comporte comme elle.

Cas d'évènements simultanés

Dans le cas de données réelles, l'hypothèse d'existence d'ex-aequos est souvent vérifiée. Supposons qu'il y ait plusieurs sujets qui subissent l'évènement en t_i.

Par exemple nous avons deux individus e_1 et e_2 de vecteurs de caractéristiques Z_1 et Z_2 respectivement subissent l'événement en t_i, la vraisemblance partielle dans le cas d'un modèle continu s'écrit en tenant compte deux ordres possibles des temps d'évènement de

ces sujets comme suit :

$$\frac{\exp\{\beta' Z_1\}}{\sum_{j\in\Re(T_i)}\exp\{\beta' Z_j\}}\times\frac{\exp\{\beta' Z_2\}}{\sum_{j\in\Re(T_i)-e_1}\exp\{\beta' Z_j\}}+\frac{\exp\{\beta' Z_2\}}{\sum_{j\in\Re(T_i)}\exp\{\beta' Z_j\}}\times\frac{\exp\{\beta' Z_1\}}{\sum_{j\in\Re(T_i)-e_2}\exp\{\beta' Z_j\}}$$

Dans le cas ou le nombre d'événements simultanés augmente, l'écriture des permutations dans la PL devient de plus en plus compliquée.
Plus généralement si k individus ont le même temps d'évènement, alors la vraisemblance partielle pour ce terme sera correspondante au $k!$ termes.

Cette difficulté nous conduit à utiliser une expression approchée en faisant l'approximation de la vraisemblance partielle en utilisant la méthode de Breslow (1974) , que nous ne allons pas introduire dans notre travail .

2.5 Comparaison de deux groupes

Nous souhaitons tester l'hypothèse nulle H_0 contre l'hypothèse alternative H_1, avec
$(H_0) : S_A(t) = S_B(t)$ vs $(H_1) : S_A(t) \neq S_B(t)$

Dans le cas où il n'y a pas de données censurées, nous pourrons utiliser les tests non paramétriques suivants
- Test de Kolmogorov Smirnov
- Test de rangs.
- Test de Savage.
- Test de Wilcoxon.

En présence des données censurées, nous généralisons les tests non paramétriques usuels. Nous obtenons
- Le test de Log-Rank est une généralisation du test de Savage
- Le test de Gehan et de Peto-Prentice est une généralisation du test de Wilcoxon

Dans cette section nous allons détailler la méthode de comparaison de Log-Rank

2.5.1 Définition du test de Log-Rank

Dit aussi test de Mantel-Haenszel, c'est le test le plus utile, le plus simple et le plus performant pour comparer deux courbes de survie à condition qu'elles ne se croisent pas et quelles soient estimées par la méthode de Kaplan-Meir.
Soit deux groupes d'individus dénotés A et B. Supposons que les instants où les événements se produisent sont fixés t_i et comparons le nombre de sujets qui ont subi l'évènement dans chaque groupe.

2.5.2 Notation

Nous construisons une statistique d'ordre qui est constituée des temps d'apparition des événements d'intérêts de deux groupes A et B simultanément ie $A \cap B$

$$t_{(1)}, t_{(2)}, \cdots t_{(i)}, \cdots t_{(k)},$$

Nous notons
m_{Ai}, m_{Bi} : nombre d'individus qui ont subi l'événement observé dans chaque groupe A et B respectivement, à l'instant t_i.
$m_{Ai} + m_{Bi} = m_i$: le nombre total d'évènements réalisés dans chaque groupe, $m_i > 0$ à l'instant t_i.
n_{Ai}, n_{Bi} : nombre de sujets exposés au risque de subir l'événement dans chaque groupe en t_i.
$n_{Ai} + n_{Bi} = n_i$: Nombre total de sujets exposés au risque de subir l'événement en t_i.
n_i : Le nombre total de sujets éxposés aux risques en t_i.

Les observations à un instant fixé peuvent être résumées dans la table de contingence observée comme suit

	Nombre d'évènements réalisés en t_i	Nombre d'évènements non réalisés après t_i	Total
Groupe A	m_{Ai}	$n_{Ai} - m_{Ai}$	n_{Ai}
Groupe B	m_{Bi}	$n_{bi} - m_{Bi}$	n_{Bi}
Total	m_i	$n_i - m_i$	n_i

2.5.3 Statistique du test

À chaque instant d'évènement t_i, nous construisons une table de contingence correspondante, nous calculons le nombre d'observations attendu.

Sous l'hypothèse nulle H_0 d'égalité de fonction de survie entre les deux groupes étudiés ,nous avons l'indépendance des lignes et des colonnes. Dans le cas idéal $m_i = 1$ c'est-à-dire là où il n'y a pas d'ex-aequos soit :
e_{ip} : Le nombre de décès attendu sou H_0 à l'instant t_i pour le groupe taille p

De manière générale

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

dans notre cas nous avons

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

e_{Ai},e_{Bi} représentent le nombre d'évènements réalisés en t_i dans les groupes A et B respectivement sous H_0.

Nous retrouvons la table de contingence attendue sous H_0 comme suit

	Nombre d'évènements réalisés en t_i	Nombre d'évènements non réalisés après t_i	Total
Groupe A	e_{Ai}	$n_{Ai} - e_{Ai}$	n_{Ai}
Groupe B	e_{Bi}	$n_{bi} - e_{Bi}$	n_{Bi}
Total	m_i	$n_i - m_i$	n_i

Posons

$$E_A = \sum_{i=1}^{k} e_{Ai}$$

$$E_B = \sum_{i=1}^{k} e_{Bi}$$

le nombre total d'événements attendus dans A et B respectivement sous H_0

$$O_A = \sum_{i=1}^{k} m_{Ai}$$

$$O_B = \sum_{i=1}^{k} m_{Bi}$$

Le nombre total d'événements observés dans A et B respectivement sous H_0

Sous H_0

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \sim \chi_1^2$$

Sous H_0, la quantité χ^2 suit une loi de Khi-Deux à 1 degré de liberté ; au seuil $\alpha = 0.05$ nous cherchons les valeurs de χ_1^2 de la table de Khi-Deux
Si $\chi^2 < \chi_1^2$ nous acceptons H_0
Si $\chi^2 > \chi_1^2$ nous rejettons H_0

Remarque

Nous pouvons comparer plusieurs groupes (k groupes) en utilisant la méthode de Log-Rank

2.6 Méthodes Bayésienne d'analyse de survie

2.6.1 L'analyse Bayésienne

Calcul de la loi a postériori $\pi(\theta \backslash x)$

La démarche de l'analyse bayésienne conduit d'abord au calcul d'une loi a posteriori $\pi(\theta \backslash x)$

Etant donnée la densité des observations (la vraisemblance)$f(x\backslash\theta)$ et la densité a priori $\pi(\theta)$, nous pouvons construire la densité jointe de $h(\theta, x)$.Avec

$$h(\theta, x) = f(x\backslash\theta)\pi(\theta) \tag{2.6}$$

La loi marginale$f(x)$ est définie comme suit :

$$f(x) = \int_{\Theta} f(x\backslash\theta)\pi(\theta)d\theta \tag{2.7}$$

En utilisant la version continue du théorème de Bayes, nous obtenons $\pi(\theta\backslash x)$ qui est la distribution a posteriori définie comme suit :

$$\begin{aligned} \pi(\theta\backslash x) &= \frac{h(\theta, x)}{\int_{\Theta} h(\theta, x)d\theta} \\ &= \frac{f(x\backslash\theta)\pi(\theta)}{f(x)} \end{aligned} \tag{2.8}$$

C'est une quantité qui contient toutes les informations disponibles sur le paramètre inconnue θ.

$$\frac{1}{\int_{\Theta} h(\theta, x)d\theta} = \frac{1}{f(x)}$$

est une constante de normalisation, donc

$$\pi(\theta\backslash x) \propto f(x\backslash\theta)\pi(\theta) \tag{2.9}$$

D'où la dépendance de cette densité de l'information a priori.

La distribution a priori des paramètres

Le choix de la loi a priori est une étape fondamentale dans l'analyse Bayésienne.

La distribution a priori conjuguée naturelle Avant de définir la distribution a priori conjuguée naturelle nous allons définir le modèle exponentiel

1- Famille exponentielle Nous appellons famille exponentielle à s paramètres toute famille de loi de distribution dont la densité a la forme suivante :

$$f(x\backslash\theta) = exp\{\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - \beta(\theta)\}g(x) \tag{2.10}$$

$\eta_i(.)$, $\beta(.)$ sont des fonctions de paramètre θ et $T_i(x)$ est une statistique.

Remarque 2.1. Soit $f(x\backslash\theta)$ appartient à une famille exponentielle, alors une famille de lois a priori conjuguée par $f(x\backslash\theta)$ est donnée par :

$$\pi(\theta\backslash\mu) = K(\mu,\lambda)exp(\theta\mu - \lambda\beta(\theta))$$

Avec $k(\mu,\lambda)$ est la constante de normalisation.

$$\pi(\theta) \propto exp(\theta\mu - \lambda\beta(\theta))$$

$$\pi(\theta\backslash x) \propto exp(\theta(\mu + x) - (\lambda+1)\beta(\theta))$$

C'est-à-dire la loi a priori et a postériori ont la même forme, comme le montre le tableau suivant :

Exemple des lois conjuguées naturelles

La vraisemblance	La distribution a priori conjuuée	La distribution a postériori
$X\backslash\theta \sim N(\theta,\sigma^2)$	$\theta \sim N(m,z^2)$	$\theta\backslash x \sim N((\frac{1}{\sigma^2}+\frac{1}{z^2})^{-1}(\frac{x}{\sigma^2}+\frac{m}{z^2}),(\frac{1}{\sigma^2}+\frac{1}{z^2})^{-1})$
$X\backslash\theta \sim N(\mu,\theta^2)$	$\theta^2 \sim IG(a,b)$	$\theta^2\backslash x \sim IG(a+\frac{1}{2},b+\frac{1}{2}(x+\mu)^2)$
$X\backslash\theta \sim Poisson(\theta)$	$\theta \sim Gamma(a,b)$	$\theta\backslash x \sim Gamma(a+x,b+1)$
$X\backslash\theta \sim exp(\theta)$	$\theta \sim Gamma(a,b)$	$\theta\backslash x \sim Gamma(a+1,b+x)$
$X\backslash\theta \sim binom(n,\theta)$	$\theta \sim Beta(a,b)$	$\theta\backslash x \sim Beta(a+x,b+n-x)$
$X\backslash\theta \sim Gamma(\alpha,\theta)$	$\theta \sim Gamma(a,b)$	$\theta\backslash x \sim Gamma(\alpha+a,b-x)$
$X\backslash\theta \sim Geom(\theta)$	$\theta \sim Beta(a,b)$	$\theta\backslash x \sim Beta(a+x,b+n-1)$

2- Lois non informatives

Dans le cas où nous n'avons pas d'informations nous utilisons les lois non informative qui ne nous donnent pas beaucoup d'informations mais elles sont utiles

a- Loi a priori impropre $\pi(\theta)$ application telle que

$$\int_\Theta f(x\backslash\theta)d\theta = +\infty \tag{2.11}$$

Nous disons que $\pi(\theta)$ est une loi a priori impropre car elle diverge donc elle ne définit pas une densité de probabilité. Nous avons donc ce que nous appellons estimateur de Bayes généralisé qui sera vérifié à condition que

$$\int_\Theta f(x\backslash\theta)\pi(\theta)d\theta < \infty$$

la densité de la loi a postériori est définie comme suit :

$$\pi(\theta\backslash x) = \frac{f(x\backslash\theta)\pi(\theta)}{\int_\Theta f(x\backslash\theta)\pi(\theta)d\theta} \tag{2.12}$$

Et que $\pi(\theta\backslash x)$ est« propre », de plus l'a posteriori n'est pas une application de la formule de Bayes.

b- Loi uniforme Quand le choix repose sur l'équiprobabilité des valeurs possible du paramètre θ, la loi n'apporte aucune information supplémentaire sur θ, nous parlons dans ce cas sur la loi uniforme :

$$\pi(\theta) = \frac{1}{k} \tag{2.13}$$

Avec k est la taille de l'ensemble de Θ Lorsqu'on choisi une loi a priori uniforme sur $[0, 1]$,

$$\pi(\theta\backslash x) \Leftrightarrow f(x\backslash\theta)$$

Donc l'approche fréquentiste n'est qu'un cas particulier de l'approche bayésien

c- La loi de Jeffrey Cette méthode utilise l'information de ficher I_θ qui représente la quantité d'information apportée par l'observation x sur le paramètre θ .

Soit θ un paramètre inconnu réel. La loi a priori de Jeffrey est définie comme suit :

$$\pi_J(\theta) = c\sqrt{I_x(\theta)}\mathbb{I}_\Theta(\theta) \tag{2.14}$$

Avec c =la constante de normalisation

$$\pi_J(\theta) \propto \sqrt{I_x(\theta)} \tag{2.15}$$

avec

$$I_X(\theta) = E[(\frac{\partial}{\partial\theta} lnf(x\backslash\theta)^2)]$$

Dans le cas ou le domaine de X ne dépend pas de θ.

$$I_X(\theta) = -E[(\frac{\partial^2}{\partial\theta^2} lnf(x\backslash\theta))]$$

Plus $I_X(\theta)$ est grand, plus l'observation apporte de l'information sur θ.
Dans le cas multidimensionnel $\theta(\theta_1, \theta_2 \cdots \theta_p)$ est un vecteur, dans ce cas nous considérons la loi de Jeffrey sous la forme suivante :

$$\pi(\theta) \propto (detI(\theta))^{\frac{1}{2}} \tag{2.16}$$

Où $I(\theta)$ est la matrice d'information de Fisher dont l'information de Fisher est donnée par :

$$I(\theta)_{ij} = E[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} lnf(x\backslash\theta)]$$

2.6.2 Analyse Bayensienne des modèles paramétriques de survie

L'inference Bayésienne :

L'inférence consiste à choisir une règle de décision $\delta \in \Delta$ concernant $\theta \in \Theta$ sur la base d'une observation $x \in \chi$, x et θ étant lié par la loi$f(x\backslash\theta)$.

Les différents espaces Soit un individu dans un environnement (nature) donné, sur la base des observations, il mène des actions et prend des décisions qui auront un coût.

Les espaces qui interviennent dans l'écriture d'un modèle de décision sont :

1- Espace des observations χ : ensemble des résultats d'une expérience x donnée.

2- Espace des états de nature Θ : l'ensemble des paramètres inconnus θ.

3- Espace des actions ou décisions $\mathcal{A}$: l'ensemble de décisions possibles à prendre, suite à l'information observée x

4- Ensemble de stratégies ou règles de décision Δ : Soit l'application δ qui définit une règle de décision comme suit

$$\begin{aligned} \delta \quad : \quad & \chi \to \mathcal{A} \\ & x \mapsto \delta(x) = a \end{aligned}$$

Avec

δ : la règle de décision, c'est un estimateur.

$\delta(x) = a$: action ou une décision, c'est l'estimation.

Le coût L Dit aussi mesure de perte, c'est une relation de préférence pour choisir une décision $\delta(x)$ et que l'espace des états de nature est Θ,la fonction coût est définie comme suit :

$$\begin{aligned} L \quad : \quad & \Theta \times \mathcal{A} \to \mathbb{R}_+ \\ & (\theta, a) \mapsto L(\theta, \delta(x)) \end{aligned}$$

Elle évalue le coût de la décision $\delta(x) = a$ quand le paramètre vaut θ , elle calcule la perte causé par une mauvaise décision ou une mauvaise évaluation de θ :

Si $L(\theta, \delta(x)) < 0 \Rightarrow$ le coût correspond au gain.

Si $L(\theta, \delta(x)) > 0 \Rightarrow$ le coût correspond à la perte.

Si $L(\theta, \delta(x)) = 0 \Rightarrow$ la décision est bonne elle correspond à la solution de cette équation.

θ etant inconnu, $L(\theta, \delta(x))$ est une équation à deux inconnues, que nous ne pouvont pas résoudre de plus elle ne prend pas en considération de l'information apportée par la vraisemblance.

1- Risque fréquentiste R **:** Pour prendre en considération la vraisemblance $f(x\backslash\theta)$, nous calculons la perte moyenne pour retrouver le risque fréquentiste :

$$\begin{aligned} \mathrm{R}(\theta, \delta) &= E^{f(.\backslash\theta)} L(\theta, \delta(x)) \\ \mathrm{R}(\theta, \delta) &= \int_{\chi} L(\theta, \delta(x)) f(x\backslash\theta) dx \end{aligned} \tag{2.17}$$

Elle nous permet de comparer deux décisions δ_1 et δ_2.
δ_1 est préférable que δ_2 si $\mathrm{R}(\theta, \delta_1) \leq \mathrm{R}(\theta, \delta_2)$.

Le risque fréquentiste permet de faire une comparaison sur l'ensemble de décision Δ , mais elle ne peut pas se faire quand nous prenons en considération aussi l'espace des paramètres θ en même temps :

$$\begin{aligned} \mathrm{R}(\theta_1, \delta_1) &< \mathrm{R}(\theta_1, \delta_2) \\ \mathrm{R}(\theta_2, \delta_1) &> \mathrm{R}(\theta_2, \delta_2) \end{aligned}$$

2- Risque intégré r : C'est la moyenne du Risque fréquentiste ou la moyenne du coût moyen suivant la loi a priori, noté $r(\pi, \delta)$.
Nous avons

$$\begin{aligned} r(\pi, \delta) &= E^{\pi}(\mathrm{R}(\theta, \delta(x)) \\ &= \int_{\Theta} \mathrm{R}(\theta, \delta(x))\pi(\theta)d\theta \\ &= \int_{\Theta}\int_{\chi} L(\theta, \delta(x))f(x\backslash\theta)dx\pi(\theta)d(\theta) \\ &= \int_{\Theta}\int_{\chi} L(\theta, \delta(x))f(x)\pi(\theta\backslash x)dxd\theta \end{aligned} \tag{2.18}$$

3- Le coût a posteriori φ : $\varphi(\pi, \delta(x))$ est la moyenne de coût par rapport à la loi a posteriori

$$\begin{aligned} \varphi(\pi, \delta(x)) &= \mathbf{E}^{\pi(.\backslash x)} L(\theta, \delta(x)) \\ &= \int_{\Theta} L(\theta, \delta(x))\pi(\theta\backslash x)d\theta \end{aligned} \tag{2.19}$$

c'est une fonction de x.
Nous pouvons alors dire que le risque intégré $r(\pi, \delta)$ est la moyenne du coût a posteriori $\varphi(\pi, \delta(x))$ suivant la loi $f(x)$

$$r(\pi, \delta) = \int_{\chi} \varphi(\pi, \delta(x))f(x)dx \tag{2.20}$$

$$= E^{f}\varphi(\pi, \delta(x)) \tag{2.21}$$

4- Le risque Bayesien

$$r(\pi, \delta_\pi) = \int_{\chi} \varphi(\pi, \delta_\pi(x))f(x)dx \tag{2.22}$$

Il contient l'estimateur Bayesien δ_π qui est déterminé en minimisant le risque intégré et minimiser ce dernier revient à minimiser le coût a posteriori. Le risque Bayesien est le plus petit des risques fréquentistes.

$$r(\pi, \delta_\pi) < r(\pi, \delta)$$

Quelques exemples de fonctions de perte usuelles et leurs estimateurs

1- Le coût quadratique

$$L(\theta, \delta) = (\theta - \delta)^2$$

L'estimateur bayésien δ_π associé à la loi a priori π et au coût quadratique est la moyenne a posteriori

$$\mathbf{E}(\theta \backslash X) = \delta_\pi(\theta)$$

2- Le coût absolu

$$L(\theta, \delta) = \begin{cases} \theta - \delta & \text{si } \theta > \delta \\ \delta - \theta & \text{si } \theta < \delta \end{cases}$$

L'estimateur de Bayes δ_π associé à la distribution prioriπ et au coût absolu est la médiane de la distribution a posteriori $\pi(\theta \backslash x)$

3- Le coût $0-1$

$$L(\theta, \delta) = \begin{cases} 0 & \text{si la décision est bonne } |\theta - \delta| < \epsilon \\ 1 & \text{si la décision est mauvaise} |\theta - \delta| > \epsilon \end{cases}$$

L'estimateur bayésien δ_π associé à la fonction du coût $0-1$ et à la distribution a priori π est le mode de la distribution a posteriori.

Méthodes Bayésienne d'analyse de survie

Dans l'approche Bayésienne de l'analyse de modèles de survie, nous retrouvons les méthodes semi paramétriques comme la méthode de Kalbfleisch, les méthodes non paramétriques comme la méthode de Florens et Rolin ainsi que la méthode paramétrique que nous détaillerons dans ce qui suit.

Détermination de la loi a posteriori de données censurées Soit la fonction de vraisemblance des données de survie censurées, nous définissons la distribution du couple d'observation comme suit

$$\Phi(t_i, d_i) = f_T(t_i)^{d_i} S_T(t_i)^{1-d_i} \tag{2.23}$$

Soit un échantillon de n individus de fonction $\Phi(t_i, d_i)$

La loi a posteriori $\pi(\theta \backslash x)$ est définie par

$$\begin{aligned} \pi(\theta \backslash x) &= \frac{f(x \backslash \theta)\pi(\theta)}{\int_\Theta f(x \backslash \theta)\pi(\theta) d\theta} \\ &= \frac{f(x \backslash \theta)\pi(\theta)}{f(x)} \end{aligned} \tag{2.24}$$

Où $f(x\backslash\theta)$ est la fonction de vraisemblace.Dans le cas des données censurées, elle s'écrit comme suit

$$L(\theta, t_1, t_2, \cdots, t_n) = f(x\backslash\theta) = \prod_{i=1}^{n} f_T(t_i, \theta)^{d_i} S_T(t_i, \theta)^{1-d_i} \tag{2.25}$$

Alors la loi a posteriori dans le cas des données censurée s'écrit comme suit :

$$\begin{aligned} \pi(\theta\backslash x) &= \frac{\pi(\theta) \prod_{i=1}^{n} f_T(t_i, \theta)^{d_i} S_T(t_i, \theta)^{1-d_i}}{\int_\Theta \pi(\theta) \prod_{i=1}^{n} f_T(t_i, \theta)^{d_i} S_T(t_i, \theta)^{1-d_i} d\theta} \\ &\propto \pi(\theta) \prod_{i=1}^{n} f_T(t_i, \theta)^{d_i} S_T(t_i, \theta)^{1-d_i} \end{aligned} \tag{2.26}$$

Cas du modèle exponentiel Nous calculons la vraisemblance , avec

$$S(t) = exp\{-\lambda t\}$$

et

$$f(t) = \lambda exp\{-\lambda t\}$$

de fonction de vraisemblance

$$f(x\backslash\lambda) = \lambda^{\sum_{i=1}^{n} d_i} \exp\{-\lambda \sum_{i=1}^{n} t_i\}$$

En prenant comme loi a priori une loi gamma définie comme suit

$$\begin{aligned} \pi(\lambda) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda), \alpha > 0, \beta > 0 \\ &\propto \lambda^{\alpha-1} \exp(-\beta\lambda) \end{aligned}$$

D'après le théorème de Bayes, nous obtenons la distribution a posteriori

$$\begin{aligned} \pi(\lambda\backslash x) &\propto \lambda^{\alpha-1} \exp(-\beta\lambda) \lambda^{\sum_{i=1}^{n} d_i} \exp\{-\lambda \sum_{i=1}^{n} t_i\} \\ &\propto \lambda^{\alpha+\sum_{i=1}^{n} d_i - 1} \exp\{-\lambda(\sum_{i=1}^{n} t_i + \beta)\} \\ &\propto \Gamma(\alpha + \sum_{i=1}^{n} d_i, \beta + \sum_{i=1}^{n} t_i) \end{aligned}$$

2.7 Méthodes de simulation

Afin de comparer des estimateurs et leurs performances et de choisir le meilleur, les méthodes de simulation répondent à cet objectif.
Dans ce qui suit nous allons voir la méthode de Monte-Carlo et méthode de Monte-Carlo par chaine de Markov (MCMC) qui sont très utiles dans la recherche.

2.7.1 Méthode de Monte-Carlo

Pour calculer les estimateurs, la méthode de simulation de Monte-Carlo est utile.
C'est une méthode algorithmique visant à calculer une valeur numérique approchée en utilisant des techniques probabilistes. Les méthodes de Monte-Carlo sont particulièrement utilisées pour le calcul des intégrales, ces méthodes ont été inventées par Nicolas Metropolis en 1947.

Soit $g(x), x \in \mathbb{R}^m$ une fonction quelconque mesurable et $f(x)$ une fonction de densité de support $\chi \subset \mathbb{R}^m$

- Nous voulons calculer une intégrale $I = \int_{\mathcal{A}\in\chi} g(x)f(x)dx$

1- Nous écrivons tout d'abord l'intégrale I sous forme d'espérance

$$\begin{aligned} I &= E^f(g(x)) \\ &= \int_{\mathcal{A}\in\chi} g(x)dx \end{aligned} \tag{2.27}$$

2- Simuler des variables aléatoires $g(x)$

Nous génèrons un certain nombre n de variables aléatoires $x_i, i = 1, \ldots, n$ iid de densité f

Nous estimons I par l'espérance empirique :

$$I = I_n = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \tag{2.28}$$

La convergence de cette méthode est justifiée par la loi forte des grands nombres et la vitesse de la convergence est précisée par le théorème central limite.

La méthode de Monte-Carlo est beaucoup plus efficace que les méthodes numériques traditionnelles mais cette méthode est limitée par la modélisation car il y a beaucoup de lois de probabilités qui ne sont pas facilement simulables, donc il est souvent nécessaire de mettre en œuvre des techniques de simulation sophistiqueés pour simuler les distributions souhaitées.

2.7.2 Méthode de Monte-Carlo par chaine de Markov (MCMC)

C'est une méthode qui permet de simuler les distributions qui ne sont pas facilement simulables à l'aide d'une chaine de Markov ergodique, elle produit une chaine de Markov X_n ergodique de distribution stationnaire qui sera la densité ciblée f.

Rappel sur les chaines de Markov :

Un processus est dit chaine de Markov, si :

$$\begin{aligned} P[X_{n+1} = j \backslash X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}, X_i = i] &= P[X_{n+1} = j \backslash X_n = i] \\ &= P_{i,j} \end{aligned}$$

où $P_{i,j}$ sont des probabilités de transition ie probabilités de passage d'un état i à l'état j en une seule étape et $P_{i,j}^{(n)}$ probabilité de passage d'un état i à l'état j en n étapes.

Dans notre cas nous allons prendre en considération une chaine de Markov homogène définie comme suit :

$$P[X_{n+1} = j \backslash X_n = i] = P[X_n = j \backslash X_{n-1} = i] = \cdots = P[X_1 = j \backslash X_0 = i]$$

Chaine de Markov est dite homogène si les $P_{i,j}$ ne dépendent pas du temps.

- la probabilité d'être à l'état i à l'instant n $P[X_n = i]$:

Une fois que nous avons défini la probabilité conditionnelle $P[X_n = 0 \backslash X_0 = 1] = P_{ij}^{(n)}$, nous définissons la probabilité d'être à l'état i à l'instant n, n $P[X_n = i]$, pour celà on définie :

- La loi initiale du système :

$P[X_0 = i]$pour $i \in 0, 1, \cdots, n$

$P[X_0 = i] = \pi_i(0) = \pi_1(0) = \pi_2(0) \cdots = \pi_n(0)$

- La loi du processus $P[X_n = k]$

$P[X_n = k] = \pi_k(n)$avec $n \geq 0, k \in E$

Donc $P_k^n = (\pi_1^n, \pi_2^n, \cdots)$

- Comportement asymptotique d'une chaine de Markov : (X_n) varie en fonction du temps et de la loi initiale $\pi(0)$

- Le régime permanant : c'est le cas où $\pi(n)$converge vers une distribution limite qui ne dépend pas de la loi initiale

- La distribution limite : c'est quand une chaine de Markov homogène converge vers π

$$\lim_{n \to +\infty} \pi(n) = \pi$$

Indépendement de π_0.Nous disons que :
La chaine est stationnaire , ergodique ; la distribution limite est l'unique distribution stationnaire.

- La distribution stationnaire π : Nous disons qu'une distribution π est stationnaire par rapport à la matrice de transition P si :

$\pi = \pi P$

Où P la matrice de transition définie par :

$P = (P_{ij}) i, j \in E$.

- Chaine de Markov irréductible : une chaine de Markov homogène et finie est dite irréductible si tous les états de E sont communicants, où E est l'espace des états.

L'algorithme de Métropolis-Hastings

Cette méthode est utilisée dans le cas de simulation des distributions unidimensionnelles.
A partir d'une loi nous allons construire une chaine de Markov ergodique,

f est la loi cible, c'est une loi suivant la quelle nous allons simuler un échantillon donné. Pour ce faite, il faut disposer d'une valeur du départ et une fonction de proposition $q(x, y)$ en utilisant sa densité conditionnelle $q(X\backslash y)$,$q(Y\backslash x)$ avec $\sup\{f\} \subset \sup\{q(.\backslash x)\}$

Nous allons générer les valeurs de $Y_{1+t} = y$ selon la densité $q(X\backslash x)$à partir des valeurs$X_t = x$

L'algorithme de Métropolies-Hastings : 1) Poser la valeur initiale X_t

2) Nous générons des variables aléatoire $Y_{1+t} = y$ selon $q(Y\backslash x)$ à partir des valeurs précédentes $X_T = x$ en utilisant la probabilité d'acceptance $\alpha(x, y)$, avec :

$$\alpha(x, y) = min\{1, \frac{f(y) - q(X\backslash y)}{f(x)q(Y\backslash x)}\}$$

3) Générer une loi uniforme $U \sim u[0, 1]$

4) Déterminer l'échantillon :

Nous avons X_t donc :

$X_{t+1} = Y_{t+1}$ si $U < \alpha(x_t, y)$,dans ce cas nous acceptons la proposition avec une probabilité $\alpha(x, y)$

$X_{t+1} = X_t$ si $U > \alpha(x_t, y)$ nous rejetons la proposition avec la probabilité $1 - \alpha(x, y)$

En répétant l'expérience nous obtenons l'échantillon de variables aléatoires $X_1, X_2, \cdots X_n$ dépendAnt de la loi f

Remarque :

Dans le cas où la densité $q(x, y)$ est symétrique c'est-à-dire $q(x, y) = q(y, x)$, nous nous retrouvons dans de l'algorithme de Métropolies simple définit comme suit :

$$\alpha(x, y) = min\{1, \frac{f(y)}{f(x)}\}$$

2.7.3 L'échantillonnage de Gibbs :

Ce sont des algorithmes spécifiques adapté à la simulation des distributions multidimensionnelles, introduites par Geman en 1984 .C'est une forme particulière de $MCMC$ qui est largement utilisée dans de nombreux domaines d'analyse statistique bayésienne comme suit :

- Nous choisissons un point de départ des d composantes des covariables θ.

Soit $\pi(\theta \backslash x)$ avec θ un vecteur c'est-à-dire $\theta = (\theta_1, \theta_2, \theta_3, \cdots, \theta_d)$, et x sont des données observées.

- Nous utilisons la densité conditionnelle $\pi((\theta_1 \backslash \theta_2, \theta_3, \cdots, \theta_d, x)$

- A la n ème étape, par la théorie de chaine de Markov, lorsque $n \to +\infty$ la densité de réalisation converge vers $\pi(\theta \backslash x)$

Dés que les lois conditionnelles sont exprimées $\pi(\theta_i \backslash \theta_j, x)$, $i \neq j$,il est possible de faire la simulation jointe $\pi(\theta) = \pi(\theta_1, \theta_2, \theta_3, \cdots, \theta_d)$ à partir de la distribution conditionnelle.

Cette méthode peut obtenir une réalisation $\theta = (\theta_1, \theta_2, \theta_3, \cdots, \theta_d)$ du paramètre suivant la loi a postériori $\pi(\theta \backslash x)$

Le principe de la méthode de l'échantillonnage de Gibbs : Soit θ un vecteur de d composantes $\theta = (\theta_1, \theta_2, \theta_3, \cdots, \theta_d)$

1- Etape 01 : Nous choisirons le point de départ des composantes du vecteur aléatoire, le vecteur initial est $\theta^0 = (\theta_1^0, \theta_2^0, \theta_3^0, \cdots, \theta_d^0)$:

2- Etape 02 : Nous considérons le vecteur initial $\theta^0 = (\theta_1^0, \theta_2^0, \theta_3^0, \cdots, \theta_d^0)$

- Générer $\theta_1^1 = \pi(\theta_1 \backslash \theta_2^0, \theta_3^0, \cdots, \theta_d^0, x)$
- Générer $\theta_2^1 = \pi(\theta_2 \backslash \theta_1^1, \theta_3^0, \cdots, \theta_d^0, x)$

  $\vdots$
- Générer $\theta_i^1 = \pi(\theta_i \backslash \theta_1^1, \theta_2^1, \cdots, \theta_{i-1}^1, \theta_{i+1}^0, \cdots, \theta_d^0, x)$

  $\vdots$
- Générer $\theta_d^1 = \pi(\theta_d \backslash \theta_1^1, \theta_2^1, \cdots, \theta_{d-1}^1, x)$

3- Etape $n+1$: Nous considérons le vecteur précédent $\theta^n = (\theta_1^n, \theta_2^n, \theta_3^n, \cdots, \theta_n^d)$

- Générer $\theta_1^{n+1} = \pi(\theta_1 \backslash \theta_2^n, \theta_3^n, \cdots, \theta_d^n, x)$
- Générer $\theta_2^{n+1} = \pi(\theta_2 \backslash \theta_1^{n+1}, \theta_3^n, \cdots, \theta_d^n, x)$

  $\vdots$
- Générer $\theta_{d-1}^{n+1} = \pi(\theta_{d-1} \backslash \theta_1^{n+1}, \theta_2^{n+1}, \cdots \theta_{d-2}^{n+1}, \theta_d^n, x)$
- Générer $\theta_d^{n+1} = \pi(\theta_d \backslash \theta_1^{n+1}, \theta_2^{n+1}, \cdots, \theta_{d-1}^{n+1}, x)$

Exemple : cas fréquentiste bivarié

Nous permet de générer x suivant $f(x)$ en utilisant la densité conditionnelle en dimension $d = 2$ Soit (x, y) vecteur aléatoire, $f(x, y)$ est une densité jointe.

Nous allons générer les couples d'observation (x_i, y_i) à partir des lois conditionnelles $X\backslash Y$ et $Y\backslash X$

Posons la valeur initiale y_0

Nous générons une observation x_1 à partir de $f(x\backslash y_0)$ comme suit :

- Nous avons $Y_0 = y_0$
- Nous génèrons $X_0 \sim f(X\backslash Y_0 = y_0)$
- $Y_1 \sim f(Y\backslash X_0 = x_0)$
  $X_1 \sim f(X\backslash Y_1 = y_1)$
  $\vdots$
  $Y_n \sim f(Y\backslash X_{n-1} = x_{n-1})$
  $X_n \sim f(X\backslash Y_n = y_n)$

Avec ce système nous aurons une chaine de Markov irréductible stationnaire de loi cible $f(x, y)$

Chapitre 3

Review

3.1 Introduction

Beaucoup de travaux de recherche ont été réalisés pour résoudre divers problèmes de survie liés à plusieurs domaines, ainsi qu'améliorer l'efficacité des modèles existant en prenant en considération une, deux , une infinité de variables, du point de vue fréquentiste et Bayésien.

Dans le cas de l'analyse de survie univariée, citons par exemple les travaux effectués par Lawless (1982), Cox et Oakes (1984),et Kalbfleisch et Prentice (2002).

Lawless a exposé ses travaux dans son ouvrage intitulé « Statistical models and Methods for Lifetime Data » dans la première édition a été publiée fois en 1982.
Il a rassemblé des documents sur l'analyse des données sur la durée de vie et donné une présentation complète des modèles et méthodes importants.
Son objectif était de donner une large couverture de la zone de l'analyse des données de survie sans se concentrer exclusivement sur un domaine d'application spécifique.
Il n'a pas travaillé dans d'autres domaines à part celui de biomédicale car la plupart des exemples qu'il a traités toutefois de l'ingénierie ou des sciences biomédicales, où ces méthodes sont largement utilisées.
Il a aussi interprété divers modèles paramétriques et leurs méthodes statistiques associées et des méthodes non paramétriques ainsi que des procédures graphiques.

Cox et David Oakes dans leur livre qui a été publié en 1984 intitulé « analysis of survival data »,ont dit que l'analyse de la survie ne se limite pas aux statistiques médicales où les avantages de l'analyse des données sur des facteurs tels que l'espérance de vie.
Les techniques de l'étude de survie trouvent également des applications importantes dans les tests de survie industrielle et une gamme de sujets allant de la physique à l'économétrie.

En 2002 , Kalbfleisch et Prentice ont publié leurs travaux dans le but de la collecte et la présentation unifiée des modèles et méthodes statistiques de l'analyse des données

de survie qui ont découlé principalement des contextes biomédicaux et des tests de survie industriels.
Les bases théoriques de ces méthodes ont été renforcées par de nombreuses généralisations importantes .
Les méthodes de comptage et les résultats de convergence des martingales associés ont permis d'obtenir des résultats asymptotiquement précis et généraux pour les tests et les estimateurs dans des classes clés de modèles de temps de défaillance, et d'importants mécanismes de censure et de troncature.
En préservant la caractéristique d'une présentation plutôt élémentaire et classique basée sur la vraisemblance des modèles et méthodes de temps de défaillance tout en intégrant la notation du processus de comptage et la théorie associée, ils ont traité l'estimation de la fonction de survie et de la comparaison de ses courbes.

Dans le cas bivarié, il existe aussi de nombreuses distributions à deux variables qui peuvent être utilisées dans l'analyse de données appariées (paired data).

Des travaux de recherche ont été réalisés dans ce modèle suivant la distribution exponentielle qui joue un rôle très important pour régler divers problèmes de survie, car elle est caractérisée par un taux de défaillance (fonction de risque) constant, et l'absence de mémoire.

Les distributions exponentielles bivariées sont largement utilisées tout en discutant de l'analyse de survie paramétrique de deux composants d'un système.
La littérature médicale considère les organes jumelés comme les reins, les yeux, les poumons, les seins, etc. d'un individu comme un système à deux composants, qui fonctionnent dans des circonstances d'interdépendance.

Weier (1981) : a proposé un estimateur Bayésien de la fonction de fiabilité en utilisant une fonction a priori non informatif et conjugué. Nous allons le voir en détail dans la première section .

Klein et Basu(1985) ont considéré l'estimation de la fonction de survie bivariée du modèle ACBVE (l'exponentielle bivariée absolument continue) en utilisant une approche classique.

Pena et Gupta (1990) ont considéré l'estimation Bayésienne des paramètres de la distribution exponentielle à deux variables de Marshall-Olkin dans deux situations à savoir : lorsque les composants sont dans un système en série et parallèle. Ils ont obtenu le mode a postérieuri dans la distribution exponentielle à deux variables de Marshall-Olkin (1967)en utilisant la distribution gamma Dirichlet comme distribution a priori.

Hanagal et Kale (1991, 1992) ont fait un travail approfondi sur les modèles exponentiels à deux variables, et ils ont obtenu l'estimateur du maximum de vraiesemblance des paramètres et développés des tests pour l'indépendance de deux composants et des tests

de symétrie de deux composants dans les modèles exponentiels bivariés basés sur les estimateurs du maximum de vraisemblance .

Hanagal et Ahmadi (2009) ont adopté une approche empirique Bayésienne du problème exponentiel à deux variables qu'on détaillera dans la suit dr notre travail.

Dans ce qui suit, nous allons définir les quatre types de distributions de survie exponentielles à deux variables, qui seront utilisés dans l'estimation dans les trois articles de recherche, le premier intitulé " estimation Bayésienne de modèle de survie bivarié basé sur les distributions exponentielles " réalisé par D. R. Weier, le deuxième est " Estimation Bayésienne des paramètres des distributions exponentielles bivariée " écrit par David D. Hanagal et K. A. Ahmadi, le dernier fait par Assia Chadli, Hamida Talhi et Hocine Fellag intitulé " Comparaison de l'estimateur de maximum de vraisemblance et l'estimateur Bayésien de la distribution exponentielle bivariée symétrique sous différentes fonctions perte ".

3.2 Les distributions exponentielle de survie bivaeiées

3.2.1 La distribution de Frend (1961)

De nombreux modèles bivariés sont dérivés du modèle exponentiel et développés par divers auteurs, par exemple,la distribution de Frend (1961) qui a donné la distribution jointe d'un système de deux composantes A et B.

Quand elles fonctionnent séparément (indépendamment), leurs fonctions de survie sont de loi exponentielle $\exp(\alpha)$ et $\exp(\beta)$ respectivement.
Quand elles fonctionnent simultanément, et quand un échec se produit en A ou B , les fonctions de survie de ses composantes devient $\exp(\grave{\alpha})$ et $\exp(\grave{\beta})$ respectivement.
Si la composante B est défectueuse alors $\grave{\alpha} > \alpha$ ou bien si A est défectueuse alors $\grave{\beta} > \beta$ c'est-à-dire le risque de défaillance de la composante restante augmente.

3.2.2 La distribution de Marshall-Olkin (1967)

La distribution exponentielle bivariée la plus populaire est la distribution exponentielle à deux variables de Marshall-Olkin (1967) qui ont obtenu une distribution exponentielle bivariée dite, distribution exponentielle à deux variables de Marshall-Olkin (MOBE) comme modèle de choc, à partir de trois distributions exponentielles indépendantes.Il a aussi une belle interprétation physique.
Les marginales de cette distribution sont des distributions exponentielles et une distribution singulière.
Par conséquent, s'il existe des liens dans les données de la distribution exponentielle à deux

variables de Marshall-Olkin peut être utilisée de manière assez efficace pour modéliser de tels ensembles de données.

3.2.3 La distribution de Block et Basu (1974)

Nous avons aussi les travaux de Block et Basu (1974) qui ont obtenu la distribution exponentielle à deux variables de Block-Basu à partir de la distribution exponentielle à deux variables de Marshall-Olkin en supprimant la partie singulière et en ne conservant que la partie absolument continue .
Contrairement à la distribution de Marshall-Olkin, la distribution exponentielle à deux variables de Block-Basu est une distribution absolument continue .
Parmi les différentes distributions exponentielles bivariées, la distribution exponentielle à deux variables de Block-Basu a reçu le plus d'attention en raison de sa flexibilité et de sa facilité d'analyse.
Ses travaux ont été largement utilisés pour l'analyse des données à deux variables, en particulier lorsqu'il n'y a pas de lien dans l'ensemble de données, même si les marginales ne sont pas des distributions exponentielles, contrairement à la distribution de Marshall-Olkin.

3.2.4 La distribution de Proschan et Sullo (1974)

son modèle exponentiel bivarié n'est pas absolument continu, est une combinaison des modèles Marshall-Olkin et Freund .

3.3 Estimation Bayésienne de modèle de survie bivarié basé sur les distributions exponentielles (D.R Weier)

D. R. Weir en 1981 a discuté dans son article, l'analyse Bayésienne du modèle de survie à deux variables régies par la distribution exponentielle, en utilisant les lois a priori vagues et conjuguées.
Ses travaux sur le modèle de survie bivariée sont liés aux distributions de Freund et Block et Basu qui ont des distributions absolument continues.Nous allons les définir comme suit :

3.3.1 La distribution de Freund (1961)

Cette distribution est absolument continue, elle est basée sur la survie conjointe de deux composants qui sont initialement indépend et sur des tests avec des distributions de survie exponentielle des paramètres α et β respectivement.

La défaillance d'un composant réduit la durée de vie moyenne supplémentaire du composant restant en augmentant le risque soit de α à $\grave{\alpha}$ ou de β à $\grave{\beta}$.

3.3.2 La distribution de Block et Basu (1974)

C'est une reparamétrisation du cas particulier de la famille Freund ayant une généralisation bivariée des distriburions qui possèdent le manque de propriété de mémoire LMP(lack of memory property).

L'auteur a considèré comme estimateur du maximum de vraisemblance (MLE) de la distribution exponentielle bivariée absolument continue.
La démarche consiste à estimer les paramètres, en utilisant la méthode de l'estimateur du maximum de vraisemblance et l'estimation Bayésienne en utilisant les lois a priori conjuguée et vague et aussi estimer la fonction de fiabilité.
L'étude de Monte-Carlo a pour objectif de comparer les résultats tels que, les estimateurs à priori vague sont essentiellement équivalents aux leurs MLE, mais les estimateurs à priori conjugué ont l'avantage d'introduire une connaissance préalable des paramètres.
Considérons un système qui fonctionne longtemps quand deux fonctions sont identiques ou très similaires.
Initialement, les deux composants doivent être testés indépendamment avec des distributions de durée de survie exponentielle avec les paramètres λ, notés $\exp(\lambda)$, l'échec de l'un modifie la distribution de la survie de l'autre en $\exp(\lambda\theta)$,$\theta > 0$, ils ont distingué les cas suivants :
- Si $\theta = 1$ implique l'indépendance des deux composants.
- Si $\theta > 1$ la charge de travail du composant restant augmente implique une diminution de sa durée de vie moyenne
- dans le cas $\theta \geq 1$, ce modèle est une reparamétrisation d'un cas particulier du modèle Freund.
- Dans le cas $1 \leq \theta < 2$ donne des reparamétrisations des distributions de la famille Block-Basu

Soit X et Y les durées de vie des deux composants.
Le modèle est plus facilement dérivé en fonction de $U = \min(X, Y)$ et $W = \max(X, Y) - \min(X, Y)$.
U suit une loi exponentielle$\exp(2\lambda)$ et par le la propriété d'absebse de mémoire univarié, W est distribué indépendamment sous la forme $\exp(\lambda\theta)$. Ainsi, la densité jointe de U et$W$ est

$$f(u, w) = 2\theta\lambda^2 \exp(-2\lambda u - \lambda\theta w), u, w > 0, \theta, \lambda > 0 \tag{3.1}$$

Des estimateurs conjoints de fiabilité du délai avant la première défaillance et du temps supplémentaire pour la seconde défaillance ont été fournis.
L'estimateur Bayésien en utilisant l'apriori vague et MLE et sont simples à calculer, mais l'étude de Monte-Carlo a indiqué la performance supérieure de l'estimateur de Bayes.
Cela implique que l'estimateur Bayes correspondant de la fiabilité du système, c'est-à-dire

le temps d'un deuxième échec, peut également être plus performant que celui de l'approche MLE.
Cependant, dans ce cas, l'estimateur de Bayes, sans forme fermée, est plus difficile à calculer.
Là encore, l'utilisation d'un conjugué a priori l'estimation de la fiabilité de Bayes aurait l'avantage d'incorporer des connaissances a priori.

3.4 Estimation Bayésienne des paramètres des distributions exponentielles bivariée (Hanagal et Ahmadi)

Cet article réalisé pa rHanagal et Ahmadi en 2009, ils ont développé une approche Bayésienne empirique, dans ce cas ils ont estimé les hypers paramètres des distributions a priori , pour obtenir les estimations Bayésiennes des paramètres dans les quatre distributions exponentielles bivariées (bivariate exponential) BVE.Cette approche empirique de Bayes est utile par rapport à la méthode traditionnelle de Bayes lorsque nous ne connaissons pas les hypers paramètres de distributions a priori. Dans les méthodes Bayésiennes traditionnelles, les hypers paramètres sont connus.

Dans cet article, l'auteur utilise une approche Bayésienne empirique dans quatre modèles BVE suivants proposés par : Marshall et Olkin (1967), Block et Basu (1974), Freund (1961), er le modèle de Proschan et Sullo (1974), nous considérons (X, Y) comme des temps de défaillance correspondant à deux composants.comme suit

3.4.1 Marshall et Olkin (M-O)(1967) :

ont proposé un BVE avec deux propriétés importantes à savoir la perte de propriété de mémoire et les fonctions exponentielles marginales. La fonction de survie de deux composants (X, Y) est définie comme suit

$$\begin{aligned} \bar{F_M}(x,y) &= P(X > x, Y > y) && (3.2)\\ &= \exp[-\lambda_1 x - \lambda_2 y - \lambda_3 \max(x,y)] avec x, y > 0 et \lambda_1, \lambda_2, \lambda_3 > 0 && (3.3) \end{aligned}$$

Les fonctions marginales de X et Y suivent une loi exponentielle de paramètre $(\lambda_1 + \lambda_3)$ et $(\lambda_2 + \lambda_3)$ respectivement.Dans le cas ou le paramètre $\lambda_3 = 0$,implique que les deux composantes sont indépendantes.
Soit $(X_i, Y_i), i = 1, 2, 3, \cdots, n$ un échantillon aléatoire de taille n de BVE de Marshall et Olkin (1967) et soit n , n le nombre des observations avec $X < Y$, $X > Y$ respectivement. La distrivbution de (n, n) est trinomiale de parametre $(n, \lambda_1 \backslash \lambda, \lambda_2 \backslash \lambda)$ avec $\lambda = \lambda_1 + \lambda_2 + \lambda_3$

La fonction de probabilité de la fonction à deux variables de M-O est

$$f_{M-O}(x,y) = \begin{cases} \lambda_1(\lambda_2+\lambda_3)\exp(-\lambda_1 x-(\lambda_2+\lambda_3)y & \text{si } 0<x<y \\ \lambda_2(\lambda_1+\lambda_3)\exp(-\lambda_2 y-(\lambda_1+\lambda_3)x & \text{si } 0<y<x \\ \lambda_1\exp(-\lambda x) & \text{si } 0<y=x \end{cases} \tag{3.4}$$

Ce modèle n'est pas absolument continu par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$

3.4.2 Block–Basu (B–B)(1974) :

Block et Basu ont proposé un BVE absolument continu comme modèle alternatif à Marshall et Olkin . La fonction de survie de BVE de Block et Basu est donnée par :

$$S(x,y) = \frac{\lambda}{\lambda_1+\lambda_2}\exp\{-\lambda_1 x-\lambda_2 y-\lambda_3\} - \frac{\lambda_3}{\lambda_1+\lambda_2}\exp\{-\lambda\max(x,y)\} \tag{3.5}$$

Avec $\lambda_1,\lambda_2,\lambda_3>0$ et $\lambda=\lambda_1+\lambda_2+\lambda_3$
La distribution marginale de X et Y n'est pas exponentielle, mais une combinaison pondérée de deux exponentielles avec des echeles $(\lambda_i+\lambda_3)$, $i=1,2$ des poids $1+\frac{\lambda_3}{\lambda_1+\lambda_2}$ et $\frac{-\lambda_3}{\lambda_1+\lambda_2}$.
La densité de probabilité de ce modèle est donné par

$$f_{B-B} = \begin{cases} \frac{\lambda\lambda_1(\lambda_2+\lambda_3)}{(\lambda_1+\lambda_2)}\exp(-\lambda_1 x-((\lambda_2+\lambda_3)y) & \text{si } 0<x<y \\ \frac{\lambda\lambda_2(\lambda_1+\lambda_3)}{(\lambda_1+\lambda_2)}\exp(-\lambda_1 y-((\lambda_2+\lambda_3)x) & \text{si } 0<y<x. \end{cases} \tag{3.6}$$

3.4.3 Freund(1961)

BVE proposé comme modèle pour la distribution du temps de défaillance d'un système avec des durées de vie (X,Y) fonctionnant de la manière suivante : X et Y sont des distributions exponentielles de taux d'échec θ_1,θ_2 respectivement avec $\theta_1,\theta_2>0$ L'interdépendance des composants est telle que la défaillance d'un composant modifie le taux de défaillance des autres composants de θ_1 à $\grave{\theta}_1$ et θ_2 à $\grave{\theta}_2$
La distribution jointe exponentielle bivarivé de Freund est donnée par

$$f_F(x,y) = \begin{cases} \theta_1\grave{\theta}_2\exp(-\grave{\theta}_2 y-(\theta_1+\theta_2-\grave{\theta}_2)x) & \text{si } 0<x<y \text{ est pair} \\ \theta_2\grave{\theta}_1\exp(-\grave{\theta}_1 x-(\theta_1+\theta_2-\grave{\theta}_1)y) & \text{si} 0<y<x \end{cases} \tag{3.7}$$

Avec $\theta_i>0,\grave{\theta}_i>0, i=1,2$ Pour $P(X=Y)=0$ ce qui donne la continuité absolue de la distribution.

3.4.4 Proschan-Sullo (P-S)(1974)

Ce modèle exponentiel bivarié est une combinaison des modèles Marshall-Olkin et Freund et le système à deux composants fonctionne de la manière suivante.

Au début X et Y suivent une loi exponentielle bivariée de Marshall-Olkin des paramétres θ_1,θ_2, et θ_3 Lorsqu'une composante échoue, implique la modification du taux

d'échec de l'autre composante de $\theta_1 + \theta_3$ à $\grave{\theta}_1 + \theta_3$ ou $\theta_2 + \theta_3$ à $\grave{\theta}_2 + \theta_3$.La densité exponentielle bivariée proposé par Proschan-Sullo est définie par

$$f_{P-S(1)} = \begin{cases} \theta_1 \grave{\eta}_2 \exp(-\grave{\eta}_2 y - (\theta - \grave{\eta}_2)x) & \text{si } 0 < x < y \\ \theta_2 \grave{\eta}_1 \exp(-\grave{\eta}_1 x - (\theta - \grave{\eta}_1)y) & \text{si } 0 < y < x \\ \theta_3 \exp(-\theta x) & \text{si } 0 < x = y. \end{cases} \tag{3.8}$$

avec
$\theta = \theta_1 + \theta_2 + \theta_3$
$\grave{\eta}_1 = \grave{\theta}_1 + \theta_3$
$\grave{\eta}_2 = \grave{\theta}_2 + \theta_3$
et
$\grave{\eta}_1 \neq \theta \neq \grave{\eta}_2$
cette fonction n'est pas absolument continue les paramètres ont été estimés sur la base de la méthode des moments et des estimations du maximum de vraisemblance (MLE).
Une étude de simulation a été menée pour calculer les estimations Bayésiennes empiriques des paramètres et de leurs erreurs.
il utilise des estimateurs de moment ou MLE pour estimer les hypers paramètres des distributions a priori. De plus, la comparaison du mode a posteriori des paramètres obtenus par différentes distributions a priori et les estimations Bayésiennes basées sur la loi gamma a priori sont très proches des vraies valeurs par rapport aux fonctions a prioris impropres.

l'auteur a utilisé la méthode MCMC pour obtenir la moyenne a posteriori et comparons la même en utilisant les a priori impropres et les estimations classiques, les MLE.

Les Bayésiens pensent que les informations sur les paramètres inconnus peuvent être exprimées sous forme de distribution de probabilités, que ce soit avant les données (en priori) ou après les données(en a posteriori).
L'approche Bayésienne des statistiques traite les paramètres inconnus comme des variables aléatoires.Les distributions a priori sont des informations de modèle des paramètres et peuvent être obtenues sur la base d'expertise et de données préalables.
En revanche, l'approche classique n'a pas besoin de distributions a priori, car elle traite les paramètres inconnus comme des constantes fixes.
Dans cet article, l'auteur a montré que les estimations Bayésiennes sont plus proches des valeurs vraies si les hypers paramètres de la distribution a priori sont sélectionnés sur la base des MLE ou des estimateurs de moments des paramètres du modèle(une approche Bayésienne empirique).

Dans la deuxième partie de l'analyse, l'auteur calcule la moyenne a posteriori par la méthode MCMC.
L'auteur a également montré que la mise en œuvre de l'analyse Bayésienne basée sur des procédures de calcul pour ces quatre modèles de BVE n'est pas difficile. Dans l'étude médicale, les modèles BVE sont très utiles quand les chercheurs travaillent sur des organes jumelés tels que les reins, les yeux, les poumons et les seins.

3.5 Comparaison de l'estimateur de maximum de vraisemblance et l'estimateur Bayésienne de la distribution exponentielle bivariée symétrique sous différentes fonctions perte (Assia Chadli, Hamida Talhi et Hocine Fellag)

Dans cet article, l'auteur s' intéresse à l'estimation des paramètres et du temps moyen de bon fonctionnement (the mean time between failure) MTBF d'un modèle exponentiel à deux variables basé sur les travaux de Freund (1961) et eux de Block et Basu (1974) , en supposant dans le modèle de Freund que les composantes sont identiques
$\alpha = \beta = \lambda$ et $\grave{\alpha} = \grave{\beta} = \lambda\theta$ la reparamétrisation de la densité $f(x, y)$ donnée par Block et Basu, supposant que $\lambda_1 = \lambda_2 = 2\lambda(2 - \theta)$ et $\lambda_{12} = 2\lambda(\theta - 1)$

La fonction de densité est définit par :

$$f(x,y) = \begin{cases} \theta\lambda^2 \exp -2\lambda x - \theta\lambda(y-x) & \text{si } x < y \\ \theta\lambda^2 \exp -2\lambda y - \theta\lambda(x-y) & \text{si } y < x \end{cases} \tag{3.9}$$

avec $\lambda > 0, \theta > 0$.
Dans le cas où $\theta \geq 1$,donne le modèle de Freund.
Quand $1 \leq \theta \leq 2$ on obtinet le modèle de Block et Basu. La fonctionn $f(x, y)$ donné en (2) est définie sur $\mathbb{R}^2$.
Ils ont effectué le changement de variable suivant

$$\begin{aligned} U &= \min(X, Y) \\ V &= \max(X, Y) \\ W &= V - U \end{aligned}$$

La densité de (U, V) est

$$f_{U,V}(u,v) = 2\theta\lambda^2 \exp(-2\lambda u - \lambda\theta(v-u)) pour 0 < u < v \tag{3.10}$$

La distribution de (U, W) est

$$f_{U,W}(u,w) = 2\theta\lambda^2 \exp(-2\lambda u - \lambda\theta w) pour u > 0, w > 0 \tag{3.11}$$

Alors U et W sont indépendants avec $U \sim Ex(2\lambda)$ et $W \sim Ex(\lambda\theta)$. Dans notre modèle, la fonction de survie correspondante pour $V = \max(X, Y)$ est

$$f_V(v \backslash \lambda, \theta) = \begin{cases} \frac{2\theta\lambda}{(\theta-2)}\{\exp(-2\lambda v) - \exp(-\lambda\theta v)\} & \text{si } \theta \neq 2 \\ 4\lambda^2 v \exp(-2\lambda v) & \text{si } \theta = 2. \end{cases} \tag{3.12}$$

dans le cas ou $\theta = 2$, V suit une loi de $Gamma(2, 2\lambda)$

Le temps moyen de bon fonctionnement (MTBF)est

$$T_0 = \begin{cases} E(V\backslash\lambda,\theta) = \int_0^\infty v f_V(v\backslash\lambda,\theta)dv = \frac{2+\theta}{2\lambda\theta} & \text{si } \theta \neq 2 \\ \frac{1}{\lambda} & \text{si } \theta = 2 \end{cases} \tag{3.13}$$

Dans ce travail, l'auteur considère l'estimation des paramètres et de MTBF noté T_0 dans le modèle bivarié donné au-dessus, et dans l'objectif de prouver avec une étude de Monte-Carlo, qu'il est possible d'améliorer l'estimation du maximum de vraisemblance en utilisant une méthode Bayésienne et une fonction perte appropriée.
Pour cela ils ont procédé comme suit

Tout d'abord l'auteur a utilisé l'estimateur de maximum de vraisemblance classique puis l'approche Bayésienne pour estimer les paramètres et du temps moyen entre les échecs(mean time between failure noté MTBF) en utilisant des lois a priori non informative et conjugué, il a considéré les fonctions pertes suivantes :

3.5.1 Les fonctions perte

Fonction perte quadratique

La fonction perte est dédinit par

$$L_1(t,d) = (t-d)^2$$

qui a été proposé par Legendre (1805) et Gauss (1810) l'estimateur Bayésien de t est la moyenne a postériori $\hat{d}_B = E(t\backslash x)$.

Fonction perte DeGroot

Cette fonction est introduite par DeGroot(1970), est définie comme suit

$$L_2(\theta,d) = (\frac{t-d}{d})$$

Son estimateur Bayésien est $\hat{d}_B = \frac{E(t^2\backslash x)}{E(t\backslash x)}$.

Fonction perte linex

C'est une fonction asymétrique introduite par Varian (1975) comme suit

$$L_3(t,d) \propto \exp(a\Delta) - a\Delta - 1$$

Avec $a \neq 0$ et $\Delta = (d-t)$.
l'estimateur Bayesien est : $\hat{d}_B = -\frac{1}{a}\ln(E(\exp(-at)))$
quand a tend vers 0 , ils obtiennent la fonction perte quadratique.

Fonction perte d'entropie

Calabria et Pulcini (1994) ont déduit une nouvelle fonction perte de Linex comme suit :

$$L_4(t,d) = \left(\frac{d}{t}\right)^p - p\ln\left(\frac{d}{t}\right) - 1$$

Son estimateur est $\hat{d}_B = [E(\theta^{-p})]^{-\frac{1}{p}}$ Dans le contexte Bayésien, lorsqu'il y a peu ou pas d'informations sur le paramètre, il faut utiliser des fonctions a prioris vagues.

Une étude de Monte-Carlo est réalisée pour faire la comparaison des estimateurs en utilisant la proximité de Pitman et les critères d'efficacité relatifs.L'auteur montre ensuite que l'estimation du maximum de vraisemblance des paramètres et MTBF de la distribution exponentielle à deux variables peuvent être améliorées en utilisant la méthode Bayésienne. L'auteur a prouvé également que cette amélioration peut être efficace en utilisant une fonction perte appropriée.
Pour les perspectives futures, l'auteur a proposé de construire un mélange des fonctions perte utilisées dans cet article pour obtenir une estimation optimale.

Chapitre 4

Application sur des données réelles

4.1 Introduction

L'analyse des modèles de survie est une branche statistique qui est née dans le domaine médical et son application a pu s'étaler dans plusieurs domaines comme par exemple le domaine d'économie, la fiabilité, les assurances, la biologie, la physique ...

Dans notre cas pratique nous ferons une étude de survie applicable dans le domaine médical et qui est l'étude de survie des femmes atteintes d'un cancer du sein.
L'événement étudié est le décès, c'est le passage entre deux états nommés vivant et décès.
L'événement terminal est le décès du patient.

Le cancer du sein est la forme de cancer la plus répandue chez la femme dans le monde entier, c'est l'une des principales causes de décès par cancer chez les femmes. Le cancer du sein peut aussi apparaître chez l'homme mais ce n'est pas fréquent.
C'est un phénomène qui a connu une croissance ces dernières 20 années .
La lutte contre ce cancer se base essentiellement sur la détection précoce associée à un traitement adapté et au dépistage.
Certaines études ont démontré que la thérapie hormonale adjuvante et les polychimiothérapies réduisent le risque de récidive et de mortalité par cancer du sein et augmentent la durée de survie.

Les facteurs pronostiques qui sont sensés être des facteurs indépendants incluent l'extension ganglionnaire et le statut lymphatique, la taille tumorale et le statut des récepteurs hormonaux d'œstrogène et de progestérone (ER/PR).Les facteurs additionnels incluent le grade, l'extension vasculaire et lymphatique, l'âge et le facteur héréditaire. Certains facteurs biologiques, incluant ER/PR et HER2, sont à la fois des facteurs pronostiques et prédictifs.
Le principal facteur pronostique pour les tumeurs diagnostiquées à un stade précoce est la présence ou l'absence d'atteinte ganglionnaire axillaire. En outre, il y a un lien direct entre

le nombre de ganglions axillaire atteint et le risque de récidive à distance.

Notre étude a été réalisé au sein de l'hôpital universitaire NEDIR MOUHAMED de Tizi-Ouzou au niveau du service d'oncologie.

L'objectif de notre travail est de faire une analyse statistique des modèles de survie sur un échantillon de 186 femmes atteintes d'un cancer du sein traitées au niveau de service d'oncologie.

4.2 Méthodologie de travail

Le suivi s'est étalé sur une période de 7 ans et 8 mois, équivalent à 104 mois. qui commence le 10 Novembre 2009 . La date de diagnostic correspond à la date d'origine et la date de point a été fixé au 10 Juillet 2017 .

Les données ont été recueillies à partir des dossiers médicaux et une enquête complémentaire pour les femmes ayant interrompu leur suivi à un moment donné. L'un des moyens de notre enquête est de joindre les patientes par téléphone .
Ensuite, pour celles que nous n'avonspas pu joindre nous avons cherché leurs données au niveau de l'association d'Al Fedjr et l'APC d'Ait Aissa Mimoune de Ouagnoun pour se renseigner sur leur état .
Les informations sur la date et la cause de décès ont été prélevées en consultant leurs actes de naissance et ceux du décès.

Les données reculées ont trait aux informations personnelles de chaque patiente à savoir l'âge, le poids, la taille et la situation familiale. Aussi les informations concernent la tumeur comme le côté de localisation de la tumeur, le type et la taille tumorale, les métastases ganglionnaires lymphatiques et régionaux et la métastase à distance et d'autres informations liées au traitement de cette maladie : chimiothérapie la radiothérapie, la chirurgie, l'hormonothérapie , thérapie ciblée,et les types des récepteurs hormonaux et HER2.

Les données ont été saisies sur l'Excel et analysées à l'aide du logiciel de statistique le R. Les tracés des courbes de survie ont été réalisés par la méthode de Kaplan-Meier. Le seuil de signification a été fixé à 0.05 .

4.3 Description des données

4.3.1 Informations personnelles

La majorité des patientes sont mariées, elles représentent 61.29%. 15.60% sont des femmes célibataires dont le reste sont des données manquantes.
La maladie survient dans 42.47% chez les femmes âgées entre 40 ans et 50 ans ce qui représente la classe médiane.Nous retrouvons des cas rares de cancer du sein chez les

malades de moins de 30 ans.
Enfin le poids moyen des individus est de $69kg$ et la taille moyenne est de $1.57m$.

La représentation graphique suivante nous résume la distribution des pourcentages de notre échantillon selon l'âge des patientes ,

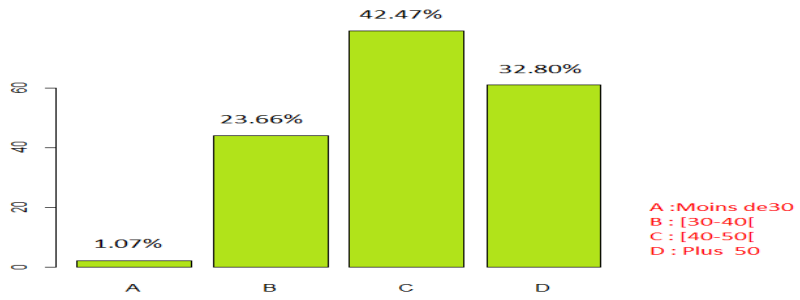


FIGURE 4.1 – Graphe représentatif de l'âge des patientes

4.3.2 Codage des variables

Afin de mieux comprendre la signification des variables qui seront données prochainement, le tableau suivant nous donne la signification de chaque code correspondant.

Variable	Code de la variable	Modalités et valeurs la variable
Type de sujet	1	Exclu vivant
	2	A subi l'événement (Le décès)
Coté de localisation de la tumeur	G	Gauche
	D	Droit
La taille tumorale (mm)	TX	Tumeur primaire non évaluable
	T0	Pas de tumeur primaire
	T1	Tumeur ≤ 20
	T2	$20 <$ Tumeur ≤ 50
	T3	Tumeur > 50
	T4A	Tumeur, quelle que soit sa taille, avec une extension directe à la paroi thoracique en excluant le muscle pectoral.
	T4B	Tumeur, quelle que soit sa taille, avec une extension directe à la peau
	T4C	T4A+T4B
	T4D	Carcinome inflammatoire
	DM	Données manquantes

Métastase ganglionnaire lymphatiques régionaux	NX	Ganglions régionaux non évaluables
	N0	Absence de métastase ganglionnaire
	N1	Métastase dans 1 à 3 ganglions axillaires ou/et Métastase supérieure à 0.2mm dans des ganglions sentinelles mammaires internes sans signe clinique
	N2	Métastase dans 4 à 9 ganglions axillaires ou Métastase mammaire interne clinique sans envahissement axillaire à l'examen microscopique
	N3	Métastase dans ≥ 10 ganglions axillaires ou Métastase mammaire interne clinique avec envahissement axillaire à l'examen microscopique ou métastase ganglionnaire su-claviculaire homolatérale
	DM	données manquantes
Métastase à distance	MX	Renseignements insuffisants
	M0	Absence de métastase
	M1	Présence de métastase
	DM	Données manquantes
Chimiothérapie	N	Néo adjuvante
	A	Adjuvante
	P	Palliative
	R	Rien
	DM	Données manquantes
Chirurgie	C	Conservatrice
	NC	Non conservatrice
	R	Rien
	DM	Données manquantes
Hormonothérapie	O	Oui
	N	Non
	DM	Données manquantes
Thérapie ciblée	O	Oui
	N	Non
	DM	Données manquantes
Les récepteurs hormonaux	P	Positif
	N	Négatif
	DM	Données manquantes
HER2	P	Positif
	N	Négatif
	DM	Données manquantes

4.3.3 Information concernant la tumeur

Puisque nous parlons de la néoplasie du sein, la tumeur est localisée soit au niveau du sein gauche ou droit ou bien bilatéral c'est-à-dire qu'il se retrouve au niveau des deux seins,

ce qui représente des cas rares.D 'après notre étude, la localisation de la tumeur au niveau du sein droit représente presque le même pourcentage que le côté gauche comme le montre ce diagramme :

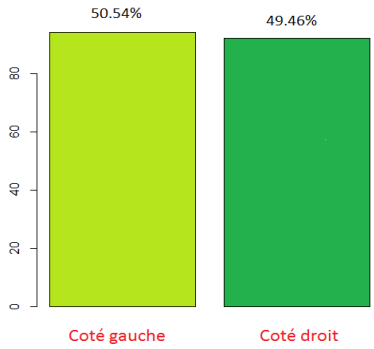


FIGURE 4.2 – Graphe représentatif de coté de localisation de la tumeur

Il existe plusieurs types de tumeurs.Il y a une catégorie très fréquente qui est le carcinome canalaire infiltrant. Dans l'échantillon nous retrouvons 167 cas qui sont équivalents à $88,71\%$ de sujets , 4.30% de carcinome lobulaire infiltrant et 6.99% qui représente d'autres types.

Classification tumorale

Selon Classification TNM du cancer du sein, 6e édition, 2002, le système TNM distingue le stade clinique pré-thérapeutique noté « cTNM » et le stade anatomopathologique post-chirurgical noté « pTNM » où
T représente la taille tumorale en millimètres linéaire (mm) du diamètre il se décompose en plusieurs cas selon la taille de la tumeur :

- TX : tumeur primaire non évaluable
- T0 : pas de tumeur primaire
- T1 : tumeur ≤ 20
- T2 : $20 <$ tumeur ≤ 50
- T3 : tumeur > 50
- T4A : tumeur, quelle que soit sa taille, avec une extension directe à la paroi thoracique en excluant le muscle pectoral
- T4B : tumeur, quelle que soit sa taille, avec une extension directe à la peau
- T4C : c'est $T4A. + T4B$
- T4D : carcinome inflammatoire

La représentation graphique de la taille tumorale, sans prendre en considération des données manquantes, nous indique que 33.87% de cas de notre échantillon ont une tumeur avec une dimension entre 20 et $50mm$ c'est ce qui représente le pourcentage le plus élevé comme le montre le graphe suivant

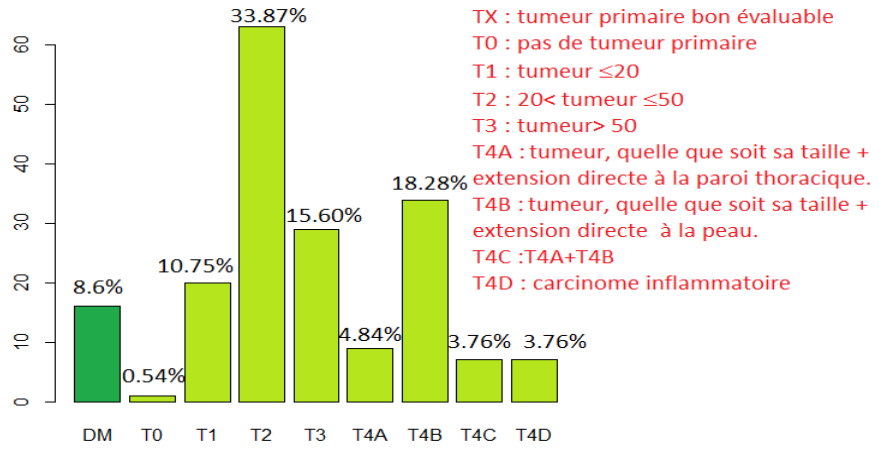


FIGURE 4.3 – La taille tumorale (diamètre (mm))

N représente la métastase ganglionnaire lymphatique régionale :

- NX indique Ganglions régionaux non évaluables

- N0 absence de métastase ganglionnaire

- N1 Métastase dans 1 à 3 ganglions axillaires ou/et Métastase supérieure à $0.2mm$ dans des ganglions sentinelles mammaires internes sans signe clinique.

- N2 Métastase dans 4 à 9 ganglions axillaires ou Métastase mammaire interne clinique sans envahissement axillaire à l'examen microscopique.

- N3 : métastase dans ≥ 10 ganglions axillaires ou Métastase mammaire interne clinique avec envahissement axillaire à l'examen microscopique ou métastase ganglionnaire sus-claviculaire homolatérale.

Les résultats de notre échantillon d'envahissement ganglionnaire sont résumés au niveau du graphe suivant :

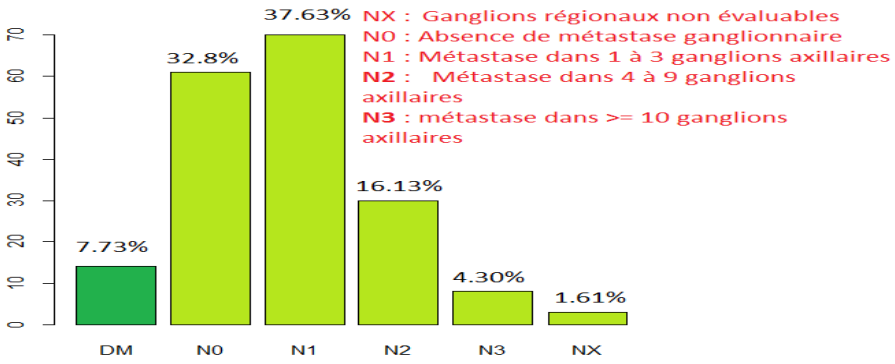


FIGURE 4.4 – L'envahissement des ganglions lymphatiques régionaux

Enfin on a M qui représente métastase à distance ; on retrouve :
- MX nous indique que les renseignements sont insuffisants.
- M0 affirme l'absence de métastases.
- M1 indique présence de métastases.

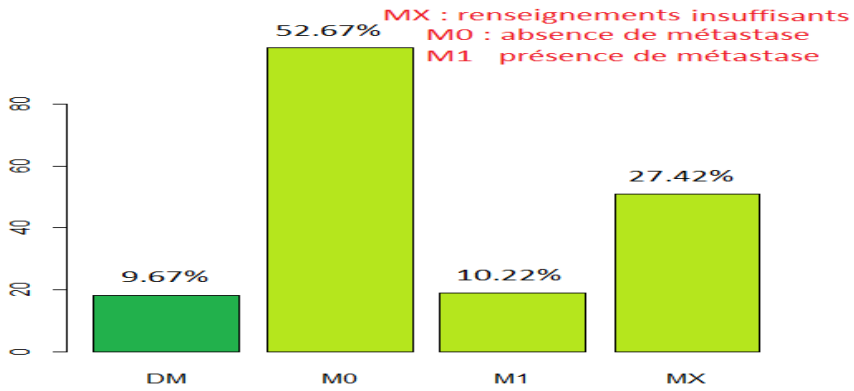


FIGURE 4.5 – Métastase à distance

Le graphe indique qu'environ la moitié de notre échantillon représente l'absence de métastase 52.69% et la présence de métastases représente le pourcentage le plus petit de 10.22 sans prendre en considération des données manquantes.

Traitement

Une fois que nous avons vu la localisation, le type et la classification de la néoplasie du sein, nous allons représenter le traitement qui contient la chimiothérapie la radiothérapie, la chirurgie, l'hormonothérapie et la thérapie ciblée.

Une fois que la maladie a été diagnostiquée , le patient passe à la phase de traitement qui commence par la chirurgie qui peut être une mastectomie qui veut dire une ablation du sien ou tractomie qui veut dire enlever juste une partie du sein et la chimiothérapie dont nous trouvons la chimiothérapie Néo adjuvante qui se fait avant la chirurgie, la chimiothérapie adjuvante qui se fait après l'opération et enfin la chimiothérapie Palliative qui est donnée pour augmenter la survie du malade ; elle est utilisée quand la tumeur est au stade final .

Ensuite il y a la radiothérapie qui se fait une fois que la chimiothérapie et la chirurgie sont faites, le traitement avec la thérapie ciblée se fait juste après la chimio il dépend de HER2(récepteur 2 du facteur de croissance épidermique humain) il est demandé uniquement dans le cas où il est positif. En dernière phase de traitement, nous avons aussi l'hormonothérapie qui dépend des récepteurs hormonaux notés RH, elle est demandée quand les récepteurs sont positifs, c'est-à-dire RE(récepteurs aux estrogènes positives),

RP(récepteurs à la progestérone positive), ou l'une d'entre elles est positive . RH prend le signe négatif dans le cas où RP et RE sont négatifs simultanément .

Dans le cas où RE et HER2 sont négatifs dits le triple négatif, ni l'hormonothérapie ni la thérapie ciblée ne seront demandées au malade, sauf dans le cas de présence de métastases, en ce moment un autre traitement sera donné après une étude médicale .

Revenant à cas pratique à étudier , la chimiothérapie est faite par la majorité des malades dont la plupart ont opté pour une chimiothérapie adjuvantée avec 82.8% ensuite la Néo adjuvante avec un pourcentage de 10.75%. La chimiothérapie palliative ne représente qu'une minorité de 3.76% de notre échantillon, tout ça sans prendre en considération des données manquantes, comme le montre le graphe suivant :

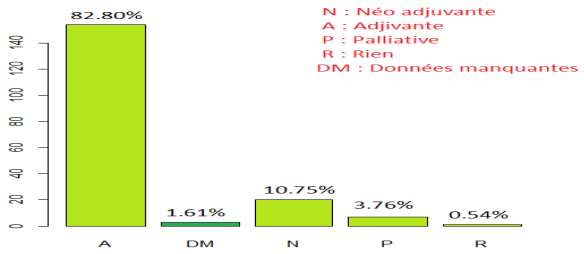


FIGURE 4.6 – Chimiothérapie

Le traitement par radiothérapie, sans prendre en compte des données manquantes, 84.95% des patientes l'ont fait, pour celles qui ne l'ont pas fait restes des cas rares dus au décès avant la fin du traitement ou pour des causes médicales.

Concernant la chirurgie nous avons trois types qui dépendent de la gravité du cas d'où la majorité des femmes soumises à ce traitement ont fait une ablation du sein avec une proportion de 76.35% par contre la chirurgie conservatrice ne présente que 17.20, rare ou la chirurgie n'est pas nécessaire pour le traitement comme le représente le graphe suivant :

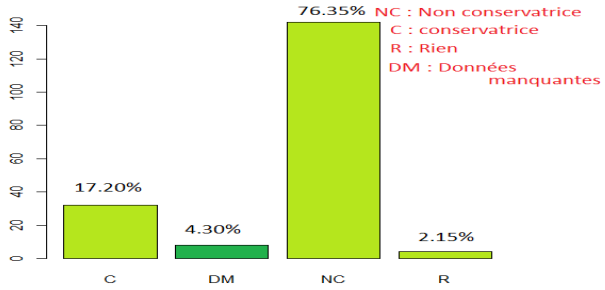


FIGURE 4.7 – Chirurgie

Le traitement par l'hormonothérapie dépend des récepteurs hormonaux comme nous l'avons dit précédemment, d'ailleurs leurs graphes sont presque similaires c'est les données manquantes qui leur donne une légère différence car parfois nous avons l'information du signe des récepteurs hormonaux et nous n'avons pas d'information si la patiente a fait l'hormonothérapie ou pas. Pour clarifier les informations nous avons les deux représentations suivantes

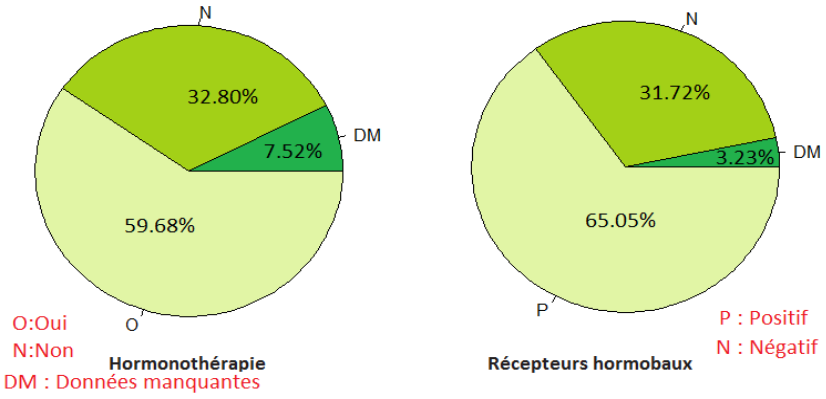


FIGURE 4.8 – comparaison entre l'hormonothérapie et les récepteurs hormonaux

Le traitement avec la thérapie ciblée dépend de HER2 comme nous l'avons expliqué dans le paragraphe précédent, d'ailleurs leurs graphes sont presque similaires la petite différence est causée par données manquantes car parfois nous avons l'information du signe de HER2 et nous n'avons pas d'information si la patiente a fait ce traitement ou pas. Les deux représentations suivantes nous donnent une observation plus claire sur la répartition des proportions de notre échantillon

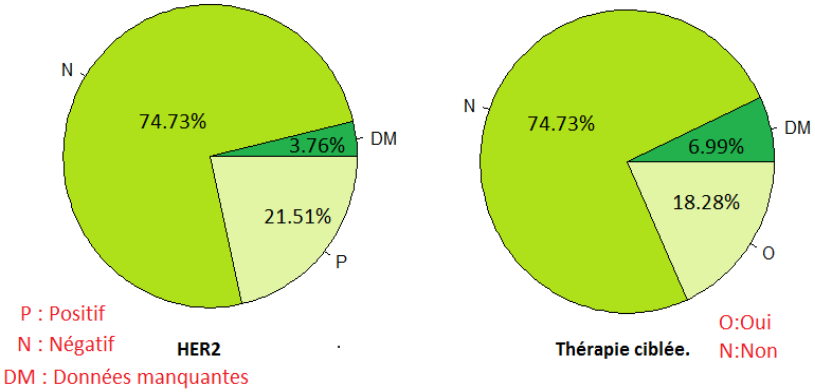
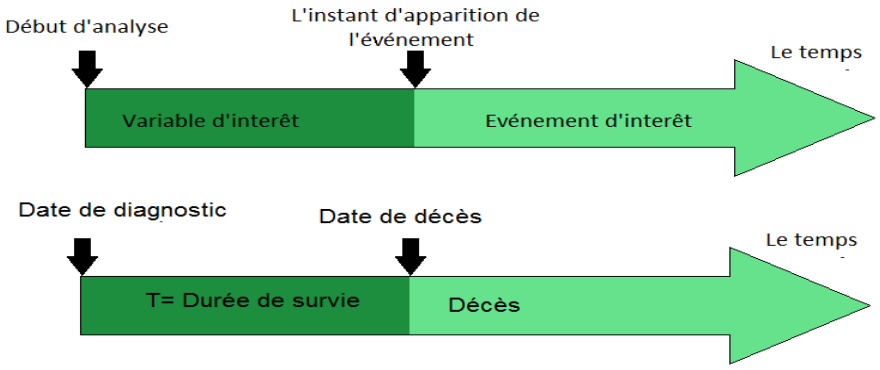


FIGURE 4.9 – comparaison entre la thérapie ciblée et RH2

4.4 Représentation de la courbe de survie

Rappelons que :
- La date de début d' obsérvation 10 Novembre 2009
- La date de fin d'obsérvation 10 Juillet 2017
- Le recul 7 ans et 8 mois, équivalent à 104 mois.
- La taille de l'eschantiloon est de 186 sujets

Nous avons les représentations suivantes



Nous avons dit que la fonction de survie théorique est représentée par une courbe de survie continue décroissante qui retrace la probabilité de survie en fonction du temps.

Pour la représentation graphique empirique de la fonction de survie S, nous avons choisi l'estimateur de Kaplan-Meier, qui est couramment utilisée .

Le nombre total des sujets exposés au décès est de 180 sujets, dans ce cas nous n'avons pas pris en considération les perdus de vue mais juste les exclus vivants qui représentent 77.96% et les décès, 22.04% dans notre étude, la majorité des sujets n'ont pas subi l'évènement d'intérêt qui est le décès,comme le montre le graphe suivant, nous nous intéressons aux décès.

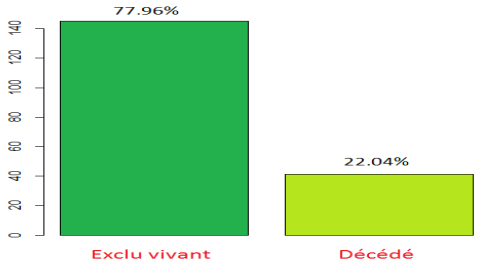


FIGURE 4.10 – Graphe représentatif des individues de notre échantillon

Comme nous travaillons sur le cas des données censurées non aléatoire à droite les perdus de vue ne sont pas pris en consideration, les exclues vivantes sont représentées par des traits verticaux. Dans notre cas ou le nombre de censures est grand, les barres ne seront pas représentées sur le graphe, pour ne pas l'encombrer et rendre la lecture difficile, comme suit :

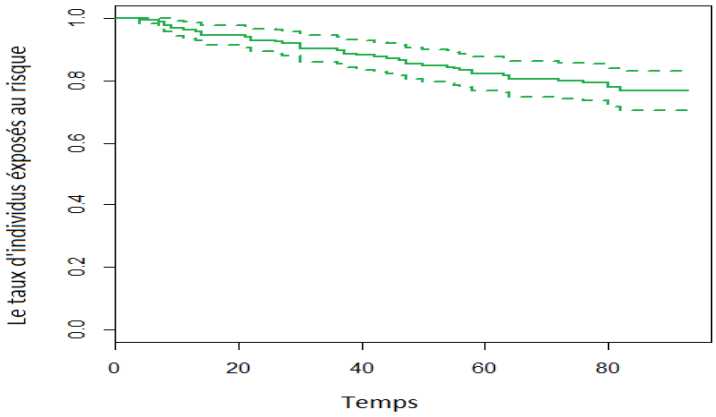


FIGURE 4.11 – Graphe représentatif de l'estimation de la fonction de survie

Nous voyons bien la fonction de survie empirique estimée par la méthode de Kaplan-Meir par une courbe décroissante sous forme d'escaliers. Chaque fois qu'un malade décède, il se traduit par une marche, sa hauteur dépend de nombres patients qui ont subi l'évènement à cet instant, nous remarquons bien que la courbe n'est pas fortement décroissante celà revient au nombre du décès qui n'est pas important. Nous remarquons qu'au début de l'observation, toutes les patientes sont encore vivantes.
La précision de l'estimation dépend de l'intervalle de confiance qui est représenté par des traits discontinus sur le graphe, qui est figuré sur le graphe avec une erreur de 0.05%.
À la fin de notre graphe, nous remarquons aussi qu'il reste un groupe de sujets qui n'a pas encore subi l'évènement, nous l'interprètons soit par l'apparence de stabilité de survie à long terme, ou bien par la disparition du risque à un certain temps.

Conclusion générale

L'objectif de notre travail est de présenter les différentes méthodes d'estimation de la fonction de survie.

Dans un premier temps, nous avons défini le concept de survie et les distributions de survie, et les caractéristiques des données de survie.

Ensuit, nous avons modélisé la survie en utilisant les trois méthodes d'estimation fréquentiste paramétrique, Semi-paramétrique du modèle de Cox qui répond aux divers problèmes qui prend en considération des modèles de risque proportionnel. Nous avons défini aussi l'estimation non paramétrique particulièrement l'estimateur de Kaplan-Meier ,le tracer et la lecture de la courbe de survie.

Puis, nous avons représenté l'estimation en utilisant la méthode Bayésienne d'analyse des modèles de survie dans le cas paramétrique, et aussi nous avons exposé les méthodes de simulation de Monte-Carlo et la méthode de Monte-Carlo par Chaîne de Markov qui sont très utiles pour la comparaison des différentes méthodes d'estimation. Et aussi une présentation de quelques travaux de recherche réalisés dans ce domaine.

Enfin de notre travail, nous avont présenté d'une application des données réelles, sur un échantillon des patients atteints d'un cancer du sien. Nous avons pu décrire les données, présenter et interpréter la fonction de survie en utilisant l'estimateur de Kaplan-Meier.

En perspectives,il serait intéressant d'étudier ces données réelles en utilisant les modèles paramétriques non paramétriques, semi-paramétriques selon l'approche fréquentiste et Bayésienne et faire une comparaison ainsi que l'impact des différentes variables sur la durée de survie des patientes.Nous pouvons aussi utiliser les méthodes de simulation pour comparer les deux approches .

Bibliographie

[1] Ancelle. T, (2002), *statistique épidémiologie*, éd MALOINE.

[2] Block, H. W., Basu, A. P, (1974), *A continuous bivariate exponential extension*, Journal ofthe American Statistical Association 69 :1031-1037.

[3] Boudjerdakhawla, (2017), *Etude de l'estimateur de Bayes sous differentes fonctions de perte*,Thèse de doctorat en Mathematiques , Université de Badji Mokhtar Annaba.

[4] Calabria. R and Pulcini. G,(1994), *An Engineering Approach to Bayes estimation for the Weibull distribution*, Micron Electron Reliab, 34. 789-802.

[5] Chadli Assia, Talhi Hamida and Fellag Hocine, (2013), *Comparison of the maximum likelihood and Bayes estimators for symmetric bivariate exponential distribution under different loss functions*,AfrikaStatistika, 2316-090X , vol.8, pages 499-514.

[6] Cox, D. R, Oakes. D (1984), *Analysis of Survival Data*,éd Chapman Hall.

[7] DeGroot .M.H, (1970), *Optimal Statistical Decisions*,éd Mc-Graw-Hill.

[8] Demengel Gilbert, Benichou Paul, Benichou Rosine, Boy Norbert, Pouget Jean-Perre, (1997), *Probabilités Statistique inférentielle Fiabilité Outils pour l'ingénieur*,éd ellipses.

[9] Freund. J.A, (1961), *bivariate Extension of the exponential distribution*, J.Amer.Statist.Assoc. Vol 56, 971-977.

[10] Gauss. C.F, (1810) , *Least Squares method for the Combinations of Observations*, Translated by Bertrand . J, (1955), Mallet-Bachelier .

[11] Genin Michael, (2015), *Introduction à l'analyse de survie*,docplayer.fr /23457390-Introduction a-l-analyse-de-survie.html

[12] Guide - Affection Longue Durée, Tumeur maligne, (2010), *affection maligne du tissu lymphatique ou hématopoïétique Cancer du sein*, urlwww.has-sante.fr et sur www.e-cancer.fr

[13] Guyon Xavier, (2001), *Statistique et économétrie du modèle liniaire aux modèles non-linéaires*, éd ellipses .

[14] Hanagal .D.D and Ahmadi. K.A, (2009), *Bayesian estimation of the parameters of bivariate exponential distribution*,Communications in Statistics - Simulation and Computation, 0361-0918, 38, Pages 1391-1413.

[15] Hanagal. D. D, Kale .B. K, (1991a), *Large sample tests of λ_3 in the bivariate exponential distribution* Statistics and Probability Letters 12 :311-313.

[16] Hanagal. D. D, Kale. B. K ,(1991b), *Large sample tests of independence in an absolutely continuous bivariate exponential distributions*, Communications in Statistics-Theory and Methods 20(4) :1301-1313.

[17] Hanagal. D. D, Kale. B. K, (1992), *Large sample tests for testing symmetry and independence in some bivariate exponential models*, Communications in Statistics-Theory and Methods 21(9) :2625-2643.

[18] Hill Catherine, Com-Nougué Catherine, Kramar Andrew, Moreau Thierry, O'Quigley John, SenoussiRachid, Chastang Claude , (1990), *Analysestatistique des données de survie*, éd INSERM.

[19] Huber Catherine, *Modèles pour des durées de survie*, www.biomedicale.parisdescartes.fr/survie/enseign/survie_sansi.pdf

[20] Jeffreys. H, (1961), *Theory of Probability*, 3rd Edition,éd Clarendon Press.

[21] John P. Klein, Melvin L. Moeschberger, (2003), *Survival Analysis Techniques for Censored and Truncated Data*, ed Springer.

[22] Jonathan Lenoir, (2013), *Analyses de Survie*http://jonathanlenoir.files.wordpress.com/2013/12/analyse-de-survie-cox.pdf

[23] Kalbfleisch. J. D, Prentice. R, (2002), *The Statistical Analysis of Failure Time Data*,éd John Wiley and Sons.

[24] Khribi, Lotfi , (2007), *L'échantillonnage de gibbs pour l'estimation bayésienne dans l'analyse de survie*, www.archipel.uqam.ca/3100/1/M9739.pdf

[25] Klein. J.P and Basu. A.P, (1985), *On estimating reliability for bivariate exponential distri- butions*, Sankhya, Series B, 47, 346-353.

[26] Lawless . J. F, (1982), *Statistical Models and Methods for Lifetime Data*, éd, John Wiley

[27] Legendre. A, (1805), *New Methods for the Determination of Orbits of Comets Courcier*, Paris.

[28] Marshall. A. W, Olkin. I, (1967), *A multivariate exponential distribution*, Journal of the American Statistical Association 62 :30-44.

[29] Paroissin Christian, (2015), *Programmation et analyse statistique avec R* éd ellipses.

[30] Pena. E. A, Gupta .A. K ,(1990) , *Bayes estimation for the Marshall-Olkin exponential distribution*, Journal of Royal Statistical Society B 52(2) :379-389.

[31] Proschan . F, Sullo. P, (1974), *Estimating the parameters of a bivariate exponential distributions in several sampling situations*. In : Proschan, F., Surfling. R. J, éd, Reliability and Biometry. Philadelphia : SIAM, pp. 423-440.

[32] Saint-Pierre. P, (2015), *Introduction à l'analyse des durées de survie*, http://www.lsta.upmc.fr/psp/Cours_Survie_1.pdf

[33] Terry .M, Themeau Patricia. M, Grambsch, (2000), *Modeling Survival Data : Extending the Cox Model*,ed Springer.

[34] Varian. H.R, (1975), *A Bayesian Approach to real estate assessment*,Amsterdam, North Hol- land. 195-208.

[35] Weier. D.R,(1981) , *Bayes estimation for bivariate survival models besed on the exponential distributions*, Communications in Statistics-Theory and Methods, 10. 1415-1427.

[36] Weier. D.R, (1981), *Bayes estimation for bivariate survival models besed on the exponential distributions*, Communications in Statistics-Theory and Methods, 10. Pages 1415-1427.