République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Mohammed Seddik Ben Yahia - Jijel

Faculté des Sciences Exacte et Informatique
Département de Mathématique

Mémoire de fin d'études

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques Appliquées

Option : Probabilités et statistique

Thème

L'utilité de l'approche bayésienne dans l'estimation de la fonction de survie.

Présenté par :

Bouzeriba Yasmina

Boumendjel Ahlam

Devant le jury :

Président	: Mr.Gherda Mebrouk	(M.A.A)	Université de Jijel
Encadreur	: M^{me} Djeridi Zahra	(M.A.A)	Université de Jijel
Examinatrice	: M^{me} Abdi Zeyneb	(M.A.A)	Université de Jijel
Examinatrice	: M^{me} Yakoubi Fatima	(M.A.A)	Université de Jijel

Promotion **2017/2018**

Remerciement

Tout d'abord, nous tenons à remercier Allah, le tout puissant et le miséricordieux, de nous avoir donné la santé, la volonté et la patience pour mener à terme notre formation de master.

C'est avec un grand honneur que nous remercions notre enseignante et promoteur M^{me} ***Djeridi Zahra*** *pour nous avoir dirigés pour la réalisation de ce travail, pour ces précieux conseils et ces encouragements.*

Nous tenons à remercier Mr ***Gherda Mebrouk*** *d'avoir accepté la présidence du jury de notre travail, qu'il trouve ici toutes nos expressions respectueuses.*

Nous remercions également M^{me} ***Abdi Zeyneb*** *de nous avoir fait l'honneur de faire partie des membres du jury et d'examiner ce travail. Nous tenons à vous remercier.*

Nous remercions également M^{me} ***Yakoubi Fatima*** *de nous avoir fait l'honneur de faire partie des membres du jury et d'examiner ce travail. Nous tenons à vous remercier.*

Nos remerciements les plus sincères s'adressent à nos familles pour leur soutien sans faille et pour l'équilibre qu'elles nous ont apporté et pour leurs encouragements.

Enfin, nous voulons remercier toutes les personnes qui ont contribué de loin ou de près à l'avancement de ce travail, nos enseignants et collègues à l'université de Jijel.

Table des matières

Notations **v**

Introduction **1**

1 Outils de bases de la statistique de survie **3**

1.1 Les données de survie 3

1.1.1 Fonction de survie et fonction de risque 5

1.1.2 Quantités associées à la loi de la survie 7

1.2 Données censurées 8

1.2.1 Censure à droite 9

1.2.2 Censure à gauche 10

1.2.3 Censure par intervalle 10

1.3 La vraisemblance pour données censurées 13

1.3.1 La censure à droite 13

1.4 Estimation de la fonction de survie 14

1.4.1 Estimateurs paramétriques 14

1.4.2 Estimateur non paramétrique 20

2 Estimation bayésienne de la fonction de survie **27**

2.1 Approche bayésienne 27

2.1.1 Modèle bayésien 28

2.2 Choix de la loi a priori 29

2.2.1 Lois a priori conjuguées 29

2.2.2 Lois a priori non informatives 31

2.3 Modèles paramétriques 33

2.3.1 Modèle exponentiel 33

2.3.2 Modèle de Weibull 35

2.3.3 Modèle Gamma 36

2.3.4 Modèle de valeur extrême 38

2.4 Méthodes non paramétriques bayésiennes 39

2.4.1 La distribution de Dirichlet 40

2.4.2 Le processus de Dirichlet 41

2.4.3 Méthodes non paramétriques bayésiennes pour estimer la fonction de survie 43

3 Applications dans les essais cliniques **47**

3.1 Essais cliniques 47

3.1.1 Les différentes phases d'un essai clinique 47

3.2 Application sur les données de VIH 48

3.2.1 Estimateur non-paramétrique de K-M 48

3.2.2 Modèle paramétrique exponentiel 49

3.2.3 Une application informatique 49

3.3 Application sur les données de Freireich 54

3.3.1 Estimateur de K-M 54

3.3.2 Estimateur avec le processus Dirichlet 55

3.4 Conclusion 57

Bibliographie **63**

Notations

n	Effectif.
$\wedge$	Minimum.
$\vee$	Maximmum.
$\mathbb{1}()$	Fonction indicatrice.
$S(.)$	Fonction de survie.
$h(.)$	Fonction de risque.
$H(.)$	Fonction de risque cumulée.
δ_i	L'indicatrice de l'évènement.
$\propto$	Proportionelle à.
$\pi(\theta)$	La loi a priori.
$\pi(\theta \mid x)$	La loi a posteriori.
$I(\theta)$	Information de Ficher.
L	Vraisemblance.
EMV	L'estimateur du maximun de vraissemblance.
IC	Intervalle de confiance.
$i.i.d$	Indépendants et identiquement distribués.
R_i	Le nombre de sujets qui sont vivants juste avant l'instant t_i.
$R(t)$	L'ensemble des sujets à risque à l'instant t^-.
VIH	Virus de l'immunodéficience humaine.
IG	La fonction gamma incomplète.
$\Gamma(\alpha)$	La fonction gamma.
DP	Processus de Dirichlet.
$\xi(\lambda)$	Distribution exponentielle de paramètre λ.

Introduction

L'analyse des données de temps d'un événement, généralement appelée analyse de la survie, est née au vingtième siècle, et a connu un développement important dans la seconde moitié du siècle [16].

Les données de survie sont généralement décrites et modélisées en fonction de deux concepts connexes ; à savoir la fonction de survie et la fonction de risque. L'analyse de survie fait référence à une famille de méthodes statistiques utilisées pour analyser la durée jusqu'à la survenue d'un événement bien défini, par exemple la durée de rémission de certaines maladies dans des essais cliniques (l'hypertonie B, les sars, etc...), les temps de défaillance de certains produits manufacturés, les durées de vie des personnes âgées dans des programmes sociaux particuliers, le temps pris par un individu pour compléter sa thèse, etc. Il se réfère aussi à l'analyse du temps qui se produit dans un certain nombre de champs d'application tels que la biologie, l'ingénierie, la médecine, la démographie, la santé publique, l'économie et les sciences sociales [11]. Bien que les méthodes que nous présentons dans ce mémoire puissent être utilisées dans toutes ces disciplines, nos applications se concentreront exclusivement sur la médecine et la santé publique [découlant de la recherche médicale], et pour cette raison une grande partie de la discussion générale sera exprimée en termes de temps de survie d'un patient individuel de l'entrée à une étude jusqu'à la mort.

Pour de nombreux statisticiens, l'analyse statistique des données sur la durée de survie est devenue un sujet d'intérêt considérable. Il y a eu plusieurs manuels écrits qui traitent de l'analyse de survie d'un point de vue freqentiste. Cela inclut Lee T.E ; Wang.W.J.(2003), Tableman M ; Kim J.S. (2004), et Commenges D ; Jacqmin-Gadda H. (2015). Bien que ces livres soient assez complets et abordent plusieurs sujets, ils n'abordent pas en profondeur l'analyse bayésienne des données de survie [Klein et Moeschberger (2003)], cependant, présentent une section sur les méthodes non paramétriques bayésiennes.

L'analyse bayésienne des données de survie a reçu beaucoup d'attention récemment en raison des progrès dans les méthodes de calcul et de modélisation. Les méthodes bayésiennes sont en train de devenir courantes pour les données de survie et ont fait leur chemin dans le domaine médical et de la santé publique [11].

Les méthodes bayésiennes paramétriques et non paramétriques dans l'analyse de survie sont récemment devenues très populaires grâce aux progrès récents de la technologie informatique et le développement d'algorithmes computationnels efficaces pour la mise en œuvre de ces méthodes. De telles méthodes sont devenues courantes et bien employés dans la pratique.

Dans chaque cas des modèles paramétriques bayésiens nous donnons un développement du processus a priori, construirons la fonction de vraisemblance, dérivons les distributions a posteriori.

Ce mémoire comporte trois chapitres :

- Le premier chapitre, est consacré à quelques concepts nécessaires à l'étude de l'analyse de survie, telles que les fonctions de survie et de risque, les différents types de données censurées. Nous parlons aussi des méthodes fréquentistes paramétriques et non paramétriques d'analyse de survie, notamment la méthode de Kaplan-Meier.
- Dans le deuxième chapitre, nous examinons l'approche bayésienne de l'analyse de la survie. Nous discutons plusieurs types de modèles, y compris les modèles paramétriques ainsi que les modèles impliquant les processus a priori non paramétriques plus précisément, le processus de Dirichlet.
- Finalement, dans le troisième chapitre, nous décrivons la dualité qui existe entre l'approche fréquentiste et l'approche bayésienne. En considérant deux cas d'applications ; le 1^{er} compare les modèles paramétriques et le second les méthodes non paramétriques.

Chapitre 1

Outils de bases de la statistique de survie

L'analyse de survie est l'étude du délai de survenue d'un évènement d'intérêt telle que le décès. Cet évènement est souvent associé à un changement d'état, communément le passage de l'état "vivant "à l'état "décédé ". Cependant, on s'intéresse souvent à d'autres types de délais que la durée de vie proprement dite : la durée jusqu'à l'apparition d'une maladie, le délai entre la prise d'un traitement et la guérison d'une maladie, la durée de séropositivité sans symptômes de patients infectés par le VIH.

Dans ce chapitre, nous allons introduire les notions de bases des données de survie, sa caractéristique principale est la présence de "données incomplètes "qui sont dites censurées. Enfin, nous exposerons, très brièvement, les principales méthodes classiques d'estimation de la fonction de survie.

1.1 Les données de survie

On donne quelques définitions couramment utilisées dans l'étude de survie.

Définition 1.1.

i. ***Date d'origine*** *: Elle correspond à l'origine de la durée étudiée. Elle peut être le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie où la date d'entrée dans l'étude. Chaque individus*

peut donc avoir une date d'origine différente.

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

iii. ***Date des dernières nouvelles*** *: C'est la date la plus récente où des informations sur un sujet ont été recueillies.*

iv. ***Temps de participation*** *: Durée de surveillance pour chaque sujet utilisée dans l'estimation de la survie. On a trois cas :*

1^{er} *cas : L'événement a lieu au cours de la surveillance* $\Rightarrow$ *Temps de participation = Date de survenue de l'évènement − Date d'origine.*

2^{eme} *cas : Le sujet est vivant à la date de point* $\Rightarrow$ *Temps de participation = Date de point − Date d'origine.*

3^{eme} *cas : Le sujet est perdu de vue* $\Rightarrow$ *Temps de participation = Date de dernière nouvelle − Date d'origine.*

v. ***L'évènement d'intérêt*** *: Pour être sur de pouvoir comparer ces délais entre différents individus il faut avoir une définition précise de l'évènement d'intérêt. Si on étudie la mortalité, la définition est généralement claire. Par contre, si on étudie l'application d'une maladie se pose la question des critères diagnostiques et du délai entre le début de la maladie et son diagnostic. Si ce délai est court, il peut être approprié de choisir comme date d'évènement, la date de diagnostic. Si ce délai peut être long, il faut éventuellement le prendre en compte dans les analyses. De plus il est courant que l'évènement étudié soit en fait un évènement composite, comme par exemple "rechute ou décès" et dans ce cas le temps d'évènement est le temps correspondant à l'évènement se produisant le premier dans l'ordre chronologique [4].*

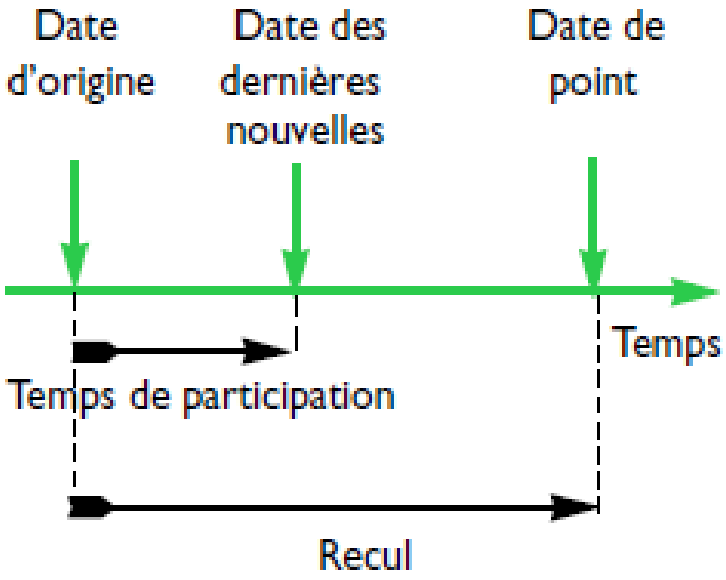

FIGURE 1.1: Déférentes dates dans l'étude des données de survie

Définition 1.2. *(Durée du survie)*
La durée de survie est une variable aléatoire X positive absolument continue représentant le temps auquel un certain évènement se produit.

Exemple 1.1. *En médecine ça peut être la durée de guérison d'un patient, où la durée de rémission d'un malade.*

1.1.1 Fonction de survie et fonction de risque

L'analyse d'un ensemble de données de survie, issues d'un essai clinique commence généralement par un résumé numérique où graphique des durées de survie des individus, dans les différents groupes de traitement.

De tels résumés peuvent être intéressants, par eux mêmes où précurseurs d'une analyse détaillée de données. Deux fonctions décrivant la distribution des temps de survie, qui sont d'importance centrale dans l'analyse des données de survie : la fonction de survie et la fonction de risque [3].

La fonction de survie $S(t)$

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

$$S(t) = P(X > t), \qquad t \geq 0$$

C'est donc une fonction continue monotone non croissante telle que :
$S(0) = 1$ et $\lim_{t\to\infty} S(t) = 0$.

Si la variable aléatoire X a une fonction de densité de probabilité $f(t)$, alors la fonction de survie est donnée par :

$$S(t) = \int_0^{+\infty} f(u)du = 1 - F(t);$$

où $F(t)$ est la fonction de distribution cumulative de X.

La fonction de risque $h(t)$

Cette fonction décrit la manière dont la probabilité instantané de mort d'un individu change avec le temps.

$$h(t) = \lim_{\Delta t\to 0} \frac{P\{t \leq X < t + \Delta t | X \geq t\}}{\Delta t} = \frac{f(t)}{S(t)}$$

Il est encore possible de définir le risque cumulé $H(t)$ selon :

$$H(t) = \int_0^t h(u)du$$

Remarque 1.1. *Si l'on connaît S, on peut calculer h. En effet,*

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

Et inversement, en intégrant cette égalité on obtient :

$$H(t) = \int_0^t h(u)du = -[lnS(u)]_0^t = -ln[S(t)]$$

puisque S(0)=1
ce qui s'écrit aussi : $S(t) = e^{-H(t)}$

Remarque 1.2. *Les fonctions (F, S, f, h, H) sont donc liées entre elles, c'est-à-dire, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également définies.*

1.1.2 Quantités associées à la loi de la survie

Quantiles de la durée de survie

La médiane de la durée de survie est le temps t pour lequel la probabilité de survie $S(t)$ est égale à 0.5, c'est-à-dire, la valeur t_m qui satisfait $S(t_m) = 0.5$.

Dans le cas où l'estimateur est une fonction en escalier (ex : Kaplan-Meier), il se peut qu'il y ait un intervalle de temps vérifiant $S(t_m) = 0.5$. Il faut alors être prudent dans l'interprétation, notamment si les deux événements encadrant le temps médian sont éloignés.

Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[B_i; B_s]$ un intervalle de confiance de niveau α de $S(t_m)$, alors un intervalle de confiance de niveau α du temps médian t_m est :

$$[S^{-1}(B_s); S^{-1}(B_i)].$$

La fonction quantile de la durée de survie est définie par :

$$\begin{aligned} q(p) &= inf\{t : F(t) \geqslant p\} \qquad 0 < p < 1, \\ &= inf\{t : S(t) \leqslant 1 - p\} \end{aligned}$$

Lorsque la fonction de répartition F est strictement croissante et continue alors :

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

Le quantile $q(p)$ est le temps où une proportion p de la population a disparu.

Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbb{E}(X)$ ainsi que sa variance $var(X)$ sont des quantités importantes :

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty S(t)dt. \\ var(X) &= 2\int_0^\infty tS(t)dt - \{E(X)\}^2. \end{aligned}$$

En effet, supposons que l'espérance existe. On écrit que

$$\int_0^\infty tdF(t) = \lim_{u\to\infty} \int_0^u tdF(t)$$

en intégrant par parties on peut écrire

$$\begin{aligned}\int_0^u tdF(t) &= -\int_0^u tdS(t) \\ &= -uS(u) + \int_0^u S(t)dt,\end{aligned}$$

l'inégalité de Markov assure alors que $tS(t) \leqslant E(X)$ et donc le terme $uS(u)$ est borné. On en déduit que l'intégrale $\int_0^\infty S(t)dt$ converge, ce qui implique $\lim\limits_{t\to\infty} tS(t) = 0$ et en passant à la limite on obtient le résultat attendu [29].

On montre de même manière que :

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

Remarque 1.3. *La moyenne et la variance peuvent être déduites de n'importe laquelle des cinq fonction ci-dessus(F,S,f,h,H), mais pas vice versa [28].*

1.2 Données censurées

On cherche à évaluer l'efficacité d'un nouveau traitement chez des personnes atteintes de maladies graves, qui peuvent donner lieu à des complications, à des rechutes, où même aboutir au décès.

Dans tous les cas, les critères d'intérêt peut s'exprimer comme la durée entre l'instauration du traitement et l'apparition de l'évènement témoignant de l'échec du traitement [30].

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie pour l'individu i, considérons :

- Son temps de survie X_i ;
- Son temps de censure C_i ;
- La durée réellement observée T_i.

Il existe trois catégories de censure qu'on nomme censure à droite, censure à gauche et censure par intervalle (lorsqu'on connait la borne supérieure et la borne inférieure d'un évènement) [16].

1.2.1 Censure à droite

Une durée de vie est dite censurée à droite si l'individu n'a pas subi l'évènement à sa date de dernières nouvelles (où à la date de point si celle-ci est antérieure à la date de dernier nouvelles).

Dans ce cas le délai n'est pas observé, on a comme seule information qu'il est supérieur au délai correspondant au temps de participation.

En analyse de survie, on peut rencontrer la censure à droite pour deux raisons :

i. Le sujet n'a pas encore subi l'évènement à la date de point, on parle alors "d'exclu vivant".

ii. Le sujet a quitté l'étude à une date à laquelle il n'avait pas encore subi l'évènement pour une raison telle qu'un déménagement, un refus de continuer de participer à une cohorte, c'est ce que l'on nomme des "perdus de vue".

Dans ce cas, on suppose l'indépendance entre la cause de la censure et l'évènement étudié. Si cette hypothèse n'était pas vérifiée, cela pourrait induire un biais dans l'estimation de la fonction de risque [4].

Une écriture mathématique courante et pratique pour représenter des données censurées à droite et d'associer à chaque individu un couple de variable aléatoire $(\tilde{T},\delta)$ avec les définitions suivantes :

$$\tilde{T} = min(X, C)$$

$$\delta = \begin{cases} 0 & X > C \\ 1 & sinon \end{cases}$$

où δ est l'indicatrice de l'évènement.

Exemple 1.2. *Dans cet exemple, sur la survie des sujets depuis leur entrées en institution, les sujets sont soit décédés soit encore en vie à la dernière visite. Nous sommes donc en présence de données censurées à droite.*

Le suivi des sujets 29,31,39 et 40 est représenté à la Figure 1.2. Les sujets 29 et 39 sont décédés respectivement 5.36 et 4.66 ans après l'entrée en institution, alors que les sujets 31 et 40 sont vivants à la date de leurs dernières nouvelles, 2.63 et 7.03 ans après l'entrées en institution. Leurs temps de décès sont censurés à droite.

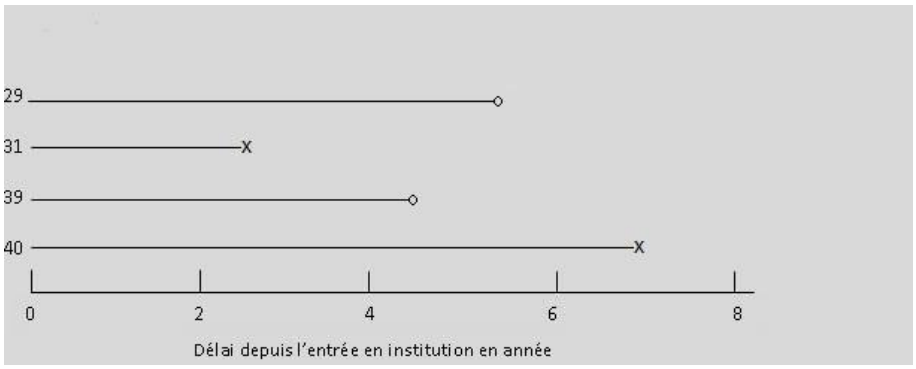

FIGURE 1.2: Durée de suivie en année de quatre sujets entrés en institution dans la cohorte paquid. ∘ est le décès ; × est le censure à droite.

1.2.2 Censure à gauche

La censure à gauche correspond au cas où l'individu a déja subi l'évènement avant que l'individu soit observé. On sait uniquement que la date de l'évènement est inférieure à une certaine date connue. Pour chaque individu, on peut associé un couple de variables aléatoires (T,δ)

$$T = X \vee C = max(X, C)$$

$$\delta = \mathbb{1}_{\{X \geq C\}}.$$

1.2.3 Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'évènement, la seule information disponible est qu'il a eu lieu entre deux dates connues. Par exemple, dans le cas d'un suivi cohorte les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'évènement s'est produit entre ces deux temps d'observation. (On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'évènement correspond au temps de la visite pour se ramener à de la censure à droite) [30].

À l'intérieur de ces trois catégories, il existe différents types de censure :

1. La censure de type I : fixée
   Si le temps de censure est fixé par le chercheur comme étant la fin d'étude. Soit C une valeur fixée, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq C$. Sinon on sait uniquement que $X_i > C$.

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles [30].

2. La censure de type II : attente

   On décide d'observer les durées de survie de n patients jusqu'à ce que r d'entre eux soient décédés et d'arrêter l'étude à ce moment là.

   Si l'on ordonne les durées de survie $X_1, ..., X_n$, soit $X_{(1)}$ la plus petite, $X_{(i)}$ la i^{eme} ... etc :

   $$X_{(1)}, X_{(2)}, ..., X_{(n)}$$

   On dit que les $X_{(i)}$ sont les statistiques d'ordre des X_i. La date de censure est alors $X_{(r)}$ et on observe

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

3. La censure de type III : aléatoire

   Soient $C_1, ..., C_n$ des variables aléatoires *i.i.d.* On observe les variables $\tilde{T}_i = X_i \wedge C_i$ l'information disponible peut être résumée par un indicateur

   • $\delta_i = 1$: Si l'évènement est observé (d'où $\tilde{T}_i = X_i$), on observe les "vraies"durées où les durées complètes.

   • $\delta_i = 0$: Si l'individu est censuré (d'où $T_i = C_i$), on observe des durées incomplètes (censurées).

   Remarque 1.4. *La censure aléatoire est la plus courante, par exemple, lors d'un essai thérapeutique [27], pour plusieurs causes :*

   a. *Perte de vue : le patient peut décider d'aller se faire soigner ailleurs et on ne le revoit plus.*

   b. *Arrêt du traitement : le traitement peut avoir des effets secondaires si désastreux que l'on est obligé d'arrêter le traitement (ces patients sont exclus de l'étude).*

   c. *Fin de l'étude : l'étude se termine alors que certains des patients sont toujours vivants. (ils ne sont pas subi l'événement).*

La Figure(1.3) illustre les situations de censure de type I et de censure aléatoire.

Dans le premier graphique (celui de gauche), les censures de types I sont déterminées à la fin de l'étude, tandis que les censures du deuxième graphique varient aléatoirement et peuvent surgir avant la fin de l'étude (les censures A et E dans le graphiques de droite).

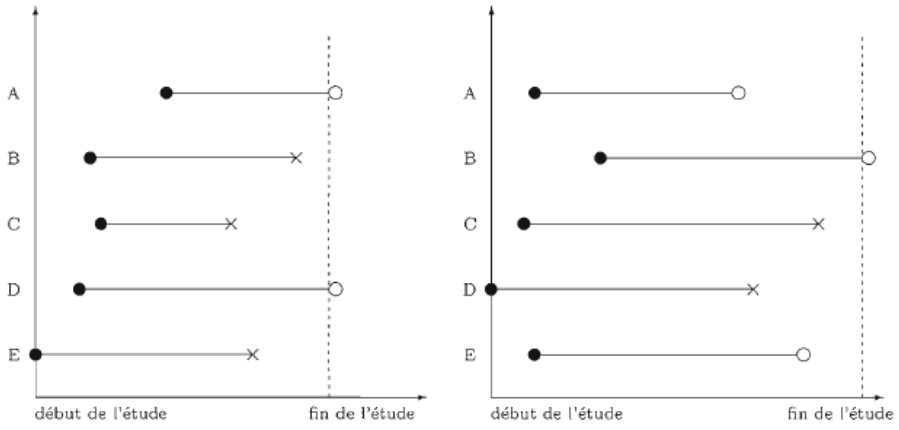

FIGURE 1.3: Illustration de censure de type l (à gauche) et de censure aléatoire (à droite). ∘ représente les données censurées ; × représente le décès.

Exemple 1.3.

La figure 1.4 représente le suivi de trois patients. Le premier est entré au début de l'étude et il est mort à la date $X_1 = 6$.

Le deuxième était toujours vivant à la fin de l'étude, qui a eu lieu au temps 10, il est donc censuré en $t = 10$.

Et le troisième patient à été perdu de vue avant la fin de l'étude. Il est donc censuré au temps $t = 7$.

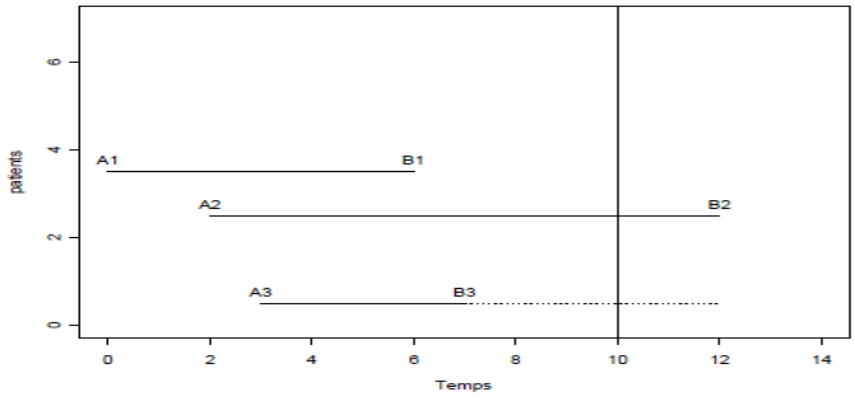

FIGURE 1.4: Exemple de 3 patients. A_1, A_2, A_3 sont les patients ; B_1 représente le décès ; B_2, B_3 représente les censures.

Remarque 1.5. *L'hypothèse d'indépendance de X_i et de C_i est utile mathématiquement. Il est important de voir si elle se justifie. Dans le cas où la censure est due à un arrêt du traitement, elle n'est pas vérifiée [28].*

1.3 La vraisemblance pour données censurées

1.3.1 La censure à droite

Dans un modèle de survie, avec des données censurées à droite par les variables C_i, nous observons $(\tilde{T}_i, \delta_i)$, *i.i.d*, $i = 1, ..., n$. Supposons d'abord que les C_i sont fixes (elles peuvent dépendre de i mais sont connues à l'avance). C'est le cas, par exemple, si les seuls sujets censurés sont des sujets exclus vivants à une date de point ou à la fin de l'étude, fixée à l'avance. Dans ce cas, la contribution à la vraisemblance d'un sujet i est :

- $f(\tilde{T}_i)$: si le sujet i a subi l'évènement $(\delta_i = 1)$;
- $S(\tilde{T}_i)$: si le délai de survie du sujet i est censuré à droite $(\delta_i = 0)$.

La vraisemblance de l'ensemble des n sujets s'écrit donc :

$$L = \prod_{i=1}^{n} f(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)^{(1-\delta_i)}, \tag{1.1}$$

que l'on peut aussi écrire :

$$\begin{aligned} L &= \prod_{i=1}^{n} (S(\tilde{T}_i)h(\tilde{T}_i))^{\delta_i} S(\tilde{T}_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} (h(\tilde{T}_i))^{\delta_i} S(\tilde{T}_i) \end{aligned}$$

1.4 Estimation de la fonction de survie

Une des première étapes possibles de l'analyse des données de survie est l'estimation de la distribution des temps de survie ; en général représentée par une courbe de survie qui est la représentation graphique de la fonction de survie. Pour estimer la fonction de survie on peut utiliser différentes approches : des estimateurs non-paramétriques où des estimateurs associés à des fonctions paramétriques.

Comme, on peut trouver aussi l'estimation semi-paramétrique pour plus de détail voir [30,4].

1.4.1 Estimateurs paramétriques

On suppose que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Les paramètres de ces distributions seront estimés à partir des données en maximisant la vraisemblance, par une méthode itérative comme l'algorithme dans R de Newton-Raphson(voir [4]).

Le modèle exponentiel

L'unique distribution continue qui admette un risque instantané constant est l'exponentielle. Pour X distribuée selon une loi exponentielle $\xi(\lambda)$ tel que $\lambda > 0$, on donne :
La fonction de densité par :

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

La fonction de survie par :

$$S(t) = \int_t^{+\infty} \lambda e^{-\lambda u} du = e^{-\lambda t}; \quad t \geq 0, \lambda > 0$$

La fonction de répartition par :

$$F(t) = 1 - S(t) = 1 - e^{-\lambda t}; \quad t \geq 0, \lambda > 0$$

La fonction de risque par :

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

L'espérance et la variance par :

$$\mathbb{E}(X) = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}$$

Ajuster les données au modèle exponentiel

i. Cas où il n'y a pas de censure.

Tous les "échecs"sont observés. Les $X_1, ..., X_n$ sont *iid*.

Vraisemblance :

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{(-\lambda x_i)} = \lambda^n e^{\left(-\lambda \sum\limits_{i=1}^{n} x_i\right)}$$

Log-vraisemblance :

$$lnL(\lambda) = nln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial lnL(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

EMV :

Soit

$$\frac{\partial lnL(\lambda)}{\partial \lambda} = 0$$

et résolvez pour λ. Donc,

$$\hat{\lambda} = \frac{n}{\sum\limits_{i=1}^{n} x_i} = \frac{1}{\bar{X}}. \tag{1.2}$$

L' EMV de la moyenne $\theta = \frac{1}{\lambda}$ est $\hat{\theta} = \bar{X}$.

ii. Cas de censure aléatoire.

Supposons que u, c et n_u désignent des observations non censurées, censurées et le nombre d'observations non censurées, respectivement. Les n valeurs observées sont maintenant représentées par les vecteurs $\tilde{T}$ et δ, où $\tilde{T} = (\tilde{T}_1, ..., \tilde{T}_n)$ et $\delta' = (\delta_1, ..., \delta_n)$. Alors :

Vraisemblance : A partir de l'expression (1.1)

$$L(\lambda) = \prod_{u} [f(\tilde{T}_i|\lambda)] \prod_{c} [S(\tilde{T}_i|\lambda)]$$

$$= \prod_u [\lambda e^{(-\lambda \tilde{T}_i)}] \prod_c [e^{(-\lambda \tilde{T}_i)}]$$
$$= \lambda^{n_u} e^{\left(-\lambda \sum\limits_u \tilde{T}_i\right)} e^{\left(-\lambda \sum\limits_c \tilde{T}_i\right)}$$
$$= \lambda^{n_u} e^{\left(-\lambda \sum\limits_{i=1}^{n} \tilde{T}_i\right)}$$

Log-vraisemblance :

$$lnL(\lambda) = ln \prod_{i=1}^{n} f^{\delta_i}(\tilde{T}_i|\lambda) S^{1-\delta_i}(\tilde{T}_i|\lambda) = \sum_n lnf(\tilde{T}_i|\lambda) + \sum_c lnf(\tilde{T}_i|\lambda)$$

$$lnL(\lambda) = n_u ln(\lambda) - \lambda \sum_{i=1}^{n} \tilde{T}_i$$

$$\frac{\partial lnL(\lambda)}{\partial \lambda} = \frac{n_u}{\lambda} - \sum_{i=1}^{n} \tilde{T}_i$$

$$\frac{\partial^2 lnL(\lambda)}{\partial \lambda^2} = -\frac{n_u}{\lambda^2} = -i(\lambda)$$

le négatif de l'information observée.

EMV :

$$\hat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} \tilde{T}_i}$$

et

$$var_a(\hat{\lambda}) = \left(-\mathbb{E}(-\frac{n_u}{\lambda^2})\right)^{-1} = \frac{\lambda^2}{\mathbb{E}(n_u)}$$

où $\mathbb{E}(n_u) = n.\mathbb{P}(X \leq C)$, et d'après la convergence asymptotique de l'EMV on aura

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda^2/\mathbb{E}(n_u)}} \sim N(0,1).$$

Nous remplaçons $\mathbb{E}(n_u)$ par n_u car nous ne connaissons généralement pas la distribution de censure $G(.)$. Notez la dépendance de la variance asymptotique sur le paramètre inconnu λ. Nous substituons à λ et obtenons

$$var_a(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{n_u} = \frac{1}{i(\hat{\lambda})}$$

où $i(\lambda)$ est juste au-dessus. L'EMV pour la moyenne
$\theta = \frac{1}{\lambda}$ est simplement $\hat{\theta} = \frac{1}{\hat{\lambda}} = \sum\limits_{i=1}^{n} \tilde{T}_i/n_u$.

Le modèle Weibull

C'est une généralisation du modèle exponentiel :

$$S(t) = e^{-(\lambda t)^{\alpha}}; \qquad t \geq 0, \alpha > 0, \lambda > 0$$

Alors on déduit :

La fonction de risque :

$$h(t) = \lambda\alpha(\lambda t)^{(\alpha-1)};$$

La fonction de densité :

$$f(t) = \lambda\alpha(\lambda t)^{(\alpha-1)}e^{-(\lambda t)^{\alpha}};$$

La fonction de répartition :

$$F(t) = 1 - e^{-(\lambda t)^{\alpha}}.$$

Ajuster les données au modèle weibull

La fonction de log-vraisemblance est proportionnelle à :

$$\begin{aligned} lnL(\alpha,\lambda) &\propto ln(\alpha\lambda)\sum_{i=1}^{n}\delta_i + (\alpha-1)\sum_{i=1}^{n}\delta_i ln(t_i) - \lambda\sum_{i=1}^{n}\delta_i t_i^{\alpha} - \lambda\sum_{i=1}^{n}(1-\delta_i)t_i^{\alpha} \\ &\propto (ln\alpha + ln\lambda)\sum_{i=1}^{n}\delta_i + \alpha\sum_{i=1}^{n}\delta_i ln(t_i) - \lambda\sum_{i=1}^{n}t_i^{\alpha} \end{aligned}$$

Prendre les dérivées par rapport à α et λ, et les mettres égal à zéro donne le système d'équations normales suivant :

$$\begin{cases} \frac{\partial lnL(\hat{\alpha},\hat{\lambda})}{\partial\alpha} = \frac{\sum_{i=1}^{n}\delta_i}{\hat{\alpha}} + \sum_{i=1}^{n}\delta_i lnt_i - \hat{\lambda}\sum_{i=1}^{n}t_i^{\hat{\alpha}}lnt_i = 0 \\ \frac{\partial lnL(\hat{\alpha},\hat{\lambda})}{\partial\lambda} = \frac{\sum_{i=1}^{n}\delta_i}{\hat{\lambda}} - \sum_{i=1}^{n}t_i^{\hat{\alpha}} = 0 \end{cases}$$

Ce système doit être résolu numériquement.

L'estimateur de la fonction de survie est :

$$\hat{S}(t) = exp\{-\hat{\lambda}t^{\hat{\alpha}}\}, \qquad t \geq 0.$$

Remarque 1.6.

-i- $h(t)$ *est croissante si* $\alpha > 1$*, décroissante pour* $\alpha < 1$ *et constante pour* $\alpha = 1$*.*

-ii- Si $\alpha = 1$*, la loi de weibull devient une exponentielle de paramètre* λ *[14].*

Le modèle gamma

La loi gamma dépend de deux paramètres strictement positifs (λ, k), c'est une généralisation de la loi exponentielle, que l'on retrouve si $k = 1$.

Elle a comme densité de probabilité :

$$f(t) = \frac{\lambda^k t^{k-1} e^{-\lambda k}}{\Gamma(k)}, \quad t > 0, \lambda, k > 0,$$

avec

$$\Gamma(k) = \int_0^\infty u^{k-1} e^{-u} du.$$

La fonction de répartition :

$$F(t) = \frac{1}{\Gamma(k)} \int_0^{\lambda t} u^{k-1} e^{-u} du, \quad t > 0, \lambda, k > 0.$$

La fonction de survie :

$$S(t) = 1 - \frac{1}{\Gamma(k)} \int_0^{\lambda t} u^{k-1} e^{-u} du, \quad t > 0, k > 0, \lambda > 0$$

La fonction de risque :

$$h(t) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda t}}{\Gamma(k) - \int_0^{\lambda t} u^{k-1} e^{-u} du}, \quad t > 0, k > 0, \lambda > 0$$

Remarque 1.7.

-i- La fonction $h(t)$ sera monotone croissante si $k > 1$ et monotone décroissante si $k < 1$.

-ii- La loi gamma est souvent employée quand on a une somme de variables aléatoires de loi exponentielle *i.i.d.* En effet, si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes et toutes distribuées selon une loi exponentielle $\xi(\lambda)$, alors $X = \sum_i^n X_i \sim (\lambda, n)$ [14].

-iii- Les différents modèles sont classés suivant que le risque instantané est croissant ou décroissant [27] comme suit :

Risque instantané	Modèle
Constant	Exponentiel
Croissant (RIC)	Weibull($\alpha > 1$) Gamma($k > 1$)
Décroissant (RID)	Weibull($\alpha < 1$) Gamma($k < 1$)

RIC est généralement noté IFR (Increasing Failure Rate).

RID est généralement noté DFR (Decreasing Failure Rate) .

Modèle de valeur extrême (minimum)

L'intérêt de cette distribution n'est pas son utilisation directe comme distribution à vie, mais plutôt à cause de sa relation avec la distribution de Weibull. Soit μ où $-\infty < \mu < +\infty$, et $\sigma > 0$ indiquent les paramètres de localisation et d'échelle, respectivement. La distribution de la valeur extrême standard a $\mu = 0$ et $\sigma = 1$.
La fonction de densité est donnée par :

$$f(x) = \sigma^{-1} e^{\left(\frac{x-\mu}{\sigma} - e^{(\frac{x-\mu}{\sigma})}\right)}$$

La fonction de survie est donnée par :

$$S(x) = e^{\left(-e^{(\frac{x-\mu}{\sigma})}\right)}$$

La moyenne de X est :

$$\mathbb{E}(X) = \mu - \gamma\sigma$$

La variance de X est :

$$Var(X) = \frac{\pi^2}{6}\sigma^2$$

γ est le constante d'Euler, avec $\gamma = 0.5772...$
le paramètre de localisation μ est le 0.632^{eme} quantile.
x peut aussi être négatif pour que $-\infty < x < +\infty$.

Remarque 1.8. *Si T est une variable aléatoire de Weibull avec paramètres α et λ, puis $X = ln(T)$ suit une distribution de valeur extrême avec $\mu = -ln(\lambda)$ et $\sigma = \alpha^{-1}$. X peut être représenté comme $X = \mu + \sigma Z$, où Z est une valeur extrême standard, car la distribution de valeur extrême est une famille de distributions de localisation et d'échelle.*

Comme les valeurs de μ et σ différentes de 0 et 1 n'affectent pas la forme de la fonction de densité de probabilité, mais seulement l'emplacement et l'échelle, il suffit d'afficher les courbes de la densité de probabilité standard et la fonction de survie (voir figure 1.5)

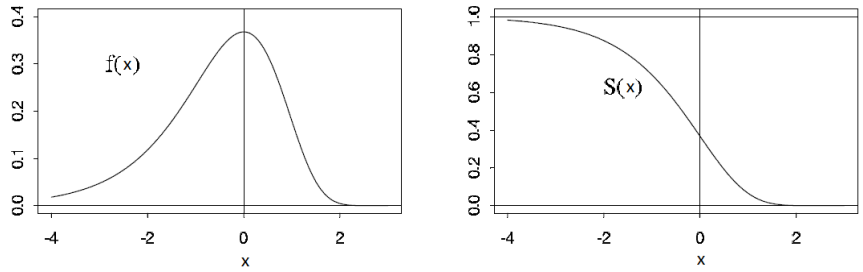

FIGURE 1.5: Densité standard extrême et fonction de survie.

1.4.2 Estimateur non paramétrique

Les méthodes non paramétriques sont souvent très faciles et simples à comprendre par rapport aux méthodes paramétriques. De plus, l'analyse non paramétrique est largement utilisée dans des situations, où il y a un doute sur la forme exacte de la distribution [32].

Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur non paramétrique de la fonction de survie. Le principe de la méthode repose sur l'idée qu'être encore en vie après un instant t, c'est être en vie juste avant cet instant t et ne pas mourir à cet instant [31]. Ainsi, la survie à un instant quelconque est le produit de probabilités conditionnelles de survie de chacun des instants précédents.

On utilise le théorème de probabilité conditionnelle :
Soit : $t_i \leqslant t_{i+1}$

$$\begin{aligned} S(t_i) &= P(X > t_i) \\ &= P(X > t_i, X > t_{i-1}) \\ &= P(X > t_i | X > t_{i-1}) P(X > t_{i-1}) \\ &= P(X > t_i | X > t_{i-1}) P(X > t_{i-1} | X > t_{i-2}) ... P(X > t_0 = 0). \end{aligned}$$

Nous supposons qu'au début de l'étude tous les sujets étaient vivants, alors $P(X > t_0 = 0) = 1$.

La probabilité conditionnelle est :

$$P(X > t_i | X > t_{i-1}) = p_i.$$

Qui est la probabilité de survivre pendant l'intervalle de temps $I_i =]t_{i-1}, t_i]$ quand on était vivant au début de cet intervalle.

Notons :

R_i : Le nombre de sujets qui sont vivants juste avant l'instant t_i, ce que l'on note : nombre des vivants à l'instant t_i, où nombre des sujets de $R(t_i)$ en désignant par R(t) l'ensemble des sujets à risque à l'instant t^-.

M_i : Le nombre des morts à l'instant t_i.

$q_i = 1 - p_i$ est la probabilité de mourir pendant l'intervalle I_i sachant que l'on était vivant au début de cet intervalle. Alors l'estimateur naturel de q_i est :

$$\hat{q}_i = \frac{M_i}{R_i}$$

A) Cas où il n'y a pas d'ex-æquo :

Si $\delta_i = 1$, c'est qu'il y a eu un mort en t_i et donc $M_i = 1$.

Si $\delta_i = 0$, c'est qu'il y a eu une censure en t_i et donc $M_i = 0$.

Par suite, $\hat{p}_i = \begin{cases} 1 - \frac{1}{R_i} & \text{en cas de mort en } t_i \\ 1 & \text{en cas de censure en } t_i \end{cases}$

L'estimateur de Kaplan-Meier est donc dans ce cas :

$$\hat{S}(t) = \prod_{t_i \leqslant t} (1 - \frac{1}{n-i+1})^{\delta_i}.$$

Exemple 1.4 (Cancer des bronches)**.**

Sur 10 patients atteints de cancer des bronches on a observé les durées de survie suivantes, exprimées en mois (avec (+) indique q'il existe une censure à droite).

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

L'estimateur de Kaplan-Meier de la fonction de survie $S(t)$ vaut :

$\hat{S}(0) = 1$ *et* $\hat{S}(t) = 1$ *pour* t *dans* $[0, 1[$

$\hat{S}(1) = (1 - \frac{1}{10})\hat{S}(0) = 0.900$

$\hat{S}(3) = (1 - \frac{1}{9})\hat{S}(1) = 0.800$

$\hat{S}(5) = (1 - \frac{1}{7})\hat{S}(3) = 0.686$

$\hat{S}(8) = (1 - \frac{1}{5})\hat{S}(5) = 0.549$

$\hat{S}(9) = (1 - \frac{1}{4})\hat{S}(8) = 0.411$

$\hat{S}(11) = (1 - \frac{1}{2})\hat{S}(9) = 0.206$

Temps	R_i	M_i	*Survie*	*Intervalle*
0	*10*	*0*	*1*	*[0,1[*
1	*10*	*1*	*0.900*	*[1,3[*
3	*9*	*1*	*0.800*	*[3,5[*
5	*7*	*1*	*0.686*	*[5.8[*
8	*5*	*1*	*0.549*	*[8,9[*
9	*4*	*1*	*0.411*	*[9,11[*
11	*2*	*1*	*0.206*	

La Figure 1.6 donne la représentation graphique de cet estimateur qui est une fonction en escalier décroissante.

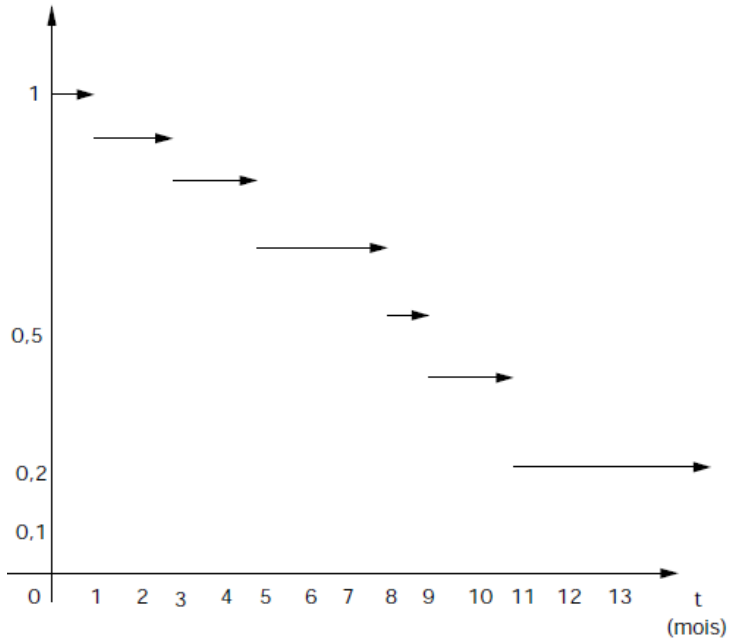

FIGURE 1.6: Estimateur de kaplan-Meier pour les temps de survie des patients

B) Cas où il y a des ex-æquo :

(a) Si ces ex-æquo sont tous des morts, la seule différence tient à ce que M_i n'est plus égal à 1 mais au nombre des morts et l'estimateur de Kaplan-Meier devient :

$$\hat{S}(t) = \prod_{t_i \leqslant t} \left(1 - \frac{M_i}{R_i}\right)$$

(b) Si ces ex-æquo sont des deux sortes, on considère que les observations non censurées ont lieu juste avant les censurées.

Propriétés de l'estimateur de Kaplan-Meier

L'estimateur de K-M possède beaucoup de propriétés analogues à celles de la fonction de répartition empirique. Nous commençons par ses propriétés de convergence [2].

Théorème 1.1. *([2]) Si pour $t > 0$, $F(t) < 1$ et $\bar{Y}_n(t) \underset{n\to\infty}{\overset{\mathcal{P}}{\to}} +\infty$ alors*

$$sup_{0\leq s\leq t}|\hat{S}(s) - S(s)| \underset{n\to\infty}{\overset{\mathcal{P}}{\to}} 0$$

On suppose que C est indépendante de X, la variable observée n'est plus X mais $\tilde{T} = X \wedge C$, de fonction de survie Ψ, vérifie $\Psi = S.G$ (G la fonction de survie de la censure droite C).

Théorème 1.2. *([2]) Dans l'espace $D([0,\tau], \|.\|_\infty, \mathcal{P})$ des fonctions continues à droite possédant des limites à gauche en tout point de $[0,\tau]$, muni de la norme infinie et de sa tribu de projection, si $\Psi(\tau^-) > 0$ (Ψ : la survie des observations $\tilde{T} = X \wedge C$) et si S est continue sur $[0,\tau]$, alors*

$$\sqrt{n}(\hat{S}(t) - S(t)) \xrightarrow{L} Z$$

avec Z est un processus gaussienne centrée de covariance

$$Cov(Z(s), Z(t)) = S(s)S(t)\int_0^{s\wedge t} \frac{dF(u)}{S^2(u)(1-G(u))}$$

Théorème 1.3. *([30]) En tout point t_0 de continuité de S ; $t_0 \in [0,\tau]$ et $S(\tau^-) > 0$,*

$$\sqrt{n}(\hat{S}(t_0) - S(t_0)) \underset{n\to\infty}{\overset{L}{\to}} \mathcal{N}(0, V^2(t_0))$$

avec

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0} \frac{S(du)}{S^2(u)(1-G(u))}$$

L'estimateur de la variance de l'estimateur de Kaplan-Meier à un temps t fixé est donnée par la formule de Greenwood (1926) [2] :

$$\widehat{var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i\leqslant t} \frac{M_i}{R_i(R_i - M_i)}$$

Démonstration.

delta méthode :

Pour estimer la variance de l'estimateur de Kaplan-Meier, nous devons introduire la delta-méthode. La delta méthode utilise l'expansion de Taylor de premier ordre

d'une fonction f d'une variable aléatoire X autour de $u = \mathbb{E}(X)$ pour estimer la variance de $f(X)$.

$$\begin{aligned} f(X) &\simeq f(u) + f'(u)(X-u) \\ var(f(X)) &\simeq var(f(u) + f'(u)(X-u)) \\ &= f'^2(u)var(X-u) \\ &= f'^2(u)var(X) \\ &= f'^2(u)\sigma^2 \end{aligned}$$

Tel que $\sigma^2 = var(X)$.

L'estimateur de delta-méthode est :

$$\widehat{var}(f(X)) = f'^2(\hat{u})\hat{\sigma}^2$$

Tel que $\hat{\sigma}^2$ est un estimateur de $var(X)$ et $\hat{u}$ est un estimateur de $\mathbb{E}(X)$.

Ici, nous montrons comment trouver la formule de Greenwood en utilisant la delta methode.

On a utiliser la delta-méthode deux fois :

$$ln(X) \simeq ln(u) + (X-u)\frac{1}{u} \Rightarrow \widehat{var}(ln(X)) \simeq \hat{\sigma}^2\frac{1}{\hat{u}^2}$$

et :

$$e^X \simeq e^u + (X-u)e^u \Rightarrow \widehat{var}(e^X) \simeq e^{2\hat{u}}\hat{\sigma}^2$$

Soit

$$ln(\hat{S}(t)) = ln\left(\prod_{t_i \leqslant t}\left[1 - \frac{M_i}{R_i}\right]\right) = \sum_{t_i \leqslant t} ln\left[1 - \frac{M_i}{R_i}\right]$$

On a : $p_i = P(X > t_i | X > t_{i-1})$ alors $\hat{p}_i = [1 - \frac{M_i}{R_i}]$ est une estimation de cette probabilité conditionnelle. Cella signifie que nous supposons que $M_i \sim B(n, 1-p_i)$. Par conséquent, la variance de $\hat{p}_i$ est estimé par $\widehat{var}(\hat{p}_i) = \frac{\hat{p}_i(1-\hat{p}_i)}{R_i}$.

De plus, les variables binomiales sont indépendantes pour tous les sujets de l'étude. Nous avons alors :

$$\widehat{var}\{\sum_{t_i \leqslant t} ln(\hat{p}_i)\} = \sum_{t_i \leqslant t} \widehat{var}\,(ln(\hat{p}_i))$$

Une première utilisation de la delta-méthode donne :

$$\widehat{var}(ln(\hat{p}_i) \simeq \frac{\hat{p}_i(1-\hat{p}_i)}{R_i}\frac{1}{\hat{p}_i^2} = \frac{1-(1-\frac{M_i}{R_i})}{R_i(1-\frac{M_i}{R_i})} = \frac{\frac{M_i}{R_i}}{R_i - M_i} = \frac{M_i}{R_i(R_i-M_i)}$$

$$\Longrightarrow ln[\widehat{var}(\hat{S}(t))] \simeq \sum_{t_i \leqslant t} \frac{M_i}{R_i(R_i - M_i)}$$

Nous utilisons la delta-méthode pour la deuxième fois et trouvons en fin :

$$\begin{aligned}\widehat{var}(\hat{S}(t)) &= \widehat{var}\{exp[ln(\hat{S}(t))]\} \\ &= exp^2[ln(\hat{S}(t))] \sum_{t_i \leqslant t} \frac{M_i}{R_i(R_i - M_i)} \\ &= \hat{S}^2(t) \sum_{t_i \leqslant t} \frac{M_i}{R_i(R_i - M_i)}\end{aligned}$$

■

On conclue que l'estimateur de K-M est consistant et asymptotique.

Construction d'intervalles de confiance pour la survie

L'estimation ponctuelle de la survie à un moment donnée doit impérativement être accompagné d'un intervalle de confiance (I.C), gage de la précision de l'estimation, habituellement au seuil de $\alpha = 0.05$ (I.C à 95%) est alors défini par les bornes suivantes [24].

$$[\hat{S}(t) - \hat{\sigma}(\hat{S}(t))z_{1-\frac{\alpha}{2}}; \hat{S}(t) + \hat{\sigma}(\hat{S}(t))z_{1-\frac{\alpha}{2}}]$$

où $z_{1-\frac{\alpha}{2}}$ est le fractile de rang $100 \times (1 - \frac{\alpha}{2})$ de la distribution normale standarisée.

Exemple 1.5.

Trente patients atteints de mélanome (stages 2 à 4). traités par Corynebacterium, ont des temps de survie comme indiqué ci-dessous.

Tous les patients ont été réséqués avant le début du traitement et n'avaient donc aucun signe de mélanome au moment du traitement. L'objectif habituel de ce type de données est de déterminer la durée de survie.

Patients	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Temps de survie	33.7^+	*3.9*	*10.5*	*5.4*	*19.5*	23.8^+	*7.9*	16.9^+
Patients	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Temps de survie	16.6^+	33.7^+	17.1^+	*8.0*	26.9^+	21.4^+	18.1^+	16.0^+
Patients	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
Temps de survie	*6.9*	11.0^+	24.8^+	23.0^+	*8.3*	10.8^+	12.2^+	12.5^+
Patients	*25*	*26*	*27*	*28*	*29*	*30*		
Temps de survie	*24.4*	*7.7*	4.8^+	8.2^+	8.2^+	7.8^+		

L'estimateur de kaplan-meier de ces données est :

```
> summary(survfit(Surv(temps,censur)~1))
Call: survfit(formula = Surv(temps, censur) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3.9     30       1    0.967  0.0328        0.905        1.000
  5.4     29       1    0.933  0.0455        0.848        1.000
  6.9     28       1    0.900  0.0548        0.799        1.000
  7.7     27       1    0.867  0.0621        0.753        0.997
  7.9     25       1    0.832  0.0686        0.708        0.978
  8.0     24       1    0.797  0.0740        0.665        0.956
  8.3     21       1    0.759  0.0796        0.618        0.933
 10.5     20       1    0.721  0.0842        0.574        0.907
 19.5      9       1    0.641  0.1064        0.463        0.888
 24.4      5       1    0.513  0.1428        0.297        0.885
```

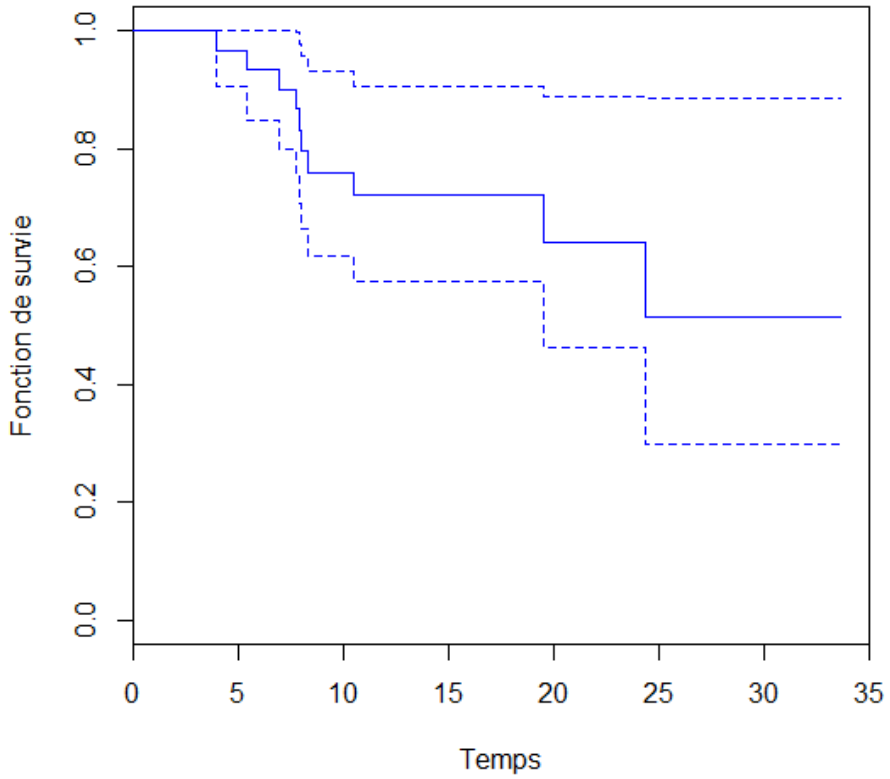

FIGURE 1.7: Estimateur de kaplan-Meier pour les temps de survie des patients

Le package de R utilisé "survival"donne aussi un intervalle de confiance à 95% *pour cet estimateur on remarque toujours que c'est une fonction en escalier décroissante (Figure 1.7).*

Chapitre 2

Estimation bayésienne de la fonction de survie

L'approche bayésienne fournit un cadre naturel pour résoudre des problèmes d'inférence statistique. Elle se distingue de la statistique classique parce qu'elle considère le(s) paramètre(s) du modèle comme des variables aléatoires.

Dans ce chapitre, nous rappelons tout d'abord quelques notions de base sur la statistique bayésienne tel que le modèle bayésien et le choix a priori, dans la section "Modèles paramétriques" nous passons en revue les modèles de survie paramétriques bayésiens comme le modèle exponentiel, weibull, gamma et le modèle de valeurs extrême. Dans la section suivante nous donnons la notion d'un processus de Dirichlet et nous discutons les méthodes bayésiennes non paramétriques pour l'estimation de la fonction de survie.

2.1 Approche bayésienne

Dans l'approche bayésienne, l'idée de base consiste à traiter le paramètre inconnu θ comme étant une variable aléatoire admettant une densité de probabilité $\pi(\theta)$ qui s'appelle densité a priori.

L'objectif est donc d'utiliser cette information supplémentaire. Sachant que l'information contenue dans les observations x est contenue dans $L(x|\theta)$ et l'information a priori sur θ dans $\pi(\theta)$, on peut utiliser la règle de Bayes pour combiner ces deux types d'informations

en définissant la densité a posteriori par :

$$\pi(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)d\theta}$$

qui contiendra donc toutes les informations sur θ .

On remarque que l'inversion de cause à effet est ici beaucoup plus naturelle. Elle se fait d'une manière cohérente, car l'état de connaissance a priori sur θ traduite par la densité a priori $\pi(\theta)$ est transformé, après les observations x, en état de connaissance a posteriori par la densité a posteriori $\pi(\theta|x)$.

Remarquons que la densité a posteriori peut s'écrire :

$$\pi(\theta|x) \propto L(x|\theta)\pi(\theta).$$

2.1.1 Modèle bayésien

Soit Θ l'espace des paramètres et χ l'espace des observations. On considère un modèle statistique de loi de probabilité P_θ de densité $f(x|\theta)$ dépendant d'un paramètre inconnu de dimension K : $\theta \in \mathbb{R}^K$. On dispose d'un échantillon aléatoire de n observations $x = (x_1, ..., x_n)$ issues de cette distribution.

Définition 2.1. *On appelle modèle bayésien la donnée d'un modèle paramétrique, $L(x|\theta)$, et d'une densité a priori $\pi(\theta)$ sur les paramètres.*

Étant donné la densité des observations $L(x|\theta)$ et la densité a priori $\pi(\theta)$ on peut construire :

i. La densité jointe de (θ, x) :

$$\varphi(\theta, x) = L(x|\theta)\pi(\theta);$$

ii. La densité marginale de x :

$$m(x) = \int \varphi(\theta, x)d\theta = \int L(x|\theta)\pi(\theta)d\theta;$$

iii. La densité a posteriori de θ, obtenue par formulation de Bayes :

$$\begin{aligned}\pi(\theta|x) &= \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)d\theta} \\ &= \frac{L(x|\theta)\pi(\theta)}{m(x)} \\ &\propto L(x|\theta)\pi(\theta).\end{aligned}$$

Autrement dit, la densité a posteriori représente une actualisation de l'information a priori au vu de l'information apportée par les observations. Voir Figure 2.1

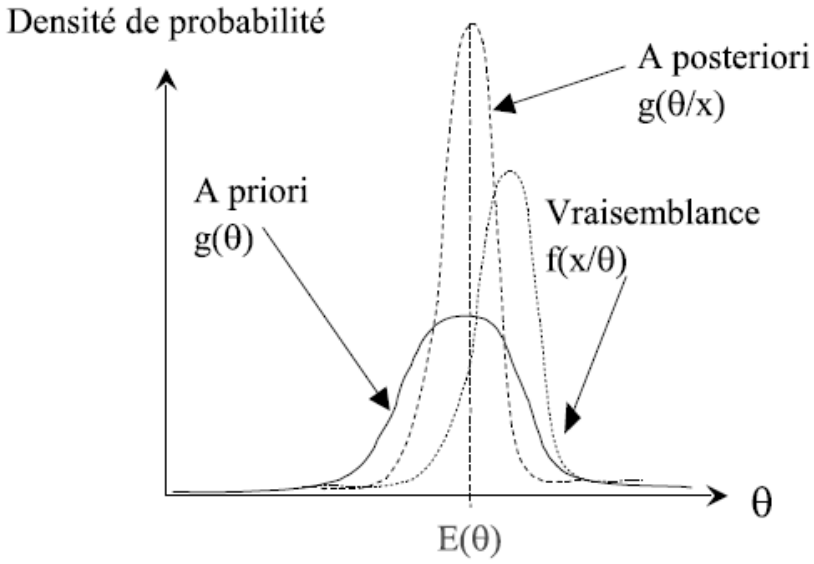

FIGURE 2.1: Représentation des distributions a priori, vraisemblance et a posteriori

2.2 Choix de la loi a priori

L'aspect de l'analyse bayésienne le plus critiqué et le plus délicat est certainement le choix de la loi a priori des paramètres. Deux types d'approches sont généralement considérés :

- Une approche dite subjective, ou informative, qui revient à tenir compte (lorsqu'elles existent) d'informations a priori sur le paramètre (expériences précédentes, avis d'experts, connaissances extérieures au processus d'observation, etc.),
- Une approche dite objective, ou non informative, qui revient à modéliser l'absence d'information a priori.

Dans la suite, nous présenterons un modèle des lois informatives (lois a priori conjuguées) et un autre des lois non informatives (lois de Jeffreys).

2.2.1 Lois a priori conjuguées

D'après [17] une famille conjuguée est une famille de lois liée au processus étudié et qui est particulièrement intéressante dans des modèles paramétriques où des statistiques

exhaustives qui peuvent être définies. Il s'agit d'une famille de lois, c-à-d d'un ensemble de lois, qui soit d'une part, assez riche et flexible pour représenter l'information a priori que l'on a sur les paramètres du modèle et, d'autre part, qui se combine remarquablement avec la fonction de vraisemblance dans la formule de Bayes, pour donner a posteriori une loi sur les paramètres du modèle qui appartienne à la même famille.

On verra que le choix de cette famille se déduit a partir du noyau de la vraisemblance considérée comme une fonction des paramètres.

Définition 2.2. *Une famille F de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) par une famille de vraisemblance $f(x|\theta)$ si, pour tout $\pi \in F$, la distribution a posteriori $\pi(.|x)$ appartient également à F.*

Famille exponentielle

Les lois a priori conjuguées sont généralement associées à un type particulier de lois d'échantillonnage qui permet toujours leur obtention ; il est même caractéristique des lois a priori conjuguées comme nous le verrons ci-dessous. Ces lois constituent ce qu'on appelle des familles exponentielles [17].

Définition 2.3. *La famille exponentielle regroupe les lois de probabilité qui admettent une densité de la forme :*

$$f(x|\theta) = h(x)e^{\alpha(\theta)T(x)-\phi(\theta)}; \theta \in \Theta.$$

Test alors une statistique exhaustive. En particulier si :

$$f(x|\theta) = h(x)e^{\theta T(x)-\phi(\theta)}$$

la famille est dite naturelle.

Proposition 2.1. *([17]) Une famille conjuguée pour $f(x|\theta)$ est donnée par*

$$\pi(\theta|\lambda, \mu) = h(x)e^{\theta\mu-\lambda\phi(\theta)}$$

La distribution a posteriori correspondante est

$$\pi(\theta|\lambda + 1, \mu + T(x))$$

Dans le tableau 2.1 on donne les lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles :

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathscr{N}(\theta, \sigma^2)$	Normale $\mathscr{N}(\mu, \tau^2)$	Normale $\mathscr{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathscr{P}(\theta)$	Gamma $\mathscr{G}(\alpha, \beta)$	Gamma $\mathscr{G}(\alpha + x, \beta + 1)$
Gamma $\mathscr{G}(\nu, \theta)$	Gamma $\mathscr{G}(\alpha, \beta)$	Gamma $\mathscr{G}(\alpha + \nu, \beta + x)$
Binomiale $\mathscr{B}(n, \theta)$	Bêta $\mathscr{B}e(\alpha, \beta)$	Bêta $\mathscr{B}e(\alpha + x, \beta + n - x)$
Binomiale Négative $\mathscr{N}eg(m, \theta)$	Bêta $\mathscr{B}e(\alpha, \beta)$	Bêta $\mathscr{B}e(\alpha + m, \beta + x)$
Multinomiale $\mathscr{M}_k(\theta_1, ..., \theta_k)$	Dirichlet $\mathscr{D}(\alpha_1, ..., \alpha_k)$	Dirichlet $\mathscr{D}(\alpha_1 + x_1, ..., \alpha_k + x_k)$
Normale $\mathscr{N}(\mu, \frac{1}{\theta})$	Gamma $\mathscr{G}(\alpha, \beta)$	Gamma $\mathscr{G}(\alpha + 0.5, \beta + \frac{(\mu - x)^2}{2})$

TABLE 2.1: Lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles

Remarque 2.1. *L'intérêt principal du caractère conjuguée devient plus évident quand F est paramétrée. Effectivement, le passage de distribution a priori à distribution à posteriori se réduit dans ce cas à une mise à jour des paramètres correspondants. Cette seule propriété peut expliquer pourquoi les lois a priori conjuguées sont si populaire, car les distributions a posteriori sont toujours calculables (au moins jusqu'à un certain degré). En revanche, une telle justification est plutôt faible d'un point de vue subjectif et d'autres familles pourraient aussi bien convenir. Notons que l'objectif d'obtenir la famille conjuguée minimale comme l'intersection de toutes les familles conjuguées est malheureusement voué à l'échec, car cette intersection est vide.*

2.2.2 Lois a priori non informatives

La section précédente a montrée que les lois conjuguées peuvent être utilisées comme approximations des véritables lois a priori, par contre lors-qu'aucune information n'est disponible sur le modèle, leur utilisation n'est justifiée que par des considérations analytiques. Dans de telles situations, il est impossible de bâtir une distribution a priori sur

des considérations subjectives. On peut alors chercher à utiliser malgré tout des techniques bayésiennes qui intègrent notre ignorance sur les paramètres du modèles, de telles méthodes sont appelées de manière évidente, non informative [11].

La loi a priori de Jeffreys

Jeffreys (1946,1961) propose une approche intrinsèque qui évite effectivement le besoin de prendre en compte une structure d'invariance potentielle [17], tout en étant souvent compatible lorsque cette structure existe. Les lois a priori non informatives de Jeffreys sont fondées sur **l'information de Fisher**, donnée par :

$$I(\theta) = E_\theta\left[\left(\frac{\partial log f(x|\theta)}{\partial \theta}\right)^2\right]$$

Dans le cas unidimensionnel, sous certaines conditions de régularité, cette information est aussi égale à :

$$I(\theta) = -E_\theta\left[\left(\frac{\partial^2 log f(x|\theta)}{\partial \theta^2}\right)\right]$$

La loi a priori de Jeffreys est :

$$\pi^*(\theta) = I^{\frac{1}{2}}(\theta)$$

Définie à un coefficient de normalisation prés quand π^* est propre.

Exemple 2.1. *Si x~B(n,p),*

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\frac{\partial^2 ln f(x|p)}{\partial p^2} = \frac{x}{p^2} + \frac{n-x}{(1-p)^2}$$

donc

$$I(p) = n[\frac{1}{p} + \frac{1}{1-p}]$$

donc la loi de Jeffreys pour ce modèle est :

$$\pi^*(p) \propto [p(1-p)]^{-\frac{1}{2}}$$

et est alors propre, car il s'agit de la distribution $Beta(\frac{1}{2}, \frac{1}{2})$.

Dans le cas où θ est un paramètre multidimensionnel, on définit la matrice d'information de Fisher, pour $\theta \in \mathbb{R}^k$, $I(\theta)$ aux éléments suivants :

$$I_{ij}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} ln f(x|\theta)\right] \quad i,j = 1,...,k$$

et la loi non informative de Jeffreys est alors définie par :

$$\pi^*(\theta) \propto [det(I(\theta))]^{\frac{1}{2}}$$

2.3 Modèles paramétriques

Les modèles paramétriques jouent un rôle important dans l'analyse de survie bayésienne, car de nombreuses analyses bayésiennes sont réalisées en utilisant ce type de modélisation parce qu'elle offre des techniques simples [11]. Dans cette partie, nous discutons des modèles paramétriques pour les données de survie censurées à droite univariées. Nous dérivons les distributions a posteriori et démontrons comment effectuer des analyses bayésiennes pour plusieurs modèles paramétriques couramment utilisés dans les essais cliniques. La littérature statistique sur l'analyse de survie paramétrique bayésienne et les tests de vie est trop énorme pour être listée ici, mais quelques références traitant d'applications à la médecine où à la santé publique incluent [Grieve (1987), Dellaportas et Smith (1993), et Kim et Ibrahim (2001)] [9 ;5 ;11].

2.3.1 Modèle exponentiel

Le modèle exponentiel est le modèle paramétrique le plus fondamental dans l'analyse de survie. Supposons que nous ayons des temps de survie indépendants identiquement distribués $i.i.d$ $x = (x_1, x_2, ..., x_n)'$, chacun ayant une distribution exponentielle avec paramètre λ dénoté par $\xi(\lambda)$. Dénoter les indicateurs de censure par $\delta = (\delta_1, \delta_2, ..., \delta_n)'$, où

$$\delta_i = \begin{cases} 0 & \text{si } x_i \text{ est censuré à droite} \\ 1 & \text{si } x_i \text{ est un temps d'échec} \end{cases}$$

Soient :

La densité de x_i :

$$f(x_i|\lambda) = \lambda e^{(-\lambda x_i)}.$$

La fonction de survie :

$$S(x_i|\lambda) = e^{(-\lambda x_i)}.$$

Les données observées :

$$D = (n, x, \delta).$$

La fonction de vraisemblance de λ est :

$$\begin{aligned} L(\lambda|D) &= \prod_{i=1}^{n} f(x_i|\lambda)^{\delta_i} S(x_i|\lambda)^{(1-\delta_i)} \\ &= \lambda^d e^{\left(-\lambda \sum_{i=1}^{n} x_i\right)}. \end{aligned}$$

Où : $d = \sum_{i=1}^{n} \delta_i$.
En prenant comme loi a priori la loi conjuguée naturelle qui est la loi gamma. Soit $G(\alpha_0, \lambda_0)$ la distribution gamma avec des paramètres (α_0, λ_0). La loi a priori est donnée par :

$$\begin{aligned}\pi(\lambda|\alpha_0, \lambda_0) &= \frac{\lambda_0^{\alpha}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}e^{(-\lambda_0\lambda)} \\ &\propto \lambda^{\alpha_0-1}e^{(-\lambda_0\lambda)}\end{aligned}$$

et d'après le théorème de Bayes on obtient comme fonction a posteriori :

$$\begin{aligned}\pi(\lambda|D) &\propto L(\lambda|D)\pi(\lambda|\alpha_0, \lambda_0) \\ &\propto \left(\lambda^{\sum_{i=1}^{n}\delta_i} e^{\left\{-\lambda \sum_{i=1}^{n} x_i\right\}}\right)\left(\lambda^{\alpha_0-1}e^{(-\lambda_0\lambda)}\right) \\ &= \lambda^{\alpha_0+d-1}e^{\left\{-\lambda(\lambda_0+\sum_{i=1}^{n} x_i)\right\}} \qquad (2.1)\end{aligned}$$

Ainsi, nous reconnaissons dans (2.1) le noyau de la distribution a posteriori comme une distribution $G(\alpha_0 + d, \lambda_0 + \sum_{i=1}^{n} x_i)$.

La moyenne et la variance a posteriori de λ sont donc données par :

$$\mathbb{E}(\lambda|D) = \frac{\alpha_0 + d}{\lambda_0 + \sum_{i=1}^{n} x_i} \qquad (2.2)$$

$$Var(\lambda|D) = \frac{\alpha_0 + d}{(\lambda_0 + \sum_{i=1}^{n} x_i)^2}$$

La distribution prédictive a posteriori d'un futur temps de défaillance x_f est donnée par :

$$\begin{aligned}\pi(x_f|D) &= \int_0^{\infty} \pi(x_f|\lambda)\pi(\lambda|D)d\lambda \\ &\propto \int_0^{\infty} \lambda^{\alpha_0+d+1-1} e^{\left\{-\lambda(x_f+\lambda_0+\sum_{i=1}^{n} x_i)\right\}} d\lambda \\ &= \Gamma(\alpha_0 + d + 1)\left(\lambda_0 + \sum_{i=1}^{n} x_i + x_f\right)^{-(d+\alpha_0+1)} \\ &\propto \left(\lambda_0 + \sum_{i=1}^{n} x_i + x_f\right)^{-(d+\alpha_0+1)} \qquad (2.3)\end{aligned}$$

La distribution prédictive a posteriori normalisée est donc donnée par :

$$\pi(x_f|D) = \begin{cases} \frac{(d+\alpha_0)(\lambda_0+\sum_{i=1}^{n} x_i)^{(\alpha_0+d)}}{(\lambda_0+\sum_{i=1}^{n} x_i+x_f)^{(\alpha_0+d+1)}} & \text{si } x_f > 0, \\ 0 & \text{sinon.} \end{cases} \tag{2.4}$$

Dans la dérivation de (2.3) ci-dessus, nous devons évaluer une intégrale gamma, ce qui a conduit à la distribution prédictive a posteriori dans (2.4). La distribution prédictive de (2.4) est connue sous le nom de distribution bêta inverse et est discutée en détail dans [1].

2.3.2 Modèle de Weibull

Le modèle de Weibull est le modèle de survie paramétrique le plus utilisé. Supposons que nous ayons des temps de survie $i.i.d\ x = (x_1, x_2, ..., x_n)'$, chacun ayant une distribution de Weibull avec le paramètre de forme α et le paramètre d'échelle γ. Il est souvent plus pratique d'écrire le modèle en termes de paramétrisation $\lambda = ln(\gamma)$, menant à :

$$f(x|\alpha, \lambda) = \alpha x^{\alpha-1} e^{(\lambda - e^{\lambda} x^{\alpha})}.$$

Tout au long, nous utiliserons la notation $W(\alpha, \lambda)$, où $\lambda = log(\gamma)$ pour dénoter la loi de weibull.
Soit :
La fonction de survie :

$$S(x|\alpha, \lambda) = e^{(-e^{\lambda} x^{\alpha})}$$

La fonction de vraisemblance de (α, λ) peut s'écrire comme suit :

$$\begin{aligned} L(\alpha, \lambda|D) &= \prod_{i=1}^{n} f(x_i|\alpha, \lambda)^{\delta_i} S(x_i|\alpha, \lambda)^{1-\delta_i} \\ &= \alpha^d e^{\left\{ d\lambda + \sum_{i=1}^{n} (\delta_i(\alpha-1) ln(x_i) - e^{\lambda} x_i^{\alpha}) \right\}} \end{aligned}$$

Lorsque α est supposé connu, on peut prendre une loi a priori conjuguée pour λ qui est $\xi(\lambda)$. Mais lorsque les deux paramètres (α, λ) sont tous les deux supposés inconnus aucun loi a priori n'est proposée voir [11].

Dans ce cas, on suppose l'indépendance entre α et λ et prendre $\alpha \sim G(\alpha_0, \kappa_0)$ et $\lambda \sim N(\mu_0, \sigma_0^2)$.

La distribution a posteriori jointe de (α, λ) est donnée par :

$$\begin{aligned}\pi(\alpha,\lambda|D) &\propto l(\alpha,\lambda|D)\pi(\alpha|\alpha_0,\kappa_0)\pi(\lambda|\mu_0,\sigma_0^2)\\ &\propto \prod_{i=1}^{n} f(x_i|\alpha,\lambda)^{\delta_i} S(x_i|\alpha,\lambda)^{1-\delta_i}\\ &= \alpha^{\alpha_0+d-1} e^{\left\{d\lambda+\sum_{i=1}^{n}(\delta_i(\alpha-1)log(x_i)-e^{(\lambda)}x_i^{\alpha})-\kappa_0\alpha-\frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right\}}\end{aligned}$$

La distribution a postériori jointe de (α, λ) n'a pas de forme explicite, mais on peut montrer que les distributions a postériori conditionnelles $\pi(\alpha|\lambda, D)$ est $\pi(\lambda|\alpha, D)$ sont log-concave [11], et donc l'échantillonnage de Gibbs est requis pour ce modèle.

2.3.3 Modèle Gamma

Le modèle gamma est une généralisation du modèle exponentiel. Pour ce modèle $x_i \sim G(\alpha, \lambda)$.
La densité de probabilité est :

$$f(x_i|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)} x_i^{\alpha-1} e^{(\alpha\lambda - x_i e^{(\lambda)})}.$$

La fonction de survie est donnée par :

$$S(x_i|\alpha,\lambda) = 1 - IG(\alpha, x_i e^{(\lambda)}),$$

où la fonction gamma incomplète est :

$$IG(\alpha, x_i e^{(\lambda)}) = \frac{1}{\Gamma(\alpha)} \int_0^{x_i e^{(\lambda)}} u^{\alpha-1} e^{(-u)} du \tag{2.5}$$

Nous pouvons donc écrire la fonction de vraisemblance de (α, λ) comme suit :

$$\begin{aligned}L(\alpha,\lambda|D) &= \prod_{i=1}^{n} f(x_i|\alpha,\lambda)^{\delta_i} S(x_i|\alpha,\lambda)^{1-\delta_i}\\ &= \frac{1}{(\Gamma(\alpha))^d} e^{\left\{d\alpha\lambda+\sum_{i=1}^{n}\delta_i(\alpha ln(x_i)-x_i e^{\lambda})\right\}}\\ &\times \prod_{i=1}^{n} x_i^{-\delta_i}(1 - IG(\alpha, x_i e^{\lambda}))^{1-\delta_i}\end{aligned} \tag{2.6}$$

Aucun conjugué a piori n'est disponible lorsque (α, λ) sont tous les deux supposés inconnus. Dans ce cas, une spécification a priori jointe typique est de prendre $\alpha \sim G(\alpha_0, \kappa_0)$ et $\lambda \sim N(\mu_0, \sigma_0^2)$ indépendamment. Sous cette formulation, la distribution a posteriori

jointe de (α, λ) est donnée par :

$$\begin{aligned}\pi(\alpha,\lambda|D) &\propto L(\alpha,\lambda|D)\pi(\alpha,\lambda|\alpha_0,\kappa_0,\mu_0,\sigma_0)\\ &= \frac{\alpha^{\alpha_0-1}}{(\Gamma(\alpha))^d} e^{\left\{d\alpha\lambda+\sum_{i=1}^{n}\delta_i(\alpha ln(x_i)-x_i e^{\lambda})\right\}}\\ &\times \prod_{i=1}^{n} x_i^{-\delta_i}(1-IG(\alpha,x_i e^{\lambda}))^{1-\delta_i}\\ &\times e^{(-\kappa_0\alpha-\frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2)}\end{aligned}$$

La distribution a posteriori jointe de (α, λ) n'a pas de forme explicite. Pour simplifier le calcul de la loi a posteriori, nous introduisons des données complètes $x^* = (x_1^*, x_2^*, ..., x_n^*)'$ tel que :

$$\begin{cases} x_i^* = x_i & \text{si } \delta_i = 1 \\ x_i^* > x_i & \text{si } \delta_i = 0 \end{cases}$$

Alors la vraisemblance complète des données de (α, λ) sachant x^* et D prend la forme :

$$\begin{aligned}L(\alpha,\lambda|x^*,D) &= \prod_{i=1}^{n} f(x_i^*|\alpha,\lambda)\\ &= \frac{1}{(\Gamma(\alpha))^n} e^{\left\{n\alpha\lambda+\sum_{i=1}^{n}((\alpha-1)ln(x_i^*)-x_i^* e^{\lambda})\right\}} \qquad (2.7)\end{aligned}$$

Ensuite nous avons :

$$L(\alpha,\lambda|D) = \int_{\{x_i^*>x_i:\delta_i=0\}} L(\alpha,\lambda|x^*,D) \prod_{i:\delta_i=0} dx_i^*,$$

Où $L(\alpha, \lambda|D)$ est donné en (2.6) [11]. En utilisant (2.7), la distribution a posteriori conjointe de (α, λ, x^*) est donc donnée par :

$$\pi(\alpha,\lambda,x^*|D) \propto L(\alpha,\lambda|x^*,D) \times \alpha^{\alpha_0-1} e^{\left(-\kappa_0\alpha-\frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right)}$$

Les distributions a posteriori conditionnelles $\pi(\alpha|\lambda, x^*, D)$ et $\pi(\lambda|\alpha, y^*, D)$ sont log-concaves aussi longtemps que $n \geqslant 2$. En outre, pour (α, λ, D), et pour $\delta_i = 0$, x^* a une distribution gamma tronquée avec densité :

$$\pi(x_i^*|\alpha,\lambda,D) = \frac{(x_i^*)^{\alpha-1}}{\Gamma(\alpha)(1-IG(\alpha,x_i e^{\lambda}))} e^{(\alpha\lambda-x_i^* e^{\lambda})}, \quad x_i^* > x_i$$

Où $IG(\alpha, x_i e^{\lambda})$ est la fonction gamma incomplète définie par (2.5). Ainsi, la mise en œuvre de l'échantillonneur de gibbs est requie.

2.3.4 Modèle de valeur extrême

Le modèle de valeur extrême est un autre modèle de survie paramétrique largement utilisé.

Supposons que nous avons des temps de survie indépendants identiquement distribués $x = (x_1, x_2, ..., x_n)'$, chacun ayant une distribution de valeur extrême, notée $\mathcal{V}(\alpha, \lambda)$, avec densité

$$f(x|\alpha, \lambda) = \alpha e^{\alpha x} e^{\lambda - e^{\lambda + \alpha x}}$$

Pour $-\infty \leq x \leq \infty$. Nous notons, que la distribution de valeur extrême peut être considérée comme une autre paramétrisation d'une distribution de Weibull. En effet, soit une variable aléatoire $T \sim \mathcal{W}(\alpha, \lambda)$; alors $x = ln(T) \sim \mathcal{V}(\alpha, \lambda)$.

Soit maintenant $S(x|\alpha, \lambda) = e^{(-e^{\lambda + \alpha x})}$ la fonction de survie. Nous pouvons écrire la fonction de vraisemblance de (α, λ) comme suit :

$$\begin{aligned} L(\alpha, \lambda|D) &= \prod_{i=1}^{n} f(x_i|\alpha, \lambda)^{\delta_i} S(x_i|\alpha, \lambda)^{(1-\delta_i)} \\ &= \alpha^d e^{\left\{\sum_{i=1}^{n} (\alpha x_i \delta_i + \lambda \delta_i - e^{\lambda + \alpha x_i})\right\}} \end{aligned}$$

Aucun conjugué a priori n'est disponible lorsque (α, λ) sont tous les deux supposés inconnus. Dans ce cas, une spécification a priori conjointe typique consiste à considérer α et λ comme indépendants, où α a une distribution gamma et λ a une distribution normale. En supposant $G(\alpha_0, \kappa_0)$ l'a priori pour α, et $N(\mu_0, \sigma_0^2)$ l'a priori pour λ, la distribution a posteriori conjointe de (α, λ) est donnée par :

$$\begin{aligned} \pi(\alpha, \lambda|D) &\propto L(\alpha, \lambda|D)\pi(\alpha|\alpha_0, \kappa_0)\pi(\lambda|\mu_0, \sigma_0^2) \\ &\propto \prod_{i=1}^{n} f(x_i|\alpha, \lambda)^{\delta_i} S(x_i|\alpha, \lambda)^{(1-\delta_i)} \\ &= \alpha^{\alpha_0 + d - 1} e^{\left\{ d\lambda + \sum_{i=1}^{n} (\alpha \delta_i x_i - e^{\lambda + \alpha x_i}) - \kappa_0 \alpha - \frac{1}{2\sigma_0^2}(\lambda - \mu_0)^2 \right\}} \end{aligned}$$

La distribution a posteriori conjointe de (α, λ) n'a pas de forme explicite, mais on peut montrer que les distributions a posteriori conditionnelles $\pi(\alpha|\lambda, D); \pi(\lambda|\alpha, D)$ sont toutes deux log-concaves, et donc L'échantillonnage de Gibbs est requis pour ce modèle.

2.4 Méthodes non paramétriques bayésiennes

La méthode bayésienne non paramétrique est une alternative à l'approche non paramétrique classique pour estimer la fonction de survie discutée dans le 1^{er} chapitre. En appliquant ces méthodes, la croyance a priori d'un investigateur sur la forme de la fonction de survie est combinée avec les données pour fournir une fonction de survie estimée. L'information a priori, qui peut être basée sur l'expérience a priori avec le processus observé ou sur l'opinion d'un expert, se reflète dans une distribution a priori pour la fonction de survie.

Typiquement, les moyennes a priori reflètent la meilleure estimation des investigateurs, avant de voir toute donnée, de la valeur des paramètres, et la variance a priori est une mesure de l'incertitude de l'investigateur de ses moyennes a priori. On peut souvent penser que la variance a priori est inversement proportionnelle à la quantité d'information d'échantillon devant être représentée par l'a priori [13].

Dans notre problème, le paramètre d'intérêt est la fonction de survie ou, de manière équivalente, la fonction de risque cumulatif. Ceci doit être traité comme une quantité aléatoire échantillonnée à partir d'un processus stochastique. On choisit une trajectoire de ce processus stochastique, et ceci est notre fonction de survie. Nous avons donc des données échantillonnées à partir d'une population avec cette fonction de survie que nous allons combiner avec notre a priori pour obtenir la distribution de la fonction de survie, compte tenu des données.

Pour obtenir une estimation de la fonction de survie, nous devons spécifier une fonction de perte sur laquelle baser la règle de décision. Analogue au cas paramétrique simple, nous utiliserons la fonction de perte quadratique

$$L(S, \hat{S}) = \int_0^\infty [\hat{S}(t) - S(t)]^2 dp(t),$$

où $p(t)$ est une fonction de poids. Cette fonction de perte est la différence intégrée pondérée entre la vraie valeur de la fonction de survie et notre valeur estimée. Pour cette fonction de perte, la valeur de $\hat{S}$, qui minimise la valeur attendue a postériori de $L(S, \hat{S})$, est la moyenne a postériori et le risque de Bayes $E[L(S, \hat{S})|D]$ est la variance a postériori.

Le processus fondateur de l'approche bayésienne des problèmes non paramétrique est le processus de Dirichlet. Rappelons tout d'abord, définition et propriétés de la loi de Dirichlet.

2.4.1 La distribution de Dirichlet

Le processus de Dirichlet est défini à partir de la distribution de Dirichlet. La plupart des propriétés du processus de Dirichlet sont analogues aux propriétés de la distribution de Dirichlet. Commençons par définir la distribution de Dirichlet.

Soit S_{K-1} le simplexe de $\mathbb{R}^{K+1}$ défini par :

$S_{K-1} = \{p = (p_1, ..., p_{K-1}) \in \mathbb{R}^{K-1} : p_i \geq 0 \text{ pour } i = 1, 2, ..., K-1, \sum_{i=1}^{K-1} p_i \leq 1\}$

La distribution de Dirichlet est définie de la façon suivante :

Définition 2.4. *La distribution de Dirichlet est une distribution sur le simplexe S_{K-1} caractérisée par la densité de $p = (p_1, ..., p_{K-1})$ par rapport à la mesure de Lebesgue dans $\mathbb{R}^{K-1}$ vérifiant :*

$$f(p) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \left(\prod_{j=1}^{K-1} p_j^{\alpha_{j-1}} \right) \left(1 - \sum_{j=1}^{K-1} p_j \right)^{\alpha_{K-1}} \mathbb{1}_{\{p \in S_{K-1}\}} \tag{2.8}$$

où $\alpha = (\alpha_1, ..., \alpha_K)$ est un jeu de paramètres avec $\alpha_j > 0$, et Γ la fonction Gamma définie par :

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \alpha > 0.$$

La moyenne de p_i est égale à :

$$\mathbb{E}(p_i) = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i},$$

et sa variance est

$$var(p_i) = \frac{((\sum_{i=1}^{K} \alpha_i) - \alpha_i)\alpha_i}{(\sum_{i=1}^{K} \alpha_i)^2 + (\sum_{i=1}^{K} \alpha_i)^3}.$$

Remarque 2.2. *Dans la suite on notera la distribution de Dirichlet de paramètre α par $Dir(\alpha)$ ou $Dir(\alpha_1, ..., \alpha_K)$. La distribution de Dirichlet est une généralisation de la loi Bêta, on a égalité de ces deux lois de probabilité si $K = 2$. Comme pour la distribution Bêta, la distribution de Dirichlet admet une représentation utile en terme de variables Gamma [6].*

Proposition 2.2. *([6]) Si $Y_1, ..., Y_K$ sont des variables aléatoires Gamma indépendantes de paramètres α_j et 1 avec $\alpha_j \geq 0$ pour tout j et $\sum_{j=1}^{K} \alpha_j > 0$, alors*

i. Le vecteur

$$\left(\frac{Y_1}{\sum_{j=1}^{K} Y_j}, ..., \frac{Y_K}{\sum_{j=1}^{K} Y_j}\right)$$

est distribué selon une loi de Dirichlet $Dir(\alpha_1, ..., \alpha_K)$;

ii.

$$(\frac{Y_1}{\sum_{j=1}^{K} Y_j}, ..., \frac{Y_K}{\sum_{j=1}^{K} Y_j})$$

est indépendant de $\sum_{j=1}^{K} Y_j$

iii. Si $p = (p_1, ..., p_K) \sim Dir(\alpha_1, ..., \alpha_K)$*, alors pour toute partition* $B_1, ..., B_m$ *de* χ*, le vecteur*

$$(\mathbb{P}(B_1), ..., \mathbb{P}(B_m)) = \left(\sum_{j \in B_1} p_j, ..., \sum_{j \in B_m} p_j \right) \sim Dir(\alpha_1^{'}, ..., \alpha_K^{'}),$$

où $\alpha_i^{'} = \sum_{j \in B_i} \alpha_j$.

En particulier, la distribution marginale de p_j *est une loi Bêta de paramètres* $(\alpha_j, \sum_{j \neq i} \alpha_i)$.

Remarque 2.3. *On peut donc interpréter le paramètre* α *comme une mesure sur* χ *en posant* $\alpha(\{j\}) = \alpha_j$*, ainsi* $\alpha(\chi) = \sum_{j=1}^{K} \alpha_j$.

Pour obtenir un échantillon selon une distribution de Dirichlet $Dir(\alpha)$, il suffit de tirer K variables indépendantes $(y_1, ..., y_K)$ de loi Gamma $G(\alpha_j, 1)$.
Le vecteur $x = (x_1, ..., x_K)$ construit de la façon suivante :

$$x_i = \frac{y_i}{\sum_{j=1}^{K} y_j} \quad , i = 1, ..., K.$$

sera donc une réalisation de la loi (2.8) [6].

On définit alors le processus de Dirichlet :

2.4.2 Le processus de Dirichlet

Soit $(\chi, \mathcal{A})$ un espace mesurable. Notons $\mathcal{P}$ l'ensemble de toutes les lois de probabilité sur $(\chi, \mathcal{A})$.

Le processus de Dirichlet est défini comme une mesure de probabilité sur l'espace des mesures de probabilités.

Définition 2.5. *(Processus de Dirichlet). Soient* $(\chi, \mathcal{A})$ *un espace mesurable,* P_0 *une mesure de probabilité sur cet espace et* α_0 *un nombre réel positif. Une mesure de probabilité* G *est distribuée selon un processus de Dirichlet de paramètres* G_0 *et* α_0 *si pour*

toutes partitions finies $(B_l)_{l=1,...,r}$ *de* χ*, la loi de* $(G(B_1), ..., G(B_r))$ *est une loi de Dirichlet* $D(\alpha_0 G_0(B1), ..., \alpha_0 G_0(B_r))$.
On note ceci par

$$G \sim DP(\alpha_0 G_0)$$

Ce processus est donc défini par deux paramètres : α_0 qui correspond à un paramètre de concentration (ou paramètre d'échelle) et G_0 qui est une mesure de probabilité de base.

Proposition 2.3. *Si* $G \sim DP(\alpha_0 G_0)$ *alors* $G(B)$ *suit une loi Bêta de paramètres* $\alpha_0 G_0(B)$ *et* $\alpha_0(1 - G_0(B))$. *Ce qui implique en particulier que*

$$\mathbb{E}[G(B)] = G_0(B)$$

pour tout ensemble mesurable B *de* χ.
De même :

$$var(G(B)) = \frac{G_0(B)(1 - G_0(B))}{\alpha_0 + 1}$$

α_0 *est un paramètre d'échelle du processus.*

Démonstration. On considère la partition (B, B^c), d'après la définition 2.5

$$G(B) \sim Bêta(\alpha_0 G_0(B), \alpha_0(1 - G_0(B))).$$

Ainsi la preuve est immédiate. ■

Le théorème suivant est très important. Il justifie le choix des processus de Dirichlet dans un modèle bayésien non paramétrique en montrant la simplicité de la mise à jour de la distribution a posteriori.

Théorème 2.1. *([14])*

1) *Si* G *est a priori distribuée suivant un processus de Dirichlet* $DP(\alpha_0 G_0)$ *et si* $\theta = (\theta_1, ..., \theta_n)$ *est un échantillon i.i.d. de loi* G *alors la loi a posteriori de* G *est un processus de Dirichlet* $DP(\alpha_0' G_0')$ *tel que :*

$$\alpha_0' = \alpha_0 + n, \quad G_0' = \frac{\alpha_0}{\alpha_0 + n} G_0 + \frac{n}{\alpha_0 + n} G_n,$$

où $G_n = (1/n) \sum \delta_{\theta_i}$ *est la loi empirique de l'échantillon.*

2) *La loi marginale de* $(\theta_1, ..., \theta_n)$ *peut être composée de la manière suivante :*

$$\theta_1 \sim G_0$$

$$\theta_{i+1} | \theta_1, ..., \theta_i \sim \frac{\alpha_0}{i + n} G_0 + \frac{i}{i + n} G_i, \quad (i = 1, ..., n - 1)$$

où G_i *est la loi empirique associée à l'échantillon* $(\theta_1, ..., \theta_n)$.

Remarque 2.4. *Ce théorème permet d'observer deux choses importantes :*

i. La distribution de Dirichlet constitue une famille fermée pour l'inférence dans un échantillonnage i.i.d. non paramétrique.

ii. On remarque que si α_0 est petit, la loi a posteriori devient le processus de Dirichlet $DP(nG_n)$. Ce processus ne dépend que de l'échantillon et il est centré sur G_n. Ainsi, dans le cas où $\alpha_0 = 0$, le processus de Dirichlet $DP(\alpha_0 G_0)$ fournirait une distribution a priori non informative. Le paramètre α_0 représente donc le degré de la connaissance a priori et G_0 reflète la connaissance a priori sous la forme de la distribution.

2.4.3 Méthodes non paramétriques bayésiennes pour estimer la fonction de survie

Pour affecter une distribution a priori à la fonction de survie, nous supposons que $S(t)$ suit une distribution de Dirichlet avec la fonction de paramètre $\alpha = (\alpha_1, ..., \alpha_K)$.

En règle générale, nous prenons la fonction de paramètre sous la forme :

$$\alpha([t, \infty)) = cS_0(t)$$

où $S_0(t)$ est notre estimation a priori de la fonction de survie et c est une mesure de combien de poids à mettre sur notre estimation préalable. Avec cette distribution a priori pour $S(t)$, la moyenne a priori est exprimée par

$$\mathbb{E}[S(t)] = \frac{\alpha(t, \infty)}{\alpha(0, \infty)} = \frac{cS_0(t)}{cS_0(0)} = S_0(t),$$

et la variance a priori est donnée par

$$V[S(t)] = \frac{[\alpha(0, \infty) - \alpha(t, \infty)]\alpha(t, \infty)}{[\alpha(0, \infty)^2 + \alpha(0, \infty)^3]} = \frac{S_0(t)[1 - S_0(t)]}{c + 1}.$$

Remarque 2.5. *Notons que la variance a priori est l'équivalent de la variance d'échantillon que l'on aurait si on avait un échantillon non censuré de taille c+1 d'une population avec une fonction de survie $S_0(t)$ [13].*

Les données que nous avons à combiner avec notre a priori sont les temps d'étude sur X_j et l'indicateur d'événement δ_j. Pour simplifier les calculs, soit $0 = t_0 < t_1 < \ldots < t_N <$

$t_{N+1} = \infty$, dénotent les N temps distincts (censurés ou non censurés). A l'instant t_i, soit R_i le nombre d'individus à risque, M_i le nombre de décès et λ_i le nombre d'observations censurées. Soit :

$$\delta_i = \begin{cases} 1 & \text{si } M_i > 0 \\ 0 & \text{si } M_i = 0. \end{cases}$$

En combinant ces données avec l'a priori, nous trouvons que la distribution a postériori de S est aussi de Dirichlet. Le paramètre de la distribution a postériori, α^* pour tout intervalle (a, b) est :

$$\alpha^*((a,b)) = \alpha((a,b)) + \sum_{j=1}^{n} \mathbb{1}_{[\delta_j > 0, a < X_j < b]},$$

L'estimateur de Bayes de la fonction de survie est :
pour $t_i \leq t \leq t_{i+1}$ $\quad ,i = 0, ..., N.$

$$\hat{S}_D(t) = \frac{\alpha(t,\infty) + R_{i+1}}{\alpha(0,\infty) + n} \prod_{K=1}^{i} \frac{\alpha(t_K,\infty) + R_{K+1} + \lambda_K}{\alpha(t_K,\infty) + R_{K+1}}$$

L'estimateur de Bayes est une fonction continue entre les temps de mort distincts et a des sauts à ces temps de mort.

Remarque 2.6. *Pour n grand, ceci se réduit à l'estimateur de Kaplan-Meier, de sorte que l'information a priori ne joue aucun rôle dans l'estimation. Pour les petits échantillons, l'a priori dominera, et l'estimateur sera proche de la supposition précédente de S.*

Remarque 2.7. *Il existe une deuxième approche pour estimer la fonction de survie en utilisant l'estimateur de risque cumulée. Ce dernier est estimer a partir du processus bêta. Pour plus de détail voir [13].*

Remarque 2.8. *Phadia et Susarla (1983) ont dérivé l'estimateur de Bayes de S en considérons le processus de Dirichlet comme a priori mais ils n'ont pas déterminé la distribution a posteriori. Arnold et al (1984) identifiait la distribution a posteriori comme un mélange de processus de Dirichlet (théorème 2.1) et démontrait l'incohérence générale des estimateurs bayésien correspondants de S ; un défaut attribuable à la non identifiabilité du modèle. L'efficacité de l'estimation bayésienne a été démontrée par Neath et Samaniego (1996a,1996b) dans le contexte de la théorie décisionnelle [10].*

Remarque 2.9. *L'estimateur a priori du processus de Dirichlet de la fonction de survie a été proposé par Ferguson (1973)[7] pour des données non censurées. Susarla et Van*

Ryzin (1976) [19] et Ferguson et Phadia (1979) [8] étendent le processus d'estimation aux données censurées à droite.

Exemple 2.2. *Nous considérons l'exemple de Susarla et Van Ryzin (1976) pour obtenir l'estimateur de Bayes de S, où $F_0(t) = 1 - e^{-\theta t}$. Nous considérons les valeurs de $\theta = 0.12$ et $c_0 = 4, 8, 16$.(la moitie des observations, le nombre des observation et le double des observation respective)*
Les données sont constituées de décès en mois à

$$0, 8, 3, 1, 5, 4, 9, 2$$

et d'observations censurées en mois à

$$1, 0, 2, 7, 7, 0, 12, 1$$

Le tableau 2.2 donne les intervalles de t et l'estimateur de Bayes de $S(t)$.
D'après le tableau 2.2, on peut observer que tous les points d'observation non censurés, à savoir 0.8, 3.1, 5.4 et 9.2, sont des points de discontinuité de $\hat{S}(.)$.
Aux points d'observation censurés, à savoir, 1.0, 2.7, 7.0 et 12.1, $\hat{S}(.)$ est continu mais les dérivées gauche et droite de $\hat{S}(.)$ sont différentes à ces points. De plus, dans la représentation donnée par Susarla et Van Ryzin (1976), l'estimateur de Bayes est plus lisse que l'estimateur K-M, les sauts aux observations non censurées n'étant pas aussi importants pour l'estimateur de Bayes.

t dans	L'estimation de bayes,$F_0 = 1 - e^{(-\theta t)}$	L'estimation de K-M
[0,0.8)	$\frac{c_0 e^{(-\theta t)}+8}{c_0+8}$	1.0
[0.8,1.0)	$\frac{c_0 e^{(-\theta t)}+7}{c_0+8}$	7/8
[1.0,2.7)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}$	7/8
[2.7,3.1)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8
[3.1,5.4)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8
[5.4,7.0)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8
[7.0,9.2)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8
[9.2,12.1)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8
[12.1,∞)	$\frac{c_0 e^{(-\theta t)}+6}{c_0+8}\ \frac{c_0 e^{(-\theta t)}+7}{c_0+6}\ \frac{c_0 e^{(-\theta t)}+6}{c_0+5}$	7/8

TABLE 2.2: L'estimation de Bayes et de Kaplan Meier de S(t).

D'autre part, si F_0 a une masse positive à l'un des points d'observation censurés, alors l'estimation de bayes $\hat{S}(.)$ serait discontinue à ce point là. La question du choix

du paramètre F_0 du processus de Dirichlet a priori est, bien entendu, importante. Dans l'exemple avec $F_0 = 1 - e^{(-\theta t)}$, θ a été pris comme étant 0.12 basé sur l'argument heuristique suivant. Si l'on devait considérer l'estimateur de Bayes dans le problème sans données, l'estimateur de Bayes serait $\hat{S}(t) = e^{(-\theta t)}$. Si on force cet estimateur à satisfaire $e^{(-\theta M)} = 0.525$ où M est le quartile 0.525 de la courbe K-M, alors le θ ainsi obtenu est approximativement 0.12.

Nous avons choisi le quartile de 0.525(21/40) comme le saut le plus proche du 50^{eme} quartile de la courbe K-M observée. Ainsi, par une double utilisation des données, une valeur raisonnable de θ est de 0.12. D'autres valeurs de θ tendent à éloigner l'estimateur de Bayes de l'estimateur K-M. Grosso modo, ils prennent la masse précédente sur $(0, \infty)$ pour être 1/2, égaux, et deux fois la taille de l'échantillon $n = 8$. Plus le c_0 est grand, plus l'estimateur de Bayes devient lisse par rapport à l'estimateur K-M. Autrement dit, les sauts aux points de mort sont plus petits. Le cas extrême serait de laisser $c_0 \longrightarrow \infty$ dans de tel cas l'estimateur de Bayes se réduit à $e^{(-\theta t)}$ avec $\theta = 0.12$ dans cet exemple.

Chapitre 3

Applications dans les essais cliniques

Nous allons maintenant appliquer la théorie sur des données réelles. Dans la première application, nous nous intéressons aux modèles paramétriques et dans la deuxième aux estimateurs non paramétriques pour les deux approches ; frequentiste et bayésienne.

3.1 Essais cliniques

Les essais cliniques sont des études menées chez l'être humain pour démontrer la validité d'un nouveau produit thérapeutique (p.ex. un nouveau médicament, une association de médicaments ou une méthode de diagnostic) [19]. Ils ont pour but d'évaluer les produits thérapeutiques et d'estimer le bénéfice de leur utilisation favorable. Au cours du développement clinique leurs conditions optimales d'emploi sont déterminées. Le développement d'un médicament entre le brevet et la commercialisation dure en moyenne 12 ans.

3.1.1 Les différentes phases d'un essai clinique

On distingue **IV** phases dans le développement d'un produit thérapeutique :

•**Phase I** : Le nouveau produit est testé pour la première fois chez l'homme, en général chez des volontaires sains. Dans certains cas, il n'est pas éthique de tester le produit sur les volontaires humains sains, comme en oncologie, par exemple où le produit sera administré à des malades ayant déjà reçu le traitement standard. Avant d'entamer une phase I, le produit a été évalué sur des cultures cellulaires en laboratoire et sur des animaux. Le

but de cette phase est de définir la tolérance du nouveau produit et de vérifier si les résultats toxicologiques obtenus lors du développement préclinique sont comparables à ceux obtenus chez l'homme. La 1^{ere} phase permet de déterminer la dose maximale tolérée chez l'être humain.
Ces essais sont souvent menés sur quelques dizaines (20 à 100) volontaires sains [9].

•**Phase II** : Lors de la 2^{eme} phase, on s'intéresse à la détermination de la posologie optimale du produit en termes d'efficacité et de tolérance sur une population de patients limitée et homogène.
Ces essais sont souvent menés sur des petits groupes de sujets (100 à 200 sujets).

• **Phase III** : Cette étape a pour objectif de prouver l'efficacité du nouveau produit par rapport à un produit de référence déjà commercialisé ou par rapport à un placebo, c'est-à-dire un traitement sans activité pharmacologique.
Les essais concernent un plus grand nombre de patients (500 à 3000) qui sont des malades volontaires.

•**Phase IV** : Il s'agit de la seule phase qui est réalisée après la commercialisation du produit. Ces essais vont approfondir les connaissances du produit dans les conditions réelles d'utilisation et de ses effets secondaires sur une population importante.
Les phases **II, III et IV** peuvent comprendre plusieurs essais cliniques.

3.2 Application sur les données de VIH

Appliquons l'estimation paramétrique sur les données d'infection au VIH fournie par David W. Hosmer et Stanley Lemeshow [10]. Nous nous intéressons au temps de survie des patients infectés par le VIH.

Ici nous donnons le temps de survie en mois de chaque patient et une information de censure. CENSOR = 1 si la donnée n'est pas censurée et CENSOR = 0 si la donnée est censurée à droite. Cet ensemble de données est fourni dans la table de l'annexe B.

3.2.1 Estimateur non-paramétrique de K-M

Il est intéressant de commencer par l'estimateur non paramétrique classique de Kaplan Meier (1.3.2) pour avoir une idée de la forme de la fonction de survie que l'on peut

comparer ensuite avec certains modèles ajustés. Nous estimons d'abord la fonction de survie en utilisant l'estimateur de Kaplan-Meier.

3.2.2 Modèle paramétrique exponentiel

Le modèle que nous adaptons est le modèle exponentiel simple :

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

Rappelons que :

i. Dans cas fréquentiste : l'estimateur de maximum de vraisemblance $\hat{\lambda}$ est défini dans (1.1).

ii. Dans cas Bayésien : l'estimateur de maximum de vraisemblance $\hat{\lambda}_b$ est défini dans (2.2).

C'est un modèle très simple prenant seulement une valeur positive ce qui est raisonnable pour expliquer le temps de survie. Afin de s'adapter au modèle, nous trouvons l'EMV en maximisant la vraisemblance logarithmique comme expliqué dans la section 1.3. Cependant, en utilisant le logiciel R, il existe un package "Survival"prêt à l'emploi qui fournit quelques outils dont nous avons besoin pour adapter notre modèle [31].

Nous dessinons le graphique du temps de survie en utilisant la formule :

$$\hat{S}(t) = 1 - F(t) = e^{-\hat{\lambda}t}$$

Nous comparons le modèle simple dans les deux cas fréquentiste et bayésien avec l'estimation non-paramétrique de K-M.

3.2.3 Une application informatique

Nous utilisons la fonction de R "survReg"pour ajuster les modèles paramétriques (avec l'approche EMV) pour les données censurées. Le programme R est destiné à dupliquer certains des calculs de la main précédente. Il adapte un modèle exponentiel aux données VIH, et donne un intervalle de confiance à 95% à cet estimateur.

Rappelons que le modèle exponentiel est juste un Weibull de paramètre $\alpha = 1$ ou, en ln(temps), est un modèle de valeur extrême (1.3.1) avec une échelle $\sigma = 1$. La fonction "survReg"ajuste ln(temps) et génère le coefficient $\hat{\mu} = -ln(\hat{\lambda})$, l'EMV de μ, le paramètre

de localisation de la distribution de valeur extrême. Par conséquent, $EMV(\lambda) = \hat{\lambda} = exp(-\hat{\mu})$ et $EMV(\theta) = \hat{\theta} = exp(\hat{\mu})$. La sortie inutile a été supprimée.

La fonction R "predict"est une fonction compagnon de "survReg". Il fournit des estimations de quantiles avec leur erreur standard (s.e). L'un des arguments de la fonction "predict"est le "type". Soit type = "uquantile" pour produire des estimations basées sur la transformation logarithmique.

Estimation non paramétrique de la fonction de survie de K-M

```
> summary(survfit(Surv(temps,stats)~1))
Call: survfit(formula = Surv(temps, stats) ~ 1)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	100	3	0.9700	0.0171	0.9371	1.000
2	83	1	0.9583	0.0205	0.9190	0.999
3	73	2	0.9321	0.0270	0.8805	0.987
4	61	1	0.9168	0.0306	0.8587	0.979
5	56	2	0.8840	0.0373	0.8139	0.960
6	49	1	0.8660	0.0406	0.7899	0.949
7	46	1	0.8472	0.0439	0.7654	0.938
8	39	1	0.8254	0.0478	0.7368	0.925
9	35	1	0.8019	0.0520	0.7062	0.910
10	32	3	0.7267	0.0627	0.6137	0.860
11	28	3	0.6488	0.0702	0.5248	0.802
12	25	2	0.5969	0.0736	0.4688	0.760
13	21	1	0.5685	0.0754	0.4384	0.737
14	20	1	0.5401	0.0768	0.4087	0.714
15	19	2	0.4832	0.0785	0.3514	0.664
22	16	1	0.4530	0.0792	0.3216	0.638
30	14	1	0.4207	0.0799	0.2899	0.610
31	13	1	0.3883	0.0800	0.2593	0.582
32	12	1	0.3559	0.0796	0.2296	0.552
34	11	1	0.3236	0.0787	0.2009	0.521
35	10	1	0.2912	0.0772	0.1732	0.490
36	9	1	0.2589	0.0751	0.1466	0.457
43	8	1	0.2265	0.0723	0.1211	0.424
53	7	1	0.1942	0.0689	0.0969	0.389
54	6	1	0.1618	0.0645	0.0740	0.354
57	4	1	0.1213	0.0598	0.0462	0.319
58	3	1	0.0809	0.0517	0.0231	0.283

Estimation paramétrique fréquentiste du modèle exponentielle

Générons le coefficient $\hat{\mu}$ en utilisant la fonction "survRreg"comme suit :

```
> exp.fit <- survreg(Surv(temps,stats)~1,dist="exponential")
> exp.fit
Call:
survreg(formula = Surv(temps, stats) ~ 1, dist = "exponential")

Coefficients:
(Intercept)
   3.424351

Scale fixed at 1

Loglik(model)= -163.7   Loglik(intercept only)= -163.7
n= 100
```

L'estimateur de μ est :

```
> coeff <- exp.fit$coeff # muhat
> coeff
(Intercept)
   3.424351
```

L'estimateur de θ est :

```
> thetahat <- exp(coeff) # exp(muhat)
>  thetahat
(Intercept)
    30.7027
```

Les données non censurées sont

```
> weeks.u <- temps[stats == 1]
> weeks.u
 [1]  1  1  1  2  3  3  4  5  5  6  7  8  9 10 10 10 11 11 11 12 12 13 14 15 15
[26] 22 30 31 32 34 35 36 43 53 54 57 58
```

Telle que le nombre des données non censurées est 37.

L'estimateur de $\hat{\lambda}$ est :

```
> scalehat <- (exp(-muhat))
> scalehat
(Intercept)
 0.03257042
```

L'estimation de la fonction de survie est données par :

```
> Sfhat <- function(t) 1 - pexp(t,scalehat)
> Sfhat(weeks.u)
 [1] 0.9679543 0.9679543 0.9679543 0.9369355 0.9069107 0.9069107 0.8778481
 [8] 0.8497168 0.8497168 0.8224871 0.7961299 0.7706173 0.7459223 0.7220187
[15] 0.7220187 0.7220187 0.6988811 0.6988811 0.6988811 0.6764850 0.6764850
[22] 0.6548065 0.6338228 0.6135115 0.6135115 0.4884348 0.3763963 0.3643344
[29] 0.3526591 0.3304188 0.3198303 0.3095811 0.2464668 0.1779536 0.1722510
[36] 0.1562162 0.1512102
```

Estimation paramétrique bayésienne de modèle exponentielle

Pour trouver l'estimateur bayésien $\hat{\lambda}_b$, on peut choisir une a priori informative conjuguée $\lambda \sim G(1,2)$ (le choix des paramètres de l'a priori était arbitraire), d'où $\hat{\lambda}_b = \mathbb{E}(\lambda|D) = 0.03339192$.

L'estimation bayésienne de la fonction de survie du modèle exponentielle est :

```
> Sbhat <- function(t) 1 - pexp(t,lamhatb)
> Sbhat(weeks.u)
 [1] 0.9671594 0.9671594 0.9671594 0.9353974 0.9046784 0.9046784 0.8749683
 [8] 0.8462338 0.8462338 0.8184430 0.7915649 0.7655695 0.7404277 0.7161117
[15] 0.7161117 0.7161117 0.6925942 0.6925942 0.6925942 0.6698490 0.6698490
[22] 0.6478508 0.6265750 0.6059979 0.6059979 0.4796867 0.3672335 0.3551733
[29] 0.3435092 0.3213176 0.3107654 0.3005597 0.2379125 0.1703719 0.1647768
[36] 0.1490700 0.1441745
```

Finalement les résultats sont représentées dans la Figure 3.1 :

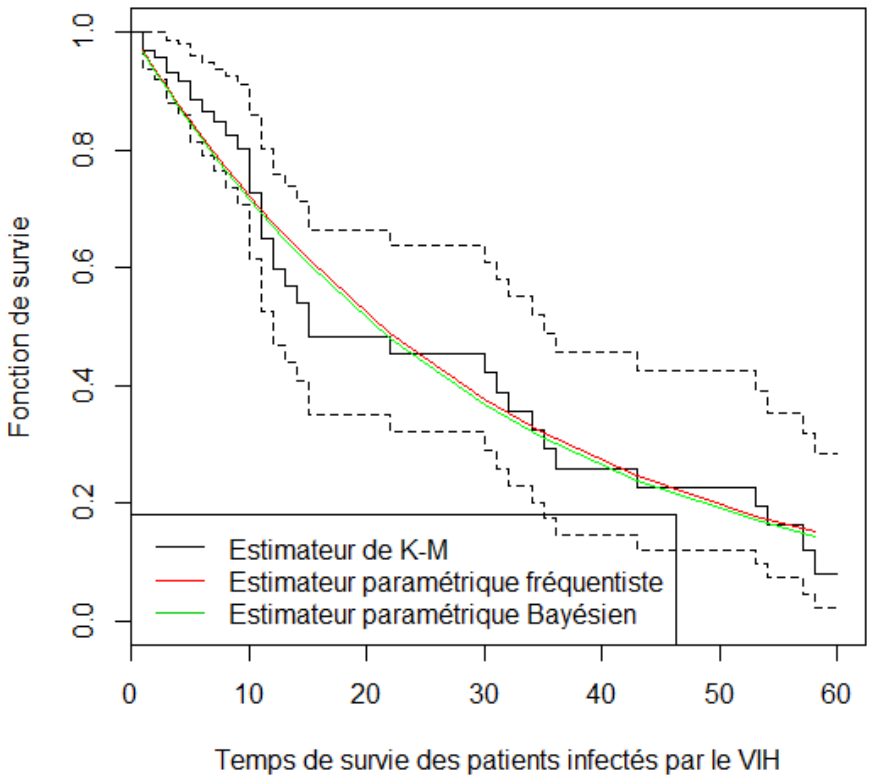

FIGURE 3.1: Comparaison entre la fonction de survie donnée par le modèle exponentiel fréquentiste et bayésienne et l'estimateur de K-M

.

Discussion : On remarque que les trois estimateurs sont proches l'un des autres car les graphes sont presque ajustés. On peut conclure que ces données peuvent être représenter par une distribution $\xi(\hat{\lambda}) = \xi(0.0326) \approx \xi(\hat{\lambda}_B) = \xi(0.0334)$. Mais pour confirmer le modèle il faut faire des testes statistiques adéquates [13 ;20].

On choisit le test de Kolmogorov et Smirnov pour tester :

$$\begin{cases} H_0 : F(t) = F_0(t) & \forall t \in \mathbb{R} \\ H_1 : F(t) \neq F_0(t) \end{cases}$$

tel que :

$$F_0(t) = 1 - e^{-0.0334t}, \quad \forall t \in \mathbb{R}^+$$

Pour cela, on a trouvé les résultats suivantes :

$t_{(i)}$	1	3	4	6	10
F_0	0.0328	0.0953	0.1250	0.1815	0.2838
F_n	0 1/10	2/10	3/10	4/10	5/10
$t_{(i)}$	11	13	30	43	57
F_0	0.3074	0.3521	0.6327	0.7620	0.8509
F_n	6/10	7/10	8/10	9/10	1

La statistique du test est :

$$\begin{aligned} D_n &= sup_{1 \leq i \leq n} \left(\left| \frac{i-1}{n} - F_0(t_{(i)}) \right| ; \left| \frac{i}{n} - F_0(t_{(i)}) \right| \right) \\ &= 0.348 \end{aligned}$$

Pour un teste bilatéral, on trouve $d_{10,0.05} = 0.4092$ On a $D_{10} = 0.348 < d_{10,0.05} = 0.4092$ donc on accepte H_0 au niveau de signification 5%.
Alors on peut conclure que ce modèle représente les données au niveau de signification 5%.

3.3 Application sur les données de Freireich

Freireich, en 1963 [27], a fait un essai thérapeutique ayant pour but de comparer les durées de rémission, en semaines, de sujets atteints de leucémie selon qu'ils ont reçu ou non du 6 M-P. Les données sont comme suit :

Traitement	Durée de rémission, en semaines
6 M-P	$6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23,$ $25^+, 32^+, 32^+, 34^+, 35^+$

Les chiffres suivis du signe + correspondent à des patients qui ont été perdus de vue à la date considérée. Ils sont donc exclus "vivants"de l'étude et on sait donc seulement d'eux que leur "durée de survie"est supérieure à celle indiquée. Par exemple, le quatrième patient traité, par 6 M-P a eu une durée de rémission supérieure à 6 semaines.

Nous nous intéressons à l'estimation de la fonction de survie associée à ces données en utilisons les deux méthodes non paramétriques classique (de K-M) et bayésienne (en utilisant le processus de Dirichlet).

À partir de ces données, nous avons les informations suivantes :

t_i	6	7	9	10	11	13	16	17	19	20	22	23	25	32	34	35
R_i	21	17	16	15	13	12	11	10	9	8	7	6	5	4	2	1
M_i	3	1	0	1	0	1	1	0	0	0	1	1	0	0	0	0
c_i	1	0	1	1	1	0	0	1	1	1	0	0	1	2	1	1
δ_i	1	1	0	1	0	1	1	0	0	0	1	1	0	0	0	0

3.3.1 Estimateur de K-M

```
> summary(survfit(Surv(temps1,status1)~1,conf.type="plain"))
Call: survfit(formula = Surv(temps1, status1) ~ 1, conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.707        1.000
    7     17       1    0.807  0.0869        0.636        0.977
   10     15       1    0.753  0.0963        0.564        0.942
   13     12       1    0.690  0.1068        0.481        0.900
   16     11       1    0.627  0.1141        0.404        0.851
   22      7       1    0.538  0.1282        0.286        0.789
   23      6       1    0.448  0.1346        0.184        0.712
```

3.3.2 Estimateur avec le processus Dirichlet

Nous illustrerons l'estimateur bayésien utilisant le processus de Dirichlet, en considérant les données sur la durée de rémission pour les patients ayant reçu le médicament 6-MP, qui a été présenté précédemment.
Pour l'a priori de Dirichlet, nous utiliserons une estimation a priori $S_0(t)$ de

$$\frac{\alpha(t,\infty)}{\alpha(0,\infty)} = e^{-0.1t.}$$

Le choix de α est heuristique (voir [13] et[19]). Notre degré de croyance dans cette estimation a priori est considéré dans deux cas ; tel que $c = 5$ (La moitié des observations non censurées) et pour $c = 9$ (Le nombre des observations non censurées), pour que $\alpha(0,\infty) = 5$ et $\alpha(t,\infty) = 5e^{-0.1t}$ et pour le deuxième cas $\alpha(0,\infty) = 9$ et $\alpha(t,\infty) = 9e^{-0.1t}$. Les calculs ont été faites pour le premier cas, tandis que pour le deuxième cas il se fait de la même manière.

La figure 3.2 présente une simulation des trajectoires pour cette a priori dans le cas où $c = 5$.

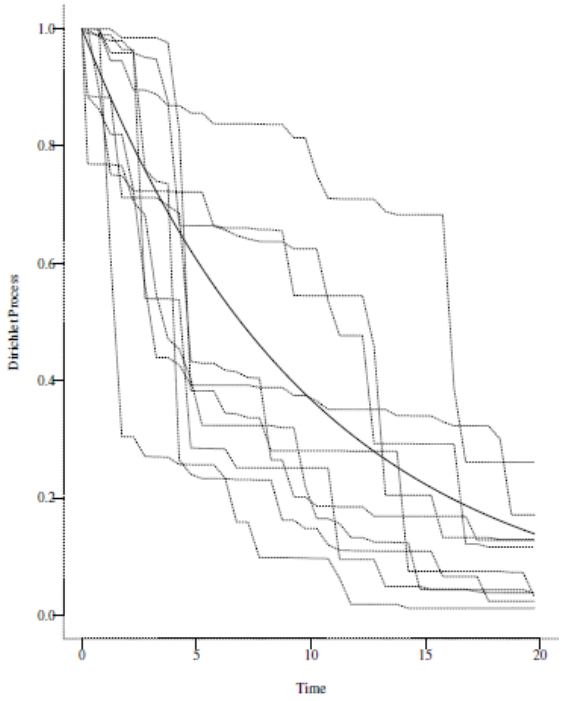

FIGURE 3.2: Simulation de 10 trajectoires (lignes pointillées) et leur moyenne (ligne continue) pour une a priori de Dirichlet avec $S_0(t) = e^{(-0.1t)}$ et $c = 5$

Pour illustrer les calculs, considérons d'abord un t dans l'intervalle $[0, 6)$. Pour l'estimateur du processus de Dirichlet de la fonction de survie on a :

$$\hat{S}_D(t) = \left[\frac{5e^{-0.1t} + 21}{5 + 21}\right] = \left[\frac{5e^{-0.1t} + 21}{26}\right]$$

Pour un t dans l'intervalle $[6, 7)$,

$$\hat{S}_D(t) = \left[\frac{5e^{-0.1t} + 17}{5 + 21}\right] \left\{\frac{5e^{-0.6} + 18}{5e^{-0.6} + 17}\right\}$$

Les calculs sont illustrés dans la table 3.1.

t dans	L'estimation de Dirichlet	L'estimation de KM
[0,6)	$\frac{5e^{-0.1t}+21}{5+21}$	0.857
[6,7)	$\frac{5e^{-0.1t}+17}{5+21}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}$	0.807
[7,10)	$\frac{5e^{-0.1t}+15}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}$	0.753
[10,13)	$\frac{5e^{-0.1t}+12}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}\frac{5e^{-1}+14}{5e^{-1}+12}$	0.690
[13,16)	$\frac{5e^{-0.1t}+11}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}\frac{5e^{-1}+14}{5e^{-1}+12}$	0.627
[16,22)	$\frac{5e^{-0.1t}+7}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}\frac{5e^{-1}+14}{5e^{-1}+12}\frac{5e^{-1.6}+10}{5e^{-1.6}+7}$	0.538
[22,23)	$\frac{5e^{-0.1t}+6}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}\frac{5e^{-1}+14}{5e^{-1}+12}\frac{5e^{-1.6}+10}{5e^{-1.6}+7}$	0.448
[23,35)	$\frac{5e^{-0.1t}+1}{26}\frac{5e^{-0.6}+18}{5e^{-0.6}+17}\frac{5e^{-0.7}+14}{5e^{-0.7}+13}\frac{5e^{-1}+14}{5e^{-1}+12}\frac{5e^{-1.6}+10}{5e^{-1.6}+7}\frac{5e^{-2.3}+5}{5e^{-2.3}+1}$	0.448

TABLE 3.1: L'estimation de Dirichlet et de Kaplan Meier de $S(t)$.

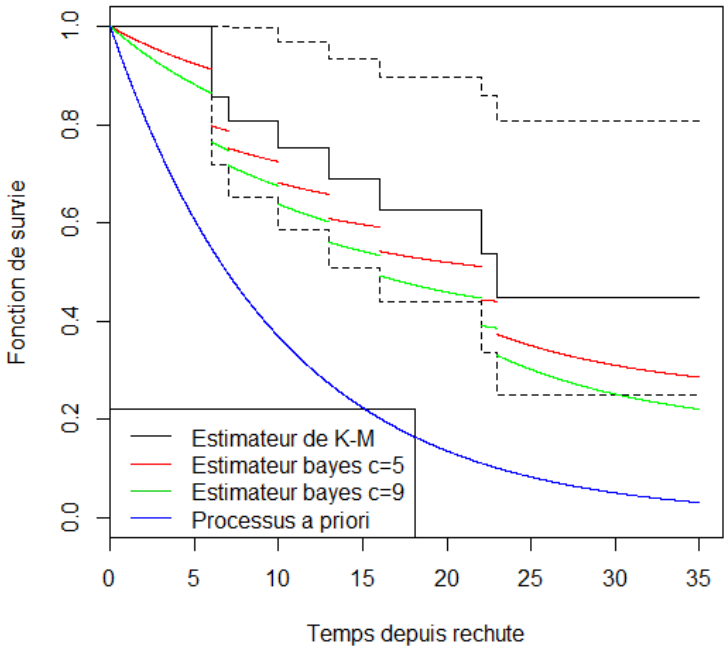

FIGURE 3.3: L' estimateur de Dirichlet, l'estimateur de Kaplan-Meier et l'a priori de la fonction de survie.

La figure(3.3) montre que l'estimateur bayésien se trouve dans l'intervalle de confiance de K-M, mais il est plus proche de l'a priori que ce dernier. De plus, les sauts de l'estimateur bayésien pour les données non censurées sont plus petits.

Finalement, on remarque que l'estimateur bayésien converge plus vite surtout lorsque $c \longrightarrow +\infty$ [19].

3.4 Conclusion

D'après cette application, on conclu que :

i. Pour calculer l'estimateur de K-M (1.4.2), il n'est pas nécessaire de connaître les observations censurées mais uniquement le nombre d'observations censurées entre deux observations non censurées, alors que l'on a besoin de toutes les observations (censurées et non censurées) pour l'estimateur Bayésien (2.4.3).

ii. Puisque l'estimateur de Bayes utilise toutes les données , il peut être préférable à l'estimateur de K-M.

iii. Si $\alpha(u, \infty)$ est continu en u, les données (δ, X) peuvent être récupérées exactement

en fonction de l'estimateur, ce qui n'est pas vrai pour l'estimateur K-M.

iv. L'estimateur de Bayes est une fonction de la statistique suffisante, c'est-à-dire les données (δ, X), alors que l'estimateur de K-M ne l'est pas.

v. L'estimation de K-M et l'estimation de Bayes sont toutes deux discontinues à des observations non censurées, à condition que l'on suppose que α est continue dans ce dernier cas.

vi. L'estimation de Bayes lisse l'estimateur à des observations censurées ce qui donne la continuité aux observations censurées à condition que α n'ait pas de masses ponctuelles.

Conclusion générale

La fonction de survie constitue une caractéristique importante des modèles de durée de vie, en particulier lorsque les données ne sont pas complètement observées. Nous nous intéressons aux données aléatoirement censurées à droite, elle est considérablement développée dans le domaine biomédical, pour les études de mortalité, l'effet de certains facteurs thérapeutiques ou pronostiques sur l'apparition d'un évènement au cours du temps (par exemple : le décès, la récidive d'une pathologie...). L'estimateur de K-M de la fonction de survie est un estimateur non paramétrique ; aucune hypothèse n'est faite sur la distribution des durées de survie. Les propriétés de cet estimateur ont suscité l'intérêt d'un grand nombre d'auteurs. Son comportement asymptotique en terme de convergence et de normalité asymptotique est encore d'actualité pour différents types de données.

L'analyse bayésienne constitue un autre développement important de ces 20 dernières années. L'utilisation de méthodes bayésiennes peut présenter certains avantages. En effet, ces méthodes utilisent des connaissances antérieures, exprimées sous forme de distribution de probabilité, afin de modifier une information nouvelle. L'utilisation de lois de distributions, et surtout de leurs propriétés, offre une certaine souplesse pour modéliser des problèmes complexes. Jusqu'à récemment, un obstacle à l'utilisation de ces méthodes venait du fait que, dans des problèmes complexes, la distribution a posteriori des paramètres étudiés est multidimensionnelle et possède rarement une écriture analytique simple.

Des études réalisées sur des données de patients atteints de VIH et de la leucémie, avec des méthodes paramétriques et d'autres non-paramétriques ont montré l'efficacité des méthodes bayésiennes. L'efficacité globale à été vérifiée dans le cadre de la théorie décisionnelle (pour les risques bayésiens), et le sujet reste ouvert pour d'autre étude.

Résumé

L'objectif de ce mémoire est d'étudier les méthodes d'estimation de la fonction de survie dans les essais cliniques. Nous comparons les deux approches : fréquentiste et bayésienne en donnant les points de différences et de ressemblances entre les deux approche. En plus, nous donnons une application du processus stochastiques dans l'inférence statistique.

Annexes

Annexe A

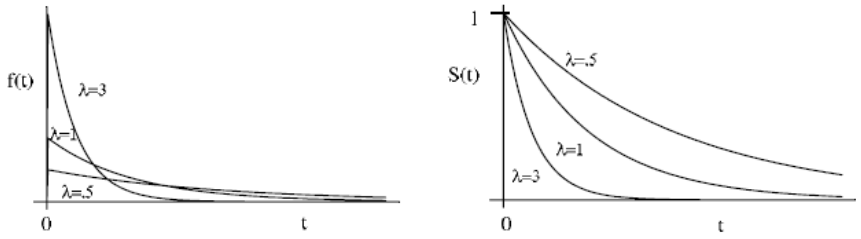

FIGURE 3.4: Densité exponentielle et courbes de survie.

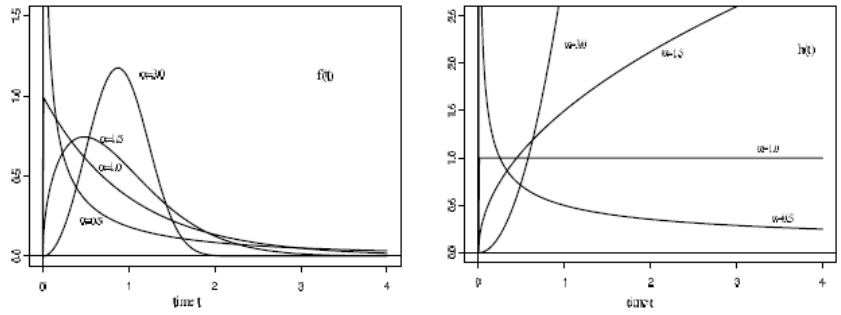

FIGURE 3.5: Weibull densité et la fonction de risque avec $\lambda = 1$.

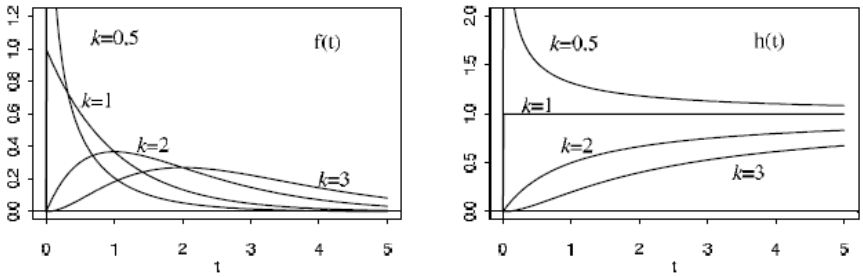

FIGURE 3.6: Les densités gamma et les risques avec $\lambda = 1$ et k = 0,5,1,2 et 3.

Annexe B

TEMPS	CENSURE	TEMPS	CENSURE	TEMPS	CENSURE
5	1	30	1	7	0
6	0	7	0	5	0
8	1	4	0	31	1
3	1	8	0	5	0
22	1	5	0	58	1
1	0	10	1	1	0
7	1	2	0	2	0
9	1	9	0	1	1
3	1	36	1	3	0
12	1	3	0	43	1
2	0	9	0	1	0
12	1	3	0	6	0
1	1	35	1	53	1
15	1	8	0	14	1
34	1	1	0	4	0
1	1	5	0	54	1
4	1	11	1	1	0
19	0	56	1	1	0
3	0	2	0	8	0
2	1	3	0	5	0
2	0	15	1	1	0
6	1	1	0	1	0
60	0	10	1	2	0
7	0	1	0	7	0
60	0	7	0	1	0
11	1	3	0	10	1
2	0	3	0	24	0
5	1	2	0	7	0
4	0	32	1	12	0
1	0	3	0	4	0
13	1	10	0	57	1
3	0	11	1	1	0
2	0	3	0	12	0
1	0				

TABLE 3.2: Les données d'infection au VIH

Bibliographie

[1] Aitchison, J ; Dunsmore, I.R.(1975).*Statistical Prediction Analysis* .New York : Compridge University Press. ISBN : 978-0-521-20692-1.

[2] Belkadi S. *Analyse par Bootstrap des données censurées*. Mémoire de Magister. Faculté de Mathématiques, Université des Sciences et de la Technologie Houari Boumediène.

[3] Brian S. E ; Andrw P. (1999). *Statistical Aspects Of The Design And Analysis Of Clinical Trials*. Imperial College Press London ; ISBN : 1-86094-153-2.

[4] Commenges D, et Jacqmin-Gadda H. (2015). *Modèles biostatistiques pour l'épidémiologie*. Louvain-la-Neuve ; ISBN : 978-2-8073-0026-2.

[5] Dellaportas, P ; Smith, A.F.M.(1993). *Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling.*
Applied Statistics 42,443-459.

[6] Denis M. (2010). *Méthodes de modélisation bayésienne et applications en recherche clinique*. Thèse de Doctorat, Université Montpellier 1, France.

[7] Ferguson, T. S.(1973). *A Bayesian Analysis of Some Nonparametric Problems*. Annals of Statistics 1 (1973) : 209–230.

[8] Ferguson, T. S ; Phadia, E. G.(1979). *Bayesian Nonparametric Estimation Based on Censored Data*. Annals of Statistics 7 (1979) : 163–186.

[9] Grieve, A.P.(1987). *Applications of Bayesian software : Tow examples.*
The Statistician 36,283-288.

[10] Hosmer.D.W, Lemeshow.S.(1999). *Applied Survival Analysis*. John Wiley and Sons, USA. ISBN : 0-471-15410-5.

[11] Ibrahim, J. G ; Chen, M. H ; Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag New York, Inc ; ISBN : 0-387-95277-2.

[12] Ibrahim.(2005).*Bayesian Survival Analysis* .Encyclopedia of Biostatistics. John Wiley et Sons, Ltd. DOI : 10.1002/0470011815.b2a11006.

[13] John, P.K, Moeschberger L.M. (2003). *Survival Analyses Techniques for Censored and Truncated Data*. Springer-Verlag New York ; ISBN : 0-387-95399-X.

[14] Khribi L. (2007). *L'échantillonnage de Gibbs pour l'estimation bayésienne dans l'analyse de survie*. Mémoire présenté comme exigence partielle de la maîtrise en mathématiques, Université du Québec, Montréal.

[15] Lee T.E ; Wang.W.J.(2003).*Statistical Methods for Survival Data Analysis* .Canada ; ISBN : 0-471-36997-7.

[16] Rabhi Y. (2006). *Modèles de survie avec un point de rupture*. Mémoire présenté comme exigence partielle de la maîtrise en mathématiques, Université du Québec, Montréal.

[17] Robert P.C. (2006). *Le choix bayésien : Principes et pratique*. Springer-Verlag France, Paris ; ISBN-10 : 2-287-25173-1.

[18] Samaniego F.J.(2010). *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer New York Dordrecht Heidelberg London. ISBN 978-1-4419-5940-9.

[19] Susarla, V ; Van Ryzin, J. (1976). *Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations*. Journal of the American Statistical Association 61 (1976) : 897–902.

[20] Tableman M ; Kim J.S. (2004). *Survival Analysis Using S*. Boca Raton London New York Washington, D.C ; ISBN : 0-203-59444-4.

[21] Touraine C. (2013). *Modèles illness-death pour données censurées par intervalle : Application à l'étude de la démence*. Thèse de Doctorat, Université Bordeaux 2, France.

[22] Venables.W.N ; Ripley B.D. (2002) *Modern Applied Statistics with S*, Springer, fourth edition.

[23] André M ; Eidelman A. *Statistique bayésienne-NOTES DE COURS*. www.crest.fr/ck finder/userfiles/files/Page paperso/MAndre/SB-cours.pdf.

[24] Colletaz G. *Modèles de survie. Notes de cours. Master 2 ESA*. (Novembre 2012). http ://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Survie-Sas.pdf.

[25] Diez M.D. *Survival Analyses in R* https ://www.openinto.org/download.php ?...Survival-analyses-in....

[26] Geffray S. *Analyse des durées de vie avec le logiciel R* . iml.univ-mvs.fr/ reboul/R-survie.pdf.

[27] Huber C. *Cours de Biostatistique er Modélisation* .
www.biomedical.parisdescartes.fr/survie/enseign/cours-stat-et-modèles.pdf.

[28] Huber C. *Modèles pour des durées de survie.*
www.biomedical.parisdescartes.fr/survie/enseign/survie-sansi.pdf.

[29] Planchet.F.(2017).*Modèles de durée*
www.ressources-actuarielles/.../.../0/.../FILE/Seance1.pdf?...

[30] Saint Pierre.P. (Octobre 2011). *Introduction à l'analyse des durées de survie.*
www.lsta.upmc.fr/psp/Cours-Survie-1.pdf.

[31] Samartzis L. (2005-2006). *Survival and censored data.*
http ://infoscience.ep.ch/record/112202/les/lafteris.pdf.

[32] Zaman Q ; Pfeiffer P.K. (2011) *Survival Analyses in Medical Research.*
interstat.stat journals.net/1105005.pdf.