الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Ref :……

Centre Universitaire de Mila

Institut des sciences et de la technologie **Département de Mathématiques et Informatique**

Introduction à l'analyse de survie

Mémoire préparé En vue de l'obtention du diplôme de Master en Mathématiques

Préparé par : **Benkouider Lemya**
Mokhnache Fatima Zohra

Encadré par : **Bououden Rabah**

Soutenue le .. juin 2013 devant le jury composé de :

M. Mehazem Allal	**C. U. Mila**	**président**
M. Elhadj Ali Thouria	**C. U. Mila**	**examinateur**
M. Bououden Rabah	**C. U. Mila**	**rapporteur**

Filière : **Mathématiques**

Spécialité : **Mathématiques Fondamentales et Appliquées**

Année universitaire : 2012/2013

Remerciements

...ӨƘƮǨǨdžǩƚ ǴɑǮūǲƛ Ǭɑơ ǻƶǩƚ Ż ƴǮƮǩƚ...

Nous tenons à remercier notre encadreur

Monsieur R. Bououden

pour l'aide, les conseils ont contribué à la réalisation de ce travail.

Nos très vifs remerciements et notre gratitude s'adressent évidement à tous nos enseignants qui ont contribué à notre formation

Nous exprimons notre très vive reconnaissance à nos familles et nos amies pour leurs soutien et encouragements.

Enfin, nous tenons à remercier toute personne qui nous a aidé de prés ou de loin afin de réaliser ce travail.

Lemya & Fatima Zohra

مدخل إلى دراسة مدة الحياة

ملخص:

تطرقنا في هذه المذكرة إلى دراسة مدة الحياة، حيث تناولنا في البداية بعض المفاهيم الأساسية التي يمكن الاعتماد عليها في هذه الدراسة.

كما تطرقنا أيضا إلى دراسة كل من: النموذج الغير وسيطي، النموذج النصف وسيطي و النموذج الوسيطي، مع إعطاء امثلة عن هذه النماذج.

Introduction à l’analyse de survie

Résumé :

Dans ce mémoire, nous avons étudié quelques notions sur l’analyse de survie. De plus nous étudions les modèles non paramétriques, les modèles semi-paramétriques et les modèles paramétriques avec quelques exemples.

Introduction to the survival analysis

Summary:

In our memory we studied some concepts of survival analysis.

At first we have dealt with some fundamental notions about survival analysis.

In addition we have spoken about non parametric models, semi-parametric models and parametric models with some examples.

Table des matières

Introduction Générale **3**

1 Notions préliminaires **6**
1.1 Définitions 6
1.2 Distributions de la durée de survie 8
1.2.1 Fonction de survie S 8
1.2.2 Fonction de répartition F 8
1.2.3 Densité de probabilité f 9
1.2.4 Risque instantané λ 9
1.2.5 Taux de hasard cumulé Λ 10
1.2.6 Quantités associées à la distribution de survie 10
1.3 Censure et troncature 11
1.3.1 Censure 12
1.3.2 Troncature 17
1.4 Fonction de vraisemblance 18

2 Estimation non paramétrique **20**
2.1 Estimateur de Kaplan-Meier de la survie 21
2.1.1 Estimateur de Kaplan-Meier 21
2.1.2 Estimation de la variance de $\hat{S}(t)$ 24
2.2 Estimateur de Nelson-Aalen du risque cumulé 25
2.2.1 Estimateur de Nelson-Aalen 25
2.2.2 Estimateur de la variance de $\hat{\Lambda}(t)$ 28
2.3 Autres estimateurs 28
2.3.1 Estimateur de Breslow du risque cumulé 28
2.3.2 Estimateur de Harrington et Fleming de la survie 29
2.3.3 Estimation de la survie par la méthode actuarielle 29

3 Modèles semi-paramétriques **32**
3.1 Les modèles à hasards proportionnels 32
3.2 Modèle de Cox 34
3.2.1 Vraisemblance partielle de Cox 34
3.2.2 Evénements simultanés 36
3.2.3 Estimation 38
3.2.4 Tests 41
3.2.5 Interprétation des coefficients de régression 44
3.2.6 Quelques extensions 45

4 Modèles paramétriques **54**
4.1 Risque instantané constant (loi exponentielle) 54
4.2 Risque instantané monotone 56
4.2.1 Loi de Weibull 56
4.2.2 Loi Gamma 58
4.2.3 Autres lois 60
4.3 Risque instantané en $\cap$ et $\cup$ 60
4.3.1 Lois log-normale 60
4.3.2 Lois de weibull généralisée 61
4.3.3 Autres lois 63
4.4 Introduction de covariables 63
4.4.1 Comparaison de deux groupes 64
4.4.2 Exemple 65
4.4.3 Modèles de vie accélérée 66

Bibliographie **68**

Introduction Générale

L'analyse des durées de survie constitue un domaine de la statistique qui s'intéresse à mesurer le temps jusqu'a événement particulier, souvent appelé temps d'échec, ou temps de survie. Les applications de telles analyses sont multiples, nous pouvons citer comme exemples de temps d'échec la durée de fonctionnement de pièces avant une défaillance en fiabilité industrielles, la durée de grèves ou de périodes de non-emploi en économie, la durée de vie de patients lors d'essais cliniques. Ce type d'analyse est particulièrement utile dans la recherche biomédicale. Des modèles peuvent être construits pour essayer de mieux comprendre le développement de certaines maladies. L'analyse de données de survie permet également d'évaluer l'efficacité de divers traitement ou la résistance du patient face à une maladie, en observant par exemple le temps entre le début d'un traitement et la guérison, ou le temps avant une rechute.

Historiquement, L'analyse des durées de survie voit le jour XVII^{e} siècle, dans le domaine de démographie. L'objectif des analystes de ce siècle est l'estimation, à partir des registres de décès, de diverses caractéristiques de la population son effectif, sa longévité, etc. Ces analyse, très générales, ne sont affinées qu'a partir du XIX^{e}

siècle, avec l'apparition de catégorisations suivant des «variables exogènes» (sexe, nationalité, ...). Durant ce siècle, apparaîssent également les premières modélisations concernant la probabilité de mourir à un certain âge, probabilité qui sera par la suite désignée sous le terme de «fonction de risque».

Enfin, l'analyse des durées de survie commence de déborder le cadre stricte de la démographie pour investir, au XX^{e} siècle, toutes les disciplines susceptibles d'avoir recours à de tels types de données : l'actuariat, la physique (avec l'apparition de la théorie de la fiabilité), l'industrie (pharmaceutique, biomédicale).

Jusqu'en 1950, la communauté des statisticiens s'intéresse peu à l'analyse des données de survie, la principale contribution étant celle de GreenWood (1926), qui propose une formule pour l'erreur standard d'une table de survie.

En 1951, Weibull conçoit un modèle paramétrique dans le domaine de la fiabilité ; à cet effet, il fournit une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de la survie : la «loi de Weibull».

En 1958, Kaplan et Meier présentent d'important résultats concernant l'estimation non paramétrique de la fonction de survie ; de l'estimateur résultant.

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

Ce mémoire constitué d'une introduction générale et de quatre chapitres. Le premier chapitre regroupe les notions de base de l'analyse de survie. Le deuxième chapitre se propose d'étudier l'estimation non paramétrique tel que l'estimateur de Kaplan-Meier, l'estimateur de Nelson-Aalen et d'autres estimateurs avec des exemples. Le troisième chapitre se propose d'étudier les modèles semi-paramétrique tel que le modèle de Cox avec un exemple et dans le dernier chapitre, nous étudions les modèles paramétriques et nous utilisons quelques lois pour étudier la durée de survie.

Chapitre 1

Notions préliminaires

1.1 Définitions

Quelques définitions sont couramment utilisées dans les études de survie :

courbe de survie : c'est une courbe qui donne une estimation de la proportion de sujets qui seront encore en vie ou n'auront pas présenté le phénomène étudié passé un certain délai après le début de l'observation du sujet.

L'estimation tient compte des sujets incomplètement suivis.

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

Date de point : c'est la date au-delà de laquelle on arrêtra l'étude et on ne tiendra plus compte des informations sur les sujets.

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

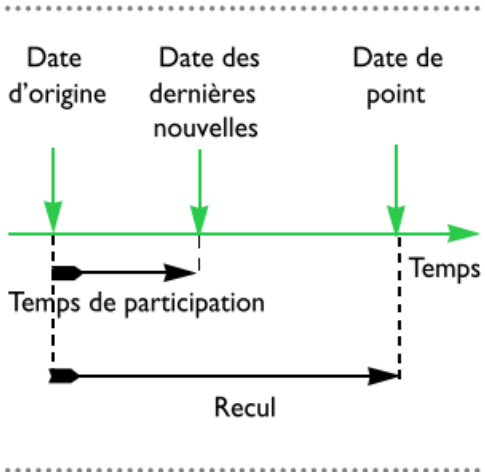
Le recul : c'est le délai qui sépare la date d'origine et la date de point. Ce délai situe le sujet dans le temps par rapport à la date de l'analyse. Les sujets qui ont un recul identique ont la même date d'origine.

Le temps de participation : c'est le temps écoulé entre la date d'origine et :

- La date de survenue de l'événement,

- ou la date des dernières nouvelles si le sujet est perdu de vue,

- ou la date de point si le sujet est présent pendant toute la durée de l'observation sans que l'événement se soit produit pendant cette période ; dans ce cas, les données sont dites censurées à droite.



1.2 Distributions de la durée de survie

Cinq fonctions équivalentes définissent la loi de la durée : Supposons que la durée de survie X soit une variable positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des fonctions suivantes :

1.2.1 Fonction de survie S

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

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

1.2.2 Fonction de répartition F

La fonction de répartition représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à-dire

$$F(t) = P(X \leq t) = 1 - S(t).$$

Remarque 1 *Il est arbitraire de décider que*

$$S(t) = P(X \geq t) \text{ ou } S(t) = P(X > t).$$

Cela n'a aucune importance quand la loi de X est continue car $P(X > t) = P(X \geq t)$. Dans les cas où F a des sauts (quand le temps est discret, par exemple, compté en mois ou semaine), on utilise les notations suivantes :

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

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

1.2.3 Densité de probabilité f

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

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

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

$$f(t) = \lim_{h \to 0} \frac{P(t \leq X < t + h)}{h} = F^{'}(t) = -S'(t).$$

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

1.2.4 Risque instantané λ

Le risque instantané (ou taux de hasard), pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq X < t + h \mid X \geq t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))^{'}.$$

$\lambda(t)$ est aussi appelé la fonction de mortalité.

1.2.5 Taux de hasard cumulé Λ

Le taux de hasard cumulé est l'intégrale du risque instantané λ :

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

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

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)du\right).$$

On en déduit que

$$f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)du\right).$$

1.2.6 Quantités associées à la distribution de survie

Moyenne et variance de la durée de survie

Le temps moyen de survie $E(X)$ et la variance de la durée de survie $V(X)$ sont définis par les quantités suivantes :

$$E(X) = \int_0^\infty S(t)dt,$$

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

Ainsi on peut déduire l'espérance et la variance à partir de n'importe laquelle des fonctions F, S, f , λ, Λ (mais pas l'inverse).

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.

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

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

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.

1.3 Censure et troncature

Une des caractéristiques des données de survie est l'existence d'observations incomplètes.

En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accés à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

1.3.1 Censure

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.

Dï£¡finition 2 *(censure)*

La durée T est dite censurée si :

- le patient est toujours vivant à la fin de l'étude (exclus-vivants).

- le patient est perdu de vue : il a quitté l'étude avant qu'on ait pu observer l'événement d'intérêt.

Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

- **Censure de type 1** : fixée

Au lieu d'observer les variables X_1, ..., X_n qui nous intéressent, on n'observe X_i que lorsque X_i est inférieur ou égal à une durée fixée C, $X_i \leq C$, sinon on sait seulement que X_i est supérieur à C. On note aussi $T_i = X_i \wedge C$. (le signe $\wedge$ signifie : $a \wedge b = \min(a, b)$, la plus petite des deux valeurs a et b).

- **Censure de type 2** : attente

On décide d'observer les durées de survie des 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 petit, $X_{(i)}$ la ième etc... :

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

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 :

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

- **Censure de type 3** : aléatoire

Soient $C_1, ..., C_n$ des variables aléatoires indépendantes et identiquement distribuées. On observe les variables

$$T_i = X_i \wedge C_i.$$

L'information disponible peut être résumée par :

- la durée réellement observée T_i,

- un indicateur $\delta_i = 1_{\{X_i \leq C_i\}}$

- $\delta_i = 1$ si l'événement est observé (d'où $T_i = X_i$). On observe les "vraies" durées ou les durées complètes.

- $\delta_i = 0$ si l'individu est censuré (d'où $T_i = C_i$). On observe des durées incomplètes (censurées).

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

- la perte de vue : le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients "perdus de vue".
- L'arrêt ou le changement du traitement : les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.
- La fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients "exclus-vivants".

Les "perdus de vue" (et les exclusions) et les "exclus-vivants" correspondent à des observations censurées mais les deux mécanismes sont de nature différente (la censure peut être informative chez les "perdus de vue").

Exemple 3

On s'intéresse au temps de survie de personnes atteints d'une maladie. On fixe le temps s'étude et à la fin de ce temps certaine personnes sont encore vivantes. Pour ces personnes tout ce que l'on sait est que leur temps de survie dépasse le temps observé, ce sont des données censurées à droite de type 1.

Censure à gauche

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

$$\begin{aligned} T &= X \vee C = \max(X, C), \\ \delta &= 1_{\{X \geq C\}}. \end{aligned}$$

Comme pour la censure à droite, on suppose que la censure C est indépendante X.

Exemple 4

Sur le même exemple précédent, on ne peut pas toujours savoir le moment exact du déclenchement de la maladie. Ainsi si on s'intéresse à l'âge du début de la maladie, pour certaines personnes, on sait seulement que leur âge est inférieur à leur âge au moment de l'étude. Ces données sont censurées à gauche.

Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues. Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (par en continu), on sait alors uniquement que l'événement

s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à de la censure à droite.

1.3.2 Troncature

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

La troncature à gauche

Soit Z une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X n'est observable que si $X > Z$. On observe le couple (X, Z), avec $X > Z$. Par exemple, si la durée de vie d'une population est étudiée à partir d'une cohorte tirée au sort dans cette population, seule la survie des sujets vivants à l'inclusion pourra être étudiée (il y a troncature à gauche car seuls les sujets ayant survécu jusqu'à la date d'inclusion dans la cohorte sont observables).

La troncature à droite

De même, il y a troncature à droite lorsque X n'est observable que si $X < Z$.

La troncature par intervalle

Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle. Par exemple, on rencontre ce type de troncature lors de l'étude des patients d'un régistre : les patients diagnostiqués avant la mise en place du registre ou répertoriés après la consultation du registre ne seront pas inclus dans l'étude.

1.4 Fonction de vraisemblance

Considérons le cas d'une censure aléatoire droite C indépendante de la durée d'intérêt X. Supposons que les variables X et C ont pour densités respectives f et g et pour survies S et G. La distribution de X est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (T_i, δ_i), où $T_i = \min(X_i, C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = 1_{\{X_i \leq C_i\}}$. Ainsi, la contribution à la vraisemblance pour l'individu i est

$$\begin{aligned} L_i &= P(T_i \in [t_i, t_i + dt],\ \delta_i = 1 \mid \theta)^{\delta_i} \times P(T_i \in [t_i, t_i + dt],\ \delta_i = 0 \mid \theta)^{1-\delta_i} \\ &= P(X_i \in [t_i, t_i + dt],\ C_i \geq X_i \mid \theta)^{\delta_i} \times P(C_i \in [t_i, t_i + dt],\ C_i < X_i \mid \theta)^{1-\delta_i} \\ &= [f(t_i \mid \theta) G(t_i^-)]^{\delta_i} \times [g(t_i) S(t_i \mid \theta)]^{1-\delta_i}. \end{aligned}$$

Par l'hypothèse (de censure non informative), le paramètre d'intérêt θ n'apparaît

pas dans la loi de la censure. La partie utile de la vraisemblance se réduit alors à

$$L = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i} S(t_i \mid \theta)^{1-\delta_i}.$$

Chapitre 2

Estimation non paramétrique

Dans ce chapitre nous nous placerons dans le cadre le plus fréquent d'une censure à droite aléatoire de type 1. Si aucun modèle n'est supposé, les principaux estimateurs sont :

- l'estimateur de Kaplan-Meier de la fonction de survie,

- l'estimateur de Nelson-Aalen du risque cumulé.

2.1 Estimateur de Kaplan-Meier de la survie

2.1.1 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t, c'est-à-dire, si $t'' < t' < t$

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

En considérant les temps d'événements (décès et censure) distincts$T_{(i)}$,

tel que $T_{(i)}$ $(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$P(X > T_{(j)}) = \prod_{k=1}^{j} P(X > T_{(k)} \mid X > T_{(k-1)}),$$

avec $T_{(0)} = 0$. Considérons les notations suivantes :

- Y_i le nombre d'individus à risque de subir l'événement juste avant le temps $T_{(i)}$,

- d_i le nombre de décès en $T_{(i)}$.

Alors la probabilité p_i de mourir dans l'intervalle $\left]T_{(i-1)}, T_{(i)}\right]$ sachant que l'on était vivant en $T_{(i-1)}$, *i.e.* $p_i = P(X \leq T_{(i)} \mid X > T_{(i-1)})$, peut être estimée par

$$\hat{p}_i = \frac{d_i}{Y_i}.$$

Comme les temps d'événements sont supposés distincts, on a

$d_i = 0$ en cas de censure en $T_{(i)}$, *i.e.* quand $\delta_i = 0$,

$d_i = 1$ en cas de décès en $T_{(i)}$, *i.e.* quand $\delta_i = 1$.

On obtient alors l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{\substack{i=1,\dots,n \\ T_{(i)} \leq t}} \left(1 - \frac{\delta_i}{Y_i}\right) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{\delta_i}{n-(i-1)}\right) = \prod_{i:T_{(i)} \leq t} \left(\frac{n-i}{n-i+1)}\right)^{\delta_i}.$$

L'estimateur $\hat{S}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance. $\hat{S}(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car le temps de décès ne sont pas connus).

Remarque 5 *Dans le cas où il y a des ex-aequo :*

- si ce sont des événements de nature différente, on considère que les observations non censurées ont lieu avant les censurées,

- s'il y a plusieurs décès au même temps $T_{(i)}$, alors $d_i > 1$ et on a

$$\hat{S}(t) = \prod_{\substack{i=1,\dots,n \\ T_{(i)} \leq t}} \left(1 - \frac{d_i}{Y_i}\right).$$

Remarque 6 *Estimation empirique :*

Pour un échantillon i.i.d. de durées non censurées $(X_i)_{i=1,\dots,n}$, un estimateur "naturel" de la survie de la variable X est la survie empirique

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i > x\}}.$$

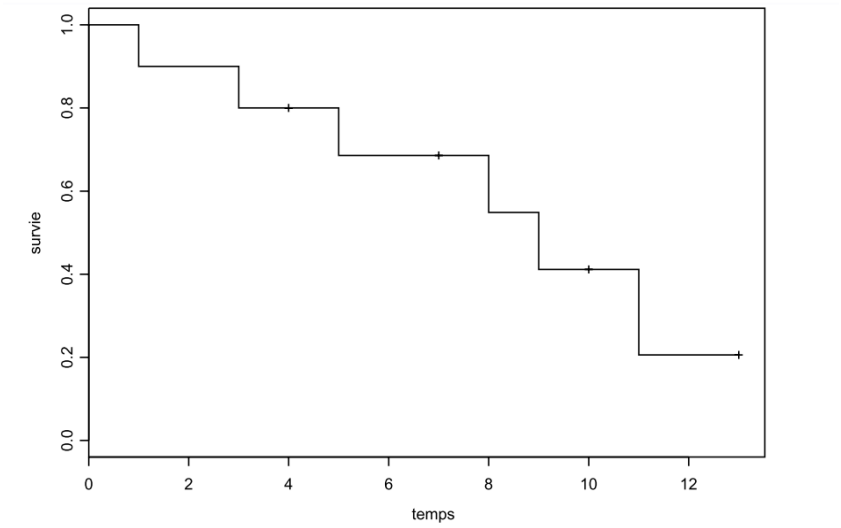
Exemple 7 *Cancer des bronches*

Sur 10 *patients atteints de cancer des bronches on a observé les durées de survie suivantes, exprimées en mois :*

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

L'estimateur de Kaplan-Meier de la fonction de survie $S(t)$ *se calcule de la manière suivante :*

temps	Y_i	d_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	$[11, \infty[$



Estimateur de kaplan-Meier de la fonction de survie pour le

cancer des bronches

2.1.2 Estimation de la variance de $\hat{S}(t)$

L'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier est

$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:T_{(i)}\leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

Il est obtenu en utilisant l'approximation suivante,

$$\widehat{Var}\left(\log\left(\hat{S}(t)\right)\right) \approx \sum_{i:T_{(i)}\leq t} \frac{d_i}{Y_i(Y_i - d_i)},$$

et on appliquant la delta-méthode ($Var(f(Z)) \approx [f'(E(Z))]^2 var(Z)$) pour montrer que

$$\widehat{Var}\left(\log\left(\hat{S}(t)\right)\right) \approx \frac{1}{\hat{S}(t)^2} Var\left(\hat{S}(t)\right).$$

Remarque 8 *Ce résultat s'obtient, de manière théorique, de la propriété de normalité asymptotique de l'estimateur de Kaplan-Meier.*

2.2 Estimateur de Nelson-Aalen du risque cumulé

2.2.1 Estimateur de Nelson-Aalen

Si la variable X admet une densité, on a par définition du risque cumulé, du risque instantané et de la densité

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

Dans le cas où X n'admet pas de dérivée en tout point de $\mathbb{R}^+$, on peut toujours définir le risque cumulé en utilisant la définition de la densité de X,

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

Considérons les quantités $H(t) = P(T > t)$ et $H_1(t) = P(T > t, \delta = 1)$ et introduisons $G(t)$ la fonction de survie de la variable C. D'après l'hypothèse d'indépendance, on obtient les égalités suivantes

$$\begin{aligned} H(t) &= P(T > t) = P(X > t, C > t) = S(t)G(t) \\ H_1(t) &= P(T > t,\ \delta = 1) = P(X > t,\ C \geq X) = E(1_{\{X>t\}}G(X^-)) \\ &= \int_t^\infty G(u^-)f(u)du = -\int_t^\infty G(u^-)S(du). \end{aligned}$$

Par conséquent, $H_1(dt) = G(t^-)S(dt)$ et on obtient l'expression suivante pour le risque cumulé :

$$\Lambda(t) = -\int_0^t \frac{H_1(du)}{H(u^-)}.$$

Un estimateur "naturel" s'obtient en remplaçant les fonctions H et H_1 par leurs équivalentes empiriques (calculables car les variables T et δ sont observées). Soient

$$\hat{H}(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i>u\}} \text{ et } \hat{H}_1(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i>u,\delta_i=1\}} = 1 - \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i\leq u,\delta_i=1\}},$$

l'estimateur de Nelson-Aalen est donné par les expressions suivantes

$$\hat{\Lambda}(t) = -\int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u^-)} = \sum_{i:T_i\leq t} \frac{\sum_{j=1}^{n} 1_{\{T_j=T_i,\delta_j=1\}}}{\sum_{j=1}^{n} 1_{\{T_j\geq T_i\}}} = \sum_{i:T_i\leq t} \frac{d_i}{Y_i},$$

où Y_i représente le nombre d'individus à risque juste avant T_i et d_i représente le nombre de décès en T_i. L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut de taille d_i/Y_i à chaque instant de décès.

Exemple 9 *Données de Nelson et Aalen*

Il s'agit de la durée de vie ventilateurs, en nombre de milliers d'heure de fonctionnement. La question qui se posait était de savoir si la fonction de risque λ était décroissante dans le temps. Les durées sont en milliers d'heures.

Durées :

04.5	04.6	11.5	11.5	15.6	16.0	16.6	18.5	18.5	18.5	18.5	18.5
20.3	20.3	20.3	20.7	20.7	20.8	22.0	30.0	30.3	30.0	30.0	31.0
32.0	34.5	37.5	37.5	41.5	41.5	41.5	41.5	43.0	43.0	43.0	43.0

46.0 48.5 48.5 48.5 48.5 50.0 50.0 50.0 61.0 61.0 61.0 61.0

63.0 64.5 64.5 67.0 74.5 78.0 78.0 81.0 81.0 82.0 85.0 85.0

85.0 87.5 87.5 87.5 94.0 99.0 101.0 101.0 101.0 115.0.

Censures :

1 0 1 1 0 1 0 0 0 0 0 0

0 0 0 1 1 1 0 0 0 0 0 0

1 0 0 0 0 0 0 0 0 1 0 0

0 0 0 0 0 0 0 0 0 0 0 0

0 0 1 0 0 0 0 0 0 0

Si on appelle t_1 *le premier instant de "mort" (ici : panne),* t_2 *le second, etc..., on calcule* $\hat{H}(t)$*, pour* t *supérieur ou égal à la plus grande valeur observée, qui est de* 87 500 *heures, comme*

$$\begin{aligned}
\hat{H}(t) &= \frac{\textit{Nombres de pannes en } t_1}{\textit{Nombres de ventilateurs à risque en } t_1} \\
&\quad + \frac{\textit{Nombres de pannes en } t_2}{\textit{Nombres de ventilateurs à risque en } t_2} \\
&\quad + \textit{ etc...} \\
&= \frac{\textit{Nombres de pannes en } 4.5}{\textit{Nombres de ventilateurs à risque en } 4.5} \\
&\quad + ... + \frac{\textit{Nombres de pannes en } 87.5}{\textit{Nombres de ventilateurs à risque en } 87.5} \\
&= \frac{1}{70} + \frac{2}{68} + ... + \frac{1}{8} \\
&= 0.3368
\end{aligned}$$

2.2.2 Estimateur de la variance de $\hat{\Lambda}(t)$

En utilisant la théorie des processus de comptage et en faisant une approximation par une loi de poisson, on montre que la variance de l'estimateur de Nelson-Aalen est,

$$\widehat{Var}(\hat{\Lambda}(t)) = \sum_{i:T_i \leq t}^{n} \frac{d_i}{Y_i^2},$$

où d_i et Y_i sont le nombre de décès et d'individus à risque en T_i.

2.3 Autres estimateurs

2.3.1 Estimateur de Breslow du risque cumulé

Un estimateur du risque cumulé peut également être obtenu à partir de l'estimateur de Kaplan-Meier en utilisant la relation $\Lambda(t) = -\log(S(t))$:

$$\begin{aligned} \hat{\Lambda}_2(t) &= -\log\left(\hat{S}(t)\right) \\ &= -\sum_{i:T_i \leq t} \log\left(1 - \frac{d_i}{Y_i}\right). \end{aligned}$$

La variance de cet estimateur est donnée par

$$\widehat{Var}(\hat{\Lambda}_2(t)) = \sum_{i:T_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

2.3.2 Estimateur de Harrington et Fleming de la survie

A partir de la relation $S(t) = \exp(-\Lambda(t))$ et de l'estimateur de Nelson-Aalen, on peut en déduire un autre estimateur de la fonction de survie :

$$\begin{aligned}
\hat{S}_2(t) &= \exp(-\hat{\Lambda}(t)) \\
&= \prod_{i:T_i \leq t}^{n} e^{-\frac{d_i}{Y_i}} \\
&\approx \prod_{i:T_i \leq t}^{n} (1 - \frac{d_i}{Y_i}), \qquad \text{si } \frac{d_i}{Y_i} \to 0,
\end{aligned}$$

où d_i et Y_i sont le nombre de décès et d'individus à risque en T_i. En appliquant un développement limité, on retrouve l'estimateur de Kaplan-Meier. En utilisant la delta-méthode ($Var(f(Z)) \approx [f'(E(Z))]^2 Var(Z)$), on peut obtenir un estimateur de la variance de cet estimateur,

$$\begin{aligned}
\widehat{Var}(\hat{S}_2(t)) &= (\hat{S}_2(t))^2 \widehat{Var}(\hat{\Lambda}(t)) \\
&= \exp\left(-2 \sum_{i:T_i \leq t}^{n} \frac{d_i}{Y_i}\right) \times \left(\sum_{i:T_i \leq t}^{n} \frac{d_i}{Y_i^2}\right).
\end{aligned}$$

2.3.3 Estimation de la survie par la méthode actuarielle

La méthode actuarielle repose sur le même principe de construction que l'estimateur de Kaplan-Meier. La différance est que la probabilités conditionnelles sont estimées sur des intervalles fixés par l'utilisateur et non déterminés par le temps d'événements. Ces intervalles sont généralement de longueur égale, par exemple, un mois, un trimestre, une année.

Considérons, k intervalles de temps $[0, t_1[$, $[t_1, t_2[$, ..., $[t_{k-1}, \infty[$, fixés a priori.

Définissons,

- d_i le nombre de décès dans le $i^{ème}$ intervalle $[t_{i-1}, t_i[$ (avec $t_0 = 0$ et $t_k = \infty$),

- n_{i-1} le nombre de sujets vivants au temps t_{i-1},

- c_i le nombre de sujets censurés dans l'intervalle $[t_{i-1}, t_i[$,

- r_i le nombre de sujets à risque dans l'intervalle $[t_{i-1}, t_i[$,

Afin de simplifier les calculs, on suppose généralement que les censures sont réparties uniformément dans l'intervalle, c'est-à-dire, que les sujets censurés sont exposés en moyenne un demi-intervalle. Dans le calcul des individus à risque, leur contribution pour l'intervalle $[t_{i-1}, t_i[$ est donc $c_i/2$. Le nombre d'individus à risque pour l'intervalle $[t_{i-1}, t_i[$ est donc

$$r_i = n_{i-1} - \frac{c_i}{2}.$$

Alors la probabilité $p_i = P(X \leq t_i \mid X > t_{i-1})$ de mourir dans l'intervalle $[t_{i-1}, t_i[$ sachant que l'on était vivant en t_{i-1} est estimée par

$$\hat{p}_i = \frac{d_i}{r_i}.$$

L'estimateur de la fonction de survie est donc,

$$\hat{S}_3(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{r_i}\right).$$

La formule de Greenwood permet d'obtenir une estimation de la variance,

$$\widehat{Var}\left(\hat{S}_3(t)\right) = \hat{S}_3(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}.$$

Comparaison Kaplan-Meier/actuarielle

	Kaplan-Meier	Actuarielle
Intervalle de temps	Variable ; déterminé par événements	Fixe, déterminé a priori
Estimation $S(t)$	À chaque événement	Par interpolation linéaire
Courbe	Marches d'escalier	segment de droite reliant bornes des intervalles.
Utilisation	Courante	En fréquente

Chapitre 3

Modèles semi-paramétriques

Le modèle de Cox est largement utilisé en analyse de données de survie. Celui-ci permet de modéliser des temps de survie avec des données censurées. Elle est très utilisée dans le domaine médical (temps de survie ou de guérison d'un patient).

Le principe du modèle de Cox est de relier la date d'arrivée d'un événement à des variables explicatives. Par exemple dans le domaine médical, on cherche à évaluer l'impact d'un prétraitement sur le temps de guérison d'un patient.

3.1 Les modèles à hasards proportionnels

Ces modèles expriment un effet multiplicatif des diverses covariables sur la fonction de hasard (modèle à structure multiplicative). On introduit une fonction de hasard de base qui donne la forme générale du hasard et qui est commune à tous les individus. Les modèles à hasards proportionnels se caractérisent par la relation suivante, pour

tout $t > 0$,

$$\lambda(t \mid Z) = \lambda_0(t)h(\beta, Z),$$

où Z est un vecteur de covariables, β le paramètre d'intérêt et h une fonction positive. La fonction de hasard est le produit d'une fonction qui ne dépend que du temps et d'une fonction qui n'en dépend pas. En général, on suppose que l'effet des covariables se résume à une quantité réelle $\beta' Z$, c'est-à-dire $\lambda(t \mid Z) = \lambda_0(t)h(\beta' Z)$.

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

$$\frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)} = \frac{h(\beta' Z_i)}{h(\beta' Z_j)}.$$

Les fonctions de hasard sont donc proportionnelles. C'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier. Le rapport des fonctions de hasard est par définition un risque relatif à l'instant t des sujets de caractéristiques Z_i par rapport aux sujets de caractéristiques Z_j.

Un cas particulier très important est le modèle de Cox, qui suppose que la fonction h est la fonction exponentielle, c'est-à-dire,

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z).$$

D'autres choix de fonctions h sont possibles, néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et $\exp(0) = 1$.

Remarque 10 *Si λ_0 et/ou h ont une forme inconnue, le modèle est dit semi-paramétrique.*

3.2 Modèle de Cox

On se place dans le cadre du modèle de Cox,

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z),$$

où Z est un vecteur de covariables de dimension $p \times 1$ et β un vecteur $(p \times 1)$ de coefficient de régression.

Considérons,

- D le nombre de décès observés parmi les n sujets à l'étude,

- $T_1 < T_2 < ... < T_D$, les temps d'événements (décès) distincts,

- $(1), (2), ..., (D)$, les indices des individus décédés respectivement en $T_1, T_2, ..., T_D$,

- Z_i la valeur des covariables de l'individu i,

- $R(T_i)$ l'ensemble des individus encore à risque à T_i^- (juste avant T_i).

3.2.1 Vraisemblance partielle de Cox

Le principe de la méthode est d'estimer uniquement le coefficient de régression β en considérant la fonction λ_0 comme un paramètre de nuisance. Par conséquent, on ne cherche pas à estimer λ_0. L'idée de Cox est qu'aucune information ne peut être donnée sur β par les intervalles pendant lesquels aucun événement n'a eu lieu, car on peut concevoir que λ_0 soit nulle dans ces intervalles (On suppose que les moments où

se produisent les censures n'apportent peu ou pas d'information sur β). On travaille alors conditionnellement à l'ensemble des instants où un décès a lieu.

Supposons, dans un premier temps, qu'il n'y qu'un seul décès à chaque temps d'événement (car le raisonnement provient du cas continu). La probabilité qu'il y ait un événement (décès) en T_i (dans l'intervalle $[T_i, T_i + \Delta t[$) est :

$$\sum_{j \in R(T_i)} \lambda_0(T_i) \exp(\beta' Z_j).$$

La probabilité que l'individu i subisse l'événement en T_i sachant qu'un événement a eu lieu en T_i vaut

$$\frac{\lambda_0(T_i) \exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \lambda_0(T_i) \exp(\beta' Z_j)} = \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)}.$$

Le point important est que cette probabilité dépend uniquement du paramètre β.

Comme il y a des contributions à la vraisemblance à chaque temps de décès, la vraisemblance partielle de Cox est définie comme le produit sur les temps de décès. La vraisemblance (partielle) totale est donc

$$L_{Cox}(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)}.$$

La vraisemblance partielle ne dépend pas de la fonction de hasard de base $\lambda_0(t)$. On peut donc estimer β, sans connaître la fonction de hasard de base, par maximisation de la vraisemblance partielle de Cox.

3.2.2 Evénements simultanés

Le raisonnement précédent suppose des temps d'événements distincts. Dans le cas des données réelles, cette hypothèse n'est pas toujours vérifiée (ex : mesures tous les mois ou trimestres).

La probabilité que l'individu j décède en T_i est

$$p_j = \frac{\exp(\beta' Z_j)}{\sum_{k \in R(T_i)} \exp(\beta' Z_k)}.$$

En présence de plusieurs événements, la méthode "exacte" consiste à admettre que les événements se produisent les uns à la suite des autres. Cependant, on ne connaît pas l'ordre des événements, il faut donc considérer toutes les possibilités. Dans le cas de deux sujets s_1 et s_2 de caractéristiques Z_1 et Z_2 qui décèdent en T_i, la contribution exacte à la vraisemblance est

$$\frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j) \times \sum_{j \in R(T_i)\setminus s_1} \exp(\beta' Z_j)} + \frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j) \times \sum_{j \in R(T_i)\setminus s_2} \exp(\beta' Z_j)}.$$

Le problème de cette méthode est que le temps de calcul devient très long quand il y a beaucoup d'événements simultanés. Ainsi, on utilise le plus souvent l'approximation de Breslow qui consiste à supposer que la contribution des d_i événements en T_i est le produit des probabilités p_j pour les unités décédées en T_i

$$\left(i.e. \sum_{j \in R(T_i)} \exp(\beta' Z_j) \approx \sum_{j \in R(T_i) \setminus k} \exp(\beta' Z_j) \right),$$

$$L_B(T_i) = \prod_{\substack{j\ :\ unités \\ décédées\ en\ T_i}} p_j = \frac{\exp\left(\beta' \left(\sum_{\substack{j\ :\ unités \\ décédées\ en\ T_i}} Z_j \right) \right)}{\left(\sum_{k \in R(T_i)} \exp(\beta' Z_k) \right)^{d_i}}.$$

L'approximation de Breslow de la vraisemblance totale est

$$\prod_{i=1}^{D} L_B(T_i),$$

où D est le nombre de décès observés. La maximisation de cette vraisemblance est rapide.

De plus, si le nombre d'événements simultanés n'est pas trop grand alors la méthode est assez précise.

3.2.3 Estimation

Estimation des coefficients de régression β

A partir de la vraisemblance partielle, on peut obtenir une estimation du vecteur de paramètre β de dimension $p \times 1$. Notons

$$\mathcal{L}(\beta) = \log(L_{Cox}(\beta)) = \sum_{i=1}^{D} \left[\beta' Z_{(i)} - \log \left(\sum_{j \in R(T_i)} \exp(\beta' Z_j) \right) \right],$$

et $U(\beta)$ la fonction score, c'est-à-dire le vecteur $p \times 1$ des dérivées premières de $\mathcal{L}(\beta)$,

$$\begin{aligned} U(\beta) &= \frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta_1}, ..., \frac{\partial \mathcal{L}(\beta)}{\partial \beta_p}\right) \\ &= \sum_{i=1}^{D} \left[Z_{(i)} - \frac{\sum_{j \in R(T_i)} Z_j \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right] \\ &= \left(\sum_{i=1}^{D} \left[Z_{(i),1} - \frac{\sum_{j \in R(T_i)} Z_{j,1} \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right], ..., \sum_{i=1}^{D} \left[Z_{(i),p} - \frac{\sum_{j \in R(T_i)} Z_{j,p} \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right] \right). \end{aligned}$$

L'estimateur de Cox $\hat{\beta}$ des coefficients de régression est solution de l'équation

$$U(\beta) = 0.$$

Il n'y a pas de solution exacte à ce problème. L'algorithme de Newton-Raphson est souvent utilisé par les logiciels pour obtenir une solution.

Un estimateur consistant de la matrice de variance-covariance de β peut se calculer à partir de l'inverse de la matrice d'information de Fisher,

$$\widehat{Var}(\hat{\beta}) = \left\{ I(\hat{\beta}) \right\}^{-1}$$

où le terme (i, j) de la matrice $I(\beta)$ est

$$[I(\beta)]_{i,j} = -\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_i \partial \beta_j}.$$

Estimation du risque cumulé de base Λ_0

Pour déduire l'estimateur du taux de hasard de base λ_0 il est nécessaire d'estimer β, puis on maximise la vraisemblance totale par rapport à λ_0. Pour construire la vraisemblance dans le cas des données censurées nous avons besoin de considérer quelle information nous donne chaque observation.

1) une observation complète donne l'information sur la probabilité que l'événement a eu lieu à ce temps, qui est approximativement égal à la fonction de la densité prise en ce temps.

2) Pour une observation censurée à droite, tout ce que nous savons c'est que le temps de l'événement est plus grand que ce temps. Donc l'information est la fonction de survie évaluée au temps de l'étude.

Construisons maintenant la vraisemblance totale pour le modèle de hasard proportionnel dans le cas où il n'y a pas d'ex-aequo et les covariables $(Z_i)_{i=1,2,...,n}$ sont constantes. Soit $X = (X_i, \delta_i, Z_i)_{1 \leq i \leq n}$ supposons qu'il y ait L morts au temps

$$t_{(1)} < t_{(2)} < ... < t_{(L)}$$

Donc la vraisemblance de notre modèle est donnée par

$$\begin{aligned} L(\beta, \lambda_0) &= \prod_{i=1}^{n} (f(t_i \mid Z_i))^{\delta_i} S(t_i \mid Z_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} (\lambda(t_i \mid Z_i))^{\delta_i} \exp\{-\Lambda(t_i \mid Z_i)\} \\ &= \prod_{i=1}^{n} (\lambda_0(t_i))^{\delta_i} (\exp \beta' Z_i)^{\delta_i} \exp\{-\Lambda_0(t_i)(\exp \beta' Z_i)\} \\ &= \left[\prod_{i=1}^{L} \lambda_0(t_{(i)}) \exp\{\beta' Z_{(i)}\}\right] \exp\left\{-\sum_{j=1}^{n} \Lambda_0(t_j) \exp\{\beta' Z_j\}\right\} \end{aligned}$$

Posons

$$\lambda_0(t_{(i)}) = \lambda_{0i} \quad , \quad i = 1, 2, ..., L.$$

On obtient

$$\Lambda_0(t_j) = \sum_{t_{(i)} \leq t_j} \lambda_{0i}.$$

Maximer $L(\beta, \lambda_0)$ revient à maximer

$$L^*(\beta, \lambda_0) = \prod_{i=1}^{L} \lambda_{0i} \exp\left\{-\lambda_{0i} \sum_{j \in R(T_i)} \exp\{\beta' Z_j\}\right\}$$

cette fonction est maximale quand

$$\lambda_{0i} = \frac{1}{\sum_{j \in R(T_i)} \exp\{\hat{\beta}' Z_j\}}$$

Ceci suggère d'estimer la fonction de hasard cumulée Λ_0 par $\hat{\Lambda}_0$, appelé l'estimateur de Breslow, défini par

$$\hat{\Lambda}_0(t) = \sum_{T_i \leq t} \frac{1}{\sum_{j \in R(T_i)} \exp\{\hat{\beta}' Z_j\}}$$

Cet estimateur se réduit à l'estimateur de Nelson-Aalen quand il n'y a aucune covariable.

On peut estimer S_0 par

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$$

et l'estimateur de la fonction de survie pour un individu qui à pour vecteur de covariables Z_i est donné par

$$\hat{S}(t \mid Z_i) = \left[\hat{S}_0(t)\right]^{\exp\{\hat{\beta}' Z_i\}}.$$

Remarque 11 *S'il y a d'ex-aequo parmi les données, l'estimateur de* λ_0 *est donné par*

$$\hat{\lambda}_0(t_{(i)}) = \frac{d_i}{\sum\limits_{j \in R(T_i)} \exp(\hat{\beta}' Z_j)}$$

et l'estimateur de $\hat{\Lambda}_0$ *est donné par*

$$\hat{\Lambda}_0(t) = \sum_{i:T_i \leq t} \frac{d_i}{\sum\limits_{j \in R(T_i)} \exp(\hat{\beta}' Z_j)}$$

où d_i *est le nombre de décès en* T_i.

3.2.4 Tests

Test global

$H_0 : \beta = \beta_0$

- Test de Wald :

Il est basé sur la statistique

$$\chi^2_W = (\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0) \overset{H_0}{\rightsquigarrow} \chi^2(p),$$

(mesure l'écart entre $\hat{\beta}$ et β_0).

- Test de rapport de vraisemblance :

Il utilise la statistique

$$\chi^2_{RV} = 2\{\log L_{Cox}(\hat{\beta}) - \log L_{Cox}(\beta_0)\} \overset{H_0}{\rightsquigarrow} \chi^2(p),$$

(mesure la distance entre $\log L_{Cox}(\hat{\beta})$ et $\log L_{Cox}(\beta_0)$).

- Test de score :

Il est fondé sur la statistique

$$\chi^2_s = U'(\beta_0)[I(\beta_0)]^{-1} U(\beta_0) \overset{H_0}{\rightsquigarrow} \chi^2(p),$$

(mesure la pente de la tangente en β_0).

Ces trois statistiques suivent, sous H_0, une loi de χ^2 à p degrés de liberté (où β est un vecteur de dimension p).

Test local

Souvent on trouve intéressant de tester une hypothèse au sujet d'une sous ensemble de β. l'hypothèse est :

$$H_0 : \beta_1 = \beta_{10},$$

où $\beta = (\beta'_1, \beta'_2)'$. Ici β_1 est un vecteur $q \times 1$ de β et β_2 est un vecteur $p - q \times 1$ de β.

- Test de Wald :

Soit $\hat{\beta} = (\beta_1', \beta_2')'$ l'estimateur de maximum de vraisemblance partielle de β. Nous divisons la matrice d'information $I(\beta_0)$ comme suit

$$I(\beta_0) = \begin{pmatrix} I_{11} & I_{12} \\ \\ I_{21} & I_{22} \end{pmatrix},$$

où I_{11} (resp I_{22}) est la $q \times q$ (resp $(p-q)(p-q)$) sous matrice de $I(\beta_0)$ associée à β_1 (resp β_2).

Ce test se base sur la statistique

$$\chi^2_W = (\hat{\beta}_1 - \beta_{10})' \left[I^{11}(\hat{\beta}) \right]^{-1} (\hat{\beta}_1 - \beta_{10})$$

où $I^{11}(\hat{\beta})$ est la matrice $q \times q$ de $\left[I(\hat{\beta}) \right]^{-1}$. Pour les grands échantillons χ^2_W suit un χ^2 avec q degrés de liberté sous H_0.

- Test de rapport de vraisemblance :

Soit $\hat{\beta}_2(\beta_{10})$ l'estimateur du maximum de vraisemblance partielle de β_2 avec β_1 fixé à la valeur β_{10}. La statistique de test est

$$\chi^2_{RV} = \left\{ \mathcal{L}(\hat{\beta}) - \mathcal{L} \left[\beta_{10}, \hat{\beta}_2(\beta_{10}) \right] \right\}$$

pour les grands échantillons χ^2_{RV} suit un χ^2 avec q degrés de liberté sous H_0.

- Test score :

Soit $U_1 \left[\beta_{10}, \hat{\beta}_2(\beta_{10}) \right]$ le $q \times 1$ vecteur de score de β_1, évalué à la valeur supposée de β_{10} et à l'estimateur de maximum de vraisemblance partielle $\hat{\beta}_2$ de β_2. Ce test est

basé sur la statistique

$$\chi^2_{SC} = U_1' \left[\beta_{10}, \hat{\beta}_2(\beta_{10})\right] \left[I^{11}(\beta_{10}, \hat{\beta}_2(\beta_{10}))\right]^{-1} U_1 \left[\beta_{10}, \hat{\beta}_2(\beta_{10})\right].$$

Pour les grands échantillons χ^2_{SC} suit un χ^2 avec q degrés de liberté sous H_0.

3.2.5 Interprétation des coefficients de régression

Par définition le risque relatif à l'instant t, pour deux vecteurs de covariables Z_i et Z_j, est égal à :

$$RR(t) = \frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)}.$$

Dans le modèle de Cox, le risque relatif est constant au cours du temps :

$$RR(t) = RR = \exp(\beta'(Z_i - Z_j)).$$

Ainsi, dans un modèle de Cox avec une seule covariable Z, le risque relatif est

- pour une variable binaire codée 0 et 1 : $RR = \exp(\beta)$,

- pour une variable binaire codée a et b : $RR = \exp(\beta(b-a))$,

- pour une variable continue, $\exp(\beta)$ correspond au risque relatif pour une augmentation d'une unité de la variable. Le risque relatif est constant pour une augmentation d'une unité de la variable quelle que soit la valeur de la covariable : c'est une hypothèse de log-linéarité.

Il y a donc deux hypothèses importantes à vérifier dans l'utilisation du modèle de Cox : l'hypothèse de risques proportionnels (risque relatif constant au cours du temps) et l'hypothèse de log-linéarité.

3.2.6 Quelques extensions

Covariables dépendantes du temps

Le modèle de Cox permet de prendre en compte des covariables dépendantes du temps (traitement, marqueur biologique,...). Il faut néanmoins que $Z(t)$ soit prédictible, c'est-à-dire connue au temps t. Le traitement statistique est identique néanmoins, on peut faire les remarques suivantes.

- Il est nécessaire de connaître la valeur des covariables pour chaque temps d'événements.

En effet, on a $L_{Cox}(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)}(T_i))}{\sum_{j \in R(T_i)} \exp(\beta' Z_j(T_i))}$. Ceci peut poser quelques problèmes dans le cas de marqueurs biologiques.

- L'interprétation devient difficile car le risque est spécifique à chaque histoire des covariables.

- L'hypothèse de hasard proportionnel est conservée. En effet, les fonctions de risque pour les différentes modalités d'une covariable restent proportionnelles et leurs rapports sont indépendants du temps. L'effet de la covariable ne varie pas au cours du temps, c'est la variable qui varie.

- L'utilisation de certaines covariables dépendantes du temps permet de tester l'hypothèse de risques proportionnels.

Modèles de Cox stratifié

Dans le cas où une variable qualitative ne vérifie pas l'hypothèse de hasards proportionnels, on peut considérer un modèle de Cox stratifié. Prenons l'exemple, d'une variable binaire Y codée 0 et 1, par exemple, le sexe (0 pour les hommes et 1 pour les femmes).

Dans ce modèle, le risque de base est différent dans les deux strates mais les covariables Z agissent de la même manière sur les deux fonctions de hasard, c'est-à-dire,

$$\lambda(t \mid Z,\ Y = 0) = \lambda_0(t)\exp(\beta' Z),$$

$$\lambda(t \mid Z,\ Y = 1) = \lambda_1(t)\exp(\beta' Z).$$

L'effet des covariables est le même dans chaque strate. Les estimations obtenues par la méthode de la vraisemblance partielle sont applicables pour obtenir les paramètres (λ_0, λ_1 et β) du modèle. La vraisemblance partielle est calculée dans chacune des strates ; la vraisemblance totale est le produit des vraisemblances de chaque strate.

Modèles de fragilité (frailty)

Le modèle de Cox suppose que la population est homogène (malgré la prise en compte de covariables). Néanmoins, cette hypothèse n'est pas toujours réaliste, notamment quand des covariables importantes ne sont pas observables ou inconnues. Par exemple, cela peut être des facteurs environnementaux ou génétique. Les modèles fragilité permettent de prendre en compte l'hétérogénéité des observations.

Considérons une nouvelle covariable non observée Z_0. On suppose, comme dans le modèle de Cox, que l'effet des covariables se résume à une quantité réelle $\exp(\beta_0 Z_0)$, alors la fonction de risque est

$$\lambda(t \mid Z, Z_0) = \lambda_0(t) e^{\beta_0 Z_0} e^{\beta' Z}.$$

En notant $\omega = e^{\beta_0 Z_0}$, la variable aléatoire réelle positive (appelée "fragilité"), la fonction de risque devient

$$\lambda(t \mid Z, \omega) = \lambda_0(t) \omega e^{\beta' Z},$$

et la fonction de survie conditionnelle est

$$S(t \mid Z, \omega) = \exp\left(-\int_0^t \lambda_0(s) \omega e^{\beta' Z} ds\right) = \exp\left(-\omega e^{\beta' Z} \Lambda_0(t)\right).$$

Comme ω est une variable aléatoire, on s'intéresse à la fonction de survie moyennée sur ω. Cette quantité correspond à la fonction de survie marginale pour un individu quelconque

$$S(t \mid Z) = \int_0^\infty \exp\left(-v e^{\beta' Z} \Lambda_0(t)\right) f_\omega(v) dv = \mathcal{L}_\omega\left(e^{\beta' Z} \Lambda_0(t)\right),$$

où f_ω représente la densité de la v.a. ω et $\mathcal{L}_\omega(s) = E(e^{-\omega s})$ est la transformée de Laplace de la distribution de la fragilité.

Le plus souvent, les modèles de fragilité sont utilisés pour prendre en compte une dépendance entre les temps d'événements de certains individus. En effet, les individus d'un même sous-groupe d'une population peuvent être liés si tous les individus de ce

groupe ont des caractéristiques communes non observées. Par exemple, des individus d'une même famille, d'une même région ou d'un même hôpital. Le terme de fragilité est alors commun à chaque individu du groupe (permet de créer la dépendance) mais différent d'un groupe à l'autre (hétérogénéité entre groupe).

Exemple 12

Nous allons étudier un exemple. Il s'agit de l'étude de la durée de survie de 26 *patients atteints de la maladie du cancer. On a les données suivantes :*

	Temps	*Ind*	*Âge*	*Tr*
1	59	1	72.33	1
2	115	1	74.49	1
3	156	1	66.47	1
4	421	0	53.36	2
5	31	1	50.34	1
6	448	0	56.43	1
7	464	1	56.94	2
8	475	1	59.85	2
9	477	0	64.18	1
10	563	1	55.18	2
11	638	1	56.76	1
12	744	0	50.11	2
13	769	0	59.63	2
14	770	0	57.05	2
15	803	0	39.27	1
16	855	0	43.12	1
17	1040	0	38.89	1
18	1106	0	44.60	1
19	1129	0	53.91	2
20	1206	0	44.21	2
21	1227	0	59.59	2
22	268	1	74.5	1
23	329	1	74.5	1
24	353	1	63.22	2
25	365	1	64.42	2
26	377	0	58.31	2

où

1. *Temps est le nombre de jours du début du traitement jusqu'à la mort ou la censure.*

2. *Ind est l'indicateur de censure.*

3. *Âge est l'âge au moment du diagnostic.*

4. *Tr est un indicateur du traitement donné au malade.*

Nous avons donc le modèle suivant

$$\lambda(t) = \lambda(t) \exp\Big\{\beta_1 \times \hat{A}ge + \beta_2 Tr\Big\}$$

ou

$$S(t) = (S_0(t))^{\exp\left\{\beta_1 \times \hat{A}ge + \beta_2 Tr\right\}}$$

1) ***L'estimation des paramètres***

La fonction "Coxph" dans le logiciel S-plus nous permet d'estimer le vecteur du paramètre $\beta = (\beta_1, \beta_2)$ *et nous obtenons les résultats suivantes*

```
Working data will be in C:\Program Files\sp2000\users\d
ounia\_Data
> car<-coxph(Surv(Temps, Ind) ~ age + Tr,cancer5)
> summary(car)
Call:
coxph(formula = Surv(Temps, Ind) ~ age + Tr, data = cancer5)

  n= 26

       coef exp(coef) se(coef)     z      p
age  0.147     1.159   0.0461  3.19 0.0014
 Tr -0.804     0.448   0.6320 -1.27 0.2000

    exp(coef) exp(-coef) lower .95 upper .95
age     1.159      0.863      1.06      1.27
 Tr     0.448      2.234      0.13      1.54
```

D'où $\beta = (0.147, -0.448)$

2) ***Test d'hypothèses***

- ***Test global***

$H_0 : \beta = 0$ *contre* $H : \beta \neq 0$ *et en prend le seuil* $\alpha = 0.05$.

On a les résultats suivants

```
Likelihood ratio test= 15.9  on 2 df,   p=0.000355
Wald test            = 13.5  on 2 df,   p=0.00119
Score (logrank) test = 18.6  on 2 df,   p=0.0000934
```

Le coefficient de régression est $\hat{\beta} = (0.147, -0.448)$*, avec un niveau de significativité (test de rapport de vraisemblance) de* 0.000355*. Le vecteur des coefficients est donc significativement différent de* 0*. Le résultat est le même si nous utilisons le test de Wald ou le test de Score.*

- ***Test local***

Si nous nous intéressons à tester l'effet du traitement sur la fonction de hasard de base λ_0 *nous obtenons les résultats(nous utilisons par exemple le test de rapport de vraisemblance).*

```
> car1<-coxph(Surv(Temps, Ind) ~ age,cancer5)
> car$loglik
[1] -34.98494 -27.04190
> car1$loglik
[1] -34.98494 -27.83815
> 2*(-27.042+27.838)
[1] 1.592
> pchisq(1.592,df=1)
[1] 0.7929594
> 1-0.793
[1] 0.207
```

Donc

$$\chi^2_{RV} = 2\left\{\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta_0)\right\} = 1.592$$

avec un niveau de significativité de 0.207*, alors on accepte* H_0 *donc le traitement n'a pas d'effet multiplicatif sur la durée de survie.*

3) ***Estimation de la fonction de covariance***

```
> car$var
             [,1]         [,2]
[1,] 0.002129550 0.004696922
[2,] 0.004696922 0.399486402
```

4) ***Estimation de la fonction de survie de base*** S_0

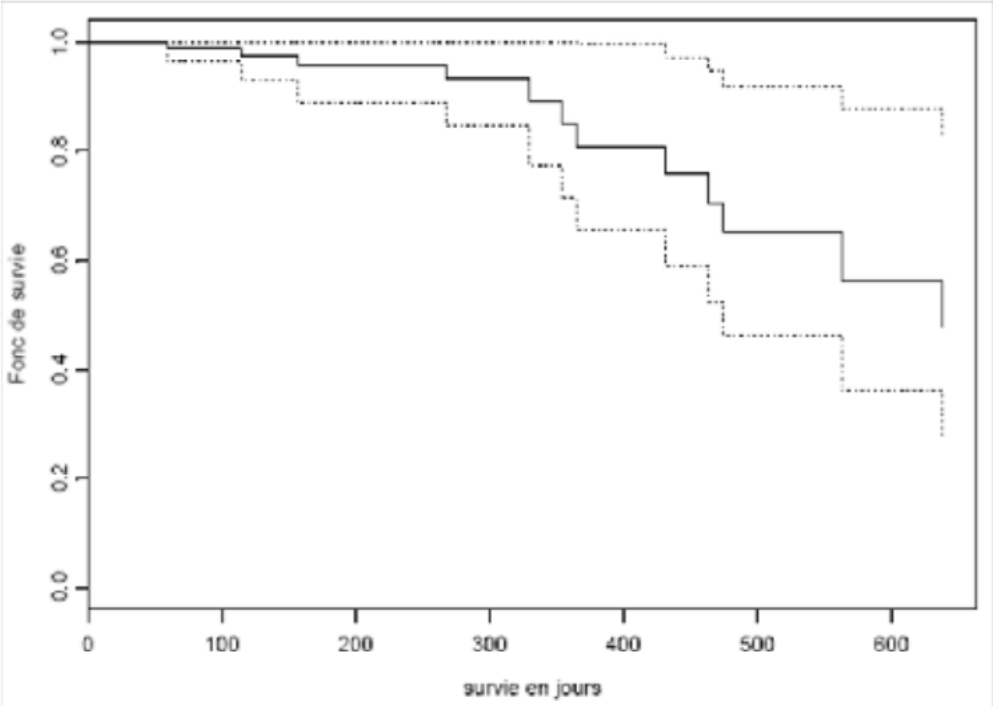
La fonction (survift) dans le logiciel S-plus nous permet d'estimer la fonction de survie de base et nous avons les résultats suivants :

```
> summary(survfit(car))
Call: survfit.coxph(object = car)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.989  0.0132        0.963        1.000
  115     25       1    0.975  0.0229        0.931        1.000
  156     24       1    0.956  0.0351        0.890        1.000
  268     23       1    0.935  0.0474        0.846        1.000
  329     22       1    0.892  0.0644        0.774        1.000
  353     21       1    0.851  0.0763        0.714        1.000
  365     20       1    0.808  0.0863        0.656        0.996
  431     17       1    0.759  0.0965        0.591        0.973
  464     15       1    0.705  0.1066        0.524        0.948
  475     14       1    0.652  0.1140        0.463        0.918
  563     12       1    0.563  0.1277        0.361        0.878
  638     11       1    0.479  0.1328        0.278        0.825
```

La figure suivante donne la courbe de l'estimateur de la fonction de survie de base S_0.

```
> plot(survfit(car),xlab="survie en jours",ylab="Fonc de survie")
```


Fonc de survie
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600
survie en jours

Chapitre 4

Modèles paramétriques

On suppose que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Ainsi, le modèle paramétrique peut être formulé en précisant la forme de l'une ou l'autre des cinq fonctions équivalentes qui définissent la loi de la durée : λ, Λ, f, S ou F. Néanmoins, on spécifie souvent la forme du risque instantané λ : constant, monotone croissant ou décroissant et en forme de $\cap$ ou de $\cup$.

Les estimateurs des paramètres du modèle sont ensuite obtenus en maximisant la vraisemblance des observations (par l'intermédiaire de méthodes itératives, par exemple l'algorithme de Newton-Raphson).

4.1 Risque instantané constant (loi exponentielle)

La loi exponentielle $\varepsilon(\theta)$, qui ne dépend que d'un paramètre θ, est la seule qui admet un risque instantané constant. Cette loi est aussi dite "sans mémoire" car la

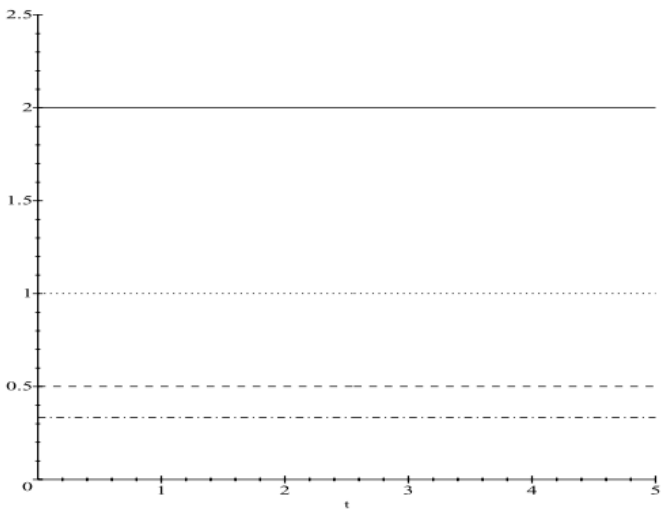


FIG. 4.1: la fonction de risque pour la loi exponentielle

probabilité de décès pour un individu dans un certain laps de temps est la même quelle que soit sa durée de vie (*i.e.* $P(X > s + t \mid X > t) = P(X > s)$).

La fonction de densité d'une loi exponentielle de paramètre $\theta > 0$ est donné par

$$f(t \mid \theta) = \theta e^{-\theta t}, \qquad t \geq 0.$$

Sa fonction de survie est $S(t \mid \theta) = e^{-\theta t}$ et sa fonction de risque est $\lambda(t \mid \theta) = \theta$, une constante indépendante de t (Voir Fig 4.1).

Soit $X_1, ..., X_n \overset{iid}{\sim} \varepsilon(\theta)$, calculons l'estimateur du maximum de vraisemblance pour cette loi.

La vraisemblance est donnée par

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

En prenant le logarithme, nous obtenons

$$\begin{aligned} l(\theta) &= \ln L(\theta) = n\ \ln\theta - \theta\sum_{i=1}^{n} x_i \\ &= n\ \ln\theta - n\theta\overline{x} \end{aligned}$$

4.2 Risque instantané monotone

Il y a beaucoup de distributions de durées de vie dont le taux est monotone.

4.2.1 Loi de Weibull

La loi de weibull est également une généralisation de l'exponentielle. Elle est caractérisée par deux paramètres, $\theta > 0$ et $\nu > 0$. La fonction de densité d'une telle loi est donnée par

$$f(t \mid \theta, \nu) = \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right), \qquad t \geq 0$$

Ainsi nous remarquons que si $\nu = 1$, on retrouve la loi exponentielle $\varepsilon(\frac{1}{\theta})$. La fonction de survie est

$$S(t|\theta, \nu) = \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)$$

et la fonction de risque vaut

$$\lambda(t|\theta, \nu) = \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1}.$$

La fonction de risque est monotone croissante si $\nu > 1$ (Voir Fig 4.2), monotone décroissante si $\nu < 1$ (Voir Fig 4.3).

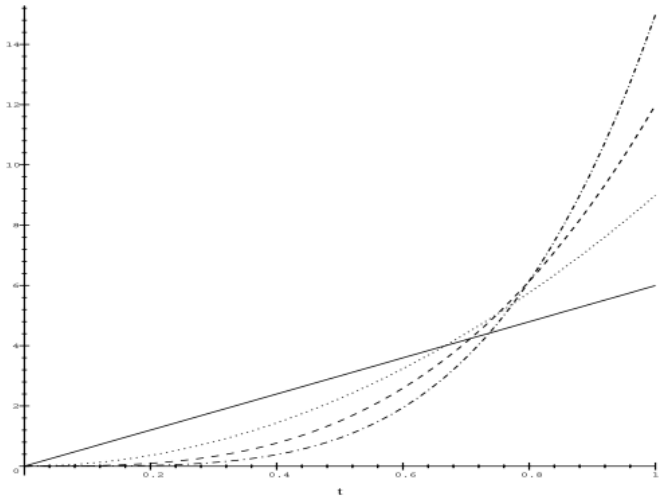


FIG. 4.2: La fonction de risque pour la loi de Weibull (croissante)

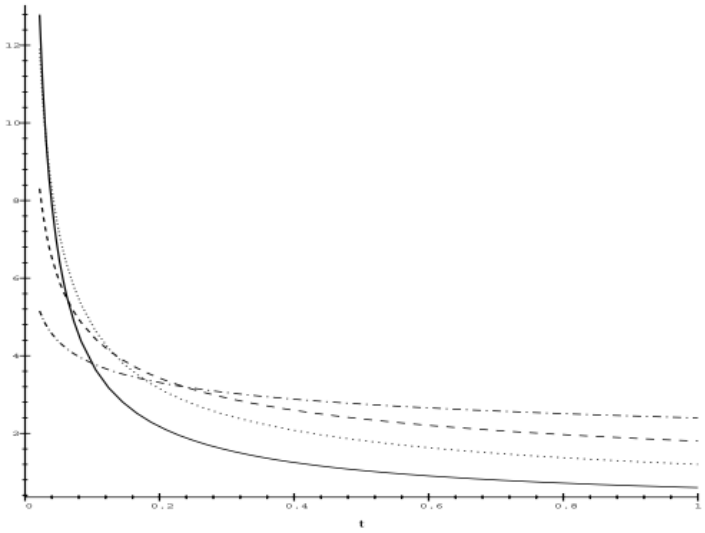


FIG. 4.3: La fonction de risque pour la loi de Weibull (décroissante)

4.2.2 Loi Gamma

La loi gamma comporte deux paramètres, $\nu > 0$ et $\theta > 0$. Le premier est appelé paramètre d'échelle alors que θ est le paramètre de forme.

La fonction de densité de cette loi est donnée par

$$f(t \mid \nu, \theta) = \frac{\nu}{\Gamma(\theta)}(\nu t)^{\theta-1}e^{-\nu t}, \qquad t \geq 0 \text{ et } \theta, \nu > 0,$$

où $\Gamma(\theta) = \int_0^\infty x^{\theta-1}e^{-x}dx$ et la fonction gamma. La fonction de survie s'exprime comme

$$S(t \mid \nu, \theta) = \int_t^\infty \frac{\nu}{\Gamma(\theta)}(\nu x)^{\theta-1}e^{-\nu x}dx$$

En choisissant le paramètre θ entier, nous obtenons la distribution dite de Erlang. Pour celle-ci, nous obtenons comme fonction de risque

$$\lambda(t \mid \nu, \theta) = \frac{\nu(\nu t)^{\theta-1}}{(\theta-1)!\sum_{k=0}^{\theta-1}\frac{1}{k!}(\nu t)^k}.$$

Si $\theta > 1$ le risque instantané $\lambda(t)$ est croissant de 0 à ν (Voir Fig 4.4). Si $0 < \theta < 1$, $\lambda(t)$ est décroissant de ∞ à $\frac{1}{\nu}$ (Voir Fig 4.5).

Si l'on choisit $\theta = 1$, nous obtenons une loi exponentielle de paramètre ν, ainsi, la loi exponentielle n'est qu'un cas particulier de la loi gamma.

Le logarithme de la vraisemblance d'un échantillon issu d'une loi gamma est donné par

$$l(\nu, \theta) = n\ \nu \ln \nu - n \ln \Gamma(\theta) + (\theta - 1)\sum_{i=1}^{n} \ln t_i - n\ \theta\ \bar{t}$$

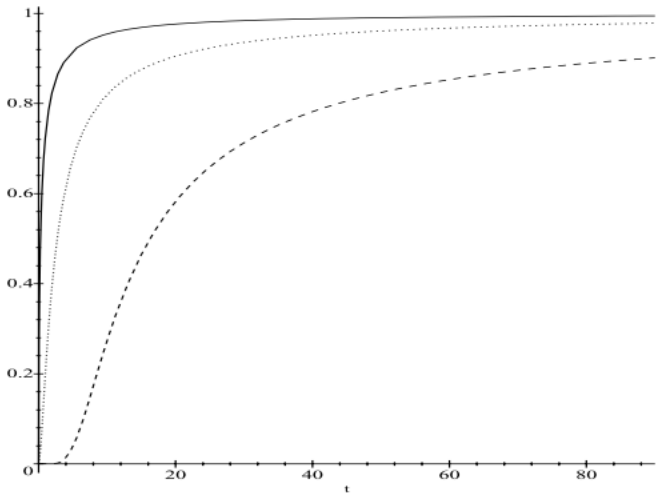


FIG. 4.4: La fonction de risque pour la loi gamma (croissante)

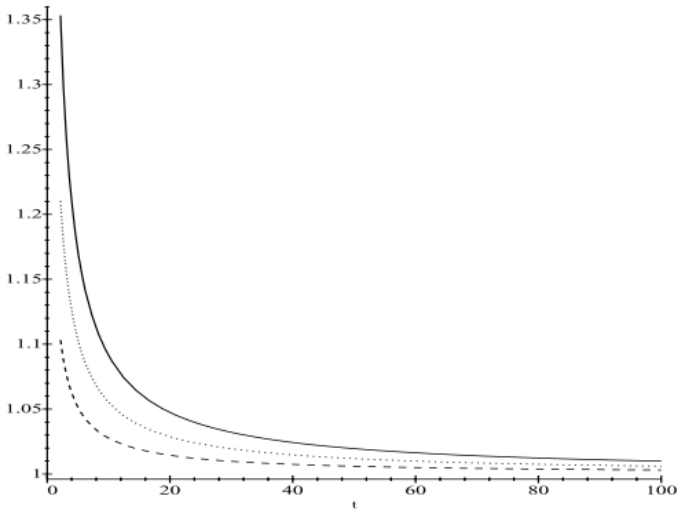


FIG. 4.5: La fonction de risque pour la loi gamma (décroissante)

En dérivant par rapport à ν et en appelant $\hat{\theta}$ l'estimateur de maximum de vraisemblance pour θ, nous obtenons $\hat{\nu} = \frac{\hat{\theta}}{\bar{t}}$.

Par contre, le calcul exact de $\hat{\theta}$ n'est pas possible, ainsi, nous pouvons seulement exprimer un estimateur en fonction de l'autre.

4.2.3 Autres lois

Il existe de nombreuses lois avec des risques monotones, citons notamment les lois de Gompertz-Makeham, les mélanges de deux distributions exponentielles, les lois de Weibull exponentiées.

4.3 Risque instantané en $\cap$ et $\cup$

4.3.1 Lois log-normale

Si le temps de survie X est tel que $\ln(X)$ suit une loi normale avec moyenne μ et variance σ^2, alors on dit que X suit une loi log-normale.

Sa fonction de densité est donnée par

$$f(t \mid \mu, \sigma^2) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\ln t - \mu)^2\right\},$$

où μ est le paramètre d'échelle et σ est le paramètre de forme. Contrairement à la loi normale les paramètres ne donnent pas la moyenne et la variance de la loi. En

posant $a = e^{-\mu}$, alors $-\mu = \ln a$ et nous obtenons

$$f(t \mid a, \sigma^2) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\ln at)^2\right\}.$$

La fonction de survie d'une variable suivant une loi log-normale est donnée par

$$S(t \mid a, \sigma^2) = 1 - \Phi\left(\ln\left(\frac{at}{\sigma}\right)\right),$$

où $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-u^2/2} du$ est la fonction de répartition de la loi normale standard . La fonction de risque est de la forme

$$\lambda(t \mid a, \sigma^2) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln at)^2}{2\sigma^2}\right\}}{1 - \Phi\left(\ln\left(\frac{at}{\sigma}\right)\right)}.$$

Pour la loi log-normale, les estimateurs de maximum de vraisemblance peuvent être calculés facilement. Nous obtenons,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln t_i.$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\ln t_i - \hat{\mu})^2.$$

Le risque instantané croît de 0 à sa valeur maximum puis décroît vers 0, *i.e.*, il est en forme de $\cap$ (Voir Fig 4.6)

4.3.2 Lois de weibull généralisée

La loi de weibull est intéressante pour modéliser des risques monotones. Cependant elle devient mal adaptée quand les risques sont en forme de cloche. Une alternative

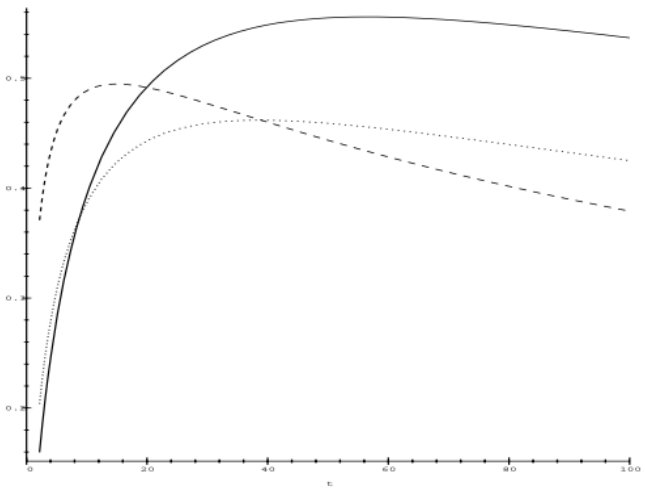


FIG. 4.6: La fonction de risque pour la loi log-normale

est l'utilisation de la loi de weibull généralisée $GW(\theta, \nu, \gamma)$:

$$\begin{aligned} \lambda(t \mid \theta, \nu, \gamma) &= \left(1 + \left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}-1} \frac{\nu}{\gamma\theta^{\nu}} t^{\nu-1}, \qquad t \geq 0 \text{ et } \theta, \nu, \gamma > 0, \\ S(t \mid \theta, \nu, \gamma) &= \exp\left[1 - \left(1 + \left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}}\right]. \end{aligned}$$

pour $\gamma = 1$, on retrouve la loi de Weibull $W(\theta, \nu)$; pour $\gamma = 1$ et $\nu = 1$, on retrouve la loi exponentielle $\varepsilon(\frac{1}{\theta})$.

Pour $\gamma > \nu > 1$, le risque instantané croît de 0 à sa valeur maximum puis décroît vers 0, *i.e.*, il est en forme de $\cap$ (Voir Fig 4.7)

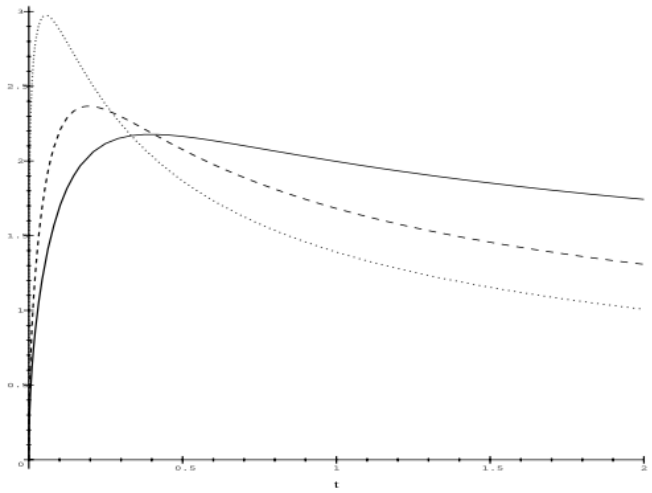


FIG. 4.7: La fonction de risque pour loi de weibull généralisée

4.3.3 Autres lois

Les lois Log-logistiques et gaussienne inverse permettent de considérer des risques instantanés en forme de $\cap$.

4.4 Introduction de covariables

Dans l'approche paramétrique, les fonctions d'intérêts peuvent dépendre de covariables explicatives succeptibles d'influencer la survie. En plus d'ajuster les fonctions de survie à différents facteurs, ceci permettra de comparer les durées de survie (l'hypothèse nulle sera l'égalité des distributions de survie).

Considérons Z un vecteur de covariables. Notons que ces covariables peuvent dépendre de temps, cependant il est nécessaire de supposer que la valeur des covariables

ne change pas entre deux mesures. Afin de simplifier les écritures on supposera dans ce qui suit que les covariables sont fixées au cours du temps. On suppose que les covariables vont modifier les fonctions de risque en suivant un modèle à risque proportionnels "de Cox" (d'autres modèles à risques proportionnels sont possibles), c'est-à-dire

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z)$$

où β est le vecteur de coefficients de régression. Les fonctions de survie et de densité correspondant à ces fonctions de risque sont données par

$$S(t \mid Z) = \exp\left(-\int_0^t \lambda(u \mid Z)du\right) = \exp\left(-\int_0^t \lambda_0(u)\exp(\beta' Z)du\right) = S_0(t)^{\exp(\beta' Z)}$$

$$f(t \mid Z) = -S'(t \mid Z) = \lambda(t \mid Z)\exp\left(-\int_0^t \lambda(u \mid Z)du\right) = \lambda_0(u)\exp(\beta' Z) \times S_0(t)^{\exp(\beta' Z)},$$

avec $S_0(t) = \exp\left(-\int_0^t \lambda_0(u)du\right)$.

Les paramètres des modèles s'obtiennent simplement par la méthode du maximum de vraisemblance.

4.4.1 Comparaison de deux groupes

Considérons la situation où l'on souhaite comparer les durées de survie de deux groupes A et B. On introduit la covariable suivante,

$Z = 0$ si l'individu appartient au groupe $A \Longrightarrow \lambda_A(t) = \lambda_0(t)$

$Z = 1$ si l'individu appartient au groupe $B \Longrightarrow \lambda_B(t) = \lambda_0(t)\exp(\beta)$.

pour comparer les deux groupes, on estime le coefficient de régression β et on teste l'hypothèse nulle $H_0 : \beta = 0$ c'est-à-dire $H_0 : \lambda_A = \lambda_B$. On peut, à cet effet, utiliser les

tests du rapport de vraisemblance, de Wald ou du score qui suivent asymptotiquement une loi de $\chi^2(1)$, sous H_0.

4.4.2 Exemple

Considérons un risque de base suivant une loi de Weibull $W(\theta, \nu)$, alors

$$\begin{aligned}
\lambda_0(t) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1}, \qquad t \geq 0 \text{ et } \theta, \nu > 0, \\
S_0(t) &= \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right), \\
f_0(t) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right).
\end{aligned}$$

D'après les résultats du début de la section, les fonctions de risque, de survie et de densité dans le cas où il y a des covariables sont

$$\begin{aligned}
\lambda(t \mid Z) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z), \qquad t \geq 0 \text{ et } \theta, \nu > 0, \\
S(t \mid Z) &= \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)}, \\
f(t \mid Z) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z) \times \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)}.
\end{aligned}$$

Pour $\nu = 1$, On retrouve la loi exponentielle $\varepsilon(\frac{1}{\theta})$. Ainsi, dans le cas d'un risque suivant une loi exponentielle avec des covariables, on obtient

$$\begin{aligned}
\lambda(t \mid Z) &= \frac{1}{\theta} \times \exp(\beta' Z), \qquad \theta > 0, \\
S(t \mid Z) &= \exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)}, \\
f(t \mid Z) &= \frac{1}{\theta} \exp(\beta' Z) \times \exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)}.
\end{aligned}$$

4.4.3 Modèles de vie accélérée

Parmi les modèles de régression, les modèles de vie accélérée sont souvent considérés notamment en fiabilité. Ces modèles peuvent être définis de deux manières. La première représentation des modèles de vie accélérée est donnée par la fonction de survie accélérée :

$$S(t \mid Z) = S_0(te^{\beta' Z}),$$

où Z est un vecteur de covariable, β le vecteur de coefficients de régression. Le terme $e^{\beta' Z}$ est un facteur d'accélération car un changement dans les covariables change l'échelle de temps. On peut obtenir une expression de la fonction de risque,

$$\lambda(t \mid Z) = [-\ln(S(tZ))]' = -\frac{[(S(t \mid Z)]'}{S(t \mid Z)} = -\frac{-e^{\beta' Z} \times \lambda_0(te^{\beta' Z}) \times S_0(te^{\beta' Z})}{S_0(te^{\beta' Z})} = e^{\beta' Z}\lambda_0(te^{\beta' Z}).$$

En effet, on a les égalités suivantes,

$$S(t \mid Z) = S_0(te^{\beta' Z}) = \exp(-\Lambda_0(te^{\beta' Z})) = \exp\left[-\int_0^t \lambda_0(ue^{\beta' Z})du\right].$$

Si on suppose que $S_0(t)$ est la fonction de survie de la variable $\exp(\mu + \epsilon)$, alors $S_0(t) = P(e^{\mu+\epsilon} > t)$. Ainsi, on obtient que

$$S(t \mid Z) = S_0(te^{\beta' Z}) = P(e^{\mu+\epsilon} > te^{\beta' Z}) = P(e^{\mu-\beta' Z+\epsilon} > t) = P(X > t),$$

est la fonction de survie de la variable X où $\log(X) = \mu - \beta' Z + \epsilon$. En considérant le changement de variable $\alpha = -\beta$, on obtient la deuxième représentation par un modèle de régression log-linéaire pour la durée de survie

$$\log(X) = \mu + \alpha' Z + \epsilon,$$

où X est la durée de survie (pas toujours observée car $T = \min(X, C)$) et ϵ est une variable aléatoire (dans le cas de plusieurs observations, les ϵ_i sont $i.i.d.$).

Plusieurs lois sont possibles pour les variables ϵ, par exemple,

- $\epsilon \sim$ loi aux valeurs extrêmes ($f_\epsilon(y) = \exp(y - e^y)$)

- $\epsilon \sim$ log-logistic

- $\epsilon \sim$ log-normal

- $\epsilon \sim$ generalized gamma

On peut déduire la loi de X et les estimations des paramètres sont obtenues par maximisation de la vraisemblance.

Remarque 13 *On peut remarquer que dans le cas des modèles de vie accélerée, pour une covariable $Z > 0$, un coefficient de régression α négatif entraîne un temps de survie plus petit est donc une survie plus faible. Alors que dans le modèle semi-paramétrique de Cox un coefficient de régression α négatif entraîne un risque d'événement plus faible et donc une survie plus grande.*

Bibliographie

[1] R. Bououden. *Modèles semi-paramétriques en analyse de survie.* Thèse de Magistère, Université Mentouri - Constantine, 2007.

[2] M. Bousseboua. *Statistique Mathématique.* Les éditions de l'université Mentouri - Constantine, 2004.

[3] J. Droesbeke, B. Fichet P. Tassi. *Analyse statistique des durées de vie : modélisation des données censurées*, Economica, Paris.

[4] J. D. Kalbfleish, R. L. Prentice, *The statistical analysis of failure time data.* Wiley and Sous, 1980.

[5] F. Massais. *Notions fondamentales de la théorie de probabilités.* Les éditions de l'université Mentouri Constantine, 2001.

[6] J. P. Klein, *Survival analysis techniques for censored and truncated data.* Springer, 1997.

[7] M. Tristan Lorino. *Modèles statistiques pour des données de survie corrélées*", Thèse de doctorat, paris, 2002.

[8] "L'analyse de survie". Traitement statistique des études cliniques, le Département Biométrie de FOVEA.

[9] "Analyse de données de survie/données censurées". D. Neveu _ MB6_2010-2011.