République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur de la Recherche **Scientifique**

Université Mohamed Seddik Ben Yahai - Jijel

Faculté des Sciences Exactes et Informatique

Département de Mathématique

N° d'ordre : ……/2020

Mémoire de fin d'études

En vue de l'obtention du diplôme de **Master**

Filière : Mathématiques

Spécialité : Probabilités et Statistique

Thème

Estimation pour des données censurées

Présenté par :

- **Abdelaziz Souad**
- **Sissaoui Amira**

Devant le jury:

BOUDJERDA Khawla	Université de Jijel	Président
MADI Meriem	Université de Jijel	Encadreur
SELLAMI Nawel	Université de Jijel	Examinateur

Année universitaire : 2019/2020

Dédicaces

Nous dédions ce travail à:

Nos chers parents, qui ont le mérite pour qui nous somme aujourd'hui

Tous ceux que nous aimons,

Nos amies,

Tous ceux qui nous ont aidées,

Soutenus

Et qui ont cru à nos capacités de réussir.

Remerciements

Tout d'abord, nous remercions Allah le tout puissant pour le courage qu'il nous a données pour faire ce mémoire de master et terminer nos études.

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui nous voudrions témoigner toute notre gratitude.

Nous adressons nos remerciements à nos chers parents qui ont tout fait pour nos éducations et nos soins.

En particulier nous remercions notre encadreur Mme. MADI Meriem pour tous ses conseils lors de la rédaction de ce mémoire.

Nous remercions les membres du jury Mme. SELLAMI Nawel et Mme. BOUDJERDA Khawla pour nous honorer de leur participation à cette soutenance.

Nous remercions nos professeurs qui nous ont accompagné dans toutes les étapes de notre formation pour leur patience, un grand merci à notre cher professeur GHERDA Mebrouk.

Nous remercions tous ceux que nous aimons et tous ceux qui nos ont soutenu, nos amis et collègues Yasmin et Meriem.

Nous n'oublions pas de remercier l'Université de Jijel, le Département de Mathématiques et notre pays l'Algérie.

Table des matières

Introduction

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant un évènement d'intérêt.

La première étude sur la mortalité en Angleterre, au 17 ème siècle, fut par Graunt (1787) qui a défini des notions d'espérance de vie et d'espérance de vie résiduelle. Puis arrive la modélisation de la probabilité de décéder par Gompertz (1825) et Makeham (1860). Weibull (1951) aborde le modèle paramétrique pour calculer la fiabilité d'un système non réparable. Il propose aussi pour le même modèle en présence de données tronquées ou censurées. En (1958), Kaplan et Meier proposent d'utiliser dans le domaine médical un estimateur non paramétrique permettant d'intégrer les données censurées en estimant la fonction de survie en présence de censure à droite.

Ce types d'analyse reste intéressante pour les chercheurs de nos jours à cause de leur utilisation dans plusieurs champs d'application, telle que la biologie médicale, la fiabilité et la démographie.

Mathématiquement, une durée de vie n'est rien d'autre qu'une variable aléatoire (v.a) non négative, lorsque certaines de ces observations sont incomplètes, nous parlons de la censure et la troncature. Nous concentrerons notre étude, sur la variable censurée.

Dans ce mémoire, nous nous intéressons aux modèles de survie paramétrique et non paramétrique.

Dans ce cadre, notre travail a été structuré comme suit :

Dans le premier chapitre nous allons introduire le concept de l'analyse de survie, les outils de probabilité nécessaires ainsi que les différentes distributions et fonctions de loi des données censurées.

Le deuxième chapitre parle de l'approche paramétrique en analyse de survie, l'estimation sur les données censurées par la méthode du maximum de vriasemblance selon les distributions théoriques (classé selon les types du risque instantané) par un calcule manuelle, ou par l'utilisation des algo-

rithmes numériques de maximisation, il donne aussi les approches de choix entre les lois, les testes de comparaison des survie et enfin le modèle paramétrique de survie dans le cas de covariable.

Le troisième chapitre parle de l'estimation non paramétrique sur les données censurées qui se base sur l'estimation des distributions de la durée de survie par différentes méthodes, dont la méthode de Kaplan-Meier est la plus connue et la plus utilisé, ensuite on introduit le teste de comparaison entre les fonctions de survie de Log-Rank.

Chapitre 1

Introduction à l'analyse des données censurées

1.1 Définitions et notations

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. On désigne par le terme durée de survie ou de vie, le temps qui s'écoule depuis un instant initial, par exemple début du traitement, diagnostic, . . . , jusqu'à la réalisation d'un évènement précis. L'évènement d'intérêt peut s'agir de la mort du patient, de l'apparition d'une maladie, d'une guérison, la panne d'une machine, la survenue d'un sinistre. Dans ces cas, la durée de vie représente respectivement, le temps avant une rechute ou un rejet de greffe, le temps entre le diagnostic et la guérison, la durée de fonctionnement d'une machine et le temps entre deux sinistres. C'est une variable aléatoire positive et souvent supposée bornée.

Les études de survie ne constituent pas un type d'enquête au même titre que, par exemple, les enquêtes cas-témoins. Ce type d'étude correspond à l'utilisation de méthodes d'analyse particulières, lorsque le critère de jugement est la survenue d'un décès ou d'un événement de santé particulier. Ces méthodes permettent d'étudier le délai de survenue d'un événement dont la survenue n'est pas constante.

Les études de survie nécessitent la connaissance d'un certain nombres de données. Lorsque le bilan de l'étude est réalisé à une certaine date, appelée date de point, on doit connaitre pour chaque sujet les données suivantes :

Date d'origine (*DO*) : si on veut faire une étude, la DO est la date

d'entrée dans l'étude, c'est à dire l'origine de l'analyse de survie. Chaque individu peut donc avoir une date d'origine différente. Ainsi dans l'étude de l'évolution d'une maladie, la DO est la date de début de la maladie. Si on s'intéresse à l'âge d'un individu jusqu'à la survenue de l'événement, la DO sera date de naissance de l'individu. Début de traitement, date d'opération...

Date de point (**DP**) : c'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets après cet instant. Commune à tous les individus.

Date des dernières nouvelles (**DDN**) : c'est la date la plus récente où des informations sur un sujet ont été recueillies. Chaque individu ou sujet dispose d'une date des dernières nouvelles. En médecine, par exemple, elle correspond à la date de la dernière consultation pour les sujet encore vivant ou à la date de décès pour les sujet décédés.

A partir de la date des dernières nouvelles, de la date d'origine et de l'état du sujet, il est possible de définir :

Durée du surveillance T_i : c'est la durée entre la date d'origine et la date des dernières nouvelles. Période de surveillance de l'instant de la survenue de l'évènement d'un sujet i.

Durée de recul L_i : appelée aussi délai de censure ; c'est la période entre la date d'origine est la date de point. C'est la période maximale d'observation. Elle peut être connue ou inconnue ; déterministe ou aléatoire, dépendante de l'individu ou non.

Durée de participation t_i : c'est le délai réellement observé pour chaque individu. C'est la durée de suivi. On distingue deux cas :

1- DDN< DP : ce qui est équivalent à $t_i \in [DO, DDN]$, dans ce cas $t_i = T_i$.

Si individu n'a pas subi l'événement à la DDN, il sera considéré comme perdu de vue et donc son observation pendant le temps de participation est incomplète. Dans le cas contraire, l'individu a subi l'événement à la DDN, le temps de participation nous donnera une observation complète de son état.

2- DDN > DP : ce qui est équivalent à $t_i \in [DO, DP]$, dans ce cas $t_i = L_i$.

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

exemple 1.1.1 *Une étude de survie débutée le 1er janvier 1977 avec une date de point le 1er juin 1978*

Tableau I : Différentes données d'une étude survie.

sujet	*DO*	*DDN*	*Etat DDN*	*Etat DP*	*T.Par*	*Recul (01/06/78-DO)*
I	*1/77*	*10/77*	*DCD*	*DCD*	*9*	*17*
II	*3/77*	*7/78*	*DCD*	*vv*	*15*	*15*
III	*5/77*	*2/78*	*VV*	*?*	*9*	*13*
IV	*5/77*	*8/77*	*DCD*	*DCD*	*3*	*13*
V	*6/77*	*5/78*	*VV*	*?*	*11*	*12*
VI	*7/77*	*1/78*	*DCD*	*DCD*	*6*	*11*
VII	*8/77*	*3/78*	*VV*	*?*	*7*	*10*

Etat DDN : Etat à la DDN
Etat DP : Etat à la date de point
T.Par : Temps de participation
DCD : décédé
VV : vivant

Figure I : Représentation graphique des données de survie.

Remarque 1.1.1 *Les observations des sujets perdus de vue et ceux exclus vivant sont censurées, mais de deux mécanismes différents de censures.*

1.1.1 Cas et types de censure

On parle de données censurées lorsque l'information disponible sur : la date d'origine, la date d'événement d'intérêt ou la présence de l'événement d'intérêt est incomplète. Généralement, les données sont censurées lorsqu'il y a des informations partielles sur la valeur d'une variable. Elles se présentent quand le chercheur ne dispose pas d'assez de temps pour attendre que toutes les observations atteignent l'événement d'intérêt.

Définition 1.1.1 *La durée de survie T est dite censurée si elle n'est pas intégralement observée.*

Pour l'individu i; considérons :

- le temps de survie T_i : c'est le temps écoulé entre la date d'origine et la data de l'événement d'intérêt.

- le temps de censure C_i : c'est à dire la durée d'observation de l'individu entre le début de l'étude et la date de visite.

Cas de censure

Censure à droite :

On parle de la censure à droite lorsque on observe la censure C (et non pas la durée de vie T) et que nous savons que $T > C$. Ce modèle est le plus fréquent en pratique, il est par exemple adapté au cas où l'évènement d'intérêt est le temps de survie à une maladie et où la date de fin de l'étude est préalablement fixée ; les patients vivants à la fin de l'étude fournissent des données censurées à droite.

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 d'intérêt est inférieure à une certaine date connue $T < C$. En fiabilité, l'exemple d'une telle situation est celui d'un composant électronique monté en parallèle avec un ou plusieurs autres composants. Une panne de ce composant n'entraîne pas nécessairement l'arrêt du système : le système peut continuer à fonctionner jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). La durée observée pour ce composant est alors censurée à gauche.

Censure double

Dans un même échantillon on peut trouver des données censurées à droite et d'autre censurées à gauche. Par exemple, l'étude qui s'intéresse à l'âge auquel les enfants apprennent a accomplir certaines tâches. Au début de l'étude, certains enfants savaient déjà effectuer les tâches étudiées, on sait seulement alors que l'âge où ils ont appris est inférieur a leur âge à la date du début de l'étude. A la fin de l'étude, certains enfants ne savaient pas encore accomplir ces tâches et on sait alors seulement que l'âge auquel ils apprendront éventuellement ont appris est supérieur à leur âge à la fin de l'étude. L'âge au début de l'étude (variable de censure à gauche L) est évidemment inférieure à l'âge à la fin de l'étude (variable de censure à droite C). L'âge d'intérêt est observé ssi il se trouve dans la période d'étude.

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 $C_1 \leq T \leq C_2$. Ceci se produit notamment si un patient se rend à l'hôpital à des dates régulières : s'il ne se présente pas à un rendez-vous, on sait seulement que son décès s'est produit dans l'intervalle entre la dernière visite et le rendez-vous.

Censure mixte

On dit qu'il y a censure mixte lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient, au lieu d'observer un échantillon de la variable d'intérêt Y, on observe un échantillon du couple (Z, A) avec

$$Z = \max(\min(T; C); L)$$

et

$$A = \begin{cases} 0, \text{ si } L < T < C \\ 1, \text{ si } L < C < T \\ 2, \text{ si } \min(T, C) \leq L, \end{cases}$$

où L et C sont des variables de censure et A est l'indicateur de censure.

Remarque 1.1.2 *Revenons au premier exemple, les données sont censurées à droite pour les sujets II, III, V et VII. Pour le sujet II, on n'a pas à tenir compte de ce qui se passe au-delà de la date de point puisque la surveillance est censée s'arrêter le 01/06/1978 date de l'analyse ; il sera considéré vivant à la date de point (exclu-vivant). L'état des sujets III, V, VII est inconnu à la date de point (perdus de vue).*

Différents types de censure

Les catégories de censure décrites ci-dessus peuvent se décliner en fonction du mode ou type de censure. On obtient alors les mécanismes suivants :

Censure de type I :(Fixe)

L'expérimentateur fixe une date (non aléatoire) de fin d'expérience. La durée de participation maximale est alors fixée (non aléatoire) et vaut, pour chaque observation, la différence entre la date de fin d'expérience, et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quand à lui, aléatoire. Ce modèle est souvent utilisé dans les études épidémiologiques. Pour une censure à date fixe. $C_i = C, i = 1, .., n$, on a :

• *Une censure à droite de type* **I :** la durée de survie n'est pas observable au delà d'une durée maximale D fixée. C'est à dire, au lieu d'observer les variables aléatoires qui nous intéressent $T_1, ..., T_n$, on observe que T_i uniquement lorsque $T_i \leq D$, sinon on sait que $T_i > D$. On utilise la notation

$$Z_i = \min(Ti, D), i = 1, ..., n$$

Censure de type II :(Attente)

L'expérimentateur fixe à priori le nombre d'événements d'intérêt à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements d'intérêt étant, quant à lui, non aléatoire. Au lieu d'observer $T_1, ..., T_n$ on observe $T_1 \leq T_2 \leq ... \leq T_k$.

• *Une censure à droite de type* **II :** Par exemple on décide d'observer les durées de survie de n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $T_{(i)}$ et $Z_{(i)}$ les statistiques d'ordres des variables aléatoires T_i et Z_i. La data de censure est donc $Z_{(k)}$ et on observe les variables suivantes

$$Z_{(1)} = T_{(1)}, Z_{(2)} = T_{(2)}, ..., Z_{(k)} = T_{(k)}, Z_{(k)} = T_{(k+1)}, ..., Z_{(k)} = T_{(n)}.$$

Censure de type III : (aléatoire de type I)

Considère que les temps de censure (C_i ou $C_g^{(i)} ou\ C_d^{(i)}$) sont des v.a. ayant une loi (connue ou inconnue) et dépendantes ou indépendantes des temps d'évènement T_i . Cela permet de construire des modèles intéressants qui s'adaptent de manière pertinente aux problèmes posés. Dans le cas de censure de type aléatoire, on considère les cas de censures à droite à gauche et par intervalle(s).

• *Une censure à droite de type* **III :** Soit D une variable aléatoire de censure, au lieu d'observer la variable aléatoire T qui nous intéresse, on observe le couple de variables aléatoires (Z, δ) avec

$$Z = \min(T, D),$$

et

$$\delta = I_{\{T \leq D\}} = \begin{cases} 1, \text{ si } T \leq D \text{ (pas de censure, on observe les données complètes)} \\ 0, \text{ si } T > D \text{ (il y a une censure à droite).} \end{cases}$$

• *Une censure à gauche de type* **III :** Soit G une variable aléatoire de censure, au lieu d'observer la variable aléatoire T qui nous intéresse, on observe pour chaque individu le couple de variables aléatoires (Z, δ) avec

$$Z = \max(T, G),$$

et

$$\delta = I_{\{T \geq G\}} = \begin{cases} 1, \text{ si } T \geq G \text{ (pas de censure, on observe les données complètes)} \\ 0, \text{ si } T < G \text{ (il y a une censure à gauche).} \end{cases}$$

• *Une censure par intervalle de type* **III**

La durée T est dite censurée par intervalle si au lieu d'observer $T_1, ..., T_n$ on observe aléatoirement $(Xi, \delta_i), i = 1 \cdots, n$ où

$$X_i = \max[\min(T_i, C_i^{(R)}), C_i^{(L)}],$$

$C_i^{(R)}, C_i^{(L)}$ sont des censures aléatoires, et δ indicatrice de l'évènement $C_i^{(R)} \leq X \leq C_i^{(L)}$ si $\delta_i = 0$ nous dirons dans ce cas que X est censuré à droite. $\delta_i = 1$ nous dirons dans ce cas que X est observé et $\delta_i = -1$ nous dirons dans ce cas que X est censuré à gauche.

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnu (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'événements observés et la durée totale de l'expérience sont aléatoires.

1.2 Distributions de la durée de survie

1.2.1 Durée de survie absolument continue

On suppose que la durée de survie T est une variable aléatoire positive, et absolument continue. Alors sa loi de probabilité peut être définie par l'une des cinq équivalentes fonctions :

Définition 1.2.1 *(fonction de répartition F)*
La fonction de répartition de T, notée F, désigne la probabilité que l'évènement d'intérêt ait lieu avant t. Elle est définie comme suit

$$F(t) = \mathbb{P}(T \leq t),\ t \geq 0.$$

Définition 1.2.2 *(densité de probabilité f)*
La densité de probabilité de T, notée f, est la fonction positive telle que pour tout $t \geq 0$,

$$F(t) = \int_0^t f(x)dx.$$

Si F admet une dérivée au point t, alors f est définie sur $[0, +\infty[$ par

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

Elle désigne que l'évènement d'intérêt ait lieu après t, dans un petit intervalle de temps.

Remarque 1.2.1 *Puisque T est une variable aléatoire absolument continue, les notations $F(t) = \mathbb{P}(T \leq t)$ et $F(t) = \mathbb{P}(T < t)$ sont équivalentes.*

Définition 1.2.3 *(fonction de survie S)*
La fonction de survie S représente la probabilité de survivre au moins jusqu'au temps t. Autrement dit, la probabilité de ne pas avoir fait l'événement d'intérêt jusqu'à l'instant t. Elle est définie comme suit

$$S(t) = \mathbb{P}(T \geq t) = 1 - F(t), t \geq 0.$$

Remarque 1.2.2 *$S(t)$ est une fonction monotone décroissante et continue telle que $S(0) = 1$ et $lim_{t\to+\infty}S(t) = 0$.*

Définition 1.2.4 *(taux de hasard ou risque instantané λ)*

Le taux de risque instantané λ est la probabilité qu'un évènement survienne dans un petit intervalle de temps après t, sachant qu'il n'a pas eu lieu avant t.

$$\lambda(t) = lim_{dt\to 0}\frac{\mathbb{P}(t \leq T < t + dt|T \geq t)}{dt}.$$

Remarque 1.2.3 *1- $\lambda(t)$ est lié à une unité de temps. Si t est en heure, alors $\lambda(t)$ mesure le risque qu'un événement survienne dans l'heure. Ce n'est pas une densité donc son intégrale ne vaut pas nécessairement 1.*

2- $\lambda(t)$ est appelée taux de défaillance ou taux instantané d'avarie en fiabilité. Et dite aussi force de mortalité ou risque instantané de décès dans le domaine de biomédical.

3- Si T représente la durée de séjour dans un état donné, $\lambda(t)$ est la probabilité des passage irréversible d'un état à un autre dans un intervalle de temps $[t, t + dt]$ sachant que le sujet était dans l'état initial jusqu'à la date t. Elle permet de mesurer la vitesse de changement d'état d'un individu à travers le temps.

Définition 1.2.5 *(taux de risque cumulé ou taux de hasard cumulé Λ)*

Le taux de risque cumulé est définit sur l'intervalle $[0, t]$ par :

$$\Lambda(t) = \int_0^t \lambda(s)ds,$$

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

Relations entre les fonctions de distribution de survie

Les fonctions de distribution de probabilité de la durée de survie T sont liées par les relations suivantes :

1-$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)} = \frac{d\Lambda(t)}{dt}$.

2- $\Lambda(t) = -\ln(S(t))$.

3- $S(t) = -\exp(\Lambda(t))$.

4- $F(t) = 1 + \exp(\Lambda(t))$.

5- $f(t) = -\frac{dS(t)}{dt}$.

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

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u) du \\
&= -\ln\left(S(t)\right).
\end{aligned}$$

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

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

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

1.2.2 Paramètres de position et de dispersion associés à la distribution de survie

Paramètres de position

On peut calculer l'espérance et la variance de la durée de survie T en utilisant n'importe quelle des ses fonctions de distribution.

Définition 1.2.6 *(temps moyen de survie)*

L'espérance de la durée de survie T, appelé aussi temps moyen de survie $\mathbb{E}(T)$, est donné par

$$\begin{aligned} \mathbb{E}(T) &= \int_0^{+\infty} t f(t)dt \\ &= -\int_0^{+\infty} t \frac{dS(t)}{dt} dt \\ &= \int_0^{+\infty} S(t)dt. \end{aligned}$$

Définition 1.2.7 *(variance de survie)*
La variance de la durée de survie T, notée $var(T)$, est donnée par

$$\begin{aligned} var(T) &= \mathbb{E}\left(T^2\right) - \left[\mathbb{E}\left(T\right)\right]^2 \\ &= 2\int_0^{+\infty} tS(t)dt - \left[\mathbb{E}\left(T\right)\right]^2. \end{aligned}$$

Paramètres de dispersion

Définition 1.2.8 *(quantiles)*
Les quantiles de la durée de survie pour $0 \leq p \leq 1$, notés Q_p, sont définis par

$$\begin{aligned} Q_p &= \inf\left(t; F(t) \geq p\right) \\ &= \inf\left(t; S(t) \leq 1-p\right). \end{aligned}$$

Q_p représente, le temps où la proportion p d'une population a subi l'évènement d'intérêt.

Définition 1.2.9 *(médiane)*
La médiane $Q_{\frac{1}{2}}$ est le quantile particulier satisfaisant $S(Q_{\frac{1}{2}}) = 0.5$. Désingne que la moitie de la population a subi l'évènement d'intérêt.

1.2.3 Durée de survie discrète

Soit T une variable aléatoire discrète prenant les valeurs ordonnées en ordre croissant $t_0, t_1, ...t_n, ..$ La fonction de survie et le taux de hasard de T sont définis comme suit :

Définition 1.2.10 *La fonction de survie S_T de la variable aléatoire discrète est donnée par :*

$$S_T(t) = \mathbb{P}(T > t) = \sum_{i, t_i > t} \mathbb{P}(T = t_i).$$

Définition 1.2.11 *Le taux de hasard λ_T, de la variable aléatoire discrète, au point t_i est donnée par :*

$$\begin{aligned}\lambda_T(t_i) &= \mathbb{P}(T = t_i | T \geq t) \\ &= \frac{\mathbb{P}(T = t_i)}{S_T(t_{i-1})} \\ &= 1 - \frac{S_T(t_i)}{S_T(t_{i-1})}.\end{aligned}$$

exemple 1.2.1 *Soit T une variable aléatoire prenant les valeurs 1, 2, 3 et 4, avec les probabilités respectives $\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{2}{8}$. Alors la fonction de survie de T est*

$$S_T(t) = \begin{cases} 1, & si\ t < 1 \\ \frac{7}{8}, & si\ 1 \leq t < 2 \\ \frac{5}{8}, & si\ 2 \leq t < 3 \\ \frac{2}{8}, & si\ 3 \leq t < 4 \\ 0, & si\ 4 \leq t. \end{cases}$$

et le taux de hasard est

$$\lambda_T(t) = \begin{cases} \frac{1}{6}, & si\ t = 0 \\ \frac{1}{3}, & si\ t = 1 \\ 1, & si\ \ t = 2. \end{cases}$$

Remarque 1.2.4 *On a pour $S_T(t_0) = 1$*

$$\begin{aligned}S_T(t) &= S_T(t_i, t_i \leq t) \\ &= \prod_{i, t_i \leq t} (1 - \lambda_T(t_i)).\end{aligned}$$

2- Pour une durée de survie T discrète, et pour tout t fixé, on a

** $F_T(t) = \mathbb{P}(T \leq t)$ représente la limite à droite de la fonction de répartition F, au point t.*

** $F_T(t_-) = \mathbb{P}(T < t)$ représente la limite à gauche de la fonction de répartition F, au point t.*

avec $F_T(t_-) < F_T(t)$.

(on a de même pour la fonction de survie $S_T(t_-) = \mathbb{P}(T \geq t)$, et $S_T(t_-) > S_T(t)$)

3- $\lambda_T(t) = \frac{\mathbb{P}(T=t)}{S_T(t_-)}$.

1.3 Fonctions de loi de probabilité pour les données censurées

Dans cette section, on représente la fonction de répartition et la fonction de densité pour les modèles de censure à droite et double.

1.3.1 Pour le modèle de censure à droite

Soient T et C deux variables aléatoires positives et absolument continue de densités de probabilités respective f_T, f_C , de fonctions de survies S_T, S_C et de fonctions de répartitions F_T, F_C.

Sous l'hypothèse que T et C sont indépendants, on considère le cas de la censure aléatoire à droite. Les observations sont les couples $(Z_1; \delta_1), ..., (Z_n; \delta_n)$ où $Z_i = \min(T_i; C_i)$ et $\delta_i = I_{\{T_i \leq C_i\}}$.

$$\delta_i = \begin{cases} 1, \text{ si } Z_i = T_i \text{ ,on observe la durée de survie (pas de censure).} \\ 0, \text{ si } Z_i = C_i \text{ ,on observe } C_i \text{ ,il y a une censure.} \end{cases}$$

La loi de probabilité des données observées est définie comme suit :
La fonction de répartition :

$$\begin{aligned} F_{(Z;\delta)}(z,0) &= \mathbb{P}(Z \leq z, \delta = 0) \\ &= \mathbb{P}(\min(T,C) \leq z, T > C) \\ &= \mathbb{P}(C \leq z, T > C) \\ &= \int_0^z \int_t^\infty dF_T(u) dF_C(t) \\ &= \int_0^z S_T(t) dF_C(t), \end{aligned}$$

donc la densité est

$$f_{(Z;\delta)}(z,0) = S_T(z) f_C(z).$$

Aussi

$$\begin{aligned} F_{(Z;\delta)}(z,1) &= \mathbb{P}(Z \leq z, \delta = 1) \\ &= \mathbb{P}(\min(T,C) \leq z, T \leq C) \\ &= \mathbb{P}(T \leq z, T \leq C) \\ &= \int_0^z \int_t^\infty dF_C(u) dF_T(t) \\ &= \int_0^z S_C(t) dF_T(t), \end{aligned}$$

et

$$f_{(Z;\delta)}(z,1) = S_C(z) f_T(z).$$

Par conséquence : la densité pour $\delta_i = \{0,1\}, \forall i = 1, ..., n$ est

$$f_{(Z_i;\delta_i)}(z_i,\delta_i) = (S_{C_i}(z) f_{T_i}(z))^{\delta_i} (S_{T_i}(z) f_{C_i}(z))^{1-\delta_i}.$$

Remarque 1.3.1 *Pour la fonction de loi dans le cas de censure à gauche on trouve la même fonction du cas de censure à droite.*

1.3.2 Pour le modèle de censure double

En plus de T et C, soit L une variable aléatoire non négative et absolument continue de densité f_L, la fonction de survies est notés S_L, La fonction de répartition est noté F_L.

Sous l'hypothèse que L est indépendante de T et C, on considère le cas de la censure aléatoire double , Les observations sont les couples $(Z_1; A_1), ..., (Z_n; A_n)$ où $Z_i = \max(\min(T_i; C_i), L_i)$.

$$A_i = \begin{cases} 0, \text{ si } L_i \leq T_i \leq C_i \text{ ,alors } Z_i = T_i \text{ , (pas de censure)} \\ 1, \text{ si } C_i \leq T_i, \text{alors } Z_i = C_i \text{ , (il y a censure à droite)} \\ 2, \text{ si } T_i \leq L_i, \text{alors } Z_i = L_i \text{ , (il y a censure à gauche)} \end{cases}.$$

La loi de probabilité des données observées est définie comme suite :

La fonction de répartition :

$$\begin{aligned} F_{(Z;A)}(z,0) &= \mathbb{P}(Z \leq z, A = 0) \\ &= \mathbb{P}(\max(\min(T;C), L) \leq z, L \leq T \leq C) \\ &= \mathbb{P}(T \leq z, L \leq T \leq C) \\ &= \int_0^z \int_0^t dF_L(u) \int_t^\infty dF_C(u) dF_T(t) \\ &= \int_0^z F_L(t) S_C(t) dF_T(t), \end{aligned}$$

donc la densité est

$$f_{(Z;A)}(z,0) = F_L(t) S_C(z) f_T(z).$$

Aussi

$$\begin{aligned} F_{(Z;A)}(z,1) &= \mathbb{P}(Z \leq z, A = 1) \\ &= \mathbb{P}(\max(\min(T;C), L) \leq z, C < T) \\ &= \mathbb{P}(C \leq z, C < T) \\ &= \int_0^z \int_t^\infty dF_T(u) dF_C(t) \\ &= \int_0^z S_T(t) dF_C(t). \end{aligned}$$

Donc

$$f_{(Z;A)}(z,1) = S_T(z) f_C(z).$$

Et

$$\begin{aligned} F_{(Z;A)}(z,2) &= \mathbb{P}(Z \leq z, A = 2) \\ &= \mathbb{P}(\max(\min(T;C),L) \leq z, T < L) \\ &= \mathbb{P}(L \leq z, T < L) \\ &= \int_0^z \int_t^\infty dF_T(u) dF_L(t) \\ &= \int_0^z S_T(t) dF_L(t), \end{aligned}$$

Alors

$$f_{(Z;A)}(z,2) = S_T(z) f_L(z).$$

En résumé on obtient la densité pour $i = \{0,1,2\}$

$$f_{(Z;a)}(z,i) = (F_L(t)S_C(z)f_T(z))^{I_{\{A=0\}}(i)}(S_T(z)f_C(z))^{I_{\{A=1\}}(i)}(S_T(z)f_L(z))^{I_{\{A=2\}}(i)}.$$

Dans ce modèle T est observée ssi $T \in [L;C]$ et une donnée censurée soit à droite soit à gauche mais pas les deux à la fois.

Chapitre 2

Estimation paramétrique sur les données censurées

Un modèle paramétrique est un modèle dans lequel les temps de survie sont supposés être distribués selon une loi paramétrique parfaitement connue. 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 probabilité de la durée de survie : 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 distributions les plus couramment utilisées sont : Exponentielle, Weibull, Gompertz, Log-logistique, Pareto.

Les estimateurs des paramètres du modèle sont ensuite obtenus par la méthode du maximum de vraisemblance.

Avantages de l'approche paramétrique

Théoriquement, une courbe de survie est une fonction valant 1 au temps 0 et 0 à l'infini : $S(0) = 1$ et $S(\infty) = 0$. Lorsque cette courbe est approximée soit par des méthodes de type Kaplan-Meier, soit par un modèle semi-paramétrique, le résultat peut être un peu différent pour deux raisons :

1- S'il y a peu de données à disposition, la "courbe" aura plutôt une forme en escalier.

2- S'il reste des sujets à risque de subir l'événement en fin d'étude, alors la courbe de survie n'atteindra pas son minimum de zéro.

Ces problèmes n'existent pas avec les modèles paramétriques.

Désavantages de l'approche paramétrique

Pour utiliser l'approche paramétrique, il faut avoir de bonnes raisons de penser que les temps de survie sont approximativement distribués selon une

certaine loi connue plutôt qu'une autre.

Les modèles paramétriques reposent sur les hypothèses suivantes :

PH (Proportional Hazard), c'est-à-dire que le rapport des risques pour deux individus est constant et indépendant du temps (signifie en pratique que l'effet des variables explicatives est multiplicatif par rapport au hasard.)

AFT (Acceleration Failure Time model), qui signifie que l'effet des variables explicatives est multiplicatif par rapport au temps de survie.

Il y a deux approches principales permettant de choisir le modèle théorique le plus adapté aux données :

1- L'approche graphique consiste à comparer la distribution effective du risque ou de la survie avec la distribution théorique suggérée par les différentes lois et à choisir le modèle dont on est le plus proche.

2- L'approche par ajustement consiste à estimer les différents modèles et à choisir celui qui s'ajuste le mieux aux données sur la base du coefficient d'information d'Akaike (AIC).

2.1 Distributions Théoriques

2.1.1 Pour un risque instantané constant (loi exponentielle)

La loi exponentielle $\mathcal{E}(\theta)$, 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 probabilité de décès pour un individu dans un certain laps de temps est la même quelle que soit sa durée de vie (i.e. $P(X > s + t|X > t) = P(X > s)$). Les quantités associées à cette loi sont :

$$\begin{aligned} f(t) &= \theta e^{-\theta t}, && t \geq 0 \text{ et } \theta > 0, \\ \lambda(t) &= \theta, \\ S(t) &= e^{-\theta t}. \end{aligned}$$

Plus le risque θ est grand, plus l'espérance de survie est faible.

Dans certaines applications, on peut découper le temps en plusieurs intervalles et considérer un θ_i différent pour chacun des intervalles (le risque est constant sur chaque période mais varie d'une période à une autre).

Ses caractéristiques sont :

$$E(T) = \frac{1}{\theta} \text{ et } Var(T) = \frac{1}{\theta^2}.$$

exemple 2.1.1 *Soit t $(t > 0)$ le temps et μ $(\mu > 0)$ la durée moyenne de survie d'un individu, alors le modèle paramétrique de survie le plus simple correspond à un modèle exponentiel dont la fonction de densité est :*

$$f(t) = \frac{1}{\mu} \exp(-\frac{1}{\mu}t).$$

Soit F la fonction de répartition associée à cette fonction de densité :

$$\begin{aligned}
F(t) &= P(T \leq t) \\
&= \int_{-\infty}^{t} f(x)dx \\
&= \int_{-\infty}^{0} f(x)dx + \int_{0}^{t} f(x)dx \\
&= \int_{0}^{t} f(x)dx \quad car\ T\ est\ positive \\
&= \int_{0}^{t} \frac{1}{\mu} \exp(-\frac{1}{\mu}x)dx \\
&= [-\exp(-\frac{t}{\mu}) - (-\exp(-\frac{0}{\mu}))] \\
&= 1 - \exp(-\frac{1}{\mu}t).
\end{aligned}$$

Soit S la fonction de survie :

$$\begin{aligned}
S(t) &= P(T > t) \\
&= 1 - P(T \leq t) \\
&= 1 - F(t) \\
&= 1 - (1 - \exp(-\frac{1}{\mu}t)) \\
&= \exp(-\frac{1}{\mu}t).
\end{aligned}$$

Soit λ la fonction de risque :

$$\begin{aligned}\lambda(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\frac{1}{\mu}\exp(-\frac{1}{\mu}t)}{\exp(-\frac{1}{\mu}t)} \\ &= \frac{1}{\mu}.\end{aligned}$$

A partir de ce modèle exponentiel simple dont la fonction λ de risque est constante au cours du temps t, on peut tracez les courbes des fonctions de densité, de répartition, de survie, et de risque pour un organisme dont la durée moyenne de survie est égale à 2 semaines ($\mu = 2$) :

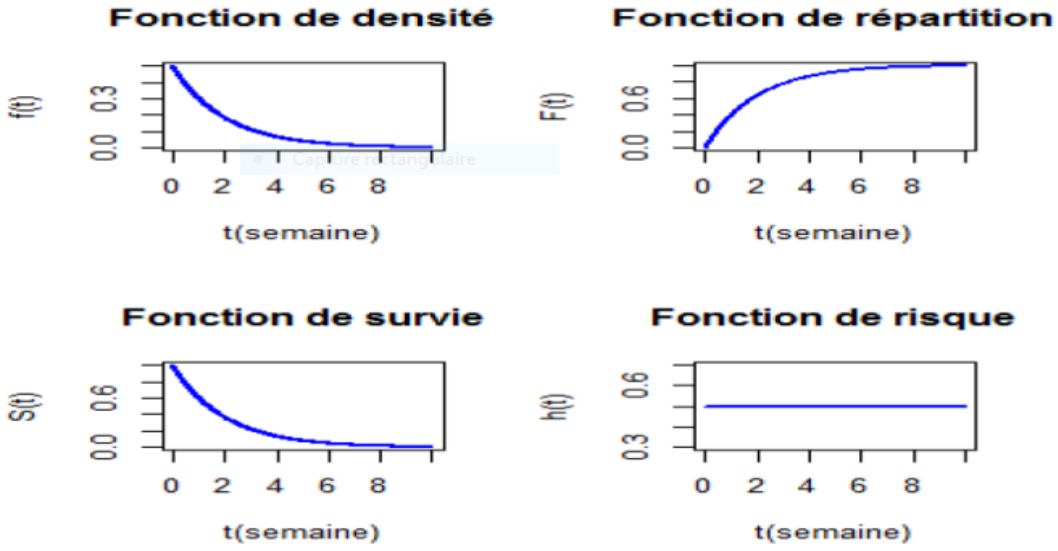

Figure II : Représentation graphique des fonctions de survie d'une loi exponentielle.

2.1.2 Pour un risque instantané monotone

Loi de *Weibull*

La distribution de *Weibull* est également une généralisation de la loi exponentielle ($\gamma = 1$). Elle est caractérisée par deux paramètres, $\gamma > 0$ et $\alpha > 0$, qui sont les paramétres de forme et d'échelle respectivement. La fonction de densité de cette loi est donnée par

$$f(t;\gamma,\alpha) = \alpha\gamma(\alpha t)^{\gamma-1}\exp(-(\alpha t)^{\gamma}), \qquad t > 0.$$

La fonction de survie est

$$S(t) = \exp(-(\alpha t)^{\gamma}),$$

et la fonction de taux de hasard vaut

$$\lambda(t) = \alpha\gamma(\alpha t)^{\gamma-1},$$

qui est donc de l'ordre de $t^{\gamma-1}$.

La fonction de risque est monotone croissante si $\gamma > 1$, monotone décroissante si $\gamma < 1$ et constante pour $\gamma = 1$, c'est pourquoi ce paramètre est appelé paramètre de forme. Comme α est un paramètre d'échelle, différentes valeurs de changent seulement l'échelle sur l'axe horizontal, et non pas la forme de base du graphe. Le modèle est assez flexible, et on a montré qu'il constitue une bonne description de plusieurs types de données de survie. Le fait que les fonctions de densité, de survie et de risque aient une forme relativement simple explique également la popularité du modèle.

Ses caractéristiques sont :

$$E(T) = \Gamma\left(\frac{1+\gamma}{\gamma}\right)/\lambda \quad \text{et} \quad Var(T) = \left(\Gamma\left(\frac{2+\gamma}{\gamma}\right) - \Gamma\left(\left(\frac{1+\gamma}{\gamma}\right)\right)^2\right)/\lambda^2,$$

où $\Gamma(y)$ est la fonction gamma :

$$\begin{aligned}\Gamma(y) &= (y-1)\Gamma(y-1)\\ &= \int_0^t x^{y-1}e^{-x}dx.\end{aligned}$$

Loi Gamma

La loi gamma comporte deux paramètres, $\beta > 0$ et $\gamma > 0$. α est appelé paramètre d'échelle et γ est le paramètre de forme. La fonction de densité de cette loi est donnée par

$$f(t, \beta, \gamma) = \frac{\beta}{\Gamma(\gamma)}(\beta t)^{\gamma-1} \exp(-(\beta t)) \ , t > 0,$$

où

$$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x} dx.$$

La fonction de survie s'exprime comme

$$S(t) = \int_t^\infty \frac{\beta}{\Gamma(\gamma)}(\beta x)^{\gamma-1} \exp(-(\beta x)) dx.$$

La fonction de taux de hasard d'une loi gamma en fonction du paramètre d'échelle β.

$$\lambda(t) = \frac{\beta(\beta t)^{\gamma-1}}{(\gamma-1)! \sum_{\kappa=0}^{\gamma-1} \frac{1}{\kappa!}(\beta t)^\kappa}.$$

En choisissant le paramètre γ entier, nous obtenons la distribution dite d'Erlang.

Ses caractéristiques sont :

$$E(T) = \lambda\gamma \ \text{ et } \ Var(T) = \lambda^2\gamma.$$

Loi de *Gompertz* (*Makeham*)

La loi de *Gompertz* Cette distribution s'obtient lorsque le risque varie de façon proportionnelle à sa valeur. Elle est très utilisée pour la distribution du taux de mortalité.

α : mortalité de base (paramètre d'échelle)

γ : influence de l'âge (paramètre de forme)

La densité de probabilité de la loi de *Gompertz* est donnée par :

$$f(t,\alpha,\gamma) = \alpha\gamma \exp(-(\alpha t)) \exp(-\gamma \exp(-(\alpha t))),$$

La fonction de taux de hasard d'une loi de Gompertz

$$\lambda(t) = \alpha \exp(\gamma t).$$

Ses caractéristiques sont :

$$E(T) = \frac{1}{\alpha}(\ln \gamma - \psi(1)) \text{ et } Var(T) = \frac{1}{\alpha^2}\psi^{(1)}(1).$$

La loi de *Gompertz -Makeham* *Makeham* ajoute un paramètre supplémentaire $\beta > 0$ pour tenir compte de la mortalité accidentelle.

La fonction de taux de hasard de la loi de *Gompertz -Makeham* devient

$$\lambda(t) = \beta + \alpha \exp(\gamma t) \qquad (t \geq 0),$$

La fonction de survie s'écrite comme

$$S(t) = \exp(-\beta t - \frac{\alpha}{\gamma}(\exp(\gamma t) - 1),$$

Et la densité de probabilité de la loi de *Gompertz -Makeham* devient

$$f(t,\alpha,\gamma) = (\beta + \alpha \exp(\gamma t))(\exp(-\beta t - \frac{\alpha}{\gamma}(\exp(\gamma t) - 1)).$$

Remarque 2.1.1 *Il existe de nombreuses lois avec des risques monotones, citons notamment les mélanges de deux distributions exponentielles, les lois de Weibull exponentielles.*

2.1.3 Risque instantané constant pour chaque individu mais variant entre individus selon une loi gamma (loi de Pareto)

Soit une loi gamma $\gamma(a,p)$ de moyenne $\lambda_0 = \frac{a}{P}$,
La densité de la loi de Pareto est

$$\begin{aligned} f(\lambda) &= \frac{p^a}{\Gamma(a)}\lambda^{a-1}\exp(-p\lambda), \\ &= \frac{(\frac{a}{\lambda_0})^a}{\Gamma(a)}\lambda^{a-1}\exp(-\frac{a\lambda}{\lambda_0}), \end{aligned}$$

$$f(t) = a\left(\frac{a}{\lambda_0}\right)^a \left(t + \frac{a}{\lambda_0}\right)^{-(a+1)}.$$

La fonction de taux de hasard de la loi s'écrite comme

$$h(t) = a\left(t + \frac{a}{\lambda_0}\right)^{-1},$$

La fonction de survie est

$$S(t) = \left(\frac{a}{\lambda_0}\right)^a \left(t + \frac{a}{\lambda_0}\right)^{-a},$$

Caractéristiques :

$$E(T) = \frac{a\lambda_0}{a-1} \qquad \text{pour } a > 1,$$

$$Var(T) = \frac{\lambda_0^2 a}{(a-1)^2(a-2)} \qquad \text{pour } a > 2.$$

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

Loi log-normale

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

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

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 $a = \exp(-\mu)$, alors $-\mu = \log a$ et nous obtenons

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

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

$$S(t) = 1 - \Phi\left(\log\left(\frac{at}{\sigma}\right)\right),$$

où

$$\Phi(y) = \frac{1}{\sqrt{2\sqrt{\pi}}} \int_{-\infty}^{y} \exp^{-u2} /^2 du,$$

est la fonction de distribution d'une loi normale standard centrée réduite. La fonction de taux de hasard est de la forme

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

Nous pouvons montrer que $\lambda(t) = 0$ pour $t = 0$, que $\lambda(t)$ croît jusqu'à un maximum et ensuite décroît et tend vers 0 lorsque $t \to \infty$. Comme la fonction de risque décroît pour de grandes valeurs de t , la distribution ne paraît pas plausible comme modèle de vie dans la plupart des situations. Malgré cela, ce modèle peut être intéressant lorsque de très grandes valeurs de t ne sont pas d'un intérêt particulier.

Ses caractéristiques sont :

$$E(T) = \exp\left(\frac{\mu + \sigma^2}{2}\right) \quad \text{et} \quad Var(T) = \exp\left(\sigma^2 - 1\right) \exp\left(2\mu + \sigma^2\right).$$

Loi log-logistique

La distribution des temps d'événement est log-logistique si la densité de T est :

$$f(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1 + \alpha t\gamma)^2}.$$

Les fonctions de survie et de risque associées sont respectivement :

$$S(t) = \frac{1}{(1 + \alpha t\gamma)} \quad \text{et} \quad \lambda(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1 + \alpha t\gamma)}.$$

On rappelle que si X est une variable aléatoire log-logistique, alors $Y = \log X$ a une distribution logistique. Rappelez-vous également que des évolutions similaires de la fonction de risque peuvent être obtenues avec la distribution log-normale.

Caractéristiques :

$$E(T) = \frac{\alpha\pi/\gamma}{\sin(\pi/\gamma)} \text{ si } \gamma > 1, \text{ sinon pas définie.}$$

posant $b = \pi/\gamma$

$$Var(T) = \alpha^2(2b/\sin 2b - b^2/\sin^2 b \text{ avec } \gamma > 2.$$

Remarque 2.1.2 *Il existe de nombreuses lois avec des risques instantané en $\cap$ et $\cup$, citons notamment Gaussienne inverse, Weibull généralisée permettent de considérer des risques instantanés en forme de $\cap$.*

2.2 Choix Entre Distributions

Il y a deux approches principales permettant de choisir le modèle théorique le plus adapté aux données :

2.2.1 Approche graphique

L'approche graphique consiste à comparer la distribution effective du risque ou de la survie avec la distribution théorique suggérée par les différentes lois et à choisir le modèle dont on est le plus proche.

Choix entre distributions théoriques

Tableau II : distributions théoriques et propriété.

Modèle	Fonction	Propriété
Exponentiel	$h(t), H(t)$	indépendante de t, linéaire en t
Weibull	$\ln(-lnS(t))$	linéaire en $\ln t$
Log-logistique	$\ln(1 = S(t) - 1)$	linéaire en $\ln t$
Gompertz	$\ln h(t), ln(ln[\Delta S(t)])$	linéaire en t
Migrant-sédentaire	$\ln[\Delta S(t)]$	linéaire en t
Pareto	$\frac{1}{h(t)}$	linéaire en t

En pratique, on fait des présentations graphiques des transformées des estimations $\widehat{S}_k$ et $\widehat{h}_k$ en fonction de t et on compare avec les propriétés attendues des différentes lois.

Si les intervalles $[t_{k-1}; t_k)$ sont de longueur variable (en particulier avec Kaplan-Meier)) $\Rightarrow$ ajuster les $\widehat{h}_k$ et $\Delta\widehat{S}_k$ pour obtenir des valeurs se rapportant à une unité de temps :

$$\widehat{h}(t_k) = \frac{2\widehat{h}_k}{(t_{k+1} - t_k)\left(2 - \widehat{h}_k\right)} \quad \text{et} \quad \Delta\widehat{S}(t_k) = \widehat{f}(t_k) = \frac{\widehat{S}_{k-1} - \widehat{S}_k}{t_k - t_{k-1}}.$$

exemple 2.2.1 *Dans cet exemple simplifié, la comparaison ciblée de modèles est utilisée pour l'approche graphique.*

Les distributions paramétriques de complexité croissante sont ajustés à un ensemble de 686 temps de survie censurés à droite de patients avec un nœud primaire positif cancer du sein (originaire de Sauerbrei et Royston (1999)).

Les modèles exponentiels, Weibull et gamma généralisé sont ajustés, qui ont un, deux et trois paramètres respectivement.

Le graphique suivant compare les courbes de survie ajustées de chaque modèle (lignes colorées), avec l'estimation de Kaplan-Meier, en noir. La courbe de survie ajustée du modèle gamma généralisé semble correspondre le plus étroitement à l'estimation de Kaplan-Meier au cours des 7 années de suivi vers le haut.

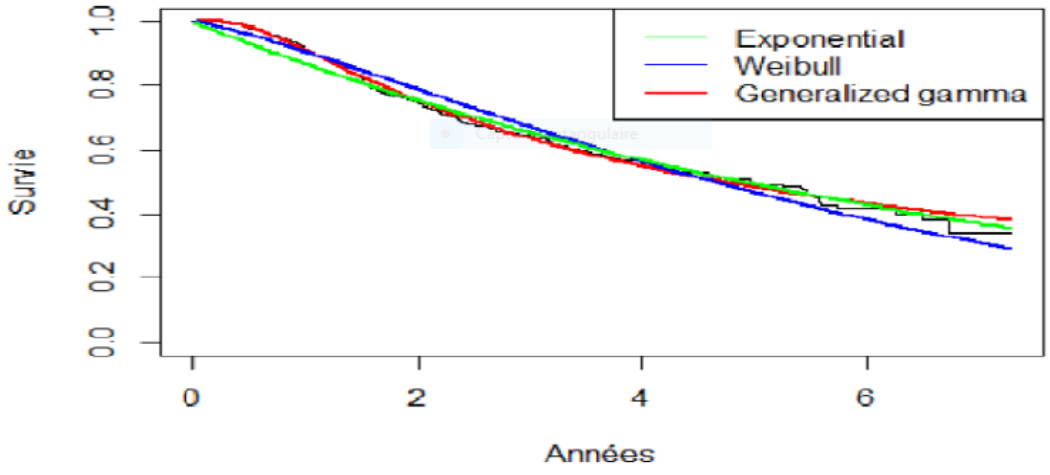

Figure III : Représentation graphique des courbes de survie.

Chaque modèle est une généralisation du précédent. Par conséquent, nous pouvons utiliser une comparaison de modèles ciblée, le modèle "large" étant considéré comme le gamma généralisée, et évaluez si l'utilisation d'un modèle plus simple conduit à des améliorations de la précision des estimations qui l'emportent sur tout biais.

2.2.2 Approche par ajustement (AIC)

L'approche par ajustement consiste à estimer les différents modèles et à choisir celui qui s'ajuste le mieux aux données sur la base du coefficient d'information d'Akaike (AIC).

Définition 2.2.1 *Le critère d'information d 'Akaike est défini comme*

$$AIC = -2LL(M) + 2k$$

Avec LL est la log-vraisemblance et k est le nombre de paramètres du modèle M.

Le modèle minimisant ce critère est le modèle offrant le meilleur ajustement aux données.

Remarque 2.2.1 *Etant donné que les vraisemblances qui sont maximisées pour l'obtention des paramètres d'un modèle de Cox et d'un modèle paramétrique sont différentes, il n'est pas possible de comparer ces deux types de modèles à l'aide d'AIC.*

exemple 2.2.2 ***Données biographiques allemandes***

Données extraites de l'enquête biographique allemande réalisée entre 1981 et 1983 (Mayer & Brückner, 1989) et utilisées notamment par (Blossfeld & Rohwer, 2002). Trois cohortes de naissance : 1929-1931 (coho1),1939-1941 (coho2), 1949-1951 (coho3). Echantillon de $n = 600$ emplois.

Comment le niveau d'éducation (edu), l'expérience sur le marché du travail (lfx), le nombre d'emplois précédents (pnoj) et le prestige de l'emploi (pres) influencent-ils : le risque de terminer un emploi ? la durée de l'em-

ploi ?. On souhaite aussi contrôler les effets par cohorte

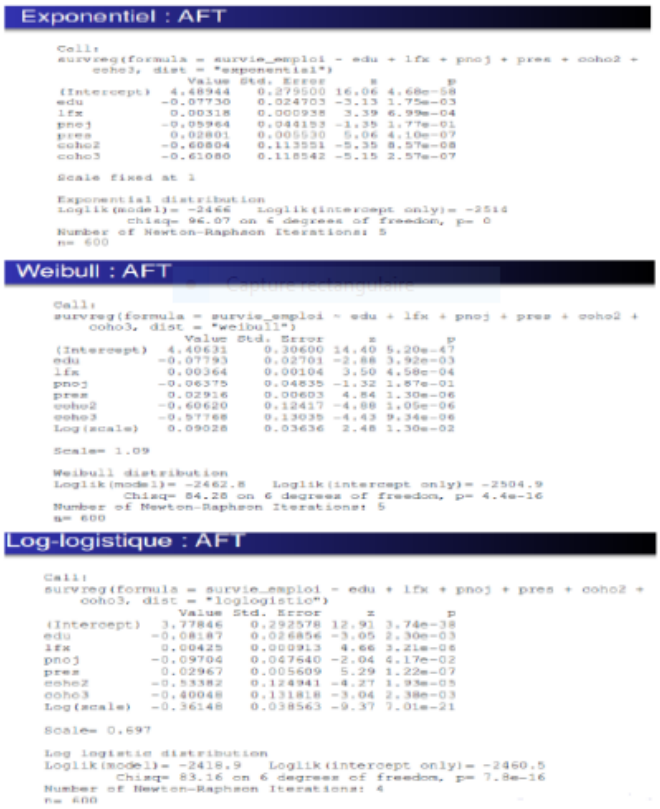

Figure IV : Modèles de vie accélérée (AFT) Exponentiel, Weibull et Log-logistique.

Modèle	*LL*	*k*	*AIC*
Exponentiel	-2466	7	4932
Weibull	-2462.8	8	4941.6
Log-logistique	-2418.9	8	4853.8

Tableau III : Comparaison des résultats des Modèles de vie accélérée (AFT).

On remarque que le modèle minimisant le critère AIC est le modèle Log-logistique car il offre le meilleur ajustement aux données.

2.3 Estimation par la méthode du maximum de vraisemblance

2.3.1 Méthode du maximum de vraisemblance dans le cas complet

Nous rappelons brièvement la méthode du maximum de vraisemblance pour estimer les paramètres réels d'un modèle d'analyse des durées de vie.

Définition 2.3.1 *Soit $(T_1; ...; T_n)$ un échantillon indépendant et identiquement distribué (i.i.d.) de densité f_θ. On appelle vraisemblance l'application L définie par*

$$L(t_1, ..., t_n, \theta) = \prod_{i=1}^{n} f_{T_i}(t, \theta).$$

Estimateur du maximum de vraisemblance

Définition 2.3.2 *Soit $(T_1; ...; T_n)$ un échantillon de T. Un estimateur du maximum de vraisemblance (EMV) du paramètre θ est un estimateur $\widehat{\theta}$ qui maximise la vraisemblance $L(t_1; ...; t_n; \theta)$, c'est-à-dire, on cherche à trouver*

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}} L(t_1; ...; t_n; \theta).$$

En général, on maximise la fonction log-vraisemblance du fait qu'il transforme le produit en somme et qu'il ne déforme pas les représentations graphiques.

$$\begin{aligned} \ln L(t_1; ...; t_n; \theta) &= \ln \left\{ \prod_{i=1}^{n} f_{T_i}(t, \theta) \right\} \\ &= \sum_{i=1}^{n} \ln f_{T_i}(t, \theta). \end{aligned}$$

Dans la suite, nous allons définir cette méthode d'estimation de paramètres d'une loi, brièvement dans le cas en présence de censure.

2.3.2 Méthode d'estimation par maximum de vraisemblance (EMV)(En présence de censure) :

Nous rappelons la méthode du maximum de vraisemblance pour estimer, au vu de données censurées, les paramètres réels d'un modèle d'analyse des durées de vie.

Nous considérons le cas de la censure aléatoire à droite. Les observations sont les couples $(Z_1, \delta_1), ..., (Z_n, \delta_n)$ où

$$Z_i = min(T_i, C_i) \text{ et } \delta_i = I(T_i \leq C_i).$$

- Si $\delta_i = 1$ alors $Z_i = T_i$: on observe la durée de vie.
- Si $\delta_i = 0$ alors $Z_i = C_i$: il y a censure.

Hypothèse Fondamentale : On suppose que le délai de censure C_i de l'individu i est une variable aléatoire indépendante de la durée de vie T_i.

Proposition 2.3.1 *Sous l'hypothèse fondamentale d'indépendance $T_i \amalg C_i$, pour $i = 1, ..., n$ la vraisemblance s'écrit :*

$$L((z_1, \delta_1), ..., (z_n, \delta_n), \theta) = \prod_{i=1}^{n} f_\theta(z_i)^{\delta_i} S_\theta(z_i)^{1-\delta_i},$$

où f_θ est la densité commune des T_i et S_θ la fonction de survie associée.

Preuve

Soit la suite des délais de censure $C_1, ..., C_n$ i.i.d. de densité commune g et G la survie associée i.e. $G(c) = P(C_1 > c)$.

Le couple de v.a. (Z_i, Δ_i) admet pour densité :

$$\begin{cases} g(z_i)S_\theta(z_i), \text{ si } \delta_i = 0 \text{ (observations censures)}; \\ f_\theta(z_i)G(z_i), \text{ si } \delta_i = 1 \text{ (dures } t_i = z_i \text{ observes)}. \end{cases}$$

que l'on peut aussi écrire de façon équivalente :

$$[f_\theta(z_i)G(z_i)]^{\delta_i}[g(z_i)S_\theta(z_i)]^{1-\delta i},$$

de plus les couples (Z_1, Δ_1),...,(Z_n, Δ_n) sont indépendants donc la vraisemblance des observations s'écrit :

$$\prod_{i=1}^{n}[f_\theta(z_i)G(z_i)]^{\delta_i}[g(z_i)S_\theta(z_i)]^{1-\delta i},$$

Comme la loi des C_i ne fait pas intervenir le paramètre θ, la partie utile de la vraisemblance se réduit à :

$$L(\theta) = \prod_{i=1}^{n} f_\theta(z_i)^{\delta i} S_\theta(z_i)^{1-\delta i}.$$

Ecriture de la log-vraisemblance en fonction du risque :

Proposition 2.3.2 *Sous l'hypothèse fondamentale d'indépendance* $T_i \amalg C_i$, *pour* $i = 1, ..., n$

$$\log L(\theta) = \sum_{i=1}^{n} \delta_i \log \lambda_\theta(z_i) + \sum_{i=1}^{n} \log S_\theta(z_i),$$

où $\lambda_\theta(\cdot)$ *désigne la fonction de risque instantané.*

En effet :

$$\begin{aligned}
\log L(\theta) &= \log[\prod_{i=1}^{n} f_\theta(z_i)^{\delta_i} S_\theta(z_i)^{1-\delta i}] \\
&= \sum_{i=1}^{n} \log[f_\theta(z_i)^{\delta_i} S_\theta(z_i)^{1-\delta i}] \\
&= \sum_{i=1}^{n} \log[f_\theta(z_i)^{\delta_i} S_\theta(z_i)^{-\delta i} S_\theta(z_i)^{1}] \\
&= \sum_{i=1}^{n} \log \left[\left(\frac{f_\theta(z_i)}{S_\theta(z_i)} \right)^{\delta_i} S_\theta(z_i)^{1} \right] \\
&= \sum_{i=1}^{n} \delta_i \log \lambda_\theta(z_i) + \sum_{i=1}^{n} \log S_\theta(z_i).
\end{aligned}$$

On note

$$L(\theta) = L((z_1, \delta_1), \cdots, (z_n, \delta_n), \theta)$$

De façon analogue au cas non censuré, on définit l'EMV $\widehat{\theta}_n$ de θ et on peut montrer sous certaines hypothèses :

Théorème 2.3.1 *L'EMV* $\widehat{\theta}_n$ *suit approximativement une loi normale de moyenne* θ *et de variance* $nI_1(\theta)$ *;* $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0; \frac{1}{\widehat{\theta}_n})$. *On peut généraliser ce résultat au cas où le paramètre* θ *est un vecteur de* $\mathbb{R}^p$*.On a alors une matrice* $p \times p$ *de variance-covariance* $nI_1(\theta)$.

Les équations de vraisemblance n'ont toutefois pas d'expression simple dans le cas général ; on utilisera les algorithmes usuels pour déterminer l'EMV de manière approchée NEWTON-RAPHSON, BHHH (BERNDT, HALL, HALL, HAUSMAN) et algorithme EM, ce dernier étant particulièrement bien adapté au cas des données incomplètes.

Cependant, dans certaines classes de modèles une approche directe reste possible : cela est notamment le cas des modèles à hasard proportionnel, étudiés ci-après.

exemple 2.3.1 *1-(Loi exponentielle)*
On a

$$\begin{aligned} f(t) &= \theta \exp(-\theta t); \\ S(t) &= \exp(-\theta t); \\ \lambda(t) &= \theta. \end{aligned}$$

Nous avons

$$\begin{aligned} \log L(\theta) &= \sum_{i=1}^{n} \delta_i \log \lambda_\theta(z_i) + \sum_{i=1}^{n} \log S_\theta(z_i) \\ &= \sum_{i=1}^{n} \delta_i \log \theta + \sum_{i=1}^{n} \log \exp(-\theta z_i) \\ &= (\sum_{i=1}^{n} \delta_i) \log \theta - \theta \sum_{i=1}^{n} z_i. \end{aligned}$$

Nous dérivons par rapport à λ

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{(\sum_{i=1}^{n} \delta_i)}{\theta} - \sum_{i=1}^{n} z_i$$

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

On en conclue d'après la solution de l'équation précédente que

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} z_i}.$$

2- (Loi de Weibull)

Dans le cas d'une censure droite. On considère donc le modèle suivant :

$$\begin{aligned} f(t) &= \gamma\alpha^{\gamma}t^{\gamma-1}\exp(-(\alpha t)^{\gamma}); \\ S(t) &= \exp(-(\alpha t)^{\gamma}); \\ \lambda(t) &= \alpha\gamma(\alpha t)^{\gamma-1}. \end{aligned}$$

La vraisemblance de ce modèle s'écrit :

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

D'où l'on déduit la log-vraisemblance

$$\begin{aligned} \log L(\theta) &= \sum_{i=1}^{n}\delta_i \log\lambda_\theta(z_i) + \sum_{i=1}^{n}\log S_\theta(z_i) \\ &= \sum_{i=1}^{n}\delta_i \log\alpha\gamma(\alpha t)^{\gamma-1} + \sum_{i=1}^{n}\log\exp(-(\alpha t)^{\gamma}) \\ &= \gamma\log\alpha\sum_{i=1}^{n}\delta_i + \log\gamma\sum_{i=1}^{n}\delta_i + (\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i - \sum_{i=1}^{n}(\alpha t_i)^{\gamma}. \end{aligned}$$

Les équations aux dérivés partielles s'écrivent donc :

$$\frac{\partial \log L(\theta)}{\partial\alpha} = \frac{\gamma\sum_{i=1}^{n}\delta_i}{\alpha} - \alpha^{\gamma-1}\gamma\sum_{i=1}^{n}t_i^{\gamma}$$

$$\frac{\partial \log L(\theta)}{\partial\gamma} = \sum_{i=1}^{n}\delta_i(\frac{1}{\gamma} - \log\alpha) + \alpha^{\gamma}\left[\log\alpha\sum_{i=1}^{n}t_i^{\gamma} - \sum_{i=1}^{n}t_i^{\gamma}\log t_i\right] + \sum_{i=1}^{n}\delta_i\log t_i.$$

On cherche donc les solutions du système suivant :

$$\widehat{\alpha} = \left(\frac{1}{\sum_{i=1}^{n}t_i^{\widehat{\gamma}}}\sum_{i=1}^{n}\delta_i\right)^{\frac{1}{\widehat{\gamma}}},$$

$$\frac{1}{\widehat{\gamma}} = \frac{\sum_{i=1}^{n} t_i^{\gamma} \log t_i}{\sum_{i=1}^{n} t_i^{\gamma}} - \frac{1}{\sum_{i=1}^{n} t_i^{\gamma}} \sum_{i=1}^{n} \log t_i.$$

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

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

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

$$l\ (\lambda, \gamma) = n\lambda \log \lambda - n \log \Gamma(\gamma) + (\gamma - 1)\sum_{i=1}^{n} \ln t_i - n\lambda \bar{t}.$$

En dérivant par rapport à λ et en appelant $\widehat{\gamma}$ l'estimateur du maximum de vraisemblance pour γ, nous obtenons

$$\widehat{\lambda} = \frac{\widehat{\gamma}}{t}$$

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

2.3.3 Les algorithmes numériques de maximisation de la vraisemblance

Comme on l'a vu ci-dessus, l'expression analytique de la log-vraisemblance ne rend que rarement possible un calcul direct de l'estimateur du maximum de vraisemblance. Bien entendu, les algorithmes standards de type Newton-Raphson peuvent être utilisés dans ce contexte. Toutefois, des méthodes spécifiques peuvent s'avérer mieux adaptées.

L'algorithme de Newton-Raphson

On utilise ici pour résoudre l'équation $f(x_0) = 0$ un algorithme construit à partir d'une linéarisation au voisinage de la solution, sur la base du développement de Taylor à l'ordre un ; en notant que

$$f(x_{k+1}) = f(x) + (x_{k+1} - x_k) - \frac{df}{dx}(x_k) + 0(x_{k+1} - x_k),$$

on propose ainsi la récurrence définie par $f(x_{k+1}) = 0$, qui conduit à :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Dans le cas d'un modèle de durée, on utilise comme fonction f la dérivée de la log-vraisemblance par rapport au paramètre (le score), ce qui conduit à l'expression :

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln L(y \mid z, c; \theta_k)\right]^{-1} \frac{\partial \ln L(y \mid z, c; \theta_k)}{\partial\theta}.$$

L'écriture ci-dessus est une écriture matricielle, valable pour un θ multidimensionnel.

Afin que cet algorithme converge il convient de partir d'une valeur initiale « proche » de la valeur théorique. Il possède une propriété intéressante : si l'on dispose d'un estimateur convergent, pas nécessairement asymptotiquement efficace, on peut l'utiliser comme valeur initiale de l'algorithme de Newton-Raphson. On obtient alors l'efficacité asymptotique dès la première itération.

Il existe une variante de l'algorithme de Newton-Raphson, appelée algorithme BHHH (BERNDT, HALL, HALL, HAUSMAN), qui consiste à remplacer dans l'expression itérative ci-dessus la matrice d'information de Fischer par son expression ne faisant appel qu'à la dérivée première de la log-vraisemblance. On obtient ainsi :

$$\theta_{k+1} = \theta_k - \left[\sum_{i=1}^{n} \frac{\partial \ln L(y_i \mid z_i, c_i; \theta_k)}{\partial\theta} \frac{\partial \ln L(y_i \mid z_i, c_i; \theta_k)}{\partial\theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\partial \ln L(y_i \mid z_i, c_i; \theta_k)}{\partial\theta}.$$

Cette version de l'algorithme de Newton-Raphson a les mêmes propriétés que la précédente.

L'algorithme Espérance-Maximisation (EM)

Cet algorithme a été imaginé plus spécifiquement dans le cadre de données incomplètes ; il s'appuie sur la remarque que, si les variables $(x_1, ..., x_n)$ étaient observables, l'estimation serait effectuée simplement en maximisant la log-vraisemblance latente $\ln L(x \mid z, c; \theta)$; comme on ne dispose pas de ces

observations, l'idée est de remplacer la fonction objectif par sa meilleure approximation connaissant les variables observables $(y_1, ..., y_n)$. Il a été proposé initialement par DEMPSTER et al. [1977].

On introduit, pour $(\theta, \widehat{\theta})$ fixé, la fonction

$$q(\theta, \widehat{\theta}) = E_{\widehat{\theta}}[\ln L^*(x \mid z, c; \theta) \mid y, z, c]$$

L'algorithme EM est alors défini par la répétition des étapes suivantes :

- calcul de $q(\theta, \theta_k)$;
- maximisation en θ de $q(\theta, \theta_k)$, dont la solution est θ_{k+1}.

En pratique cet algorithme est intéressant lorsque le calcul de $q(\theta, \theta_k)$ est sensiblement plus simple que le calcul direct de $\ln L(y \mid z, c; \theta)$; dans le cas contraire, on peut être conduit à utiliser un algorithme de Newton-Raphson pour l'étape d'optimisation de $q(\theta, \theta_k)$, ce qui alourdit la démarche.

L'algorithme EM possède sous certaines conditions de régularité qui ne seront pas détaillées ici les « bonnes propriétés » suivantes :

Proposition 2.3.3 *: L'algorithme EM est croissant, au sens où*

$$\ln L(y \mid z, c; \theta_{k+1}) \geq \ln L(y \mid z, c; \theta_k)$$

de plus toute limite θ_∞ d'une suite de solutions (θ_k) satisfait la condition du premier ordre :

$$\frac{\partial \ln L(y \mid z, c; \theta_\infty)}{\partial \theta}$$

exemple 2.3.2 *Une situation de données manquantes est illustrée à la figure suivante :*

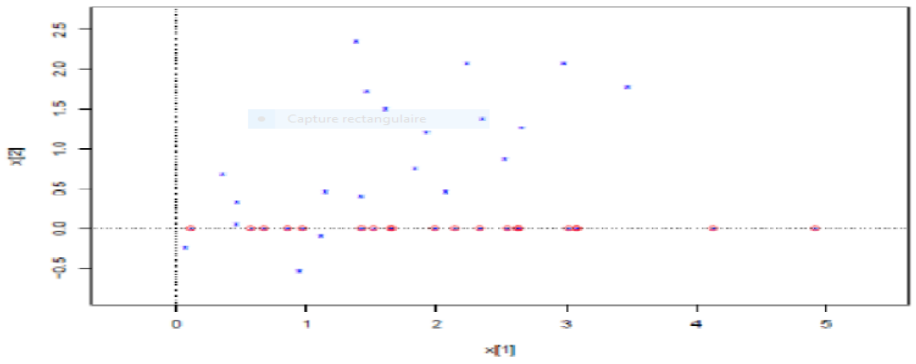

Figure VI 40 point d'une distribution normale binomiale, les 20 derniers avec X(2) manquant (encerclé).

$n = 40$ points ont été indépendamment échantillonné à partir d'une distribution normale bivariée de moyennes $(\mu_1; \mu_2)$ et de variances $(\sigma_1^2; \sigma_2^2)$ et la corrélation ρ

$$\begin{pmatrix} x_{1_i} \\ x_{2_i} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \right)$$

Cependant, les secondes coordonnées des 20 derniers points ont été perdues. Avec leurs valeurs x_2 arbitrairement fixées à 0

Nous souhaitons trouver l'estimation du maximum de vraisemblance du paramètre vecteur $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

L'estimation standard du maximum de vraisemblance

$$\begin{aligned} \widehat{\mu_1} &= \sum_1^{40} \frac{x_{1_i}}{40} \quad ; \quad \widehat{\mu_2} = \sum_1^{40} \frac{x_{2_i}}{40} \quad ; \quad \widehat{\sigma_1} = \left[\sum_1^{40} \frac{(x_{1_i} - \widehat{\mu_1})}{40} \right]^{\frac{1}{2}} ; \\ \widehat{\sigma_2} &= \left[\sum_1^{40} \frac{(x_{2_i} - \widehat{\mu_2})}{40} \right]^{\frac{1}{2}} \text{ et } \widehat{\rho} = \frac{\left[\sum_1^{40} \frac{(x_{1_i} - \widehat{\mu_1})(x_{2_i} - \widehat{\mu_2})}{40} \right]}{\widehat{\sigma_1}\widehat{\sigma_2}}. \end{aligned}$$

ne sont pas disponibles avec des données manquantes.

L'algorithme EM d'estimation des moyennes, écarts types, corrélation de

la bivariée distribution normale des données du figure précédente :

Tableau IV : L'algorithme EM d'estimation de la bivariée distribution normale des données du figure précédente.

Step	μ_1	μ_2	σ_1	σ_2	ρ
1	1.86	.463	1.08	.738	.162
2	1.86	.707	1.08	.622	.394
3	1.86	.843	1.08	.611	.574
4	1.86	.923	1.08	.636	.679
5	1.86	.971	1.08	.667	.736
6	1.86	1.002	1.08	.694	.769
7	1.86	1.023	1.08	.716	.789
8	1.86	1.036	1.08	.731	.801
9	1.86	1.045	1.08	.743	.808
10	1.86	1.051	1.08	.751	.813
11	1.86	1.055	1.08	.756	.816
12	1.86	1.058	1.08	.760	.819
13	1.86	1.060	1.08	.763	.820
14	1.86	1.061	1.08	.765	.821
15	1.86	1.062	1.08	.766	.822
16	1.86	1.063	1.08	.767	.822
17	1.86	1.064	1.08	.768	.823
18	1.86	1.064	1.08	.768	.823
19	1.86	1.064	1.08	.769	.823
20	1.86	1.064	1.08	.769	.823

L'algorithme EM commence par remplir les données manquantes d'une manière ou d'une autre, par exemple en définissant $x_{2_i} = 0$ *pour les* 20 *valeurs manquantes, donnant un ensemble de données artificiellement complet données°.*

Puis : La méthode standard est appliquée à data° pour produire

$$\widehat{\theta}^{\circ} = (\widehat{\mu}_1^{\circ}, \widehat{\mu}_2^{\circ}, \widehat{\sigma}_1^{\circ}, \widehat{\sigma}_2^{\circ}, \widehat{\rho}^{\circ})$$

c'est l'étape M (maximisation).

Chacune des valeurs manquantes est remplacée par son espérance conditionnelle (en supposant $\theta = \widehat{\theta}^{\circ}$ *) étant donné les données non manquantes ; c'est l'étape E (espérance).*

Dans notre cas, le les valeurs x_{2_i} sont remplacées par :

$$\widehat{\mu}_2^{(0)} + \frac{\widehat{\sigma}_1^{(0)}}{\widehat{\sigma}_2^{(0)}} \left(x_{1_i} - \widehat{\mu}_1^{(0)} \right),$$

Les étapes E et M sont répétées, à la $j^{\text{ème}}$ étape donnant un nouveau données d'ensemble de données$^{(j)}$ et une estimation mise à jour $\widehat{\theta}^{(j)}$. L'itération s'arrête lorsque $\left\| \widehat{\theta}^{(j)+1} - \widehat{\theta}^{(j)} \right\|$ est convenablement petit.

Remarque 2.3.1 *dans certains cas, l'algorithme EM peut ne converge que vers un point-selle ou un maximum local de la vraisemblance... si elle en possède un, naturellement. La dépendance en la condition initiale θ_0 choisie arbitrairement est forte : pour certaines mauvaises valeurs, l'algorithme peut rester gelé en un point selle, alors qu'il convergera vers le maximum global pour d'autres valeurs initiales plus pertinentes.*

L'algorithme EM peut donc parfois nécessite plusieurs initialisations différentes.

En pratique on reproche parfois à l'algorithme EM une convergence lente (convergence linéaire) contrairement à l'algorithme de Newton qui a une convergence rapide (convergence quadratique). Toutefois l'algorithme EM a les propriétés intéressantes suivantes :

- il fait croître la vraisemblance à chaque étape,

-les contraintes sont naturellement vérifiées,

-il est peut coûteux en mémoire.

Ceci n'est pas le cas de l'algorithme de Newton qui, si la fonction à optimiser n'est pas convexe, ne permet pas de faire croître la vraisemblance à chaque étape et nécessite la recherche d'une solution améliorant la solution précédente le long de la ligne de plus fort gradient. De plus, les contraintes ne sont souvent pas automatiquement vérifiées, et il faut donc avoir recours à une reparamètrisation pour que ceci soit le cas. Enfin l'algorithme de Newton peut nécessiter le calcul du gradient et de la matrice jacobéenne, ce qui est coûteux en espace mémoire, notamment en grande dimension.

Les autres méthodes

D'autres méthodes peuvent s'avérer utiles dans le cas d'échantillons fortement censurés ; en effet dans ce cas, l'estimation « fréquentielle » usuelle utilisée jusqu'ici peut s'avérer mal adaptée ; on peut alors se tourner vers des

algorithmes d'échantillonnage pondéré bayésiens, notamment les algorithmes MCMC.

Cette situation étant peu courante en assurance ne sera pas développée ici.

2.4 Tests de comparaison

En statistiques, un test d'hypothèse est une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (échantillon).

Il s'agit de statistique inférentielle : à partir de calculs réalisés sur des données observées, nous émettons des conclusions sur la population, en leur rattachant des risques de se tromper. Pour tester l'hypothèse nulle d'égalité des survies dans les deux groupes, on dispose trois tests asymptotiquement équivalents :

1. le test de Wald
2. le test du rapport de vraisemblance
3. le test de Rao ou test du score

2.4.1 Comparaison de deux groupes (dans un modèle exponentiel)

Pourquoi le modèle exponentiel ? Ce modèle suppose que la fonction de risque instantané h(t) est une constante indépendante du temps. Son avantage est l'existence de solutions explicites au maximum de vraisemblance : on dispose alors d'estimateurs et de tests faciles à calculer, ce qui permet une première approche des données.

Nous allons voir comment comparer deux échantillons exponentiels.

Cadre :

Soient A et B deux groupes d'individus dont on veut comparer la survie. On suppose que les durées de vie T_A et T_B suivent une loi exponentielle dans chacun des deux groupes.

Les densités de probabilités de T_A et T_B s'écrivent :

$$f_A(t) = h_A \exp(-h_A t) \quad \text{et} \quad f_B(t) = h_B \exp(-h_B t),$$

où h_A et h_B sont les risques instantanés de décès dans les deux groupes supposés constants au cours du temps.

On note $\exp(b)$ le rapport des risques instantanés, donnant le risque relatif du groupe B par rapport au groupe A :

$$h_B = h_A \exp(b)$$

Pour comparer les deux groupes (c.à.d. les survies S_A et S_B)), il faut estimer b et tester l'hypothèse nulle : $H_0 : h_A = h_B$ ou de façon équivalente $H_0 : b = 0$:

Soient n_A et n_B, le nombre d'individus dans chacun des groupes ($n = n_A + n_B$). Dans le groupe A, on observe $(Z_{A,i}, \delta_{A,i})_{i=1}^{n_A}$; et dans le groupe B, on observe $(Z_{B,i}, \delta_{B,i})_{i=1}^{n_B}$.

La vraisemblance des observations s'écrit :

$$\begin{aligned} L(h_A, h_B) &= L(h_A)L(h_B) \\ &= \prod_{i=1}^{n_A} f_A^{\delta_i} S_A^{1-\delta_i} \prod_{i=1}^{n_B} f_B^{\delta_i} S_B^{1-\delta_i}. \end{aligned}$$

et la log-vraisemblance :

$$\log L(h_A, h_B) = \log(h_A) \sum_{i=1}^{n_A} \delta_{A,i} - h_A \sum_{i=1}^{n_A} z_{A,i} + \log(h_B) \sum_{i=1}^{n_B} \delta_{B,i} - h_B \sum_{i=1}^{n_B} z_{B,i},$$

telle que :

$$r_A = \sum_{i=1}^{n_A} \delta_{A,i} \text{ et } r_B = \sum_{i=1}^{n_B} \delta_{B,i}$$

On reparamètre la log-vraisemblance en remplaçant (h_A, h_B) par (h_A, b) où $b = \log(\frac{h_B}{h_A})$. La log-vraisemblance s'écrit alors :

$$r \log(h_A) + b r_B - h_A \sum_{i=1}^{n_A} z_{A,i} + \exp(b) \sum_{i=1}^{n_B} z_{B,i},$$

où $r = r_A + r_B$ représente le nombre total d'individus non censurés (ou le nombre de décès observés).

On calcule le vecteur de score (dérivées partielles de $\log L$) :

$$U(h_A, b) = \begin{pmatrix} \frac{\partial \log L}{\partial h_A} \\ \frac{\partial \log L}{\partial b} \end{pmatrix},$$

Les estimateurs du maximum de vraisemblance $\widehat{h}_A$ de h_A et $\widehat{b}$ de b sont les solutions du système d'équations $U(h_A, b) = \overrightarrow{0}$. On obtient :

$$\widehat{h}_A = \frac{r_A}{\sum_{i=1}^{n_A} z_{A,i}} \quad \text{et} \quad \widehat{\exp b} = \frac{\frac{r_B}{n_B}\sum_{i=1}^{} z_{B,i}}{\frac{r_A}{n_A}\sum_{i=1}^{} z_{A,i}}.$$

Remarque 2.4.1 *On retrouve bien*

$$\begin{aligned} \widehat{h}_B &= \widehat{h}_A \widehat{\exp b} \\ &= \frac{r_B}{\sum_{i=1}^{n_B} z_{B,i}}. \end{aligned}$$

2.4.2 Le test de Wald

Proposition 2.4.1 *Sous $H_0 : b = 0$, la loi de l'EMV $\widehat{b}$ est asymptotiquement normale de moyenne nulle et*

$$\widehat{V}(\widehat{b}) = \frac{r}{r_A r_B},$$

(c'est le terme ($U(h_A, b)$) de l'inverse de la matrice d'Information de Fisher $I^{-1}(\widehat{h}_A; \widehat{b})$.)

Proposition 2.4.2 *La statistique du test de Wald pour comparer b à 0 est*

$$\chi^2_W = \frac{r_A r_B}{r}\widehat{b}^2.$$

suit asymptotique une loi $\chi^2(1)$ sous H_0.

2.4.3 Test du rapport de vraisemblance

Proposition 2.4.3 *La statistique du logarithme du rapport de vraisemblance est définie par :*

$$\chi^2_L = 2\log\left(\frac{L(\widehat{h}_A, \widehat{b})}{L(\widehat{h}, 0)}\right),$$

$L(\widehat{h}_A, \widehat{b})$ est la vraisemblance maximisée c.à.d calculée sans restriction en :$h_A = \widehat{h}_A et\ \ b = \widehat{b}$ et $L(\widehat{h}, 0)$ est la vraisemblance restreinte sous H_0, maximisée en :

$$h = \widehat{h} = \frac{r}{\sum_{i=1}^{n_A} z_{A,i} + \sum_{i=1}^{n_B} z_{B,i}} \quad et \quad b = 0.$$

Proposition 2.4.4 *La statistique du rapport de vraisemblance s'écrit :*

$$\chi^2_L = 2\left[r_A \log\left(\frac{r_A}{\sum_{i=1}^{n_A} z_{A,i}} \right) + r_B \log\left(\frac{r_B}{\sum_{i=1}^{n_A} z_{B,i}} \right) - r \log\left(\frac{r}{\sum_{i=1}^{n_A} z_{A,i} + \sum_{i=1}^{n_B} z_{B,i}} \right) \right].$$

Sous $H_0 : b = 0, \chi^2_L$ suit asymptotiquement une loi $\chi^2(1)$.

2.4.4 Test de Rao ou test du score

Proposition 2.4.5 *La statistique de test du score est donnée par*

$$\chi^2_S = \frac{r_A r_B}{r} \frac{(\exp \widehat{b} - 1)^2}{\exp \widehat{b}}.$$

Sous $H_0 : b = 0, \chi^2_S$ suit asymptotiquement une loi $\chi^2(1)$.

La pratique des procédures de tests

– Les procédures des tests conduisent à rejeter $H_0 : b = 0$ pour des valeurs élevées de la statistique de test, c. à. d. pour des valeurs supérieures au quantile d'ordre $1 - \alpha$ de $\chi^2(1)$ pour un risque de première espèce α.

– En pratique, le test de Wald peut donner parfois des résultats assez différents du test du score ou du rapport de vraisemblance qui sont assez proches en général.

– Le test du rapport de vraisemblance est le plus robuste et donc le plus fiable des trois.

2.5 Introduction de covariables

Dans l'approche paramétrique, les fonctions d'intérêts peuvent dépendre de covariables explicatives susceptibles 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 du 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 risques proportionnels "de Cox" (d'autres modèles à risques proportionnels sont possibles), c'est-à-dire

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

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

$$\begin{aligned}
S(t|Z) &= \exp(-\int_0^t \lambda(u|Z)du) \\
&= \exp(-\int_0^t \lambda_0(u)\exp(\beta' Z)du) \\
&= S_0(t)^{\exp(\beta' Z)}, \\
f(t|Z) &= -S'(t|Z) \\
&= \lambda(t|Z)\exp(-\int_0^t \lambda(u|Z)du) \\
&= \lambda_0(u)\exp(\beta' Z)S_0(t)^{\exp(\beta' Z)},
\end{aligned}$$

avec

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

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

exemple 2.5.1 *Considérons un risque de base suivant une loi de Weibull* $W(\gamma, \alpha)$*, alors*

$$\begin{aligned} \lambda_0(t) &= \gamma\alpha(\alpha t)^{\gamma-1}, \qquad t \geq 0 \text{ et } \gamma, \alpha > 0, \\ S_0(t) &= \exp(-(t\alpha)^{\gamma}), \\ f_0(t) &= \gamma\alpha(\alpha t)^{\gamma-1}\exp(-(t\alpha)^{\gamma}). \end{aligned}$$

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 :

$$\begin{aligned} \lambda(t|Z) &= \gamma\alpha(\alpha t)^{\gamma-1} \times \exp(\beta' Z), \qquad t \geq 0 \quad \text{et } \alpha\ , \gamma > 0, \\ S(t|Z) &= \exp(-(t\alpha)^{\gamma})^{\exp(\beta' Z)}, \\ f(t|Z) &= \gamma\alpha(\alpha t)^{\gamma-1} \times \exp(\beta' Z) \times \exp(-(t\alpha)^{\gamma})^{\exp(\beta' Z)}. \end{aligned}$$

Pour $\gamma = 1$*, on retrouve la loi exponentielle* $\mathcal{E}(\alpha)$*. Ainsi, dans le cas d'un risque suivant une loi exponentielle avec des covariables, on obtient*

$$\begin{aligned} \lambda(t|Z) &= \alpha \times \exp(\beta' Z), \qquad \alpha > 0, \\ S(t|Z) &= \exp(-t\alpha)^{\exp(\beta' Z)}, \\ f(t|Z) &= \alpha \times \exp(\beta' Z) \times \exp(-t\alpha)^{\exp(\beta' Z)}. \end{aligned}$$

2.5.1 Modèles de vie accélérée (Accelerated Failure Time model)

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 :

$$S(t|Z) = S_0(t\exp(\beta' Z)),$$

où Z est un vecteur de covariable, β le vecteur des coefficients de régression. Le terme $\exp(\beta' Z)$ 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,

$$\begin{aligned}\lambda(t|Z) &= [-\ln(S(t|Z))]' \\ &= -\frac{[S(t|Z)]\prime}{S(t|Z)} \\ &= -\frac{-\exp(\beta' Z) \times \lambda_0(t\exp(\beta' Z))S_0(t\exp(\beta' Z))}{S_0(t\exp(\beta' Z))} \\ &= \exp(\beta' Z)\lambda_0(t\exp(\beta' Z)).\end{aligned}$$

En effet, on a les égalités suivantes,

$$\begin{aligned}S(t|Z) &= S_0(t\exp(\beta' Z)) \\ &= \exp(-\Lambda_0(t\exp(\beta' Z))) \\ &= \exp\left[-\int_0^t \lambda_0(u\exp(\beta' Z))du\right].\end{aligned}$$

Si on suppose que $S_0(t)$ est la fonction de survie de la variable $\exp(\mu+\epsilon)$, alors

$$S_0(t) = P(\exp(\mu+\epsilon) > t).$$

Ainsi, on obtient que

$$\begin{aligned}S(t|Z) &= S_0(t\exp(\beta' Z)) \\ &= P(\exp(\mu+\epsilon) > t\exp(\beta' Z)) \\ &= P(\exp(\mu-\beta' Z+\epsilon) > t) \\ &= P(X > t),\end{aligned}$$

est la fonction de survie de la variable X où

$$\log(X) = \mu - \beta\prime Z + \epsilon.$$

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

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

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, par exemple,
– $\epsilon \sim$ loi aux valeurs extrêmes ($f_\epsilon(y) = exp(y - e^y)$)
– $\epsilon \sim$ log-logistique
– $\epsilon \sim$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 2.5.1 *On peut remarquer que dans le cas des modèles de vie accélérée, 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*

Chapitre 3

Estimation non paramétrique sur les données censurées

La statistique non paramétrique regroupe l'ensemble des méthodes statistiques qui permettent de tirer de l'information pertinente de données, sans faire l'hypothèse que la loi de probabilité de ces observations appartient à une famille paramétrée connue comme c'est le cas de la statistique paramétrique.

L'estimation est une branche des mathématiques statistiques qui permet, à partir de mesures effectuées sur un système, d'estimer la valeur de différents paramètres de ce système.

L'estimation non paramétrique est basées sur un tableau des données contient à la fois des données censurés.

On s'intéresse dans ce chapitre à l'estimation non paramétrique de la fonction de survie ainsi que la fonction de risque cumulé et celle de densité de probabilité. On se base sur les relations entre ces fonctions.

3.1 Estimation de la fonction de survie

3.1.1 L'estimateur de Kaplan-Meier

Kaplan et Meier ont proposé un estimateur de la fonction de survie S nommé aussi estimateur produit-limite. Il repose sur l'idée suivante : un "individu" est en vie après l'instant t, c'est être en vie juste avant l'instant t et ne pas "mourir" en t. Cette idée se traduit comme suite : pour $0 < t_1 < ... < t_i < t_{i+1} < ... < t_n = t$

$$\begin{aligned} S(t) &= \mathbb{P}(T > t) \\ &= \mathbb{P}(T > t, T > t_{n-1}) \\ &= \mathbb{P}(T > t/T > t_{n-1})\mathbb{P}(T > t_{n-1}) \\ &= ... \\ &= \mathbb{P}(T > t/T > t_{n-1})...\mathbb{P}(T > 1/T > 0)\mathbb{P}(T > 0). \end{aligned}$$

Si l'on choisit les instants de conditionnement où il se produit un évènement t_i (mort, panne ou censure . . .) on aura à estimer des quantités de la forme :

$$\mathbb{P}(T > t_{(i)}/T > t_{(i-1)}) = p_i$$

où les $t_{(i-1)} < t_{(i)}$ et p_i est la probabilité de survivre pendant l'intervalle de temps $I_i = [t_{(i-1)}, t_{(i)}[$ sachant qu'on était "vivant" au début de cet intervalle.

Notons n_i le nombre des sujets qui sont "vivants" (donc à risque) juste avant l'instant $t_{(i)}$ et m_i le nombre de décès à l'instant $t_{(i)}$.

or

$$q_i = 1 - p_i$$

est la probabilité de la survenue de l'évènement durant I_i. Un estimateur naturel de q_i est la fréquence

$$\widehat{q_i} = \frac{m_i}{n_i}.$$

Si on suppose qu'il n'y ait pas d'ex-æquo (plusieurs pannes en $t_{(i)}$) :

$$\delta_i = \begin{cases} 1, \text{ c'est qu'il y a eu un "évènement" en } t_{(i)},\ m_i = 1, \\ 0, \text{ c'est qu'il y a eu une censure en } t_{(i)}, m_i = 0. \end{cases}$$

Par suite

$$\begin{aligned} \widehat{p_i} &= \begin{cases} 1 - \widehat{q_i}, \text{ si } \delta_i = 1 \\ 1, \text{ sinon} \end{cases} \\ &= \begin{cases} 1 - \frac{1}{n_i}, \text{ si } \delta_i = 1 \\ 1, \text{ sinon.} \end{cases} \end{aligned}$$

On a $n_i = n - (i-1)$ (car il y'a eu $i-1$"évènements" ou censures avant $t_{(i)}$ et il y'a n individus dans l'étude). L'estimateur de Kaplan-Meier est

$$\widehat{S}_{KM(t)} = \prod_{t_{(i)} \leq t} (1 - \frac{1}{n-i+1})^{\delta_i}.$$

Remarque 3.1.1 *L'estimateur de Kaplan-Meier est une fonction en escaliers qui fait des sauts à chaque instant t_i. La valeur du saut dépend du nombre d'évènements au temps $t_i$et aussi du nombre de censures à ce temps là.*

Traitement des d'ex-æquo : Dans le cas d'existence des d'ex-æquo on n'a plus m_i égale à 1 en $t_{(i)}$ mais au nombre de décès en $t_{(i)}$ et aussi on n'a plus $n_i = n - (i-1)$. Dans ce cas on doit garder r_i et m_i et l'estimateur de Kaplan-Meier dans ce cas est

$$\widehat{S}_{KM(t)} = \prod_{t_{(i)} \leq t} (1 - \frac{m_i}{n_i})^{\delta_i}.$$

exemple 3.1.1 *: Sur 10 patients atteints de cancer des bronches, on a observé les durées de survie suivantes exprimées en mois : 1 3 4^+ 5 7^+ 8 9 10^+ 11 13^+. Les données suivies du signe + correspondent à des patients qui ont été perdus de vue à la date fournie ainsi que l'existence (statut=0) ou non (statut=1) d'une censure à droite.*

Patient i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Durées t_i	*1*	*3*	*4*	*5*	*7*	*8*	*9*	*10*	*11*	*13*
Statut m_i	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*0*	*1*	*0*

Temps t_i	r_i	m_i	$\widehat{S}(t_i)$	*Intervalle*
0	*0*	*0*	*1*	$[0;1[$
1	*10*	*1*	$(1-\frac{1}{10})\widehat{S}(0)=0.9$	$[1;3[$
3	*9*	*1*	$(1-\frac{1}{9})\widehat{S}(1)=0.8$	$[3;4[$
4	*8*	*0*	$(1-\frac{0}{8})\widehat{S}(3)=0.8$	$[4;5[$
5	*7*	*1*	$(1-\frac{1}{7})\widehat{S}(4)=0.7$	$[5;7[$
7	*6*	*0*	$(1-\frac{0}{6})\widehat{S}(5)=0.7$	$[7;8[$
8	*5*	*1*	$(1-\frac{1}{5})\widehat{S}(7)=0.6$	$[8;9[$
9	*4*	*1*	$(1-\frac{1}{4})\widehat{S}(8)=0.5$	$[9;10[$
10	*3*	*0*	$(1-\frac{0}{3})\widehat{S}(9)=0.5$	$[10;11[$
11	*2*	*1*	$(1-\frac{1}{2})\widehat{S}(10)=0.25$	$[11;13[$
13	*1*	*0*	$(1-\frac{0}{1})\widehat{S}(11)=0.25$	$[13;\infty[$

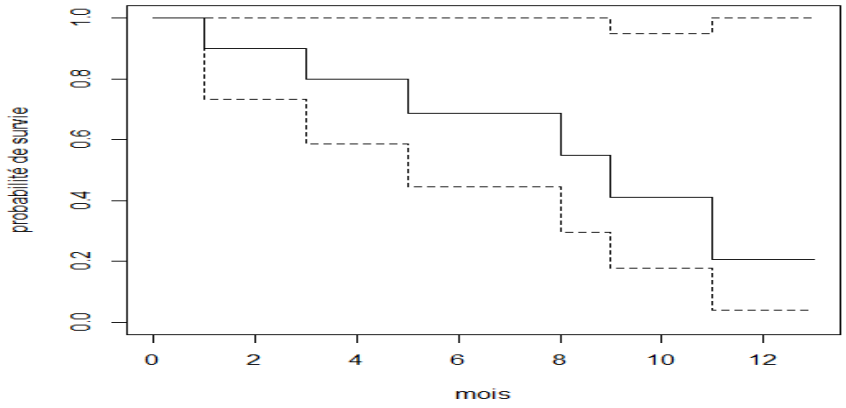

Figure VII : La courbe de Kaplan-Meier.

où la ligne continue représente la courbe de Kaplan-Meier et les lignes discontinues représentent l'intervalle de confiance.

Propriétés de l'estimateur de Kaplan-Meier :

L'estimateur de Kaplan-Meier possède un certain nombre de « bonnes propriétés » qui en font la généralisation naturelle de l'estimateur empirique de la fonction de répartition en présence de censure : il est convergent, asymptotiquement gaussien, cohérent et est également un estimateur du maximum de vraisemblance généralisé. Toutefois, cet estimateur est biaisé positivement.

L'estimateur de Kaplan-Meier est l'unique estimateur cohérent de la fonction de survie.

Propriété 1

Si aucune donnée n'est censurée, i.e. $T_i = X_i$ pour $i = 1, ..., n$ alors

$$\hat{S}_{KM}(t) = \hat{S}_n(t) = 1 - F_n(t),$$

où

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{T_i \leq t},$$

est la fonction de répartition empirique.

Propriété 2

Soit $S_C(t) = P(C > t)$ et $\tau_x = \inf\{x \geq 0 \backslash S(x)S_C(x) = 0\}$:

1) Si S et S_C n'ont pas de points de discontinuités en commun, on a, pour tout $\tau < \tau_x$:

$$\sup_{0 \leq t \leq \tau} \left| \hat{S}_{KM}(t) - S(t) \right| \underset{n \to \infty}{\overset{p.s.}{\to}} 0.$$

2) En tout point $t \in [0; \tau]$,

$$\sqrt{n}(\hat{S}_{KM}(t) - S(t)) \underset{n \to \infty}{\overset{L}{\to}} N(0, V^2(t));$$

avec

$$V^2(t) = -S^2(t) \int_0^t \frac{S(u)}{S^2(u)S_C(u)} du$$

L'estimateur de Kaplan-Meier est un estimateur fortement consistant, $\sqrt{n}-$consistant et asymptotiquement gaussien de $S(t)$.

Paramètres de position et de dispersion

Une durée de survie est une variable quantitative continue, il est donc possible de calculer les paramètres de position et de dispersion habituels : moyenne, médiane, écart-type, étendue, etc. Cependant, les distributions de ces variables sont généralement asymétriques et on préfèrera utiliser des paramètres robustes tels que médiane et quartiles plutôt que moyenne et écart-type. Pour tenir compte de la présence de données censurées, le calcul des

paramètres usuels (moyenne, médiane, quartiles) est basé sur le calcul de l'estimateur de Kaplan-Meier de la fonction de survie $S(t)$.

·Le quantile d'ordre α est obtenu par :

$$q_\alpha = inf\{t : 1 - \widehat{S}(t) \geq \alpha\},$$

$\widehat{S}(t)$ est l'estimateur de Kaplan-Meier de $S(t)$.

·Notons que l'on peut utiliser l'estimateur de la fonction de survie pour estimer une durée moyenne : puisque l'espérance de la durée peut généralement s'écrire :

$$E(X) = \int_0^\infty uf(u)du = \int_0^\infty S(u)du,$$

on peut utiliser l'estimateur suivant :

$$\widehat{\mu} = \sum_{i=1}^{n} \widehat{S}(t_{i-1})(t_i - t_{i-1}).$$

·Il est intéressant de pouvoir calculer la variance ou l'écart-type de l'estimateur pour apprécier sa qualité (comme toujours dés que l'on fait de l'estimation ponctuelle).

Certains estimateurs sont utilisés pour approcher la variance de l'estimateur de Kaplan-Meier. Un de ces estimateurs les plus courants est la formule de Greenwood :

$$\widehat{Var}(\widehat{S}(t)) = \widehat{S}^2(t) \sum_{i=1}^{n} \frac{m_i}{n_i(n_i - m_i)},$$

où m_i : nombre de décès observés au temps t_i

et n_i : nombre de sujets exposés au risque de décès au temps t_i

En effet :

$$\begin{aligned}
\widehat{Var}(\widehat{S}(t)) &= Var\left[\prod_i(1-\frac{m_i}{n_i})\right] \\
&= E\left[\prod_i(1-\frac{m_i}{n_i})^2\right]^2 - E^2\left[\prod_i(1-\frac{m_i}{n_i})\right] \\
&= \prod_i E\left[(1-\frac{m_i}{n_i})\right]^2 - \prod_i E^2\left[(1-\frac{m_i}{n_i})\right] \\
&= \prod_i Var\left[\left(1-\frac{m_i}{n_i}\right)\right] + E^2\left[\prod_i(1-\frac{m_i}{ni})\right] - \prod_i E^2\left[\left(1-\frac{m_i}{n_i}\right)\right] \\
&= \prod_i\left(\frac{p_iq_i}{n_i}+p_i^2\right) - \prod_i p_i^2 \\
&= S^2(t)\prod_i(\frac{q_i}{r_ip_i}+1) - S^2(t) \\
&= S^2(t)\prod_i(\frac{q_i}{r_ip_i}) \\
&= \widehat{S}^2(t)\sum_{i=1}^{n}\frac{m_i}{n_i(n_i-m_i)}.
\end{aligned}$$

L'erreur standard de l'estimateur de Kaplan-Meier se calcule selon la formule de Greenwood, au temps de décès $t_i, i = 1, ..., n$ par :

$$\widehat{\sigma}(\widehat{S}(t_i)) = \widehat{S}(t_{_i})\sqrt{\sum_{i=1}^{n}\frac{m_i}{n_i(n_i-m_i)}}.$$

Variance de l'estimateur de Kaplan Meier

On propose ici une justification heuristique d'un estimateur de la variance de l'estimateur de Kaplan-Meier, l'estimateur de Greenwood.

L'expression

$$\widehat{S}\ (t) = \prod_{T_i\leq t}\left(1-\frac{m_i}{n_i}\right),$$

permet d'écrire :

$$\ln\left(\widehat{S}(t)\right) = \sum_{T_{(i)} \leq t} \ln\left(1 - \frac{m_i}{n_i}\right) = \sum_{T_{(i)} \leq t} \ln\left(1 - \widehat{q}_i\right),$$

Avec l'indépendance des variables $\ln\left(1-\widehat{q}_i\right)$, comme la loi de $n_i\widehat{p}_i$ est binomiale de paramètres (n_i, p_i), on a par la méthode delta :$V\left(F\left(X\right)\right) = \left(\frac{df}{dx}\left(E\left(X\right)\right)\right)^2 V\left(X\right)$,

où p_i est la probabilité de survivre pendant l'intervalle de temps $I_i = [t_{(i-1)}, t_{(i)}[$ sachant qu'on était "vivant" au début de cet intervalle, n_i le nombre des sujets qui sont "vivants" (donc à risque) juste avant l'instant $t_{(i)}$ et

$$V\left(\ln \widehat{p}_i\right) = V\left(\widehat{p}_i\right)\left[\frac{d}{dp}\ln\left(\widehat{p}_i\right)\right]^2 = \frac{\widehat{q}_i\left(1-\widehat{q}_i\right)}{n_i}\frac{1}{\left(1-\widehat{q}_i\right)^2} = \frac{\widehat{q}_i}{n_i - \left(1-\widehat{q}_i\right)},$$

ce qui conduit à proposer comme estimateur de la variance de $\ln \widehat{S}\left(t\right)$:

$$\widehat{V}\left(\ln \widehat{S}\left(t\right)\right) = \sum_{T_i \leq t} \frac{\widehat{q}_i}{n_i\left(1-\widehat{q}_i\right)} = \sum_{T_i \leq t} \frac{m_i}{n_i\left(n_i - m_i\right)},$$

En appliquant de nouveau la méthode delta avec pour f la fonction logarithme, on obtient finalement :

$$\widehat{V}\left(\widehat{S}\left(t\right)\right) = \widehat{S}\left(t\right)^2 \gamma\left(t\right)^2,$$

avec :

$$\gamma\left(t\right) = \sqrt{\sum_{T_{(i)} \leq t} \frac{m_i}{n_i\left(n_i - m_i\right)}}.$$

Cet estimateur est l'estimateur de Greenwood. Il est convergeant pour la variance asymptotique de l'estimateur de Kaplan-Meier. Il permet avec la normalité asymptotique de l'estimateur de Kaplan-Meier de calculer des intervalles de confiance (asymptotiques) dont les bornes sont, pour la valeur de la survie en $T_{(i)}$:

$$S_t \times \left(1 \pm u_{1-\frac{\alpha}{2}}\gamma\left(T_{(t)}\right)\right) = S_t \times \left(1 \pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{m_1}{n_1\left(n_1-m_1\right)} + \frac{m_2}{n_2\left(n_2-m_2\right)} + ... + \frac{m_i}{n_i\left(n_i-m_i\right)}}\right).$$

On construit de la sorte des intervalles ponctuels, à t fixé.

Propriétés asymptotiques

L'estimateur de Kaplan-Meier est asymptotiquement gaussien ; précisément on a le résultat suivant :

Proposition 3.1.1 *si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :*

$$\sqrt{n}\left(\widehat{S}-S\right)\rightarrow W_s$$

avec W_s un processus gaussien centré de covariance :

$$\rho\left(s,t\right)=S(s)S(t)\int_0^{s\curlywedge t}\frac{dF(u)}{(1-F(u))^2(1-G(u))}.$$

En particulier lorsque le modèle n'est pas censuré (i.e.$G(u)=0$) on retrouve un résultat de convergence, L'intérêt de résultats de convergence au niveau du processus lui-même plutôt que pour un instant fixé est que l'on peut en déduire des bandes de confiance asymptotique pour l'estimateur de Kaplan-Meier.

On peut trouver dans GILL [1980] une démonstration de la normalité asymptotique de $\widehat{S}_{KM}$, fondée sur la théorie des processus ponctuels. En notant $F=1-\widehat{S}$ et $\widehat{F}=1-\widehat{S}_{KM}$, la bande de confiance qu'il obtient s'écrit :

$$\lim_{n\rightarrow\infty}\inf P\left\{\sup_{s=[0,t]}\left|\frac{\widehat{F}(s)-F(s)}{1-\widehat{F}(s)}\right|\leq\frac{\sqrt{\widehat{V}(t)}}{1-\widehat{F}(t)}x\right\}\geq\sum_{k=-\infty}^{\infty}(-1)^k[\Phi((2k+1)x)-\Phi((2k-1)x)],$$

où

$$\widehat{V}(t)=\widehat{S}_{KM}^2\int_0^t\frac{d\overline{N}^1(u)}{\overline{R}(u)(\overline{R}(u)-\Delta\overline{N}^1(u))},$$

estime la variance du processus gaussien limite W_s.

La représentation et lecture d'une courbe de survie

Nous avons dit que la fonction de survie théorique est représentée par une courbe continue décroissante qui retrace la probabilité de survie en fonction du temps, avec $S(0) = 1$et $lim_{t\to+\infty}S(t) = 0$

Nous savons qu'au début d'observation, tous les sujets sont encore dans l'état initial, c'est- a- dire 100% de la population n'ont pas encore subi l'événement. La fonction de survie empirique estimée par la méthode de Kaplan-Meier est représentée par une courbe décroissante sous forme d'escaliers.

A chaque fois qu'un évènement d'intérêt se produit, il se traduit par une marche, sa hauteur dépend du nombre d'individus qui ont subi l'événement à cet instant.

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

Pour que le graphe soit complet, il faut y faire apparaitre le nombre total des sujets expos es a l'évènement ainsi que le nombre total d'évènements observé et défini les axes de coordonnées.

L'axe des abscisses représente le temps, le taux d'individus exposés aux risques est représenté sur l'axe des ordonnées. Comme la montre la figure ci-dessus.

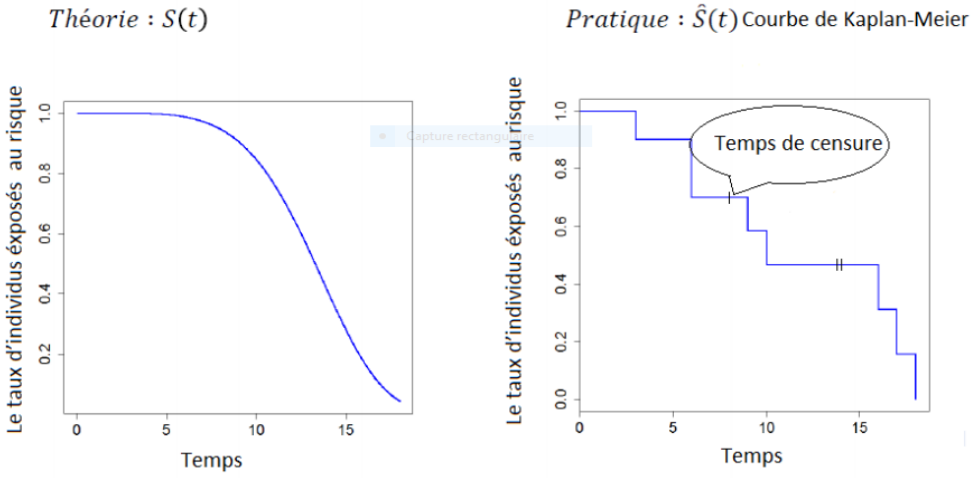

Figure VIII : Courbes de survie.

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

exemple 3.1.2 *Dans une étude de survie après le diagnostic d'une maladie, si nous observons :*

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

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

La précision de l'estimateur dépend du nombre de patients disponibles, ou d'écart-type de la fonction $S(t)$ ou par intervalle de confiance qui peut ne pas être figuré sur le graphe car il surcharge la représentation, ce qui rend la lecture difficile. À la fin du graphe, il reste toujours un groupe de sujets qui n'ont pas encore subi l'évènement, nous l'interprétons soit par l'apparence de stabilité de survie a long terme, ou bien par la disparition du risque à un certain temps.

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

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X>x\}},$$

Cet estimateur a de bonnes propriétés en termes de convergence : convergence p.s (Glivenko cantelli), convergence en loi du processus empirique associé vers un pont brownien.

Néanmoins, dans le cas des données censurées, la variable d'intérêt n'est plus la variable observée. Ainsi estimer la survie S par la survie empirique des données observées $(T_i)_{i=1,\dots,n}(S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i<x\}})$ fournit une estimation biaisée de S (les censures (qui ne sont pas des décès) sont considérées comme des décès : il y a une sous estimation de la survie) : Il en est de même si on estime la fonction de survie par la survie empirique des données observées non censurées (échantillon tronqué). Notons que quand il n'y a pas de censure, l'estimateur de Kaplan-Meier se réduit à la fonction de survie empirique.

3.1.2 Estimateur de Harrington et Fleming de la survie

Nous rappelons la relation entre le taux de hasard cumulé Λ et la fonction de survie S :

$$\Lambda(t) = -\ln(S(t)), \forall t > 0,$$

L'estimateur de la fonction de survie de Harrington et Fleming $\widehat{S}_{HF}$est dérivé de l'estimateur de Nelson Aalen ($\widehat{\Lambda}(t) = \sum_{i,T_i \leq t} \frac{mi}{ni}$). Il est donné par :

$$\begin{aligned} \widehat{S}_{HF}(t) &= \exp(-\widehat{\Lambda}(t)) \\ &= \prod_{i,T_i \leq t}^{n} e^{-\frac{m_i}{n_i}} \\ &\approx \prod_{i,T_i \leq t}^{n} (1 - \frac{m_i}{n_i}), \text{ si } \frac{m_i}{n_i} \to 0, \end{aligned}$$

- m_i : nombre de décès observés au temps t_i
- n_i : nombre de sujets exposés au risque de décès juste avant t_i.

En appliquant un développement limité, on retrouve l'estimateur de Kaplan-Meier.

3.1.3 Estimation de la survie par la méthode actuarielle

Le terme « actuarielle » vient du latin actuarius qui signifierait littéralement secrétaire aux comptes. C'est la première méthode d'analyse de survie à voir en 1912 en tant que théorie statistique. Elle a été introduite pour la $1^{ère}$ fois dans le champ des applications médicales. C'était alors la seule méthode disponible pour estimer la survie. Elle suppose a priori une analyse univariée, c'est-à-dire une situation où seul un unique facteur influence la survie. Elle fait le bilan des occurrences de survenue de l'événement étudié à intervalles fixes.

Le principe de la méthode est d'estimer la fonction de survie en des temps (des délais donc) définis à l'avance, par exemple tous les mois.

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

· m_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 $[t_{i-1}, t_i[$,

·n_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

$n_i = n_{i-1} - c_i/2$,

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

L'estimateur de la fonction de survie est donc,

$$\hat{S}_{AC}(t) = \prod_{i, t_i \leq t} (1 - \frac{m_i}{n_i}),$$

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

$$\widehat{Var}(\hat{S}_{AC}(t)) = (\hat{S}_{AC}(t))^2 \sum_{i, T_i \leq t} \frac{m_j}{n_j(n_j - m_j)}.$$

3.1.4 Comparaison des méthodes actuarielle et de Kaplan Meier

Dans les études de type populationnel, la méthode actuarielle est souvent préférée à la méthode de Kaplan-Meier et ce, même si les estimations de $S(t)$ produites par ces deux méthodes sont semblables. La principale raison de cela est que la méthode actuarielle utilise des données groupées, alors que la méthode de Kaplan-Meier ne s'appuie sur aucun regroupement de données. Puisque les données issues d'études de type populationnel sont souvent groupées, le calcul de la survie observée repose généralement sur la méthode actuarielle.

3.2 Estimation du risque cumulé

3.2.1 Estimateur de Nelson-Aalen du risque cumulé

On rappelle que le risque cumulé est défini par

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

Dans le cas où T 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 T ($f(t) = -\frac{dS(t)}{dt}$),

$$\Lambda(t) = -\int_0^t \frac{dS(u)}{S(-u)},$$

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 :

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

par conséquent,

$$H_1(dt) = G(t^-)S(u)dt,$$

et on obtient l'expression suivante pour le risque cumulé :

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

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

soient

$$\begin{aligned}\widehat{H}(u) &= \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i>u\}} \text{ et } \widehat{H}_1(u) \\ &= \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i>u,\delta_i=1\}}.\end{aligned}$$

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

$$\begin{aligned}\widehat{\Lambda}(t) &= -\int_0^t \frac{\widehat{H}_1(u)du}{\widehat{H}(u^-)} \\ &= \sum_{i,T_i\leq t} \frac{\sum_{j=1}^{n} I_{\{Tj=T_i,\delta j=1\}}}{\sum_{j=1}^{n} I_{\{Tj\geq T_i\}}} \\ &= \sum_{i,T_i\leq t} \frac{mi}{ni}.\end{aligned}$$

où n_i représente le nombre d'individus à risque juste avant T_i et m_i représente le nombre de décès en T_i : L'estimateur de Nelson-Aalen est une fonction en escalier décroissante, continue à Droite, qui ne saute qu'aux instants de morts réelles et qui a un saut de taille $m_i = n_i$ à chaque instant de décès.

Propriétés : Si les lois de T et de C n'ont pas de discontinuité commune (toujours vrai lorsque ces lois sont continues),

Absence de biais	$\forall t > X_{(1)}, E(\widehat{H}_n(t)) = H(t)$
Consistance	$\sup \left\lvert \widehat{H}_n(t) - H(t) \right\rvert \xrightarrow{ps} 0$
Normalité asymptotique	$\sqrt{n}(\widehat{H}_n(t) - H(t)) \xrightarrow{l} 0$

Estimation de la variance de $\widehat{\Lambda}(t)$

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,

$$Var(\widehat{\Lambda}(t)) = \sum_{i,T_i \leq t} \frac{m_i}{n_i^2}.$$

où m_i et n_i sont le nombre de décès et d'individus à risque en T_i.

3.2.2 L'estimateur de Breslow du risque cumulé

L'estimateur de Breslow (ou de Peterson) pour le risque cumulé est obtenu à partir de l'estimateur de Kaplan-Meier, de la fonction de survie en utilisant la relation

$$\Lambda(t) = -\ln(S(t)),$$

L'estimateur de Breslow s'écrit

$$\begin{aligned} \widehat{\Lambda}_2(t) &= -\ln(\hat{S}_{KM}(t)) \\ &= -\ln(\prod_{t_{(i)} \leq t} (1 - \frac{m_i}{n_i})^{\delta_i}) \\ &= -\sum_{t_{(i)} \leq t} \ln(1 - \frac{m_i}{n_i}). \end{aligned}$$

où m_i : nombre de décès observés au temps t_i
et n_i : nombre de sujets exposés au risque de décès au temps t_i
La variance de cet estimateur est donnée par :

$$Var(\widehat{\Lambda}_2(t)) = \sum_{i,T_i \leq t} \frac{m_j}{n_j(n_j - m_j)}.$$

Remarque 3.2.1 *L'estimateur de Breslow possède de meilleures propriétés que l'estimateur de Nelson-Aalen.*

Les estimateurs de Nelson-Aalen et Breslow sont équivalents.

3.3 Estimation de la densité

Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau réel, c'est-à-dire une fonction intégrable, d'intégrale 1, on supposera que K est continue, symétrique, à support compact, et à variations bornées. Soit, de plus une suite de paramètres positifs $(h_n)_{n \geq 1}$, dite "fenêtres". Elle vérifie $h_n \to 0$.soit enfin $\tau < \inf\{t \mid p(T > T) = 0\}$.

S'il n'y a pas de censures, l'estimateur à noyau de f au point t est la convolution du noyau K avec la fonction de survie empirique S_n, i.e.

$$\begin{aligned}\widetilde{f}_n(t) &= -\int \frac{1}{h_n} K(\frac{t-u}{h_n}) S_n \, dt \\ &= \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{t-X_i}{h_n}).\end{aligned}$$

En présence de données censurées, l'estimateur empirique naturelle de la survie est $\hat{S}_{KM}$ ce qui nous donne:

$$\begin{aligned}\widehat{f}_n(t) &= -\int \frac{1}{h_n} K(\frac{t-u}{h_n}) \hat{S}_{KM} \, dt \\ &= \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{t-X_{(i)}}{h_n}) \frac{\delta_{(i)}}{n-i+1} \hat{S}_{KM}(-X_{(i)}).\end{aligned}$$

3.4 Maximum de vraisemblance et estimation non paramétrique

La méthode du maximum de vraisemblance n'est pas réservée à l'estimation paramétrique : on peut montrer que l'estimateur de Kaplan & Meier est l'estimateur non paramétrique maximisant la vraisemblance.

Ordonnons les k valeurs détectées x_i distinctes par ordre croissant et posons $x_0 = 0$ et $x_{k+1} = \infty$.

Notons n_i le nombre de valeurs τ_j supérieures ou égales à x_i, m_i le nombre de détections égales à x_i, γ_i le nombre de valeurs censurées dans $[x_i, x_{i+1}[$ et $y_i^{(j)}$ $(j \in [1, \gamma_i])$ les valeurs censurées comprises dans cet intervalle.

Pour une détection,

$$\mathbb{P}(T_i = x_i) = (S[x_i^-] - S[x_i]).$$

(Maximiser la vraisemblance de manière non paramétrique va clairement rendre $S(t)$ discontinue, ce qui n'est pas possible de manière paramétrique quand la loi de probabilité adoptée est continue.)

La vraisemblance vaut donc

$$L = \prod_{i=1}^{\gamma_0} S[y_0^{(j)}] \times ([S(x_1^-) - S(x_1)]^{m_1} \prod_{i=1}^{\gamma_1} S[y_1^{(j)}]) \times ... \times ([S(x_k^-) - S(x_k)]^{m_k} \prod_{i=1}^{\gamma_k} S[y_k^{(j)}]),$$

S étant une fonction décroissante comprise entre 0 et 1, le maximum est obtenu en prenant $S(y_0^{(j)}) = 1$ et, pour tout $i > 0$, $S(x_i^-) = S(x_{i-1})$ et $S(y_i^{(j)}) = S(x_{i+1}^-)$.

Posons

$$\begin{aligned} P_i &= S(x_i) = S(y_i^{(j)}) = S(x_{i+1}^-)) \\ \text{et } p_i &= P_i / P_{i-1}. \end{aligned}$$

Il faut désormais maximiser

$$L = \prod_{i=1}^{k} (P_{i-1} - P_i)^{m_i} P_i^{\gamma_i},$$

On a $P_i = p_1 ... p_i$, donc

$$\begin{aligned} L &= \prod_{i=1}^{k} (p_1 ... p_{i-1})^{m_i} (1 - p_i)^{m_i} (p_1 ... p_i)^{\gamma_i} \\ &= \prod_{i=1}^{k} (p_1 ... p_i)^{m_i + \gamma_i} p_i^{-m_i} (1 - p_i)^{m_i}. \end{aligned}$$

et

$$\prod_{i=1}^{k} (p_1 ... p_i)^{m_i + \gamma_i} = \prod_{i=1}^{k} p_i^{\sum_{j=1}^{k} (m_i + \gamma_i)}.$$

$d_i + \gamma_i = n_j - n_{j+1}$, donc

$$\sum_{j=1}^{k} (m_i + \gamma_i) = n_i - n_{k+1} = n_i,$$

et

$$L = \prod_{i=1}^{k} p_i^{n_i - m_i} (1 - p_i)^{m_i},$$

soit

$$L = -\sum_{i=1}^{k}([n_i - m_i]\ln p_i + m_i \ln[1 - p_i]).$$

Le minimum de L est obtenu en cherchant les $\widehat{p}_i$ solutions du système $\{(\partial L/\partial p_i = 0)_{i\in[1,k]}\}$, soit

$$\widehat{p}_i = (n_i - m_i)/n_i,$$

c'est-à-dire le résultat obtenu pour l'estimateur de Kaplan & Meier. Celui-ci est donc un estimateur de maximum de vraisemblance.

3.5 Comparaison de deux groupes

Nous souhaitons tester l'hypothèse nulle H_0contre l'hypothèse alternative H_1, avec $(H_0) : S_A(t) = S_B(t)$ vs $(H_1) : S_A(t) \neq S_B(t)$.

En présence des données censurées, nous généralisons les tests non paramétriques usuels.

Nous obtenons

- Le test de Log-Rank est une généralisation du test de Savage

- Le test de Gehan et de Peto-Prentice est une généralisation du test de Wilcoxon

(Les testes de Savage et de Wilcoxon sont des tests non paramétriques).

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

Définition du test de Log-Rank

Dit aussi test de Mantel-Haenszel, c'est le test le plus utile, le plus simple et le plus performant pour comparer deux courbes de survie à condition qu'elles ne se croisent pas et quelles soient estimées par la méthode de Kaplan-Meir.

Soit deux groupes d'individus dénotés A et B. Supposons que les instants où les évènements se produisent sont fixés t_i et comparons le nombre de sujets qui ont subi l'évènement dans chaque groupe.

Notation

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

$$t_{(1)}, t_{(2)}, ...t_{(i)}, ...t_{(k)},$$

Nous notons

m_{A_i}, m_{B_i} : nombre d'individus qui ont subi l'événement observé dans chaque groupe A et B respectivement, à l'instant t_i.

$m_{A_i} + m_{B_i} = m_i$: le nombre total d'événements réalisés dans chaque groupe, $m_i > 0$ à l'instant t_i.

n_{A_i}, n_{B_i} : nombre de sujets exposés au risque de subir l'événement dans chaque groupe en t_i.

$n_{A_i} + n_{B_i} = n_i$: Nombre total de sujets exposés au risque de subir l'événement en t_i.

n_i : Le nombre total de sujets exposés aux risques en t_i.

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

	NER en t_i	NENR aprés t_i	Total
Groupe A	m_{A_i}	$n_{A_i} - m_{A_i}$	n_{A_i}
Groupe B	m_{B_i}	$n_{B_i} - m_{B_i}$	n_{B_i}
Total	m_i	$n_i - m_i$	n_i

NER=Nombre d'événements réalisés en t_i.

NENR=Nombre d'événements non réalisés après t_i.

Statistique du test

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

Sous l'hypothèse nulle H_0 d'égalité de fonction de survie entre les deux groupes étudiés, nous avons l'indépendance des lignes et des colonnes. Dans le cas idéal $m_i = 1$ c'est-à-dire là où il n'y a pas d'ex-æquo soit :

e_{ip} : Le nombre de décès attendu sous H_0 à l'instant t_i pour le groupe taille p.

De manière générale

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

dans notre cas nous avons

$$e_{A_i} = \frac{m_{A_i} n_{A_i}}{n_i} \quad \text{et} \quad e_{B_i} = \frac{m_{B_i} n_{B_i}}{n_i},$$

e_{A_i}, e_{B_i} représentent le nombre d'événements réalisés en t_i dans les groupes A et B respectivement sous H_0.

Nous retrouvons la table de contingence attendue sous H_0 comme suit :

	NER en t_i	NENR après t_i	total
Groupe A	e_{A_i}	$n_{A_i} - e_{A_i}$	n_{A_i}
Groupe B	e_{B_i}	$n_{B_i} - e_{B_i}$	n_{B_i}
Total	m_i	$n_i - m_i$	n_i

NER=Nombre d'événements réalisés en t_i.

NENR=Nombre d'événements non réalisés après t_i.

Posons

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

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

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

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

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

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

Sous H_0

$$\chi^2_{LR} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \sim \chi^2_1.$$

Sous H_0, la quantité χ^2suit une loi de Khi-Deux à 1 degré de liberté ; au seuil $\alpha = 0.05$

nous cherchons les valeurs de χ^2_1 de la table de Khi-Deux

Si $\chi^2_{LR} < \chi^2_1$ nous acceptons H_0.

Si $\chi^2_{LR} > \chi^2_1$ nous rejetons H_0.

Remarque 3.5.1 *Nous pouvons comparer plusieurs groupes (k groupes) en utilisant la méthode de LogRank.*

exemple 3.5.1 *Soient deux groupes de patients atteints d'un cancer du poumon. On propose deux types de chimiothérapie (A/B) aléatoirement et on étudie le temps entre le traitement et le décès du patient.*

Voici les observations :

Chimio	*Mois (ti)*	*Evt (i)*
A	*24*	*1*
A	*26*	*0*
A	*30*	*1*
A	*31*	*0*
B	*22*	*0*
B	*28*	*1*
B	*30*	*0*
B	*32*	*0*

t_i	m_{A_i}	n_{A_i}	m_{B_i}	n_{B_i}	m_i	n_i	e_{A_i}	e_{B_i}
24	1	4	0	3	1	7	$4 * 1/7 = 0.57$	$3 * 1/7 = 0.43$
28	0	2	1	3	1	5	$2 * 1/5 = 0.40$	$3 * 1/5 = 0.60$
30	1	2	0	2	1	4	$2 * 1/4 = 0.50$	$2 * 1/4 = 0.50$

$$o_A = 2, o_B = 1, e_A = 1.47, e_B = 1.53$$

$$\chi^2_{LR} = \frac{(o_A - e_A)^2}{e_A} + \frac{(o_B - e_B)^2}{e_B} \sim \chi^2_{1\ ddl}.$$

.

Annex

exemple 2.1.1

Sous R la représentation graphique des fonctions ce fait par les instructions suivantes :

> t <- c(seq(0, 10, 0.01))
> f <- (1/2)*(exp(-(1/2)*t))
> F <- 1-exp(-(1/2)*t)
> S <- exp(-(1/2)*t)
> h <- f/S
> par(mfrow=c(2, 2))
> plot(t, f, type="l", col="blue", lwd=2, ylab=c("f(t)"),xlab=c("t(semaine)"), main="Fonction de densité")
> plot(t, F, type="l", col="blue", lwd=2, ylab=c("F(t)"),xlab=c("t(semaine)"), main="Fonction de répartition")
> plot(t, S, type="l", col="blue", lwd=2, ylab=c("S(t)"),xlab=c("t(semaine)"), main="Fonction de survie")
> plot(t, h, type="l", col="blue", lwd=2, ylab=c("h(t)"),xlab=c("t(semaine)"), main="Fonction de risque")

exemple 2.2.1

> library(flexsurv)
> ex <- flexsurvreg(Surv(recyrs, censrec) ~1, data=bc, dist="exponential")
> we <- flexsurvreg(Surv(recyrs, censrec) ~1, data=bc, dist="weibull")
> gg <- flexsurvreg(Surv(recyrs, censrec) ~1, data=bc, dist="gengamma")
> plot(gg, ci=FALSE, conf.int=FALSE, ylab="Survie", xlab="Années")
> lines(we, col="blue", ci=FALSE)
> lines(ex, col="green", ci=FALSE)
> legend("topright", lty=c(1,1,1), lwd=c(0.5,0.5,0.5),
+ col=c("green", "blue","red"),c("Exponential","Weibull","Generalized gamma"))

exemple 3.1.1

L'application des formules ci-dessus sous le logiciel R à ces données conduit à :

> x=c(1,3,4,5,7,8,9,10,11,13) ; d=c(1,1,0,1,0,1,1,0,1,0)
> i=1 :10 ; ksp=d[i]/(10-i+1) ; km=cumprod(1-ksp)
> library(survival) ; s=survfit(Surv(x,d) ~1)
> plot(s,xlab="mois",ylab="probabilité de survie")
> summary(s)

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	10	1	0.900	0.0949	0.7320	1.00
2	9	1	0.800	0.1265	0.5868	1.00
5	7	1	0.686	0.1515	0.4447	1.00
8	5	1	0.549	0.1724	0.2963	1.00
9	4	1	0.411	0.1756	0.1782	0.95
11	2	1	0.206	0.1699	0.0408	1.00

exemple 3.5.1

Pour notre exemple, on a :

>stat<-((2-1.47)^2)/1.47+((1-1.53)^2)/1.53
>1-pchisq(3.84,df=1)#pourcomprendre...
[1]0.05004352
>1-pchisq(stat,df=1)
[1]0.540462

L'etude ne permet pas de montrer une relation significative le type de chimiothérapie et la survie des patients (p > 0 :05).

Bibliographie

[1] **André Berchtold,** Données longitudinales et modèles de survie, 6. Modèles paramétriques, le 17/09/2018.

[2] **Christopher Jackson,** Examples of focused model comparison : parametric survival models.

[3] **Elisa T.lee,** Statistical Methodes for Survival data Analysis, Third Edition, 2003.

[4] **Frédéric Planchet,** Modèles de Durées, Statistique des modèles paramétriques et semi-paramétriques, 2020-2021.

[5] http ://statweb.stanford.edu/~ckirby/brad/other/CASI_Chap9_Nov2014.pdf.

[6] https ://www.univ-saida.dz/busc/doc_num.php ?explnum_id=194.

[7] https ://www.univ-saida.dz/busc/doc_num.php ?explnum_id=192.

[8] **Jonathan Lenoir**, Analyses de Survie, 2013.

[9] **Julien Mancini, Stéphane Robitail,** Etudes de survie – Etudes Pronostiques & Lecture Critique, Cours 7 – 2008/2009.

[10] **Laroussi Ilhem,** Données censurées-Analyse de survie, 2020.

[11] **Lyasmine Harrouche**, Analyse statistique des modèles de survie, univercité Mouloud Mammeri de TIZI-OUZOU, mémoire de master, 17/09/2018.

[12] **Philippe Saint Pierre,** Introduction à l'analyse des durées de survie, Février 2015.

[13] **Souad Belkadi,** Analyse par bootstrap des données censurées, université Houari Boumediène, mémoire de magister.

Résumé

L'analyse de survie est un domaine active de recherche, il utilise des variables de durées qui sont dans la plupart des cas incomplètes: censurées ou tronquées, on veut présenter les approches paramétrique et non paramétrique de survie en cas de censure.

Le mécanisme du travaille dans l'approche paramétrique ce base sur l'estimation par EMV en présence de censure des distributions théorique choisi graphiquement ou par le crétaire AIC, et on peut ensuite faire des testes de comparaison et précisé le cas des covariable. En trouve que l'estimation n'est pas toujours facile manuellement alors on utilise des algorithmes de maximum de vraisemblance telle que l'algorithme EM.

Pour l'approche non paramétrique on ce base sur l'estimation des fonctions de survie : L'estimation de la fonction de survie de Kaplan Meier qui est la plus utiliser, la méthode actuarielle et la méthode de Harrington et Fleming. L'estimation du risque cumulé par la méthode de Breslow et la méthode de Nelson-Aalen et l'estimation de densité. La comparaison dans l'approche non paramétrique en présence de censure se fait par le test du Log-Rank construite à partir de l'estimateur de Kaplan Meier.

L'analyse de survie nous donne des statistiques importants par la modélisation des phénomènes de plusieurs domaines comme la biologie médicale et la fiabilité pour améliorer leurs résultats.

Mots clés :
Données censurées, Analyse de survie, Estimation paramétrique, Estimation non paramétrique.

AIC : Critère d'information d 'Akaike.
EMV : Estimation par Maximum de Vraisemblance.
EM: Espérance-Maximisation.

Abstract

Survival analysis is an active area of research, it uses duration variables that are in most cases incomplete: censored or truncated, we want to present the parametric and non-parametric approaches to survival in the event of censorship.

The mechanism of the works in the parametric approach this base on the estimation by EMV in the presence of censorship of the theoretical distributions chosen graphically or by the cretary AIC, and we can then make comparison tests and specify the case of the covariates. In fact, the estimation is not always easy manually so we use maximum likelihood algorithms such as the EM algorithm.

For the non-parametric approach, we base it on the estimation of survival functions: The estimation of the Kaplan Meier survival function which is the most used, the actuarial method and the method of Harrington and Fleming. The estimation of the cumulative risk by the method of Breslow and the method of Nelson-Aalen and the estimation of density. The comparison in the nonparametric approach in the presence of censorship is done by the Log-Rank test constructed from the Kaplan Meier estimator.

Survival analysis gives us important statistics by modeling phenomena from several fields like medical biology and reliability to improve their results.

Keywords :

Censored data, Survival analysis, Parametric estimation, Nonparametric estimation.

AIC: Akaike Information Criterion.

EMV: Maximum Likelihood Estimation.

EM: Hope-Maximization.

الملخص

يعد تحليل البقاء مجالا نشطا للبحث، ويستخدم متغيرات المدة التي تكون في معظم الحالات غير مكتملة: خاضعة للرقابة أو مبتورة، نريد تقديم المقاربتين الحدودية واللامعلمية للبقاء في حالة الرقابة.

تعتمد آلية العمل في المقاربة الحدودية على التقدير بواسطة EMV في ظل وجود رقابة على التوزيعات النظرية المختارة بيانياً أو بواسطة الخاصية AIC، ويمكننا بعد ذلك إجراء اختبارات المقارنة وتحديد حالة المتغيرات المشتركة. في الواقع، ليس من السهل دائمًا التقدير يدويًا لذلك نستخدم خوارزميات الاحتمالية القصوى مثل خوارزمية EM.

بالنسبة للنهج اللامعلمي، فإننا نعتمده على تقدير وظائف البقاء على قيد الحياة: تقدير دالة البقاء على قيد الحياة في كابلان ماير وهي الأكثر استخدامًا، والطريقة الاكتوارية وطريقة هارينغتون وفليمينغ. تقدير المخاطر التراكمية بطريقة Breslow و طريقة Nelson-Aalen وتقدير الكثافة. تتم المقارنة في النهج اللامعلمي في ظل وجود رقابة عن طريق اختبار Log-Rank الذي تم إنشاؤه من مقدرKaplan Meier .

يمنحنا تحليل البقاء إحصائيات مهمة عن طريق نمذجة الظواهر من عدة مجالات مثل علم الأحياء الطبي والموثوقية لتحسين نتائجها.

الكلمات المفتاحية:

البيانات الخاضعة للرقابة، تحليل البقاء، التقدير الحدودي، التقدير اللامعلمي.