Matière : Modèles de durées et analyse de survie.

Objectifs de l'enseignement : A l'issue de ce cours, l'étudiant sera familiarisé avec les concepts et modèles de base en analyse de survie. En outre, l'étudiant sera capable d'analyser des données réelles à l'aide de logiciels.

Connaissances préalables recommandées : Des connaissances de base de statistique mathématique et de probabilités qui sont nécessaires.

Références bibliographiques : *(Livres et polycopiés, sites internet, etc)* :

 Cox, D.R. et Oakes, D. (1984). Analysis of survival data, Chapman and Hall, New York.

 Hougaard, P. (2000). Analysis of multivariate survival data. Springer, New-York.

 Klein, J.P. et Moeschberger, M.L. (1997). Survival analysis, techniques for censored and
truncated data, Springer, New

Contenu de la matière :

Chapitre 0 : Introduction

Chapitre 1 : Concepts et spécificités de l'analyse de survie

Chapitre 2 : Phénomènes de censure et de troncature

Chapitre 3 : Estimateur non paramétrique de Kaplan-Meier : Construction

Chapitre 4 : Estimateur non paramétrique de Kaplan-Meier : Propriétés

Chapitre 5 : Table de mortalité et lissage

Chapitre 6 : Lissage et estimation paramétrique

Chapitre 0 : Introduction

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant la mort pour des organismes biologiques -l'espérance de vie- ou le temps restant avant l'échec ou la panne dans les systèmes artificiels**,** Il repose souvent sur des séries temporelles de données longitudinales. Dans les cas où les événements d'intérêt ne se sont pas produits avant la fin de la période d'observation, par exemple la maladie n'est pas apparue chez un malade, on parle de censure de la série de données.

Chapitre 1 : Concepts et spécificités de l'analyse de survie

Ce chapitre est présenté sous forme d'un exposé aux étudiants.

Chapitre 2 : Phénomènes de censure et de troncature

Définition:

Une donnée est dite "censurée" si on n'en connaît pas la valeur exacte, mais seulement une estimation, inférieure ou supérieure, c'est-à-dire une information grossière, du type $X \geq c$ ou $.X \leq c$ Une telle information est très pauvre, plus pauvre que de dire "X est entre a et b", puisqu'une seule des deux bornes est connue.

Exemples de données censurées:

I-Les données existent:

1- Données liées à l'environnement :

Il s'agit par exemple de données de radioactivité, dans l'air, dans des fûts ou containers quelconques. On a mesuré l'activité de certains radionucléides, mais lorsque cette activité est inférieure au seuil réglementaire, on ne publie pas la valeur réelle, mais seulement l'information "inférieure au seuil". Cette situation se rencontre aussi dans bon nombre de mesures liées à l'environnement (niveaux de pollution, etc.).La véritable valeur de la mesure est généralement perdue ; plus exactement, elle n'est pas conservée, parce que les responsables n'en voient pas l'intérêt.

2- Données liées aux assurances :

Pour pratiquement tous les types de sinistre, les compagnies d'assurance souscrivent une "réassurance" auprès d'une compagnie spécialisée. Cela signifie que si un montant de remboursement dépasse un certain seuil, il est transféré à la compagnie de réassurance, et la compagnie d'assurance elle-même, dans ses bilans, ne fait figurer que la partie inférieure au seuil. Ceci est tout à fait légitime sur le plan comptable, puisque c'est la seule partie qui la concerne ; le vrai coût du sinistre (coût total assurance plus réassurance) est cependant conservé, dans des fichiers séparés.

II- Les données n'existent pas :

1- Dans le milieu médical :

Il est fréquent de voir un médicament essayé sur un panel (malades ou bien-portants), mais, pour une raison ou une autre, certaines personnes quittent le panel avant la fin de l'expérience et sont perdues de vue : imaginons que 15 000 personnes dans une ville soient suivies pour un traitement contre l'obésité. Si quelqu'un quitte la ville avant la fin de l'observation, les données le concernant ne sont pas complètes. Si l'observation porte sur la survie d'un patient, on pourra dire seulement dans ces conditions que sa durée de vie a été supérieure ou égale à ce qui a été observé.

2- Dans le milieu industriel :

Imaginons un industriel qui fait une expérience quant à la résistance de certaines pièces; elles sont insérées dans une machine spéciale qui les soumet à une forte pression. Cette expérience doit normalement durer plusieurs jours. Mais il arrive que, même au bout d'une semaine, la pièce n'ait pas été détruite. L'industriel ne souhaite pas poursuivre l'expérience, pour des raisons qui peuvent être diverses : la durée observée lui convient ; il a besoin de la machine pour autre chose, etc. Dans ces conditions, on dira seulement que la durée de vie de la pièce est supérieure ou égale à la durée observée. Il s'agit d'un exemple très semblable à celui du paragraphe précédent, dans son principe. Mais outre ces situations très bien décrites, on rencontre beaucoup de cas où l'information disponible est de la forme "au moins telle valeur" ou bien "au plus telle valeur".

Exemples :

- Si on cherche à dénombrer les fraudeurs (fraude fiscale, fraude aux documents administratifs,…), l'estimation dont on dispose ne tient compte que des fraudeurs démasqués, les autres sont en nombre inconnu.

- Si on veut recenser les bénéficiaires potentiels de systèmes d'aide sociale (restaurants du cœur, allocations diverses), on comptera les ayants-droit qui se manifestent effectivement, mais on ne connaît pas le nombre de ceux qui ne se manifestent pas.

Terminologie :

Parmi les censures, nous distinguerons entre :
– Censure à droite : de la forme $X \geq C$;
– Censure à gauche : de la forme $X \leq C$.

Remarque :

- Nous ne ferons pas de différence théorique entre inégalité stricte et inégalité au sens large ; dans la pratique, il suffit de déplacer légèrement la borne pour passer de l'un à l'autre.
- Pour passer d'une variable censurée à droite à une variable censurée à gauche, ou inversement, il suffit bien sûr de remplacer X par $-X$, Mais ce peut être incommode : si les valeurs mesurées pour X étaient entièrement des nombres positifs (par exemple des durées de vie, des concentrations, etc.), on se retrouve avec des valeurs entièrement négatives, ce qui est déplaisant. Il vaut mieux procéder comme suit.

Conversion d'une censure à droite en une censure à gauche, et réciproquement :

- Si B est un nombre qui majore toutes les valeurs censurées à gauche, $X \leq C \leq B$, alors $B - X \geq B - C$ et la variable $Y = B - X$ est positive et censurée à droite.

- Inversement, si B est un nombre qui majore toutes les valeurs que peut prendre X, $B \geq X$ alors si, $X \geq C$, $B - X \leq B - C$ et la variable $Y = B - X$ est positive et censurée à gauche.

On constate donc que, dans tous les cas, on peut se ramener à une variable positive et censurée à gauche, $X \leq C$, et nous nous limiterons à cette situation dans la suite. Mais attention ! Comme nous le verrons par la suite, le choix de B peut n'être pas neutre.

Remarque :

Il est clair qu'aucune méthode mathématique, si sophistiquée soit-elle, ne peut remplacer des données manquantes. Lorsqu'une information est censurée, la valeur exacte est irrémédiablement perdue. On peut néanmoins, par un travail approprié et en faisant certaines hypothèses, obtenir des résultats de deux types :

- Reconstituer une loi de probabilité pour le phénomène, qui prenne en compte les données censurées ;

- Reconstituer un "tableau d'occurrences", pour le phénomène, qui "ressemble" à celui qu'on aurait eu sans censure.

A- Loi de probabilité :

Notre objectif est ici l'obtention d'une loi de probabilité, ce qui ne poserait aucun problème si toutes les données étaient réelles, mais nous avons un certain nombre de données censurées.

Rappelons que la définition d'une loi de probabilité repose sur un découpage des données en classes (appelées "bins" en anglais), de manière à constituer un histogramme. Ce découpage doit être fait en fonction des objectifs, et non en fonction des données.

Exemple :

Pour une durée de vie, on se demandera : ai-je besoin de l'information heure par heure, jour par jour, ou bien simplement par année ? C'est le besoin qui va conditionner le découpage ; peu importe que les données soient fournies ou non avec vingt chiffres après la virgule.

En d'autres termes encore, le découpage en classes ne peut se faire seulement sur critères mathématiques, mais doit incorporer une analyse du besoin. Ceci est très important et est trop souvent ignoré.
Comme pour nous la valeur des classes n'a aucune importance, nous dirons qu'elles sont représentées par des entiers $1, 2, \ldots, K$.
Le découpage en classes répond à la question des incertitudes sur les données : toutes les données à l'intérieur d'une même classe sont considérées comme ayant la même valeur (le centre de la classe). En d'autres termes, on "grossit" artificiellement l'incertitude sur les données.

Exemple :

On identifie les données correspondant à la même journée, quelle que soit la mesure de la précision horaire.

B- Tableau d'occurrences

Définition:

Un tableau d'occurrences est le résultat d'une expérience : à telle date, ou sur telle personne, la variable X a pris telle valeur ; il y a autant de lignes dans le tableau que de répétitions de l'expérience. Un tel tableau, en soi, n'a rien de probabiliste : c'est un compte-rendu d'une expérience. Il n'y a aucune difficulté si toutes les données sont réelles, malheureusement il se peut se trouver que certaines (voire toutes) sont censurées ; on voudrait néanmoins disposer d'un tableau d'occurrences qui reflète l'expérience réalisée, mais sans contenir de censure.

Chapitre 3 : Estimateur non paramétrique de Kaplan-Meier : Construction

Méthode de Kaplan Meier :

Elle a été introduite, dans le milieu médical, pour des durées de vie, qui sont de la forme $X \geq c$ ou $X > c$. Ces durées de vie pouvant prendre une valeur quelconque, une phrase du type "l'évènement n'a pas encore eu lieu à l'instant t équivaut à dire qu'il n'a pas eu lieu juste avant t et n'a pas lieu en t", qui est à la base de la méthode, comme on va le voir, paraît acceptable. Mais, une fois les classes définies, le "juste avant" pose problème. En fait, cette méthode, quoique largement utilisée dans le milieu médical, n'est pas correcte.

Dans ce manuel, nous avons développé la théorie pour des données sous la forme $X \leq c$ et c'est ainsi que nous allons poursuivre la présentation.

Le passage de l'un à l'autre est facile : il suffit de remplacer X par, $B-X$ où B est la plus grande valeur que peut prendre la variable X.

Présentation théorique :

Comme précédemment, nous avons K classes, numérotées de 1 à K; la k–ème classe est l'ensemble $k-1 < x \leq k$. Pour chaque $k=1,...,K$ nous notons m_K le nombre de données censurées et n_K le nombre de données exactes dans la k–ème classe. Nous notons $M=m_1+...+m_k$ et $N=n_1+...+n_k$ ce sont,

respectivement, le nombre total de données censurées et le nombre total de données exactes. Nous notons aussi $M_k=m_1+\ldots+m_k$ et $N_k=n_1+\ldots+n_k$, $k=1,\ldots,K$.

Commençons, pour présenter la méthode de **Kaplan-Meier**, par supposer que toutes les classes contiennent des données exactes (peu importe en quel nombre). Alors on écrit :

$$P(X \leq j)=P(X \leq j \mid X \leq j+1) \times P(X \leq j+1)$$

Et en réitérant :

$$P(X \leq j)=\Pi_{k=j}^{K-1} P_k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Où l'on note :

$$P_k = P(X \leq k \mid X \leq k+1),\ k=1,\ldots,K-1.$$

Remarque :

Pour la dernière, $P_{k-1} = P(X \leq k \mid X \leq k+1)$, le conditionnement est automatique, puisque la condition $X \leq K$ est toujours satisfaite.

Evaluation de P_k :

Le nombre total de données satisfaisant $X \leq k+1$ est, par définition, $M_{k+1}+ N_{k+1}$; le nombre de données satisfaisant $X \leq$ k est estimé par $M_k+ N_k$ cela revient à dire que toutes les données censurées vérifiant $X \leq k+1$ doivent automatiquement vérifier $X \leq k$ (ou, en d'autres termes, qu'il n'y en a pas entre k et $k+1$) ; cela peut sembler correct pour des durées de vie, si les classes sont à la seconde près(très bref intervalle de temps), mais c'est évidemment faux en général.

Avec cette approche, on obtient l'estimation :

$$P_k = (M_{k+1}+ N_k)/(M_{k+1}+ N_{k+1})$$

D'où il résulte par (1):

$$P(X \leq j) \approx \Pi_{k=j}^{K-1} ((M_{k+1}+ N_k)/(M_{k+1}+ N_{k+1}))$$

et par différence :

$$P(X \in C_j)=P(X \leq j)-(X \leq j-1)$$

$$= (n_j/(M_j + N_j))\Pi_{k=j}^{K-1}((m_{k+1}+ n_k)/(m_{k+1}+ n_{k+1}))$$

et le nombre de données par classe, après répartition selon cette méthode, est

$$e_k=(N+M)P(X=k)$$

Exemple :

On relève 70 observations réelles ou censurées à gauche, comme ci-dessous. Les données censurées $(X \leq a)$ sont notées a*. Les voici, mises dans l'ordre croissant :

0.09, 0.11, 0.13, 0.15*, 1.32, 1.33, 1.50*,1.70,1.65,1.77*, 1.78*, 1.81, 2.34*,2.55*, 2.59*,2.59, 2.63*, 2.87, 2.96*, 3.01, 3.07*,3.15, 3.19,3.23, 3.27, 3.41, 3.50*, 3.60, 3.81*, 3.96, 4.05, 4.15*,4.22*4.34, 4.55, 4.60*,4.77, 4.69*, 4.72*, 4.80, 4.82*,5.10*, 5.25, 5.30, 5.44, 5.50, 5.50, 5.51, 5.66, 6.02, 6.03, 7.10*, 7.15,7.42, 7.44*, 7.50*, 7.62*, 8.03, 8.10*, 8.15, 8.20*,8.96, 9.03,9.16, 9.27, 9.39*, 9.47, 9.72, 9.96*, 9.97*.

On choisit de ranger ces données dans 10 classes, d'amplitude 1. Le tableau ci-dessous récapitule le nombre de données selon la classe. Par exemple, 0.09 est compté dans la classe 1car $0< 0.09\leq 1$.

Tableau 1: Données réelles et censurées**.**

Classe	Nombre de valeurs réelles	Nombre de valeurs censurées $X \leq j$
1	3	1
2	5	3
3	2	5
4	8	3
5	5	6
6	7	1
7	2	0
8	2	4
9	3	2
10	5	3

Tableau 2 : Calculs intermédiaires de la méthode de **Kaplan-Meier**

Classe	N_k	M_k	P_K	$P(X \leq j)$	$P(X \leq j)$	e_j
1	2	1	0.58	0.24	0.24	16.52
2	8	4	0.89	0.40	0.17	11.80
3	10	9	0.73	0.45	0.05	3.33
4	18	12	0.88	0.62	0.16	11.51
5	23	18	0.86	0.70	0.09	5.99
6	30	19	0.96	0.82	0.12	8.19
7	32	19	0.96	0.85	0.03	2.34
8	34	23	0.95	0.88	0.03	2.17
9	37	25	0.93	0.93	0.04	3.15
10	42	28		1.00	0.07	5.00

Tableau 3 : Répartition des données d'après la méthode de **Kaplan-Meier.**

Classe	Nombre de valeurs réelles
1	17
2	12
3	3
4	12
5	6
6	8
7	2
8	2
9	3
10	5

Chapitre 4 : Estimateur non paramétrique de Kaplan-Meier : Propriétés

Présentation générale :

L'estimateur de Kaplan-Meier (KAPLAN et MEIER [1958]) peut être introduit via les processus ponctuels, en remarquant que la fonction de survie de base du modèle est l'unique solution de l'équation intégrale suivante :

$$S(t) = 1 - \int_0^t S(u-)h(u)du.$$

L'équation ci-dessus exprime simplement le fait que la somme des survivants en *t* et des individus sortis avant *t* est constante.

Lorsque la fonction de survie est continue, la démonstration est immédiate en effectuant le changement de variable

$$v = lnS(u) \Rightarrow dv = -h(u)du$$

En remplaçant $h(u)du$ par son estimateur $\frac{d\bar{N}^1(u)}{\bar{R}(u)}$ introduit à la section précédente on peut proposer un estimateur de la fonction de survie en cherchant une solution à l'équation :

$$\hat{S}(u) = 1 - \int_0^t \hat{S}(u-)\frac{d\bar{N}^1(u)}{\bar{R}(u)} \text{.....(2)}$$

Théorème et définition :

L'équation (2) admet une solution unique, cette solution est appelée l'estimateur de **Kaplan-Meier** de la fonction de survie.

Propriétés :

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.

(Voir DROESBEKE et al. [1989] pour la démonstration de cette propriété).