RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DES FRÈRES MENTOURI – CONSTANTINE 1
FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT DE MATHÉMATIQUES

MASTER 2 SPA

POLYCOPIÉ DE COURS

DONNÉES CENSURÉES

Par Dr. Sarra LEULMI

Année 2022/2023

Á la mémoire de mon cher père
Prof. LEULMI Salah.

Table des matières

Introduction **1**

1 Introduction à l'analyse de survie **6**

1.1 Différents types de censure et de troncature 7

1.1.1 Censures 7

1.1.2 Troncatures 13

1.2 Fonctions essentielles en analyse de survie 15

1.2.1 Fonction f_T 15

1.2.2 Fonction de répartition 16

1.2.3 Fonction de Survie 16

1.2.4 Taux de hasard (Risque instantané) 17

1.2.5 Fonction de hasard cumulé (Risque cumulé) 19

1.3 Quantités associées à la distribution de survie 23

1.3.1 Paramètres de position associés à la distribution de survie 23

1.3.2 Paramètres de dispersion associés à la distribution de survie 24

1.4 Exercices 26

2 L'approche paramétrique dans un modèle de censure **28**

2.1 Modèles paramétriques pour l'analyse de survie 29

2.1.1 Risque instantané constant 29

2.1.2 Risque instantané monotone 32
2.2 Méthode du maximum de vraisemblance dans le cas complet . 35
2.2.1 Fonction de vraisemblance 35
2.2.2 Estimateur du maximum de vraisemblance 35
2.2.3 Application sur les lois 36
2.3 Méthode du maximum de vraisemblance dans le cas de censure à droite 38
2.3.1 Fonctions de loi de probabilité 38
2.3.2 Fonction de vraisemblance 41
2.3.3 Estimateur du maximum de vraisemblance 43
2.3.4 Application sur les lois 43
2.4 Méthode du maximum de vraisemblance dans le cas de censure double 45
2.4.1 Fonctions de loi de probabilité 46
2.4.2 Fonction de vraisemblance 49
2.4.3 Estimateur du maximum de vraisemblance 50
2.5 Méthode du maximum de vraisemblance dans le cas de censure générale 51
2.6 Tests paramétriques 52
2.6.1 Récapitulation des tests d'hypothèses 53
2.6.2 Test du rapport de vraisemblance 53
2.6.3 Test de Wald 58
2.6.4 Test de Rao 60
2.7 Exercices 63

3 L'approche non paramétrique dans un modèle de censure à droite 69
3.1 Estimation de la fonction de survie et de risque cumulé 69
3.1.1 Estimateur de Kaplan-Meier pour la fonction de survie 69

3.1.2 Estimateur du risque cumulé de Breslow 74
3.1.3 Estimateur du risque cumulé de de Nelson-Aalen . . . 74
3.1.4 Estimateur de Harrington et Fleming pour la fonction de survie 77
3.2 Propriétés de l'estimateur de Kaplan-Meier 77
3.2.1 Convergence presque complète 78
3.2.2 Normalité asymptotique 85
3.3 Tests non paramétriques de comparaison de deux fonctions de survie pour des échantillons censurés 86
3.3.1 Tests Gehan 87
3.3.2 Tests log-rang 90
3.4 Exercices 92

4 Estimation des fonctions de densité et de régression pour des données censurées à droite 95
4.1 Estimation de la fonction de densité pour des données censurées à droite 95
4.1.1 Principe de l'estimation 95
4.1.2 Convergence presque sûre 96
4.1.3 Convergence presque complète 97
4.2 Estimation de la fonction de régression pour des données censurées à droite 102
4.2.1 Principe de l'estimation 102
4.2.2 Convergence presque complète 106
4.3 Exercices 120

5 Introduction à l'estimation non paramétrique pour des données doublement censurées 122
5.1 Estimation de la fonction de survie 122
5.1.1 Principe de l'estimation 123

5.2 Estimation de la fonction de régression 127
5.2.1 Principe de l'estimation 127
5.2.2 Convergence presque complète ponctuelle 131
5.3 Exercices 138

Bibliographie **140**

Introduction

La théorie d'estimation joue un rôle important dans nombreuses domaines comme la médecine, l'assurance, la fiabilité, la démographie, etc... La recherche dans ces domaines n'a à première vue rien en commun. Toutefois, dans chacun de ces domaines, on s'intéresse à des variables aléatoires (v.a.) représentant la durée de temps jusqu'à l'apparition d'un certain événement, appelées durées de survie. En médecine, elle est la durée de survie ou le temps de rémission d'un patient, alors qu'en industrie, cette variable peut être la durée séparant deux pannes successives d'une machine. Dans le domaine économique, elle est la durée de chômage ou le temps passé dans un emploi, la durée de vie d'une entreprise, etc...

La difficulté rencontrée en statistique est le manque d'information durant les différentes observations sur diverses échelles d'études et d'analyses. Ce manque d'information introduit un empêchement lors de l'observation et l'obtention du résultat. La censure et la troncature sont deux causes de données manquées (incomplètes) les plus répandues. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt. On sait bien que ce délai appartient à un certain intervalle de temps. La troncature survient qu'on ne peut pas observer les individus de l'échantillon dont le délai de survenue appartient à un certain intervalle de temps, on observe donc un sous-échantillon.

Dans les études de survie, les problématiques abordées sont principalement

l'estimation des fonctions de survie et les analyses uni-variées et multivariées des relations entre temps de survie et variable explicatives. La théorie de l'estimation est divisée en deux parties principales ; l'estimation paramétrique et non paramétrique.

L'estimation paramétrique souvent presque comme l'une des méthodes d'estimation les plus fiables, mais elle nécessite un effort important dès le départ. Les estimateurs des paramètres du modèle sont ensuite obtenus par plusieurs méthodes par exemple la méthode des moindres carrés qui consiste à trouver le paramètre qui minimise la somme des carrés des erreurs et la méthode du maximum de vraisemblance cherche par contre à trouver la valeur la plus vraisemblable du paramètre d'une population partant d'un échantillon donné. Cette méthode a été développée par le statisticien et généticien Ronald Fisher entre 1912 et 1922.

Par ailleurs, l'estimation non paramétrique connaît un grand essor chez de nombreux auteurs et dans différents domaines. Celle-ci ne fait aucune hypothèse sur la loi, ni sur ces paramètres, nos connaissances sur le modèle ne sont pas précises, ce qui est souvent le cas dans la pratique. Nombreux auteurs ont introduit des techniques nouvelles pour aborder les modèles de censure. Le travail de Kaplan et Meier [1958] présente d'importants résultats concernant l'estimation non paramétrique de la fonction de survie pour des variables aléatoires censurées à droite. Plusieurs chercheurs étudient l'espérance, la variance et les propriétés asymptotiques de cet estimateur. Ce résultat a permis l'estimation de caractéristiques fonctionnelles associées à la loi des observations concernant, par exemple, l'estimateur à noyau de la densité, introduit par Foldes et Rejto [1981a], du taux de hasard, élaboré par Foldes et al. [1981] et Diehl et Stute [1988], ou de la régression, donné dans Kohler et al. [2002], dédiés à l'étude de modèles impliquant des variables aléatoires censurées à droit. Ces estimateurs à noyau restent parmi les plus utilisés, leurs propriétés asymptotiques ont beaucoup enrichi la litté-

rature statistique, ils ont été étudiés de manière intensive pour des données indépendantes ainsi que pour des données soumises à un certain type de dépendance.

Il est vrai que la censure à droite est la plus fréquente dans la pratique, mais un phénomène de censure à gauche peut aussi intervenir dans certaines situations. Généralement, la censure à gauche s'accompagne de la censure à droite ce qui donne des données doublement censurées. Parmi les modèles qui traitent ce genre de données deux ont été largement étudiés dans la littérature statistique, à savoir le modèle de Turnbull (1974) connu par le modèle de censure double et le modèle de Patilea et Rolin (2006) que nous nommons par le modèle de censure mixte. Dans ce contexte, Patilea et Rolin (2006) donnent un estimateur produit limite de la fonction de survie de la durée d'intérêt qui généralise le célèbre estimateur de Kaplan et Meier (1958) et démontrent sa convergence uniforme presque sûre ainsi que sa normalité asymptotique sous certaines conditions d'identifiabilité. Messaci (2010) a introduit des estimateurs à poids de la régression, dont l'estimateur à noyau. Kebabi et Messaci (2012) ont donné un taux de convergence presque complète ponctuelle et uniforme de l'estimateur à noyau.

La théorie des tests est une autre branche de la statistique mathématique. Ils permettent de prendre une décision sur la loi d'une population donnée par rapport à un facteur étudié. Ils permettent aussi de comparer entre deux ou plusieurs populations. C'est pour cela qu'ils représentent un outil très utilisé dans la pratique. Cette théorie se subdivise en deux volets principaux, les tests paramétriques et les tests non paramétriques. Ainsi, en présence de censure, il est impossible d'utiliser des tests de comparaison classiques comme le test de comparaison des moyennes de Student, sous peine d'obtenir des résultats biaisés. Il apparait indispensable d'avoir recours à une méthodologie spécifique aux données censurées.

Parmi les tests non paramétriques les plus usuels dans le cas de données cen-

surées, les tests de rangs (voir Gehan (1965), Tarone et Ware (1977), Peto et Peto (1973) et puis la généralisation de Fleming et Harrington (1981)). D'autre part, nous distinguons trois tests paramétrique asymptotiquement équivalents qui sont test du rapport de vraisemblance, test de Weibul et test de Rao.

Ce fascicule est un support au cours de la statistique non paramétrique deuxième année du Master de l'option Statistique des Processus Aléatoires (SPA) de l'université Constantine 1. L'objectif de ce cours est de présenter aux étudiants les principales techniques statistiques (paramétrique et non paramétrique) utilisées pour l'analyse des durées de réalisation d'un ou de plusieurs évènements d'intérêt. Nous commençons par donner les différentes notions des caractéristiques particulières souvent présentes dans les données de survie ainsi nous allons définir les cinq fonctions de base qui jouent un rôle très important en analyse de survie et nous allons voir comment elles sont interdépendantes. Ensuite, dans le deuxième chapitre, nous présentons la fonction de vraisemblance dans le cas des données censurées et nous étudions le problème de l'estimation des paramètres en utilisant la méthode du maximum de vraisemblance. Puis, nous disposons trois tests de comparaison dans le cadre des données censurées. L'approche non paramétrique est traitée dans le chapitre trois pour des données censurées à droite.

Le quatrième chapitre est consacré essentiellement à présenter les résultats sur les estimateurs non paramétriques de la densité et de la régression dans le cas des données censurées à droite. L'estimation non paramétrique de la fonction de survie et de la régression pour des données doublement censurées sont étudiés dans le dernier chapitre. Une série d'exercices est fournie à la fin de chacun de ces chapitres.

Étant donné le temps alloué à cette unité (2 cours + 1 TD) par semaine. Le lecteur intéressé peut se référer à la littérature pour plus de détails. Lors de la préparation de ce polycopié, nous avons utilisé à bon escient un ensemble de

références présentés à la fin du document pour approfondir les sujets abordés. Enfin, merci de me communiquer toute erreur éventuelle dans le fond ou dans la forme de ce polycopié.

Chapitre 1

Introduction à l'analyse de survie

Le problème des données incomplètes est très vaste et a suscité beaucoup d'intérêt parmi les statisticiens ces dernières années. Dans le domaine des durées de survie, les données sont souvent incomplètes à cause du phénomène de la censure.

Par définition, le temps de survie d'un individu est dit censuré, lorsque sa valeur exacte n'est pas observée ; seules des bornes supérieures ou inférieures pour cette valeur sont disponibles.

La censure peut se manifester pour différentes raisons : l'événement d'intérêt n'est pas survenu au moment de l'analyse, un sujet peut être perdu de vue avant d'avoir expérimenté l'événement d'intérêt, un événement concurrent peut être survenu avant l'événement d'intérêt, un sujet peut être exclu de l'étude sans avoir expérimenté l'événement d'intérêt, ... etc.

Il existe plusieurs types de censure dont la censure à droite, la censure à gauche et la censure par intervalle. La censure à droite est la forme de censure la plus commune dans les études médicales.

Dans ce chapitre nous énonçons des définitions et des notations intervenant dans la suite, et nécessaires pour une meilleure compréhension de ce cours. Nous commençons par définir le concept général de la censure, ensuite nous

évoquerons par la suite les cinq fonctions essentielles en analyse de survie.

1.1 Différents types de censure et de troncature

Une durée de vie (la durée de survie) est dite censurée si tout ce que l'on sait est qu'elle commence ou se termine dans un intervalle de temps particulier, et souvent en dehors de la période de suivi. La durée exacte n'est donc pas connue. C'est une variable aléatoire positive et souvent supposée bornée. Cela peut être la durée de vie d'un patient après un traitement, la durée de chômage, le temps de panne d'un appareil, l'âge auquel un enfant apprend à accomplir une tâche donnée, ... etc.
Il arrive souvent, pour diverses raisons, que la durée d'intérêt ne puisse pas être observée. Cela peut être dû à la perte de vue d'un patient, au début ou à la fin de la période d'étude, ... etc. Ces valeurs sont censurées. On distingue usuellement plusieurs types de censure.

1.1.1 Censures

Censure à droite

On dit que les durées de survie des sujets sont censurées à droite , qu'on n'a pas d'information sur le fait que ces sujets ont eu ou non l'événement au delà de la période d'observation entière.
Pour les sujets pour lesquels l'évènement d'intérêt a lieu pendant la période d'observation, on dispose du délai exact d'apparition de cet évènement d'intérêt, mesuré depuis une date initiale qu'il faut spécifier sans ambiguïté. Cependant, à la fin de la période d'observation, l'évènement d'intérêt n'est pas réalisé chez certains sujets. On n'aura alors pour ces sujets qu'une information incomplète, à savoir que le délai d'apparition de l'évènement est plus

grand que la durée d'observation. De telles données sont dites censurées à droite.

Une autre cause de censure à droite, qu'on essaie de limiter au maximum dans les études, correspond aux sujets dits "perdus de vue", qui sont les individus qui ont quitté l'étude avant l'apparition de l'événement d'intérêt et dont on n'a plus de nouvelles à la date de fin d'étude.

De façon formelle, Il y a censure à droite lorsque à la fin de la période de suivi, l'évènement d'intérêt ne s'est pas encore produit, c'est-à-dire que la seule information dont on dispose est que la durée de survie du i^{eme} sujet, T_i est strictement supérieure à sa date de dernière nouvelle R_i. Dans ce cas on observe le couple $Y = (\min(T, R), \delta)$ où $\delta = 1_{\{T \leq R\}}$ est appelé l'indicateur de censure. Remarquons que δ vaut 1 quant l'observation est complète, ce qui veut dire qu'elle correspond à une vraie valeur de la variable d'intérêt et vaut 0 sinon (donnée censurée).

Ce modèle est le plus fréquent en pratique, un exemple classique de censure droite est celui où l'étude porte sur la durée de survie T de patients atteints d'une certaine maladie. Pour les patients perdus de vue au bout du temps R alors qu'ils étaient encore vivants, R censure T à droite puisque, pour eux, T est inconnue mais supérieur à R, c'est-à-dire $T > R$.

Un autre exemple de ce genre de phénomène est études de fiabilité quand la panne d'un appareil ne permet pas de poursuivre l'observation pour l'appareil objet de notre étude.

A l'intérieur de ce modèle, il existe différents types de censure.

Censure de type I : Non-aléatoire (ou fixe)

Ce modèle décrit la situation où un test qui se termine à une certaine période, et nous savons que les individus restants n'ont pas échoué à ce moment. Par exemple, nous démarrons l'étude avec 100 ampoules électriques, et terminons l'expérience après un certain temps. Dans ce cas, le temps de censure est souvent fixé, et le nombre d'individus ayant échoué est une variable aléatoire.

Soit un échantillon de durées de survie $(T_1, ..., T_n)$. La variable de censure R ici est fixe, les observations deviennent alors (Y_1, δ_1), ..., (Y_n, δ_n) où pour $i \in \{1, ..., n\}$,

$$Y_i = T_i \wedge R = \min(T_i, R)$$

et

$$\delta_i = 1_{\{T_i \leq R\}}.$$

Ce modèle, souvent utiliser dans les études épidémiologiques, peut correspondre à l'observation de la durée de survie de n patients au cours d'une expérience de durée prédéterminée R.

Censure de type II : Attente

Pour ce modèle, l'expérience se poursuit jusqu'à ce qu'un nombre fixe d'individus k ait échoué. Nous reprenons l'exemple des ampoules, où dans ce cas nous arrêtons l'expérience une fois exactement 50 ampoules électriques ont échoué. Ici, le nombre d'individus ayant échoué est fixe, et le temps est une variable aléatoire.

Soit un échantillon de durées de survie $(T_1, ..., T_n)$ et $k > 0$ fixé. On dit qu'il y a censure de type II pour cet échantillon si au lieu d'observer directement $(T_1, ..., T_n)$ on observe (Y_1, δ_1), ..., (Y_n, δ_n) où, pour $i \in \{1, ..., n\}$,

$$Y_i = T_i \wedge T_{(k)} = \min(T_i, T_{(k)})$$

et

$$\delta_i = 1_{\{T_i \leq T_{(k)}\}}$$

Ce modèle souvent utilisé dans les études de fiabilité, correspond à l'observation de la durée de fonctionnement de n machines tant que k d'entre elles ne sont pas tombées en panne.

Censure de type III : Censure aléatoire

Ce type de censure généralise la censure de type I au cas où R est une v.a. Soient un échantillon de durées de survie $(T_1, ..., T_n)$ et un second échantillon indépendant composé de variables positives $(R_1, ..., R_n)$. On dit qu'il y

a censure de type III pour cet échantillon si au lieu d'observer directement $(T_1, ..., T_n)$ on observe les couples $(Y_1, \delta_1),\ ...,\ (Y_n, \delta_n)$ où pour $i \in \{1, ..., n\}$,

$$Y_i = T_i \wedge R_i = \min(T_i, R_i)$$

et

$$\delta_i = 1_{\{T_i \leq R_i\}}.$$

Dans les modèles classiques d'analyse de survie (cas de censures aléatoires), une hypothèse fondamentale sur les censures consiste à supposer que la censure est indépendante de l'événement. Cela signifie que les perdus de vus ne le sont pas pour des raisons liées à l'évènement étudié. Cette hypothèse est importante d'un point de vue mathématique et règle le problème d'indéfectibilité.
Un autre exemple, nous nous intéressons à une cause de décès ayant lieu au bout d'un temps T et nous désirons connaitre la loi de T ; cependant, une autre cause aléatoire (décès dû à une autre cause aléatoire, abandon, ...) peut survenir auparavant et empêcher par conséquent l'observation de T.

Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'évènement avant que l'individu soit observé. On sait uniquement que la date de l'événement d'intérêt est inférieure à une certaine date connue L, c'est-à-dire $T < L$. Les observations sont des répliques du couple (Y, δ) où

$$Y = T \vee L = \max(T, L)$$

et

$$\delta = 1_{\{T \geq L\}}.$$

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.
Un autre exemple sur la censure à gauche, un ethnologue étudie la durée d'apprentissage d'une tâche. Cette durée est une variable aléatoire T est l'âge de l'enfant. Pour les enfants qui savent déjà accomplir la tâche, L censure T à gauche car pour eux T est inconnu mais inférieur à $L : T < L$.

Censure double

Il y a des situations où les données d'un même échantillon peuvent être censurées soit à droite, soit à gauche. Par exemple, Leiderman (1973) ont étudié l'âge auquel les enfants d'une communauté africaine apprennent à 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 à 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 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 R). L'âge d'intérêt est observé ssi il se trouve dans la période d'étude. Dans cet exemple, on trouve dans un même échantillon des données censurées à gauche aussi bien que des données censurées à droite.

Censure mixte

Nous parlons de modèle mixte lorsque la censure à droite et la censure à gauche apparaissent dans un même échantillon simultanément, sans qu'on

puisse nécessairement déterminer un intervalle auquel il appartient.. Pour cela notons par, T le temps de survie, R la variable de censure droite et L la variable de censure gauche. Ce modèle est données par Patilea et Rolin (2006), dans le modèle I. Nous observons un échantillon du couple $(Z,\ A)$ où la variable observée est $Z = max(Y,\ L)$ avec $Y = min(T,\ R)$ (les variables T , R et L sont des variables aléatoires positives et indépendantes) et l'indicateur de censure A, est une variable discrète à valeur $\{0,\ 1,\ 2\}$ où 0 correspond à l'observation de T , tandis que 1 et 2 indique l'observation de R et L, respectivement. Plus précisément,

$$Z = max(min(T, R); L)$$

et

$$A \begin{cases} 0 & si \quad L < T \leq R \\ 1 & si \quad L < R < T \\ 2 & \min(T, R) \leq L \end{cases}$$

où L et R sont des variables de censure et A est l'indicateur de censure.
Un exemple d'application de ce modèle donné par Patilea et Rolin (2006) consiste à considérer un système de trois composants C_1, C_2 et C_3, avec C_1 et C_2 en série et C_3 en parallèle avec le système composé de C_1 et C_2. Les variables T , R et L représentent les temps de survie de C_1, C_2 et C_3 respectivement, et nous pouvons déterminer quel composant tombe en panne au même temps que le système. Donc au lieu d'observer T , on peut seulement observer le couple $(Z,\ A)$.

Censure par intervalle

La durée de survie X_i est dite censurée par intervalle si au lieu de l'observer de façon exacte, la seule information dont on dispose est qu'elle soit comprise entre deux dates connues $(C_1 < X_i < C_2)$. La censure par intervalle

se rencontre généralement lorsque les sujets ne sont pas observés en temps continu mais par intermittence lors de visites. Par exemple, si l'on s'intéresse à l'âge de survenue d'une maladie et que le sujet i est diagnostiqué malade au cours d'une visite, on sait seulement que $X_i \in [C_1,\ C_2]$ où C_1 est l'âge à la visite de diagnostique et C_2 est l'âge à la visite précédente.
En particulier, la censure gauche peut être considérée comme une censure par un intervalle tel que $C_1 = -\infty$ et la censure droite par une censure par intervalle tel que $C_1 = +\infty$.
Une autre exemple, pour détecter les composants défectueux d'un processus de production industriel, on effectue des contrôles selon des dates aléatoires. Lorsqu'on constate qu'un composant est à changer, on sait seulement qu'il est tombé en panne entre les dates de deux contrôles successifs.

1.1.2 Troncatures

Une autre situation dans laquelle les données incomplètes apparaissent est celle des données tronquées. Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui même. Ainsi, une variable T est tronquée par un sous ensemble éventuellement aléatoire $\mathcal{A}$ de $\mathbb{R}$ si au lieu de T, on observe T uniquement si $T \in \mathcal{A}$. Les points de l'échantillon "tronqué" n'appartiennent pas tous à $\mathcal{A}$. Il ne faut pas confondre censure et troncature. S'il y a troncature, une partie des individus ne sont pas observables et on n'étudie qu'un sous échantillon.
Il y a différentes types de troncature

Troncature à gauche

Troncature à gauche lorsque T n'est observable que si elle est supérieure à un seuil Y positif fixé ou aléatoire. Autrement dit, soient T et Y deux variables, T indépendante à Y, si $T > Y$ on n'observe la variable T dans ce

cas on dit que T est tronquée à gauche.
Par exemple, Une étude de survie des résidents du centre de retraite en Californie est décrite. On enregistre l'âge au décès et l'âge d'entrée au centre. Un individu doit survivre à un âge suffisant pour entrer dans le centre, tous les individus qui sont morts ne seront pas entrés dans le centre et sont donc hors de la connaissance de l'enquêteur, ces personnes n'ont aucune chance d'être dans l'étude et sont considérées comme tronquées à gauche.
Une autre exemple de ces données, lorsque on souhaite étudier combien de temps les gens qui ont été hospitalisés pour une crise cardiaque survivent en prenant certains traitements domiciles. L'heure de début est prise à l'époque de la crise cardiaque. Seuls les gens qui survivent pendant leur séjour à l'hôpital sont susceptibles d'être inclus dans l'étude.

Troncature à droite

lorsque T n'est observable que si elle est inférieure à Y.
Lagakos en (1998)(voir Klein et Moeschberger (1997)) présentent des données sur les temps d'infection et l'induction pour 258 adultes et 37 enfants qui ont été infectés par le virus de SIDA. Ici, le nombre de personnes infectés est inconnu et l'information est disponible seulement pour ceux qui ont été infectés et développés le SIDA dans un certain laps de temps. Ainsi, les personnes qui n'ont pas encore développé le SIDA ne sont pas connues à l'enquêteur et ne sont pas inclus dans l'échantillon.

Troncature par Intervalle

On dit qu'une durée est tronquée par intervalle si elle est tronquée à droite et à gauche.
On rencontre ce type de troncature lors de l'étude des patients d'un registre : les patients diagnostiqués avant la mise en place du registre ou répertoriés

après la consultation du registre ne seront pas inclus dans l'étude.

1.2 Fonctions essentielles en analyse de survie

Soit T une variable aléatoire (v.a.) positive (appelé temps de survie), La loi de probabilité de T peut être définie par cinq fonctions. Ces fonctions jouent un rôle important en analyse de survie, nous allons les introduire et donner les liens existant entre elles dans deux cas particulièrement intéressants dans la pratique, qui sont le cas discret et le cas absolument continu.

• Dans le cas discret : Supposons que T une variable aléatoire discrète prenant les valeurs ordonnées, par ordre croissant t_1, ..., t_n, ...

• Dans le cas absolument continu : Supposons que T a une densité f_T (par rapport à la mesure de Lebesgue).

1.2.1 Fonction f_T

Dans le cas discret

La fonction de masse d'une variable aléatoire discrète T est la fonction $f_T : \mathbb{R} \to [0, 1]$ donnée par

$$f_T(t) = P(T = t).$$

Dans le cas absolument continu

Une variable aléatoire T est à densité s'il existe une fonction Lebesgue intégrable positive f_T telle que

$$P(T \in \mathcal{A})) = \int_{\mathcal{A}} f_T(t)dt,$$

où $\mathcal{A} \in \mathcal{B}(\mathbb{R})$ avec $\mathcal{B}(\mathbb{R})$ est une tribu borélienne.

Notons que la valeur $f_T(x)$ peut être plus grande que 1 alors $f_T(x)$ n'est pas

une probabilité. Par contre, f_t est la mesure la probabilité que l'événement d'intérêt se produise dans un petite intervalle de temps infinitésimal.

1.2.2 Fonction de répartition

Pour un t fixé, elle représente la probabilité que l'événement d'intérêt se produise avant la date t.

$$F_T(t) = P(T \leq t).$$

Il est clair que F_T est une fonction croissante, continue à droite avec

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

Dans le cas discret

Nous avons

$$F_T(t) = P(T \leq t) = \sum_{t_i \leq t} P(T = t_i).$$

Dans le cas absolument continu

Nous avons

$$F_T(t) = P(T \leq t) = \int_0^t f_T(s)ds.$$

1.2.3 Fonction de Survie

Pour un t fixé, c'est la probabilité de survivre au moins jusqu'à la date t.

$$S_T(t) = P(T > t).$$

Il est clair que S_T est une fonction décroissante, continue à droite avec

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

et

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

Dans le cas discret

Nous avons

$$S_T(t) = P(T > t) = \sum_{t_i > t}^{P} (T = t_i).$$

Dans le cas absolument continu

Nous avons

$$S_T(t) = P(T > t) = \int_t^{+\infty} f_T(s)ds.$$

1.2.4 Taux de hasard (Risque instantané)

Dans le cas discret

On appelle taux de hasard du temps de survie T au point t_i, la fonction

$$\lambda_T(t_i) = P(T = t_i / T \geq t_i).$$

Proposition 1.1. *Supposons que T une variable aléatoire discrète, nous avons*

$$\lambda_T(t_i) = 1 - \frac{S_T(t_i)}{S_T(t_{i-1})}$$

et

$$S_T(t) = \prod_{i/t>t_i} \left(1 - \lambda_T(t_i)\right).$$

Preuve 1.1. *Nous avons*

$$\begin{aligned} \lambda_T(t_i) &= P(T = t_i / T \geq t_i) \\ &= \frac{P(T = t_i)}{P(T \geq t_i)} \\ &= \frac{P(T > t_{i-1}) - P(T > t_i)}{P(T > t_{i-1})} \\ &= 1 - \frac{S_T(t_i)}{S_T(t_{i-1})} \end{aligned}$$

D'autre part

$$S_T(t) = \prod_{i/t>t_i} \frac{S_T(t_i)}{S_T(t_{i-1})}$$

et comme

$$\lambda_T(t_i) = 1 - \frac{S_T(t_i)}{S_T(t_{i-1})} \Rightarrow \frac{S_T(t_i)}{S_T(t_{i-1})} = 1 - \lambda_T(t_i),$$

alors, nous obtenons

$$\begin{aligned} S_T(t) &= \prod_{i/t>t_i} \frac{S_T(t_i)}{S_T(t_{i-1})} \\ &= \prod_{i/t>t_i} (1 - \lambda_T(t_i)) . \end{aligned}$$

Dans le cas absolument continu

On appelle taux de hasard du temps de survie T ; la fonction

$$\lambda_T(t) = \begin{cases} \dfrac{f_T(t)}{S_T(t)} & si \;\; S_T(t) \neq 0 \\ 0 & sinon \end{cases}$$

Proposition 1.2. *Si f est continue sur $\mathbb{R}_+^*$ alors pour tout $t > 0$ tel que $S_T(t) \neq 0$, on a*

$$\lambda_T(t) = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta / T > t)}{\Delta}.$$

Autrement dit, pour un t donné, le taux de hasard est la probabilité que l'événement se produise dans un petit intervalle de temps après t sachant qu'il n'a pas eu lieu jusqu'à t.

Preuve 1.2. *Par la continuité de f_T, nous avons*

$$\begin{aligned}
\lim_{\Delta\to 0}\frac{P(t<T\leq t+\Delta/T>t)}{\Delta} &= \lim_{\Delta\to 0}\frac{P(t<T\leq t+\Delta)}{\Delta P(T>t)}\\
&= \lim_{\Delta\to 0}\frac{F_T(t+\Delta)-F_T(t)}{\Delta S_T(t)}\\
&= \frac{1}{S_T(t)}\lim_{\Delta\to 0}\frac{F_T(t+\Delta)-F_T(t)}{\Delta}\\
&= \frac{F_T'(t)}{S_T(t)}\\
&= \lambda_T(t).
\end{aligned}$$

Dans le cas général (Mesure de hasard)

Si T est une v.a.r. de fonction de répartition quelconque F_T, on généralise la notion de taux de hasard en la notion de mesure de hasard $d\Lambda_T$, donnée par

$$d\Lambda_T(t)=\frac{dF_T(t)}{S_T(t-)}.$$

où $S_T(t-)$ est la limite à gauche de S en t et dF_T est la loi de probabilité de T.

Notons que

- Dans le cas discrète, nous avons $dF_T(t)=P(T=t)$ alors $\lambda_T(t)=\frac{P(T=t)}{S_T(t-)}=\frac{P(T=t)}{P(T\geq t)}$.
- Dans le cas absolument continu, nous avons $dF_T(t)=f_T(t)dt$ alors $d\Lambda_T(t)=\frac{f_T(t)dt}{S_T(t)}\Rightarrow\lambda_T(t)=\frac{f_T(t)}{S_T(t)}$.

1.2.5 Fonction de hasard cumulé (Risque cumulé)

Dans le cas discret

On appelle taux de hasard cumulé du temps de survie T, la fonction

$$\Lambda_T(t)=\sum_{i/t_i\leq t}\lambda_T(t_i).$$

Dans le cas absolument continu

On appelle taux de hasard cumulé est l'intégrale du taux de hasard λ_T.

$$\Lambda_T(t) = \int_0^t \lambda_T(s)ds.$$

Proposition 1.3. *Si f_T est continue sur $\mathbb{R}_+^*$ et si $S_T(t) \neq 0\}$, $t > 0$ alors les propositions suivantes sont équivalentes.*

a. $\lambda_T(t) = \dfrac{f_T(t)}{S_T(t)}$

b. $\lambda_T(t) = (-\log S_T(t))'$.

c. $S_T(t) = \exp(-\Lambda_T(t))$.

d. $f_T(t) = \lambda_T(t)\exp(-\Lambda_T(t))$.

Preuve 1.3. $(a) \Leftrightarrow (b)$. *Nous avons*

$$\begin{aligned} \lambda_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{-S_T'(t)}{S_T(t)} = (-\log S_T(t))'. \end{aligned}$$

$(b) \Leftrightarrow (c)$. *Puisque $S_T(0) = 1$ et comme $\lambda_T(t) = (-\log S_T(t))'$, nous obtenons*

$$\begin{aligned} \int_0^t \lambda_T(s)ds &= -\log S_T(t)]_0^t \\ &= -\log S_T(t). \end{aligned}$$

$(c) \Rightarrow (d)$ *Comme $S_T(t) = \exp\{-\Lambda_T(t)\}$, nous trouvons*

$$\begin{aligned} S_T'(t) &= -f_T(t) \\ &= -\lambda_T(t)\exp\{-\Lambda_T(t)\}. \end{aligned}$$

$(d) \Rightarrow (c)$ $f_T(s) = \lambda_T(s) \exp\{-\Lambda_T(s)\}$ *implique*

$$\begin{aligned} F_T(t) &= \int_0^t f_T(s)ds \\ &= \int_0^t \lambda_T(s) \ \exp\{-\Lambda_T(s)\} \\ &= -\exp\{-\Lambda_T(s)]_0^t\} \\ &= 1 - \exp\{-\Lambda_T(t)\}, \end{aligned}$$

alors nous avons $S_T(t) = 1 - F_T(t) = \exp(-\Lambda_T(t))$.

En fiabilité, si T représente le temps de fonctionnement d'un matériel, λ_T est appelé le taux de défaillance (ou panne) et en général c'est une fonction décroissante pendant la première période de fonctionnent (période de rodage), puis elle est approximativement constante (période de vie utile), enfin elle croit (période de vieillissement).

Remarque

1. Toutes ces fonctions caractérisent la loi de T.
2. La connaissance d'une seule de ces fonctions permet de connaître les autres.
3. Un modèle/une méthode statistique pour analyser les données de survie peut donc porter sur n'importe laquelle de ces quantités.

Exemple dans le cas discret

Soit T une v.a. prenant les valeurs 1, 3 et 5 avec probabilité 2/7, 1/7 et 4/7 pour chacune.

Fonction de répartition

$$F_T(t) = \begin{cases} 0 & si \;\; t < 1 \\ \frac{2}{7} & si \;\; 1 \leq t < 3 \\ \frac{3}{7} & si \;\; 3 \leq t < 5 \\ 1 & si \;\; t \geq 5. \end{cases}$$

Fonction de Survie

$$S_T(t) = 1 - F_T(t) = \begin{cases} 1 & si \;\; t < 1 \\ \frac{5}{7} & si \;\; 1 \leq t < 3 \\ \frac{4}{7} & si \;\; 3 \leq t < 5 \\ 0 & si \;\; t \geq 5. \end{cases}$$

Taux de hasard

$$\lambda(t) = \begin{cases} 1 - \frac{S_T(1)}{S_T(1_-)} = \frac{2}{7} & si \;\; t = 1 \\ 1 - \frac{S_T(3)}{S_T(1)} = \frac{1}{5} & si \;\; t = 3 \\ 1 - \frac{S_T(5)}{S_T(3)} = 1 & si \;\; t = 5. \end{cases}$$

1.3 Quantités associées à la distribution de survie

1.3.1 Paramètres de position associés à la distribution de survie

Temps moyen de survie

Le temps moyen de survie est l'espérance de la durée de survie T, il donné par • Dans le cas discret

$$E(T) = \sum_{i=1}^{n} t_j \; P(T = t_j)$$

• Dans le cas absolument coninu

$$\begin{aligned} E(T) &= \int_0^{+\infty} t f_T(t) dt && (1.1) \\ &= \int_0^{+\infty} t \left(-S'_T(t)\right) dt && (1.2) \\ &= \int_0^{+\infty} S_T(t) dt. && (1.3) \end{aligned}$$

Mode

Définition 1.1. *Le mode est la valeur représentée d'une variable quelconque dans une population donnée.*

Donc, le mode est donné par

• Dans le cas discret

$$\theta = arg \sup_t P(t = T)$$

et on a $P(\theta) \geq P(t)$ pour tout $t \neq \theta$.

• Dans le cas absolument continu

$$\theta = arg \sup_t f_T(t)$$

et on a $f_T(\theta) \geq f_T(t)$ pour tout $t \neq \theta$.

Quantiles

Définition 1.2. *Les quantiles sont les valeurs qui divisent une population en intervalles réguliers de même probabilité égale.*

Donc, les quantiles de la durée de survie pour $0 < \alpha \leq 1$; notés Q_α ; sont définis par

$$\begin{aligned} Q_\alpha &= \inf\{t, \quad F_T(t) \geq \alpha\} \\ &= \inf\{t, \quad S_T(t) \leq 1 - \alpha\}. \end{aligned} \tag{1.4}$$

Ils représentent, le temps où la proportion α d'une population a subi l'évènement d'intérêt tel que $Q_\alpha = F^{-1}(\alpha)$.

1.3.2 Paramètres de dispersion associés à la distribution de survie

Variance de survie

La variance de la durée de survie T, est donnée par

$$Var(T) = E(T^2) - [E(T)]^2$$

tel que • Dans le cas discret

$$E(T^2) = \sum_{i=1}^{n} t_j^2 \; P(T = t_j)$$

• Dans le cas absolument coninu

$$\begin{aligned} E(T^2) &= \int_0^{+\infty} t^2 f_T(t) dt \\ &= 2 \int_0^{+\infty} t S_T(t) dt. \end{aligned} \tag{1.5}$$

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

1.4 Exercices

Exercice. 1

Quelques définitions. Donnez les vraies.

1. L'analyse de la survie est l'estimation de la probabilité de survenue d'un décès.

2. Une étude longitudinale correspond à une étude rétrospective.

3. On dit d'un individu qu'il est censuré lorsqu'on n'a plus de nouvelles ou qu'il est mort pendant la période de l'étude.

4. La date de point est la date choisie pour faire le point.

5. Aucune de ces réponses n'est correcte.

Exercice. 2

Soit T une v.a. prenant les valeurs 1, 5, 9 et 11 avec probabilité 1/4 pour chacune.

1. Calculer les cinq fonctions de base de la v.a. T.

2. Calculer l'espérance et le variance de T.

Exercice. 3

Prouver que la loi gamma de densité

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

a une risque instantané croissant pour $\alpha > 1$ et décroissant pour $\alpha < 1$.

Exercice. 4

1. Soit Z une variable aléatoire de fonction de densité

$$f_Z(z) = \frac{\exp(z)}{(1+\exp(z))^2}.$$

Donner les fonctions : de répartition, de survie, du risque instantané

2. La même question pour θZ ?

3. La distribution log-logistique de paramètres α et β est la distribution de

$$Y = \alpha T + \beta Z.$$

Donner la fonction du risque instantané et étudier leur monotonie à partir les valeurs de α et β.

Chapitre 2

L'approche paramétrique dans un modèle de censure

On suppose que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Ainsi, le modèle paramétrique peut être formulé en précisant la forme de l'une ou l'autre des cinq fonctions équivalentes qui définissent la loi de la durée (voir section 1.2). Néanmoins, on spécifie souvent la forme du risque instantané constant (loi exponentielle) ou monotone (loi de Weibul).

Les estimateurs des paramètres du modèle sont ensuite obtenus par plusieurs méthodes par exemple la méthode des moindres carrés qui consiste de trouver le paramètre qui minimise la somme des carrés des erreurs et la méthode du maximum de vraisemblance cherche par contre à trouver la valeur la plus vraisemblable du paramètre d'une population partant d'un échantillon donné.

Dans ce chapitre nous nous intéressons à l'estimation du maximum de vraisemblance. Nous introduisons tout d'abord l'estimation du maximum de vraisemblance dans le cadre complète, En suite dans de cadre de la censure pour les types de censure à droite, double et générale.

2.1 Modèles paramétriques pour l'analyse de survie

2.1.1 Risque instantané constant

*** La loi exponentielle**

la distribution exponentielle est la distribution la plus simple et la plus importante dans les études de survie. Elle joue un rôle dans les études de vie analogue à celui de la distribution normale dans d'autre domaines de la statistique. Les applications dans les études animales et humaines des maladies chroniques et infectieuses se trouvent dans Zelen (1966), Feigl et Zelen (1965), Shulz et al (1986).
Lorsque le temps de survie T suit la distribution exponentielle avec un paramètre γ,

La fonction de densité de probabilité est définie comme suit :

$$f_T(t) = \begin{cases} \gamma e^{-\gamma t} & t \geqslant 0,\ \gamma > 0 \\ 0 & t < 0. \end{cases}$$

La fonction de répartition est

$$F_T(t) = 1 - e^{-\gamma t} \qquad t \geqslant 0.$$

La fonction de survie est

$$S_T(t) = e^{-\gamma t} \qquad t \geqslant 0.$$

Taux de hasard est

$$\lambda_T(t) = \gamma \qquad t \geqslant 0.$$

Notons que ce taux de hasard est constant et nous pouvons montrer que la loi exponentielle est la seule loi, à densité, qui soit sans mémoire.

La moyenne de T est

$$E(T) = \frac{1}{\gamma}.$$

La variance de T

$$Var(T) = \frac{1}{\gamma^2}.$$

*** La loi exponentielle à deux paramètres**

Une forme plus générale de la distribution exponentielle est la distribution exponentielle à deux paramètres (γ, G).

La fonction de densité de probabilité

$$f_T(t) = \begin{cases} \gamma e^{-\gamma(t-G)} & t \geq G \\ 0 & t < G. \end{cases}$$

La fonction de répartition

$$F_T(t) = \begin{cases} 1 - e^{-\gamma(t-G)} & t \geq G \\ 0 & t < G. \end{cases}$$

La fonction de survie

$$S_T(t) = \begin{cases} e^{-\gamma(t-G)} & t \geq G \\ 1 & t < G. \end{cases}$$

Taux de hasard

$$\lambda_T(t) = \begin{cases} 0 & 0 \leq t < G \\ \gamma & t \geq G. \end{cases}$$

Le terme G est un temps de garantie pendant lequel aucun décès ou échec ne peut se produire, ou un temps de survie minimal.
Si $G = 0$, En trouve la loi exponentielle à un paramètre λ.

La moyenne de T

$$E(T) = \frac{G+1}{\gamma}.$$

2.1.2 Risque instantané monotone

*** La loi de Weibull**

La distribution de Weibull est une généralisation de la distribution exponentielle, elle ne suppose pas un taux de hasard constant. Cette distribution a été proposés par Weiball (1939) et son application à diverses situations de défaillance discutée à nouveau par Weibull (1951). Il a été utilisé, entre autres applications, pour décrire les défaillances du tube électronique pour Kao (1958) et pour analyser la cancérologue par Pike(1966), Peto et al(1972) et Williams (1978).

La distribution de Weibull est caractérisée par deux paramètres, β et γ. La valeur de β détermine la forme de la courbe de distribution et la valeur de γ détermine son échelle. Par conséquent, β et γ sont appelés les paramètres de forme et d'échelle, respectivement.

La fonction de densité

$$f_T(t) = \frac{\beta}{\gamma}\left(\frac{t}{\gamma}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\gamma}\right)^{\beta}\right] 1_{]0,+\infty[}(t)$$

La fonction de répartition

$$F_T(t) = 1 - \exp\left[-\left(\frac{t}{\gamma}\right)^{\beta}\right].$$

La fonction de survie

$$S_T(t) = \exp\left[-\left(\frac{t}{\gamma}\right)^{\beta} \right].$$

Le taux de hasard

$$\lambda_T(t) = \frac{\beta}{\gamma}\left(\frac{t}{\gamma}\right)^{\beta-1}.$$

La moyenne de T

$$E(T) = \frac{1}{\gamma}\Gamma(1+\frac{1}{\beta}).$$

La variance de T

$$Var(T) = \frac{1}{\gamma^2} - \frac{1}{\gamma^2}\Gamma^2(1+\frac{1}{\beta}).$$

où $\Gamma(\gamma)$ est définie par

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

Notons que $\Gamma(\gamma) = (\gamma-1)!$ quand γ est un entier positif.

En conséquence,

- Lorsque $\beta > 1$, le taux de hasard est croissante.
- Lorsque $\beta < 1$, le taux de hasard est décroissante.
- Si $\beta = 1$, le taux de hasard est constant (loi exponentielle de paramètre $\frac{1}{\gamma}$).

Cette flexibilité explique l'utilisation de cette distribution dès lors que l'on soupçonne une évolution monotone du risque avec la durée.

*** La loi de Weibull à trois paramètres**

La distribution de Weibull peut également être généralisée pour prendre en compte un temps de garantie G.

La fonction de densité

$$f_T(t) = \frac{\beta}{\gamma}\left(\frac{t}{\gamma}\right)^{\beta-1} \exp\left[-\left(\frac{t-G}{\gamma}\right)^{\beta}\right] 1_{]0,+\infty[}(t)$$

La fonction de répartition

$$F_T(t) = 1 - \exp\left[-\left(\frac{t-G}{\gamma}\right)^{\beta}\right].$$

La fonction de survie

$$S_T(t) = \exp\left[-\left(\frac{t-G}{\gamma}\right)^{\beta}\right].$$

Le taux de hasard

$$\lambda_T(t) = \frac{\beta}{\gamma}\left(\frac{t-G}{\gamma}\right)^{\beta-1}.$$

En général, les lois utilisées dans l'analyse de survie sont la loi exponentielle et la loi de Weibull.

Dans la suite nous allons les combiner à la méthode du maximum de vraisemblance pour obtenir des estimateurs des paramètres.

2.2 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.

2.2.1 Fonction de vraisemblance

Soit $(T_1, ..., T_n)$ un échantillon indépendant et identiquement distribué (i.i.d.) de densité f_T. On appelle vraisemblance l'application L définie par

$$L(t_1, ..., t_n, \theta) = \prod_{i=1}^{n} f_T(t_i, \theta).$$

2.2.2 Estimateur du maximum de vraisemblance

Pour trouver l'estimateur de θ, on maximise la fonction log-vraisemblance suivante

$$\mathcal{L}(t_1, ..., t_n, \theta) = \sum_{i=1}^{n} \ln[f_T(t_i, \theta)]. \tag{2.1}$$

Ensuite, l'estimateur du maximum de vraisemblance $\hat{\theta}$ est la solution du système suivant

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta} = 0 \\ \dfrac{\partial^2 \mathcal{L}}{\partial \theta^2} < 0. \end{cases}$$

2.2.3 Application sur les lois

Loi exponentielle à un paramètre

Soit T une variable aléatoire de loi exponentielle et de paramètre γ et soit $(T_1, ..., T_n)$ un échantillon de T. Nous avons

$$f_T(t) = \gamma \exp(-\gamma t)\, 1_{[0,+\infty[}(t).$$

Alors, la fonction log-vraisemblance (2.1) est

$$\begin{aligned} \mathcal{L}(\underline{t}, \gamma) &= \sum_{i=1}^{n} \ln\{\gamma \exp(-\gamma t_i)\} \\ &= n \ln(\gamma) - \gamma \sum_{i=1}^{n} t_i. \end{aligned}$$

En dérivant par rapport à γ et en annulant, nous obtenons

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \gamma} = 0 &\Rightarrow \frac{n}{\gamma} - \sum_{i=1}^{n} t_i = 0 \\ &\Rightarrow \hat{\gamma} = \frac{n}{\sum_{i=1}^{n} t_i} \end{aligned}$$

Loi de Weibull

Soit T une variable aléatoire de loi de Weibull et de paramètres (β, γ) de fonction de densité

$$f_T(t) = \frac{\beta}{\gamma} \left(\frac{t}{\gamma}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\gamma}\right)^{\beta}\right] 1_{]0,+\infty[}(t).$$

Alors, la fonction log-vraisemblance (2.1) est

$$\begin{aligned}
\mathcal{L}(\underline{x},\gamma) &= \sum_{i=1}^{n} \ln\left(\frac{\beta}{\gamma}\left(\frac{t_i}{\gamma}\right)^{\beta-1}\exp\left[-\left(\frac{t_i}{\gamma}\right)^{\beta}\right]\right) \\
&= n\ln\frac{\beta}{\gamma} + (\beta-1)\sum_{i=1}^{n}\ln\left(\frac{t_i}{\gamma}\right) - \sum_{i=1}^{n}\left(\frac{t_i}{\gamma}\right)^{\beta} \\
&= n\ln(\beta) - n\ln(\gamma) + (\beta-1)\sum_{i=1}^{n}\ln(t_i) - (\beta-1)\sum_{i=1}^{n}\ln(\gamma) \\
&- \sum_{i=1}^{n}(t_i)^{\beta} + n\left(\frac{1}{\gamma}\right)^{\beta}
\end{aligned}$$

En dérivant par rapport à β et γ et en annulant, nous trouvons

$$\frac{\partial\mathcal{L}}{\partial\gamma} = 0 \;\Rightarrow\; \frac{1}{\hat{\gamma}} = \left(\frac{n}{\sum_i^n t_i^{\hat{\beta}}}\right)^{\frac{1}{\hat{\beta}}}$$

et

$$\frac{\partial\mathcal{L}}{\partial\beta} = 0 \;\Rightarrow\; \frac{1}{\hat{\beta}} = \frac{\left(\frac{1}{\hat{\gamma}}\right)^{\hat{\beta}}}{n}\left[\ln\left(\frac{1}{\hat{\gamma}}\right)\sum_i^n t_i^{\hat{\beta}} + \sum_i^n t_i^{\hat{\beta}}\ln(t_i)\right] - \ln\left(\frac{1}{\hat{\gamma}}\right) + \frac{1}{n}\sum_i^n\ln(t_i).$$

Reste à résoudre le système pour avoir les estimateurs recherchés.

Remarque

L'expression analytique de la log-vraisemblance dans la loi Weibull (et pour autre lois) ne rend que rarement possible un calcul direct de l'estimateur du maximum de vraisemblance. Pour résoudre ce problème, nous utilisons les algorithmes numériques standards de type Newton-Raphson et l'algorithme Espérance-Maximisation

(EM). Ce dernier algorithme a été imaginé plus spécifiquement dans le cadre de données incomplètes.
Dans la suite, nous allons utiliser cette méthode d'estimation dans le cas de présence de censure.

2.3 Méthode du maximum de vraisemblance dans le cas de censure à droite

Dans cette section, nous présentons la méthode du maximum de vraisemblance pour estimer les paramètres réels d'un modèle d'analyse des durées de vie T_i censurées par C_i.
Nous considérons le cas de la censure à droite. Alors, les observations sont les couples $(Y, \Delta)_{\{i=1,\ldots,n\}}$ où

$$Y_i = \min(T_i, C_i) \; et \; \Delta_i = 1_{\{T_i \leq C_i\}}.$$

Supposons que T et C sont indépendants et que la fonction de survie S_C de C est continue.

2.3.1 Fonctions de loi de probabilité

Lemme 2.1. *La loi conjointe de* (Y, Δ) *est donnée par*

$$f_{(Y,\Delta)}(y, \delta) = [f_T(y)S_C(y)]^{\delta}[f_C(y)S_T(y)]^{1-\delta}$$

Preuve

Notons que la loi marginale de Y et Δ soit de type continue et

discrète respectivement. Alors, nous avons

$$f_{(Y,\Delta)}(y,\delta) = f_{Y|\Delta}(y \mid \delta) P_{\Delta}(\delta).$$

✓ $1^{ère}$ cas :

Si $\delta = 0$ alors $Y = C$ (c-à-d, il y a censure), alors pour $Y = y$, nous avons

$$f_{(Y,\Delta)}(y,0) = f_{Y|\Delta}(y \mid 0) P_{\Delta}(0). \tag{2.2}$$

Or, pour $Y = y = C$,

$$\begin{aligned} P_{\Delta}(0) &= P(\Delta = 0) \\ &= P(T > C) \\ &= P(T > y) \\ &= S_T(y). \end{aligned}$$

D'autre part, remarquons que

$$Y \mid \Delta = \begin{cases} C & \Delta = 0 \\ T & \Delta = 1. \end{cases} \tag{2.3}$$

alors, nous trouvons

$$f_{Y|\Delta}\ (y \mid 0) = f_C(y).$$

Donc, d'après (2.8), nous obtenons

$$f_{(Y,\Delta)}(y,0) = f_C(y) S_T(y). \tag{2.4}$$

✓ $2^{ème}$ cas :

Si $\delta = 1$ alors $Y = T$ (c-à-d, on observe la durée de vie), alors pour $Y = y$, nous avons

$$f_{(Y,\Delta)}(y,1) = f_{Y|\Delta}(y \mid 1)P_{\Delta}(1). \tag{2.5}$$

Pour $Y = y = T$ et comme S_C est continue, nous obtenons

$$\begin{aligned} P_{\Delta}(1) &= P(\Delta = 1) \\ &= P(T \leq C) \\ &= (C \geq y) \\ &= S_C(y). \end{aligned}$$

D'autre part, d'aprés (2.9),

$$f_{Y|\Delta}\,(y \mid 1) = f_T(y).$$

Alors, d'après (2.12), nous trouvons

$$f_{(Y,\Delta)}\,(y,1) = S_C(y)f_T(y).$$

Donc, nous conclurons

$$f_{(Y,\Delta)}\,(y,\delta) = \begin{cases} S_T(y)f_C(y) & \delta = 0 \\ S_C(y)f_T(y) & \delta = 1 \end{cases}$$

Ainsi, l'expression de la densité conjointe du couple (Y,Δ) est donne comme suit

$$f_{(Y,\Delta)}\,(y,\delta) = (S_T(y)f_C(y))^{1-\delta}\,(S_C(y)f_T(y))^{\delta}.$$

2.3.2 Fonction de vraisemblance

Posons $X_i = (Y = \min(T, C), \Delta)_i$ et $\underline{X} = (X_1, ..., X_n)$.

Proposition 2.1. *Sous l'hypothèse fondamentale d'indépendance $T_i \amalg C_i$ et nous supposons que les T_i (resp. C_i) sont i.i.d. de même loi de T (resp. C). ALors, la fonction de vraisemblance s'écrit*

$$f_{\underline{X}}(\underline{x}, \theta) = \prod_{i=1}^{n} [S_T(y_i, \theta) f_C(y_i)]^{1-\delta_i} \, [S_C(y_i) f_T(y_i, \theta)]^{\delta_i},$$

Preuve

Les T_i (resp. C_i) sont i.i.d. de même loi de T (resp. C) alors les $X_i = (Y_i = \min(T_i, C_i), \Delta_i)$ sont aussi i.i.d. et nous avons

$$\begin{aligned} f_{\underline{X}}(\underline{x}, \theta) &= \prod_{i=1}^{n} f_X(x_i, \theta) \\ &= \prod_{i=1}^{n} f_{(Y,\Delta)}\,(y_i, \delta_i, \theta) \\ &= \prod_{i=1}^{n} (S_T(y_i, \theta) f_C(y_i))^{1-\delta_i} \, (S_C(y_i) f_T(y_i, \theta))^{\delta_i}. \end{aligned}$$

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

$$L(\underline{x}, \theta) = \prod_{i=1}^{n} (f_T(y_i, \theta))^{\delta_i} \, (S_T(y_i, \theta))^{1-\delta_i}.$$

Ecriture de la vraisemblance en fonction du risque

Proposition 2.2. *Sous les hypothèses du proposition 2.3, nous avons*

$$f_{\underline{X}}(\underline{x},\theta) = \prod_{i=1}^{n}[\lambda_T(y_i,\theta)]^{\delta_i}\ S_T(y_i,\theta)\ [\lambda_C(y_i)]^{-\delta_i}\ f_C(y_i).$$

Preuve

D'après la propositions 2.3, nous avons

$$f_{\underline{X}}(\underline{x},\theta) = \prod_{i=1}^{n}[S_T(y_i,\theta)f_C(y_i)]^{1-\delta_i}\ \left[S_C(y_i)f_T(y_i,\theta)\right]^{\delta_i},$$

avec $\underline{X} = (X_1, ..., X_n)$.
Or, puisque $\lambda_T(y_i) = \frac{f_T(y_i)}{S_T(y_i)}$ et $\lambda_C(y_i) = \frac{f_C(y_i)}{S_C(y_i)}$, nous trouvons

$$\begin{aligned} f_{\underline{X}}(\underline{x},\theta) &= \prod_{i=1}^{n}[S_T(y_i,\theta)]^{1-\delta_i}[f_C(y_i)]^{1-\delta_i}\ [S_C(y_i)]^{\delta_i}[f_T(y_i,\theta)]^{\delta_i} \\ &= \prod_{i=1}^{n} S_T(y_i,\theta)\left(\frac{f_T(y_i,\theta)}{S_T(y_i,\theta)}\right)^{\delta_i}\ f_C(y_i)\left(\frac{f_C(y_i)}{S_C(y_i)}\right)^{-\delta_i} \\ &= \prod_{i=1}^{n}[\lambda_T(y_i,\theta)]^{\delta_i}\ S_T(y_i,\theta)\ [\lambda_C(y_i)]^{-\delta_i}\ f_C(y_i). \end{aligned}$$

De même, comme la loi des C ne fait pas intervenir le paramètre θ, la partie utile de la vraisemblance se réduit à

$$L(\underline{x},\theta) = \prod_{i=1}^{n}[\lambda_T(y_i,\theta)]^{\delta_i}\ S_T(y_i,\theta). \tag{2.6}$$

2.3.3 Estimateur du maximum de vraisemblance

Pour trouver l'estimateur de θ, on maximise la fonction log-vraisemblance suivante

$$\mathcal{L}(\underline{x}, \theta) = \sum_{i=1}^{n} \delta_i \ln[f_T(y_i, \theta)] + \sum_{i=1}^{n} (1 - \delta_i) \ln[S_T(y_i, \theta)].$$

Ensuite, l'estimateur du maximum de vraisemblance $\hat{\theta}$ est la solution du système suivant

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta} = 0 \\ \dfrac{\partial^2 \mathcal{L}}{\partial \theta^2} < 0. \end{cases}$$

Notons que la fonction log-vraisemblance en fonction du risque s'écrit

$$\mathcal{L}(\underline{x}, \theta) = \sum_{i=1}^{n} \delta_i \ln[\lambda(y_i, \theta)] + \sum_{i=1}^{n} \ln[S_T(y_i, \theta)]. \tag{2.7}$$

2.3.4 Application sur les lois

Dans cette section nous appliquons la méthode du maximum de vraisemblance pour obtenir des estimateurs des paramètres pour loi Exponentielle et la loi de Weibull dans le cas de censure à droite.

Loi exponentielle à un paramètre

Soit $Y = (\min(T, C))$ une variable aléatoire de loi exponentielle et de paramètre γ et soit $(Y_1, ..., Y_n)$ un échantillon de Y.

Nous avons

$$\begin{cases} f_Y(y) = \gamma \exp(-\gamma y)\, 1_{[0,+\infty[}(y) \\ S_Y(y) = \exp(-\gamma t)\, 1_{[0,+\infty[}(y) \\ \lambda(y) = \gamma. \end{cases}$$

Alors, la fonction log-vraisemblance (2.7) est

$$\begin{aligned} \mathcal{L}(\underline{x}, \gamma) &= \sum_{i=1}^{n} \delta_i \ln[\lambda(y_i, \gamma)] + \sum_{i=1}^{n} \ln[S_T(y_i, \gamma)] \\ &= \sum_{i=1}^{n} \delta_i \ln(\gamma) + \sum_{i=1}^{n} \ln[\exp(-\gamma y_i)] \\ &= \ln(\gamma) \sum_{i=1}^{n} \delta_i - \gamma \sum_{i=1}^{n} y_i. \end{aligned}$$

En dérivant par rapport à γ et en annulant, nous obtenons

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \gamma} = 0 \;&\Rightarrow\; \frac{1}{\gamma} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} y_i = 0 \\ &\Rightarrow\; \hat{\gamma} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i} \end{aligned}$$

Loi de Weibull

Soit $Y = (\min(T, C))$ une variable aléatoire de loi de Weibull de paramètres (β, γ et soit $(Y_1, ..., Y_n)$ un échantillon de Y. Nous

avons

$$\begin{cases} f_Y(Y) = \frac{\beta}{\gamma}\left(\frac{y}{\gamma}\right)^{\beta-1} \exp\left[-\left(\frac{y}{\gamma}\right)^{\beta-1}\right] \\ S_Y(y) = \exp\left[-\left(\frac{y}{\gamma}\right)^{\beta-1}\right] \\ \lambda_Y(t) = \frac{\beta}{\gamma}\left(\frac{y}{\gamma}\right)^{\beta-1}. \end{cases}$$

Alors, la fonction log-vraisemblance (2.7) est

$$\begin{aligned} \mathcal{L}(\underline{x}, \beta, \gamma) &= \sum_{i=1}^{n} \delta_i \ln[\lambda(y_i, \gamma)] + \sum_{i=1}^{n} \ln[S_T(y_i, \gamma)] \\ &= \sum_{i=1}^{n} \delta_i \ln\left[\frac{\beta}{\gamma}\left(\frac{y_i}{\gamma}\right)^{\beta-1}\right] + \sum_{i=1}^{n} \ln\left[\exp\left(-\left(\frac{y_i}{\gamma}\right)^{\beta-1}\right)\right] \end{aligned}$$

En dérivant par rapport à β et γ et en annulant, nous trouvons

$$\frac{\partial \mathcal{L}}{\partial \gamma} = 0 \;\Rightarrow\; \frac{1}{\hat{\gamma}} = \left(\frac{\sum_i^n \delta_i}{\sum_i^n t_i^{\hat{\beta}}}\right)^{\frac{1}{\hat{\beta}}}$$

et

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0 \;\Rightarrow\; \frac{1}{\hat{\beta}} = \frac{\left(\frac{1}{\hat{\gamma}}\right)^{\hat{\beta}}}{\sum_i^n \delta_i}\left[\ln\left(\frac{1}{\hat{\gamma}}\right)\sum_i^n t_i^{\hat{\beta}} + \sum_i^n t_i^{\hat{\beta}} \ln(t_i)\right] - \ln\left(\frac{1}{\hat{\gamma}}\right) + \frac{\sum_i^n \delta_i \ln(t_i)}{\sum_i^n \delta_i}.$$

2.4 Méthode du maximum de vraisemblance dans le cas de censure double

Dans cette section, nous présentons la méthode du maximum de vraisemblance pour estimer les paramètres réels d'un modèle d'analyse des durées de vie T_i doublement censurées par C_i et L_i,

où L est une variable aléatoire non négative et absolument continue.

Nous considérons le cas de la censure double. Alors, les observations sont les couples $(U, A)_{\{i=1,\ldots,n\}}$ où

$$Z_i = \max(\min(T_i, C_i), L)$$

et

$$A = \begin{cases} 0 & si\ L \leq T \leq C \\ 1 & si\ C < T \\ 2 & si\ T < L \end{cases}$$

Supposons que T est indépendant de C et L mais que $P(L \leq C) = 1$ et que les fonction F_T et S_C sont continue.

2.4.1 Fonctions de loi de probabilité

Lemme 2.2. *La loi conjointe de (U, A) est donnée par*

$$f_{(U,A)}(u, i) = [f_T(u)F_L(u)S_C(u)]^{1_{A=0}(i)}[f_C(u)S_T(u)]^{1_{A=1}(i)}[f_L(u)F_T(u)]^{1_{A=2}(i)}$$

Preuve

Notons que la loi marginale de U et A soit de type continue et discrète respectivement. Alors, nous avons

$$f_{(U,A)}(u, i) = f_{U|A}(u \mid i)P_A(i).$$

✓ $1^{ère}$ cas :

Si $i = 0$ alors $U = T$ (c-à-d, on observe la durée de vie), alors

pour $U = u$, nous avons

$$f_{(U,A)}(u, 0) = f_{U|A}(u \mid 0)P_A(0). \tag{2.8}$$

Or, pour $U = u = T$, et comme la fonction S_C est continue, nous avons

$$\begin{aligned} P_A(0) &= P(A = 0) \\ &= P(L \leq T \leq C) \\ &= P(L \leq u \leq C) \\ &= P(L \leq u)P(C \geq u) \\ &= F_L(u)S_C(u) \end{aligned}$$

D'autre part, remarquons que

$$U \mid A = \begin{cases} T & A = 0 \\ C & A = 1 \\ L & A = 2. \end{cases} \tag{2.9}$$

alors, nous trouvons

$$f_{U|A}\ (u \mid 0) = f_T(u).$$

Donc, d'après (2.8), nous obtenons

$$f_{(U,A)}(u, 0) = F_L(u)S_C(u)f_T(u). \tag{2.10}$$

✓ $2^{ème}$ cas :

Si $i = 1$ alors $U = C$ (c-à-d, il y a censure à droite), alors pour

$U = u$, nous avons

$$f_{(U,A)}(u, 1) = f_{U|A}(u \mid 1)P_A(1). \tag{2.11}$$

Pour $U = u = C$, nous obtenons

$$\begin{aligned} P_A(1) &= P(A = 1) \\ &= P(C < T) \\ &= P(T > u) \\ &= S_T(u). \end{aligned}$$

D'autre part, d'aprés (2.9),

$$f_{U|A}\,(u \mid 1) = f_C(u).$$

Alors, d'après (2.12), nous trouvons

$$f_{(Y,\Delta)}\,(y, 1) = S_T(u)f_C(u).$$

✓ $3^{ème}$ cas :

Si $i = 2$ alors $U = L$ (c-à-d, c-à-d, il y a censure à gauche), alors pour $U = u$, nous avons

$$f_{(U,A)}(u, 2) = f_{U|A}(u \mid 2)P_A(2). \tag{2.12}$$

Pour $U = u = L$ et comme F_T est continue, nous obtenons

$$\begin{aligned} P_A(2) &= P(A = 2) \\ &= P(T < L) \\ &= P(T < u) \\ &= F_T(u). \end{aligned}$$

D'autre part, d'aprés (2.9),

$$f_{U|A}\ (u \mid 2) = f_L(u).$$

Alors, d'après (2.12), nous trouvons

$$f_{(U,\Delta)}\ (u, 2) = F_T(u) f_L(u).$$

Donc, nous conclurons

$$f_{(U,A)}\ (u, i) = \begin{cases} F_C(u) S_C(u) f_T(u) & i = 0 \\ S_T(u) f_C(u) & i = 1 \\ F_T(u) f_L(u) & i = 2 \end{cases}$$

Ainsi, l'expression de la densité conjointe du couple (U, A) est donne comme suit

$$f_{(U,A)}(u, i) = [f_T(u) F_L(u) S_C(u)]^{1_{A=0}(i)} [f_C(u) S_T(u)]^{1_{A=1}(i)} [f_L(u) F_T(u)]^{1_{A=2}(i)}.$$

2.4.2 Fonction de vraisemblance

Posons $X_i = (Z = \max(\min(T, C), L), A)_i$ et $\underline{X} = (X_1, ..., X_n)$.

Proposition 2.3. *Sous l'hypothèse que T est indépendantes de C et L et que les T_i (resp. C_i, L_i) sont i.i.d. de même loi de T (resp. C, L). ALors, la fonction de vraisemblance s'écrit*

$$f_{\underline{X}}(\underline{x}, \theta) = \prod_{i=1}^{n} [f_T(u, \theta) F_L(u) S_C(u)]^{1_{A=0}(i)} [f_C(u) S_T(u, \theta)]^{1_{A=1}(i)} [f_L(u) F_T(u, \theta)]^{1_{A=2}(i)}.$$

Preuve

Les T_i (resp. C_i, L_i) sont i.i.d. de même loi de T (resp. C, L) alors les $X_i = (U_i = \max(\min(T_i, C_i), L_i), A_i)$ sont aussi i.i.d. et nous avons

$$\begin{aligned} f_{\underline{X}}(\underline{x}, \theta) &= \prod_{i=1}^{n} f_X(x_i, \theta) \\ &= \prod_{i=1}^{n} f_{(U,A)}\ (u_i, i, \theta) \\ &= \prod_{i=1}^{n} [f_T(u,\theta)F_L(u)S_C(u)]^{1_{A=0}(i)} [f_C(u)S_T(u,\theta)]^{1_{A=1}(i)} [f_L(u)F_T(u,\theta)]^{1_{A=2}(i)}. \end{aligned}$$

Comme les lois des C et L ne fait pas intervenir le paramètre θ, la partie utile de la vraisemblance se réduit à

$$L(\underline{x}, \theta) = \prod_{i=1}^{n} [f_T(u,\theta)]^{1_{A=0}(i)} [S_T(u,\theta)]^{1_{A=1}(i)} [F_T(u,\theta)]^{1_{A=2}(i)}.$$

2.4.3 Estimateur du maximum de vraisemblance

Pour trouver l'estimateur de θ, on maximise la fonction log-vraisemblance suivante

$$\mathcal{L}(\underline{x}, \theta) = \sum_{i=1}^{n} 1_{A=0}(i)[f_T(u,\theta)] + \sum_{i=1}^{n} 1_{A=1}(i)[S_T(u,\theta)] + \sum_{i=1}^{n} 1_{A=2}(i)[F_T(u,\theta)] \quad (2.13)$$

Ensuite, l'estimateur du maximum de vraisemblance $\hat{\theta}$ est la solution du système suivant

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta} = 0 \\ \dfrac{\partial^2 \mathcal{L}}{\partial \theta^2} < 0. \end{cases}$$

2.5 Méthode du maximum de vraisemblance dans le cas de censure générale

Si le temps de survie t_1, t_2, ..., t_n observées pour n individus ($n = \mathcal{D} + \mathcal{L} + \mathcal{R} + \mathcal{I}$), constitués d'observation non censurées (complète) ($i \in [\mathcal{D}]$), censurées à gauche ($i \in \mathcal{L}$), censurées à droite ($i \in \mathcal{R}$) et censurées par intervalle ($i \in \mathcal{I}$). Supposons que les temps de survies suivant une distribution de fonction de densité $f_T(t, \theta)$ et de fonction de survie $S_T(t, \theta)$ étant les paramètres inconnus de la distribution. Alors la fonction de vraisemblance est donnée par

$$L(\underline{x}, \theta) = \prod_{d \in \mathcal{D}} [f_T(x_d, \theta)] \prod_{r \in \mathcal{R}} [S_T(x_r, \theta)] \prod_{l \in \mathcal{L}} [1 - S_T(x_l, \theta)] \prod_{i \in \mathcal{I}} [S_T(U_i, \theta) - S_T(V_i, \theta)].$$

où $\mathcal{I}$ est l'ensemble des observations censurées par intervalle avec la seule connaissance que le temps de survie réel T est dans l'intervalle $[U_i, V_i]$.

Cette formule généralise le cas de censure à droite. Notons que nous pouvons écrire la fonction de vraisemblance (2.6) comme

suit

$$L(\underline{x}, \theta) = \prod_{d \in \mathcal{D}} [f_T(x_d, \theta)] \prod_{r \in \mathcal{R}} [S_T(x_r, \theta)].$$

2.6 Tests paramétriques

La comparaison de deux échantillons ou plus est un objectif majeur de nombre d'études cliniques. 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).

Pour faciliter notre approche, on utilise des échantillons de loi exponentielle, du fait que l'estimateur du maximum de vraisemblance de sa fonction de risque instantané est une constante indépendante du temps. Alors, pour tester l'hypothèse nulle d'égalité des fonctions de survies dans les deux groupes, on dispose trois tests asymptotiquement équivalents :

— Le test du Rapport de Vraisemblance.

— Le test de Wald.

— Le test de Rao ou test du Score.

Supposons qu'il y ait des individus n_1 et n_2 dans les groupes 1 et 2, respectivement.

- Dans le groupe 1, il y a .

• $y_1, ...y_{r_1}$ observations non censurées.

• $y_{r_1+1}, ..., y_{n_1}$ observations censurées.

- Dans le groupe 2, il y a .

• $z_1, ...z_{r_2}$ observations non censurées.

• $z_{r_2+1}, ..., z_{n_2}$ observations censurées.

Tester l'égalité de deux distributions exponentielles équivaut à tester l'hypothèse $H_0 : \gamma_1 = \gamma_2$.

2.6.1 Récapitulation des tests d'hypothèses

On peut récapituler les tests d'hypothèses de la façon suivante :

1. Définir H_0.
2. Choisir un test statistique ou une statistique pour contrôler H_0.
3. Définir la distribution de la statistique sous H_0.
4. Définir le seuil de signification du test ou région critique α.
5. Calculer, à partir des données fournies par l'échantillon, la valeur de la statistique.
6. Prendre une décision concernant H_0 et faire une interprétation.

2.6.2 Test du rapport de vraisemblance

La statistique du rapport de vraisemblance est le rapport de deux fonctions de vraisemblablement (ou de leurs logarithmes) es-

timées à partir des données de l'échantillon.

La procédure pour faire un test du rapport de vraisemblance est :

— Calculer les estimateurs de maximum de vraisemblance des paramètres γ_1 et γ_2.

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n_1} \delta_{1,i}}{\sum_{i=1}^{n_1} y_i} = \frac{r_1}{\sum_{i=1}^{n} y_i} \tag{2.14}$$

et

$$\hat{\gamma}_2 = \frac{\sum_{i=1}^{n_2} \delta_{2,i}}{\sum_{i=1}^{n} z_i} = \frac{r_2}{\sum_{i=1}^{n_2} z_i}. \tag{2.15}$$

— Calculer la fonction de vraisemblance pour les deux groupes combinés, notée $L(\hat{\gamma_1}, \hat{\gamma_2})$.

$$L(\hat{\gamma}_1, \hat{\gamma}_2) = \hat{\gamma}_1^{r_1} \hat{\gamma}_2^{r_2} \exp\left(-\hat{\gamma}_1^{r_1} \sum_{i=1}^{n_1} y_i - \hat{\gamma}_2 \sum_{i=1}^{n_2} z_i \right). \tag{2.16}$$

— Calculer l'estimateur de maximum de vraisemblance pour l'échantillon combiné selon l'hypothèse nulle, c'est-à-dire $\gamma_1 = \gamma_2 = \gamma$, noté $\hat{\gamma}$.

$$\hat{\gamma} = \frac{r_1 + r_2}{\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} z_i}. \tag{2.17}$$

— Calculer la fonction de vraisemblance pour l'échantillon combiné selon l'hypothèse nulle, c'est-à-dire $\gamma_1 = \gamma_2 = \gamma$

$$L(\hat{\gamma}, \hat{\gamma}) = \hat{\gamma}^{(r_1 + r_2)} \exp\left(-\hat{\gamma} \left[\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} z_i \right] \right). \tag{2.18}$$

— Calculer la statistique du rapport de vraisemblance S_{RV}.

$$S_{RV} = -2\log_e \Lambda, \tag{2.19}$$

où

$$\Lambda = \frac{L(\hat{\gamma}, \hat{\gamma})}{L(\hat{\gamma_1}, \hat{\gamma_2})}.$$

Notons que la statistique S_{RV} suit asymptotiquement une loi Chi-deux $\mathcal{X}^2$ un degré de liberté.

— Faire le test.

La région de rejet du test est donnée par

$$R_c = \{S_{RV} > q_{1-\alpha}\},$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{X}^2$.

C-à-d, si S_{RV} est supérieur à $\mathcal{X}^2_{1;1-\alpha}$, on rejette H_0 et concluez que les deux distributions exponentielles de survie ne sont pas égales. Sinon, on accepte H_0.

Remarque

S'il n'y a pas d'observations censurées dans les données, sont également applicables simplement en laissant $n_1 = r_1, n_2 = r_2$ et en omettant les termes des censures .

Exemple :

Soit le temps T (semaine) de rémission chez deux groupes de patients atteints de leucémie :

• Le premier groupe est soumis à un traitement

6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+,

25+, 32+, 32+, 34+, 35+.

• Le second groupe est soumis à un placebo

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

Pourriez-vous conclure sur la signification de cette différence entre les deux groupes ?

- Supposons que les deux distributions sont exponentielles avec les paramètres γ_1 et γ_2, respectivement. En utilisant le test du rapport de probabilité, nous testons l'hypothèse nulle.

$$H_0 : \gamma_1 = \gamma_2 = \gamma$$

contre

$$H_1 : \gamma_1 \neq \gamma_2$$

Nous avons

$$n_1 = n_2 = 21 \qquad r_1 = 9 \qquad r_2 = 21$$

La procédure pour faire un test du rapport de vraisemblance est :

— Nous calculons $\hat{\gamma}_1$ et $\hat{\gamma}_2$, donnés par (2.14) et (2.15).

$$\begin{aligned} \hat{\gamma}_1 &= \frac{r_1}{\sum_{i=1}^{n} y_i} \\ &= 0.025. \end{aligned}$$

et

$$\begin{aligned} \hat{\gamma}_2 &= \frac{r_2}{\sum_{i=1}^{n_2} z_i} \\ &= 0.115. \end{aligned}$$

— Nous calculons $L(\hat{\gamma_1}, \hat{\gamma_2})$, donné par (2.16).

$$\begin{aligned} L(\hat{\gamma}_1, \hat{\gamma}_2) &= \hat{\gamma}_1^{r_1} \hat{\gamma}_2^{r_2} \exp\left(-\hat{\gamma}_1^{r_1} \sum_{i=1}^{n_1} y_i - \hat{\gamma}_2 \sum_{i=1}^{n_2} z_i \right) \\ &= 7.388 \ \ 10^{-48}. \end{aligned}$$

— Nous calculons $\hat{\gamma}$, donnée par (2.17).

$$\begin{aligned} \hat{\gamma} &= \frac{r_1 + r_2}{\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} z_i} \\ &= 0.055. \end{aligned}$$

— Nous calculons $L(\hat{\gamma}, \hat{\gamma})$, donnée par (2.18).

$$\begin{aligned} L(\hat{\gamma}, \hat{\gamma}) &= \hat{\gamma}^{(r_1 + r_2)} \exp\left(-\hat{\gamma} \left[\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} z_i \right] \right) \\ &= 1.944 \ \ 10^{-51}. \end{aligned}$$

— Nous calculons la statistique S_{RV}, donnée par (2.19).

$$\begin{aligned} \Lambda &= \frac{L(\hat{\gamma}, \hat{\gamma})}{L(\hat{\gamma_1}, \hat{\gamma_2})} \\ &= 0.263 \ \ 10^{-3}. \end{aligned}$$

donc,

$$\begin{aligned} S_{RV} &= -2 \log_e \Lambda \\ &= 7.16 \end{aligned}$$

À partir du tableau $\mathcal{X}^2$, nous avons $X^2_{1,0.05} = 3.84$.
Remarquons que S_{RV} est supérieur à $\mathcal{X}^2_{1;\alpha}$, on rejette H_0 au niveau de $\alpha = 0,05$ et concluez que les deux distributions exponentielles de survie ne sont pas égales.

2.6.3 Test de Wald

La statistique de Wald est basée sur la différence entre les estimateurs des paramètres.

La procédure pour faire un test de Wald est :

— Calculer les estimateurs de maximum de vraisemblance de les paramètres λ_1 et λ_2 donnés par (2.14) et (2.14).

— Calculer la statistique de Wald.

$$S_W = \frac{r_1 r_2}{r_1 + r_2} \left[\log \left(\frac{\hat{\gamma_2}}{\hat{\gamma_1}} \right) \right]^2 . \tag{2.20}$$

Notons que la statistique S_W suit asymptotiquement une loi Khi-deux $\mathcal{X}^2$ un degré de liberté.

— Faire le test.

La région de rejet du test est donnée par

$$R_c = \{S_W > q_{1-\alpha}\},$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{X}^2$.

C-à-d, si S_W est supérieur à $\mathcal{X}^2_{1;1-\alpha}$, on rejette H_0 et concluez que les deux distributions exponentielles de survie ne sont pas égales. Sinon, on accepte H_0.

Exemple Prenons en compte les données de rémission des deux groupes de traitement données dans l'exemple précédent.

- Supposons que les deux distributions sont exponentielles avec les paramètres γ_1 et γ_2, respectivement. En utilisant le test du

rapport de probabilité, nous testons l'hypothèse nulle.

$$H_0 : \gamma_1 = \gamma_2 = \gamma$$

contre

$$H_1 : \gamma_1 \neq \gamma_2$$

Nous avons

$$n_1 = n_2 = 21 \qquad r_1 = 9 \qquad r_2 = 21$$

La procédure pour faire un test de Wald est :

— Nous calculons $\hat{\gamma}_1$ et $\hat{\gamma}_2$, donnés par (2.14) et (2.15).

$$\begin{aligned} \hat{\gamma}_1 &= \frac{r_1}{\sum_{i=1}^{n} y_i} \\ &= 0.025. \end{aligned}$$

et

$$\begin{aligned} \hat{\gamma}_2 &= \frac{r_2}{\sum_{i=1}^{n_2} z_i} \\ &= 0.115. \end{aligned}$$

— Nous calculons la statistique S_W donnée par (2.20).

$$\begin{aligned} S_W &= \frac{r_1 r_2}{r_1 + r_2} \left[\log \left(\frac{\hat{\gamma_2}}{\hat{\gamma_1}} \right) \right]^2 \\ &= 2.772. \end{aligned}$$

* À partir du tableau $\mathcal{X}^2$, nous avons $X^2_{1,0.05} = 3.84$.
Remarquons que S_W est supérieur à $\mathcal{X}^2_{1;1-\alpha}$, on rejette H_0 au niveau de $\alpha = 0,05$ et concluez que les deux distributions exponentielles de survie ne sont pas égales.

2.6.4 Test de Rao

La procédure pour faire un test de Rao est :

— Calculer les estimateurs de maximum de vraisemblance de les paramètres γ_1 et γ_2 donnés par (2.14) et (2.14).

— Calculer la statistique de Rao.

$$S_R = \frac{r_1 r_2}{r_1 + r_2} \frac{\left(\log\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1}\right) - 1\right)^2}{\log\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1}\right)}. \tag{2.21}$$

Notons que la statistique S_R suit asymptotiquement une loi Khi-deux $\mathcal{X}^2$ un degré de liberté.

— Faire le test.

La région de rejet du test est donnée par

$$R_c = \{S_R > q_{1-\alpha}\},$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{X}^2$.

C-à-d, si S_R est inférieur à $\mathcal{X}^2_{1;1-\alpha}$, on accepte H_0.

Exemple Prenons en compte les données de rémission des deux groupes de traitement données dans l'exemple précédent.

- Supposons que les deux distributions sont exponentielles avec les paramètres γ_1 et γ_2, respectivement. En utilisant le test du rapport de probabilité, nous testons l'hypothèse nulle.

$$H_0 : \gamma_1 = \gamma_2 = \gamma$$

contre

$$H_1 : \gamma_1 \neq \gamma_2$$

Nous avons

$$n_1 = n_2 = 21 \qquad r_1 = 9 \qquad r_2 = 21$$

La procédure pour faire un test de Rao est :

— Nous calculons $\hat{\gamma}_1$ et $\hat{\gamma}_2$, donnés par (2.14) et (2.15).

$$\begin{aligned} \hat{\gamma}_1 &= \frac{r_1}{\sum_{i=1}^{n} y_i} \\ &= 0.025. \end{aligned}$$

et

$$\begin{aligned} \hat{\gamma}_2 &= \frac{r_2}{\sum_{i=1}^{n_2} z_i} \\ &= 0.115. \end{aligned}$$

— Nous calculons la statistique S_W donnée par (2.20).

$$\begin{aligned} S_R &= \frac{r_1 r_2}{r_1 + r_2} \frac{\left(\log\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1} \right) - 1 \right)^2}{\log\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1} \right)} \\ &= 1.077. \end{aligned}$$

À partir du tableau $\mathcal{X}^2$, nous avons $X^2_{1,0.05} = 3.84$.

Remarquons que S_R est inférieur à $\mathcal{X}^2_{1;1-\alpha}$, on accepte H_0..

Remarques

- Les règles de décision pour les trois test : considérant la statistique S ($S \in \{S_{MV}, S_W, S_R\}$) et les hypothèses du test, les règles

de décision sont :

• Accepter l'hypothèse nulle H_0, si $S \leq \mathcal{X}^2_{1;1-\alpha}$ ou si la probabilité associée à $S \geq \alpha$.

• Rejeter l'hypothèse nulle H_0, si $S \geq \mathcal{X}^2_{1;1-\alpha}$ ou si la probabilité associée à $S \leq \alpha$.

- Généralement, le test du score ou du rapport de vraisemblance donnent des résultats très proche . Mais ce n'est pas le cas du test de Wald.

- En pratique, le test du rapport de vraisemblance est le plus robuste et donc le plus fiable des trois.

2.7 Exercices

Exercice.1

Soit T une variable aléatoire suit une distribution exponentielle à un paramètre α de fonction de densité

$$f(t) = \alpha \exp(-\alpha t)\, 1_{\{t \geq 0\}}.$$

Est- T-est adapté à d'écrire la mortalité humaine ?

2. Soit C la variable de censure suit une distribution exponentielle à un paramètre μ et soit $Y = \min(T, C)$ avec C et T sont indépendants. Déterminer la distribution de Y.

3. Vers quelle quantité $\hat{m}_1$ converge, où

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

avec Y_i est une suite des variable aléatoires i.i.d. de même loi de Y ?

4. Soit $\delta_i = 1_{\{T_i \leq C_i\}}$. Calculer $M(\Phi) = E\left[\delta \Phi(Y)\right]$.

5. Vers quelle quantité $\hat{m}_2$ converge, où

$$\hat{m}_2 = \frac{1}{\sum_{i=1}^{n} \delta_i} \sum_{i=1}^{n} \delta_i Y_i.$$

Exercice 2.

Les temps de survie suivent une distribution exponentielle à un paramètre θ, si leur fonction de densité a pour écriture

$$f(t) = \theta \exp(-\theta t)\, 1_{]0,+\infty[}(t), \quad \theta > 0$$

1. Donner l'expression de la fonction de survie associée et la fonction de risque correspondante.

2. Soient $t_1, t_2, ..., t_n$ les temps observés dans un échantillon de n individus et $\delta_1, \delta_2, ..., \delta_n$ les indicatrices de censure associées avec $\delta_i = 1$ si t_i est un temps d'événement observé et $\delta_i = 0$ si t_i est un temps censuré à droite. Quelle est l' écriture de la fonction de log-vraisemblance $l(\theta)$?

3. Donnez l'expression de $\hat{\theta}$, estimateur du maximum de vraisemblance de θ.

4. Soit le temps T (semaine) de rèmission chez deux groupes de patients atteints de leucèmie :

* Le premier groupe est soumis à un traitement

(9, 12, 20, 20, 30+, 32+) (où x+ indique que x est censurèe à droite).

* Le second groupe est soumis un placebo

(6, 8, 10, 10, 12, 14).

- Calculer $\hat{S}_1(12)$ et $\hat{S}_2(12)$, où $\hat{S}_1(t)$ est $\hat{S}_2(t)$ sont les estimateurs du maximum de vraisemblance de la fonction de survie correspondant à les groupes 1 et 2.

- Peut-on conclure à l'efficacitè du mèdicament ? (tester l'hypothèse de rèaction identique) avec $\alpha = 0.05$ et $Z_\alpha = 1.65$. (test du rapport de vraisemblance).

Exercice 3.

Les temps de survie suivent une distribution exponentielle à un paramètre θ, si leur fonction de densité a pour écriture

$$f(t) = \theta \exp(-\theta t)\, 1_{]0,+\infty[}(t), \quad \theta > 0$$

1. Donner l'expression de la fonction de survie associée et la fonction de risque correspondante.

2. Soient $t_1, t_2, ..., t_n$ les temps observés dans un échantillon de n individus et $\delta_1, \delta_2, ..., \delta_n$ les indicatrices de censure associées avec $\delta_i = 1$ si t_i est un temps d'événement observé et $\delta_i = 0$ si t_i est un temps censuré à gauche. Quelle est l' écriture de la fonction de log-vraisemblance $l(\theta)$?

3. Donnez l'expression de $\hat{\theta}$, estimateur du maximum de vraisemblance de θ.

4. Soit le temps T (semaine) de rèmission chez deux groupes de patients atteints de leucèmie :

* Le premier groupe est soumis à un traitement

(6, 7, 8, 13, 14+, 16+, 19) (où x+ indique que x est censurèe à gauche).

* Le second groupe est soumis un placebo

(7, 9, 11, 12, 12, 14, 18).

- Calculer $\hat{S}_1(14)$ et $\hat{S}_2(14)$, où $\hat{S}_1(t)$ est $\hat{S}_2(t)$ sont les estimateurs du maximum de vraisemblance de la fonction de survie correspondant à les groupes 1 et 2.

- Peut-on conclure à l'efficacitè du mèdicament ? (tester l'hypothèse de rèaction identique) avec $\alpha = 0.05$ et $Z_\alpha = 1.65$. (test de Rao).

Exercice 4. Soit le temps T (semaine) de rémission chez deux groupes de patients atteints de leucémie :

- Le premier groupe est soumis à un traitement

(9, 12, 15, 20, 30+, 30+) (où $x+$ indique que x est censurée à droite).

- Le second groupe est soumis à un placebo

(8, 8, 10, 12, 12, 13).

1. Tracer les courbes des estimateurs de Kaplan-Meier des fonctions de survie pour chacun des deux groupes.

2. Par lecture graphique, donnez le temps médian de rémission des patients atteints de leucémie dans chacun des deux groupes.

3. Peut-on conclure à l'efficacité du médicament ? (tester l'hypothèse de réaction identique).

Exercice5. Les temps de survie suivent une distribution exponentielle à deux paramètres λ et μ, si leur fonction de densité a pour écriture

$$f(t) = \lambda \exp\left(-\lambda(t-\mu)\right), \quad si \;\; t \geq \mu$$

$$f(t) = 0, \quad si \;\; t < \mu$$

1. Donner l'expression de la fonction de survie associée et la fonction de risque correspondante, lorsque $t \geq \mu$.

2. Soient $t_1, t_2, ..., t_n$ les temps observés dans un échantillon de n individus et $\delta_1, \delta_2, ..., \delta_n$ les indicatrices de censure associées avec $\delta_i = 1$ si t_i est un temps d'événement observé et $\delta_i = 0$ si t_i est un temps censuré à droite. Quelle est l' écriture de la fonction de log-vraisemblance $l(\lambda, \mu)$?

3. En fonction du précédent résultat et de la définition initiale, pouvez-vous préciser les domaines de définition de λ et μ.

4. Expliquez en 2 ou 3 lignes pourquoi l'estimateur du maximum de vraisemblance de μ est $\hat{\mu} = \min(t_1, t_2, ..., t_n)$?

5. Donnez l'expression de $\hat{\lambda}$, estimateur du maximum de vraisemblance de λ.

6. Les données suivantes correspondent à des temps d'événement. Les données pour lesquelles il y a une censure à droite sont indiquées positivement : 2, 5, +7, 4, 6, +9, 3, +5, 8, 10, +13, 6, 11, +4, +12, 4.

- Calculer $\hat{S}(3)$, où $\hat{S}(t)$ est l'estimateur de du maximum de vrai-

semblance de la fonction de survie correspondant à ce jeu de données.

Exercice 6.

Soit le temps T (semaine) de rémission chez deux groupes de patients atteints de leucémie :

- Le premier groupe est soumis à un traitement

(24, 30, 42, 15+, 40+, 42+) (où $x+$ indique que x est censurée à droite).

- Le second groupe est soumis à un placebo

(10, 26, 28, 30, 41, 12+).

Supposer que les distributions pour les deux groupes suivent une distribution exponentielle à un paramètres λ_1 et λ_2.

Peut-on conclure à l'efficacité du médicament ? (tester l'hypothèse de réaction identique).

(Utiliser le test de Wald avec $\alpha = 0.05$ et $Z_\alpha = 3.841$).

Chapitre 3

L'approche non paramétrique dans un modèle de censure à droite

3.1 Estimation de la fonction de survie et de risque cumulé

3.1.1 Estimateur de Kaplan-Meier pour la fonction de survie

Soient $T_1,\ T_2,\ ...,\ T_n$ une suite de n variables aléatoires positives i.i.d. de même loi de T et de fonction de survie $S_T(t) = P(T > t)$ représentant les durées de survie. En l'absence de censures, il est naturel et usuel d'estimer $S_T(t)$ par la fonction de

survie empirique donnée par

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i>t\}}.$$

Notons que $\sup_{t\in\mathbb{R}} |\hat{S}(t) - S_T(t)| \to 0$ presque sûrement par le théorème de Glivenko-Cantelli.

On suppose ici que certaines données T_i sont soumises á une censure á droite par des variables aléatoires positives C_i.

Pour des résons d'identifiabilité de modèle, on suppose que la suite (T_i) est indépendante de la suite (C_i) et on observe la suite $\{(Y_i, \delta_i)\}$ oú $Y_i = \min(T_i,\ C_i)$ et $\delta_i = 1_{\{T_i\leq C_i\}}$.

Malheureusement, dans ce cas, il est impossible d'utiliser la fonction $\hat{S}(t)$ puisqu'elle fait intervenir des quantités non observées (tous les T_i censurés ne sont pas observés). On estime alors généralement S_T en utilisant l'estimateur de Kaplan-Meier (EKM).

L'ideé de la construction de l'EKM est la suivante, pour $t_1 < t$

$$\begin{aligned} S_T(t) &= P(T > t) \\ &= P(T > t/T > t_1)P(T > t_1) \\ &= P(T > t/T > t_1)S_T(t_1). \end{aligned}$$

Si l'on renouvelle l'opération en choisissant une date $t_2 < t_1$, on obtient

$$\begin{aligned} S_T(t_1) &= P(T > t_1) \\ &= P(T > t_1/T > t_2)P(T > t_2) \\ &= P(T > t_1/T > t_2)S_T(t_2). \end{aligned}$$

D'où

$$S_T(t) = P(T > t/T > t_1)P(T > t_1/T > t_2)S_T(t_2). \qquad (3.1)$$

En réitérant, nous distinguons des produits d'éléments en $P(T > t/T > t_1)$. En effet, pour $t_0 < t_1 < t_2 < .... < t_n < t$, on a

$$\begin{aligned} S_T(t) &= P(T > t/T > t_n)P(T > t_n/T > t_{n-1})...P(T > t_1/T > t_0)S_T(t_0) \\ &= \prod_{i=1}^{n} P(T > t_i/T > t_{i-1}). \end{aligned}$$

où $S_T(t_0) = 1$. Soit donc $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ la statistique d'ordre associée à $Y_1, Y_2, ..., Y_n$ et notons $\delta_{(i)}$ l'indicateur de censure associé à $Y_{(i)}$. Si on choisit pour dates où l'on conditionne celles où il s'est produit un évènement (décès ou censure) c-à-d $Y_{(i)}$ on estime seulement des quantité de la forme

$$p_i = P(T > Y_i/T > Y_{i-1}),$$

par example p_i est la probabilité de survivre pendant l'intervalle $I_i =]Y_{(i-1)},\ Y_{(i)}]$ quand on est vivant au début de cet intervalle.

On peut alors estimer $q_i = 1 - p_i$ par

$$\hat{q}_i = \frac{d_i}{n_i}, \tag{3.2}$$

où,

* d_i le nombre d'événements d'intérêt observés au temps $Y_{(i)}$, en ne comptant pas les censures, c-à-d, le nombre d'individus ayant connu l'évènement en $Y_{(i)}$, avec

$$d_i = \sum_{l=1}^{n} \delta_l 1_{\{Y_l = Y_{(i)}\}} \tag{3.3}$$

* n_i le nombre d'individus à risque au temps $Y(i)$, c-à-d tous ceux pour lesquels $Y_l > Y_{(i)}$, individus censurés inclus (le nombre d'observation restantes non censurées juste avant $Y_{(i)}$, avec

$$n_i = \sum_{l=1}^{n} 1_{\{Y_l \geq Y_{(i)}\}}. \tag{3.4}$$

Par example q_i est la probabilité de mourir pendant l'intervalle I_i sachant qu'on était vivant au début de l'intervalle, d_i le nombre de morts exactes au l'instant $Y_{(i)}$ et n_i le nombre des sujets qui sont vivants(donc "à risque" de mourir) juste avant l'instant $Y_{(i)}$. Alors, on a

$$\hat{q}_i = \frac{Nombre\ de\ mort\ \text{à}\ l'instant\ Y_{(i)}}{Nombre\ de\ sujets\ \text{à}\ risque}.$$

Donc, l'estimateur de Kaplan-Meier est donné par

$$\widehat{S}_n(t) = \prod_{Y_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{3.5}$$

Cet estimateur est aussi appelé estimateur Produit-Limite (Product Limit(PL)) car il s'obtient comme limite d'un produit.
Notons que si $Y_{(i)}$ est un instant où ne se produisent que des censures alors $d_i = 0$, donc $\widehat{S}_n(t)$ ne varie qu'aux points où il y a effectivement une mort (donnée complète).
Supposons qu'il n'y ait pas d'ex-æquo (c-à-d tous les $Y_{(i)}$ sont différents). Si $\delta_{(i)} = 1$ il n'y a pas de censure à l'instant $Y_{(i)}$, implique $d_i = 1$. Et si $\delta_{(i)} = 0$ il y a censure à l'instant $Y_{(i)}$, implique $d_i = 0$.
On a alors

$$\begin{aligned}
\hat{p}_i &= 1 - \hat{q}_i \\
&= \begin{cases} 1 - \frac{1}{n_i} & si \;\; \delta_{(i)} = 1 \\ 1 & si \;\; \delta_{(i)} = 0 \end{cases} \\
&= \left(1 - \frac{1}{n - i + 1}\right)^{\delta_{(i)}}.
\end{aligned}$$

Il est claire que $n_i = n - i + 1$, car il y a eu $i - 1$ sujets ou censures avant $Y_{(i)}$ et il y n individus dans l'étude.
Donc l'EKM pour la fonction de survie de la variable durée de vie T

$$\widehat{S}_n(t) = \begin{cases} \prod \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} & si \;\; t < Y_{(i)} \\ 0 & si \; t \geq Y_{(i)}. \end{cases}$$

3.1.2 Estimateur du risque cumulé de Breslow

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

$$\Lambda(t) = -\ln\left(S_T(t)\right).$$

Alors, l'estimateur de Breslow s'écrit

$$\begin{aligned}\widehat{\Lambda}(t) &= -\ln\left(\widehat{S}_{KM}(t)\right) \\ &= -\sum_{i/Y_{(i)}\leq t} \log\left(1-\frac{d_i}{n_i}\right).\end{aligned}$$

3.1.3 Estimateur du risque cumulé de de Nelson-Aalen

Introduisons les statistiques suivantes

$$N_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i\leq t,\ \delta_i=1\}}$$

et

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i\geq t\}}.$$

Notons que $N_n(t)$ et $Y_n(t)$ sont respectivement les fonctions empiriques associées à $F_{1,Y}(t) = P(Y \leq t,\ \delta = 1)$ et $S_Y(t_-) = P(Y \geq t)$.

Comme T et C sont indépendants, alors on a

$$\begin{aligned}
E(N_n(t)) &= E\left(1_{\{Y\leq t,\,\delta=1\}}\right) \\
&= P\left(Y \leq t,\ \delta = 1\right) \\
&= P\left(T \leq t,\ T \leq C\right) \\
&= \int_0^t \int_y^{+\infty} dP_T(y) \otimes dP_C(c) \\
&= \int_0^t \left(\int_y^{+\infty} dP_C(c)\right) dP_T(y) \\
&= \int_0^t P(C \geq y) dF_T(y) \\
&= \int_0^t S_C(y_-) dF_T(y)
\end{aligned}$$

et

$$\begin{aligned}
E(Z_n(t)) &= E\left(1_{\{Y\geq t\}}\right) \\
&= P\left(Y \geq t\right) \\
&= P\left(T \geq t,\ C \geq t\right) \\
&= P\left(T \geq t\right) P\left(C \geq t\right) \\
&= S_T(t_-) S_C(t_-) \\
&= S_Y(t_-).
\end{aligned}$$

D'autre par, la fonction de hasard cumulé de T est donnée pour tout $t < T_{F_Y}$ par

$$\Lambda(t) = \int_0^t \frac{dF_T(y)}{S_T(y_-)},$$

alors

$$\begin{aligned}\Lambda(t) &= \int_0^t \frac{dF_Y(t)}{S_Y(t)} \\ &= \int_0^t \frac{S_C(y_-)dF_T(y)}{S_C(y_-)S_T(y_-)} \\ &= \int_0^t \frac{dE\left(N_n(y)\right)}{E\left(Z_n(y)\right)}.\end{aligned}$$

Cela suggérer d'estimer par

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{dN_n(y)}{Z_n(y)},$$

qui est l'estimateur de Nelson-Aalen, il s'écrit

$$\begin{aligned}\widehat{\Lambda}_n(t) &= \sum_{i/Y_{(i)}\leq t} \frac{\Delta N_n(Y_{(i)})}{Z_n(Y_{(i)})} \\ &= \sum_{i/Y_{(i)}\leq t} \frac{d_i}{n_i}.\end{aligned}$$

De plus, comme $\Delta\widehat{\Lambda}_n(Y_{(i)}) = \frac{d_i}{n_i}$,il est facile de voir qu'il est lié à l'EKM par la relation

$$\widehat{S}_n(t) = \prod_{i/Y_{(i)}\leq t} \left(1 - \Delta\widehat{\Lambda}_n(Y_{(i)})\right), \tag{3.6}$$

où $\Delta\widehat{\Lambda}_n(s) = \widehat{\Lambda}_n(s) - \widehat{\Lambda}_n(s_-)$ est le saut de $\widehat{\Lambda}_n$ au point s.

3.1.4 Estimateur de Harrington et Fleming pour la fonction de survie

Á partir la relation entre le taux de hasard cumulé Λ_T et la fonction de survie S_T.

$$S_T(t) = \exp\left(-\Lambda_T(t)\right).$$

On peut déduire un autre estimateur de la fonction de survie

$$\begin{aligned}
\widehat{S}_n(t) &= \exp\left(-\widehat{\Lambda}_n(t)\right) \\
&= \prod_{i/Y_{(i)}\leq t} \exp\left(-\frac{d_i}{n_i}\right) \\
&\simeq \prod_{i/Y_{(i)}\leq t} \left(1-\frac{d_i}{n_i}\right), \quad \frac{d_i}{n_i} \to 0.
\end{aligned}$$

3.2 Propriétés de l'estimateur de Kaplan-Meier

Nous observons un échantillon de variables aléatoires i.i.d. $(Y_i = \min(T_i, C_i),\ \delta_i = 1_{\{T_i \leq C_i\}})_{1\leq i\leq n}$ de même loi que $(Y = \min(T, C),\ \delta)$ où C est indépendant de T. Dans toute la suite de ce document, pour toute variable aléatoire réelle U, de fonction de répartition F_U , notons $T_{F_U} = \sup\{t, F_U(t) < 1\}$ le point terminal du support de U.

3.2.1 Convergence presque complète

Földes A. et Rejtö L. ont trouvé le taux de convergence qui est de l'ordre de $\sqrt{\log n/n}$ (voir Foldes et al. (1981) preuve du théorème 3.2.) sous la continuité de F_T et F_C. En, 2013 Boukeloua a retrouvé ce même taux sans exiger la continuité de F_T ni celle de F_C ; dans le théorème suivant

Théorème 3.1. *Pour tout* $\theta \in]0, T_{F_Y}[$, *on a :*

$$\sup_{t\leq\theta}|S_n(t) - S_T(t)| \xrightarrow{p.co.} 0$$

et

$$\sup_{t\leq\theta}|S_n(t) - S_T(t)| = O_{p.co.}\left(\sqrt{\frac{\log n}{n}}\right)$$

Pour montrer ce théorème, nous avons besoin du lemme suivant dont la preuve est donnée dans Rosenblatt (1956) (lemme 1 page 302).

Lemme 3.1. *Si* A *et* B *sont deux fonctions croissantes et continues à droite sur* $[0, +\infty[$ *avec* $A(t) = B(t)$ *pour* $t < 0$ *et* $\Delta A \leq 1$ *et* $\Delta B \leq 1$ *sur* $[0, +\infty[$ *et si* $\theta_B = \inf\{t \in \mathbb{R}/B(t) = +\infty\}$, *alors la seule solution locale bornée* Z *de l'équation*

$$Z(t) = Z(0) - \int_{]0,t]} \frac{Z(x^-)}{1 - \Delta B(x)} d(A(x) - B(x))$$

sur $[0, \theta_B[$ *est donnée par*

$$Z(t) = Z(0)\exp(B^c(t) - A^c(t))\frac{\prod\limits_{0\leq x\leq t}(1-\Delta A(x))}{\prod\limits_{0\leq x\leq t}(1-\Delta B(x))}.$$

Donc, la démonstration de théorème 3.1 est comme suit.

Preuve 3.1. *Soit* $\theta \in]0, T_{F_Y}[$,

• *Pour tout* $t < 0$, *nous avons*

$$S_n(t) = S_T(t) = 1 \Rightarrow S_n(t) - S_T(t) = 0,$$

donc

$$\sup_{t\leq\theta}|S_n(t) - S_T(t)| = \sup_{0\leq t\leq\theta}|S_n(t) - S_T(t)|.$$

• *Pour tout* $t \in [0, \theta]$, *nous avons*

$$S_T(t) = 1 - P(T \leq t) = 1 - \int_{]0,t]} dF(x) = 1 - \int_{]0,t]} S_T(x_-)\, d\Lambda(x),$$

donc, d'après le lemme 3.1, nous obtenons

$$S_T(t) = \prod_{x\leq t}(1 - \Delta\Lambda(x))\exp(-\Lambda^c(t)). \tag{3.7}$$

D'autre part, nous avons

$$S_n(t) = \prod_{i/Y_{(i)}\leq t}(1 - \Delta\Lambda_n(Y_{(i)}). \tag{3.8}$$

Les deux relations ci-dessus (3.7) *et* (3.8) *montrent, d'après le*

lemme 3.1, que $\frac{S_n(t)}{S_T(t)}$ *vérifie*

$$\frac{S_n(t)}{S_T(t)} = 1 - \int_{]0,t]} \frac{S_n(x_-)d(\Lambda_n(x) - \Lambda(x))}{S_T(x_-)(1 - \Delta\Lambda(x))}$$

$$\Rightarrow S_n(t) - S_T(t) = -S_T(t) \int_{]0,t]} \frac{S_n(x_-)d(\Lambda_n(x) - \Lambda(x))}{S_T(x_-)(1 - \Delta\Lambda(x))}$$

$$\Rightarrow |S_n(t) - S_T(t)| \leq \left| \int_{]0,t]} \frac{S_n(x_-)}{S_T(x_-)} dK_n(x) \right| \text{ où } K_n(t) = \int_{]0,t]} \frac{d(\Lambda_n(x) - \Lambda(x))}{1 - \Delta\Lambda(x)}.$$

Par l'application la formule d'intégration par parties, nous déduisons

$$
\begin{aligned}
|S_n(t) - S_T(t)| &\leq \frac{S_n(t)}{S_T(t)}|K_n(t)| + \left| \int_{]0,t]} K_n(x) d\left(\frac{S_n(x)}{S_T(x)}\right) \right| \\
&\leq \frac{1}{S_T(\theta)} \sup_{0\leq u\leq\theta} |K_n(u)| + \left| \int_{]0,t]} K_n(x) S_n(x_-) d\left(\frac{1}{S_T(x)}\right) \right| \\
&+ \left| \int_{]0,t]} \frac{K_n(x)}{S_T(x)} dS_n(x) \right| \\
&\leq \frac{1}{S_T(\theta)} \sup_{0\leq u\leq\theta} |K_n(u)| + \sup_{0\leq u\leq\theta} |K_n(u)| \left(\frac{1}{S_T(t)} - 1\right) \\
&+ \frac{1}{S_T(\theta)} \sup_{0\leq u\leq\theta} |K_n(u)||S_n(t) - 1| \\
&\leq \left(\frac{3}{S_T(\theta)} - 1\right) \sup_{0\leq u\leq\theta} |K_n(u)|
\end{aligned}
$$

or $S_T(x) = S_T(x_-)(1 - \Delta\Lambda(x)) \Rightarrow \frac{1}{1-\Delta\Lambda(x)} = \frac{S_T(x_-)}{S_T(x)} = 1 - \frac{\Delta S_T(x)}{S_T(x)}$

$$\Rightarrow K_n(u) = \int_{]0,u]} (1 - \frac{\Delta S_T(x)}{S_T(x)}) d(\Lambda_n(x) - \Lambda(x))$$

$$\Rightarrow |K_n(u)| \leq \left| \int_{]0,u]} d(\Lambda_n(x) - \Lambda(x)) \right| + \left| \int_{]0,u]} \frac{\Delta S_T(x)}{S_T(x)} d(\Lambda_n(x) - \Lambda(x)) \right|$$

$$\leq \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \sum_{\substack{x \in]0,u] \\ \Delta S_T(x) > 0}} \left| \frac{\Delta S_T(x)}{S_T(x)} \right| |\Delta\Lambda_n(x) - \Delta\Lambda(x)|$$

$$\leq \frac{1}{S_T(\theta)} \sup_{0 \leq u \leq \theta} |\Delta\Lambda_n(u) - \Delta\Lambda(u)| \sum_{\substack{x \in]0,u] \\ \Delta S_T(x) > 0}} |\Delta S_T(x)|$$

$$+ \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)|$$

$$\leq \left(1 + \frac{1}{S_T(\theta)}\right) \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{1}{S_T(\theta)} \sup_{0 \leq u \leq \theta} |\Lambda_n(u_-) - \Lambda(u_-)|$$

et ceci pour tout $u \in [0, \theta]$, *alors*

$\sup_{0 \leq u \leq \theta} |K_n(u)| \leq \left(1 + \frac{1}{S_T(\theta)}\right) \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{1}{S_T(\theta)} \sup_{0 \leq u \leq \theta} |\Lambda_n(u_-) - \Lambda(u_-)|$

$\Rightarrow |S_n(t) - S_T(t)| \leq \frac{(3 - S_T(\theta))(1 + S_T(\theta))}{S_T(\theta)^2} \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{3 - S_T(\theta)}{S_T(\theta)^2} \sup_{0 \leq u \leq \theta} |\Lambda_n(u_-) - \Lambda(u_-)| \; \forall\, t \in [0, \theta]$, *donc :*

$$\sup_{0\le t\le\theta}|S_n(t)-S_T(t)| \;\le\; \frac{(3-S_T(\theta))(1+S_T(\theta))}{S_T(\theta)^2}\sup_{0\le t\le\theta}|\Lambda_n(t)-\Lambda(t)| + \frac{3-S_T(\theta)}{S_T(\theta)^2}\sup_{0\le t\le\theta}|\Lambda_n(t_-)-\Lambda(t_-)|. \qquad (3.9)$$

De plus, nous avons pour tout $t\in[0,\theta]$

$$\begin{aligned}
|\Lambda_n(t)-\Lambda(t)| &= \left|\int_{]0,t]}\frac{dN_n(x)}{Z_n(x)}-\int_{]0,t]}\frac{dF_{1,Y}(x)}{S_Y(x_-)}\right| \\
&= \left|\int_{]0,t]}\frac{dN_n(x)}{Z_n(x)}-\int_{]0,t]}\frac{dF_{1,Y}(x)}{S_Y(x_-)}+\int_{]0,t]}\frac{dN_n(x)}{S_Y(x_-)}-\int_{]0,t]}\frac{dN_n(x)}{S_Y(x_-)}\right| \\
&\le \left|\int_{]0,t]}\left(\frac{1}{Z_n(x)}-\frac{1}{S_Y(x_-)}\right)dN_n(x)\right| \\
&+ \left|\int_{]0,t]}\frac{1}{S_Y(x_-)}d(N_n(x)-F_{1,Y}(x))\right|
\end{aligned}$$

$$
\begin{aligned}
&= \left| \int_{]0,t]} \frac{S_Y(x_-) - Z_n(x)}{Z_n(x) \times S_Y(x_-)} dN_n(x) \right| + \left| \int_{]0,t]} \frac{1}{S_Y(x_-)} d(N_n(x) - F_{1,Y}(x)) \right| \\
&\leq \sup_{0 \leq u \leq \theta} \left| \frac{S_Y(u_-) - Z_n(u)}{Z_n(u) \times S_Y(u_-)} \right| N_n(t) + \left| \int_{]0,t]} (N_n(x) - F_{1,Y}(x)) d\left(\frac{1}{S_Y(x)} \right) \right| \\
&+ \left| \frac{N_n(t) - F_{1,Y}(t)}{S_Y(t)} \right| \\
&\leq \frac{1}{Z_n(\theta) \times S_Y(\theta_-)} \sup_{0 \leq u \leq \theta} |Z_n(u) - S_Y(u_-)| + \frac{1}{S_Y(\theta)} \sup_{0 \leq u \leq \theta} |N_n(u) - F_{1,Y}(u)| \\
&\quad + \sup_{0 \leq u \leq \theta} |N_n(u) - F_{1,Y}(u)| \left(\frac{1}{S_Y(t)} - 1 \right) \\
&\leq \frac{1}{Z_n(\theta) \times S_Y(\theta_-)} \sup_{0 \leq u \leq \theta} |Z_n(u) - S_Y(u_-)| + \left(\frac{2}{S_Y(\theta)} - 1 \right) \sup_{0 \leq u \leq \theta} |N_n(u) - F_{1,Y}(u)|
\end{aligned}
$$

donc

$$
\begin{aligned}
\sup_{0 \leq t \leq \theta} |\Lambda_n(t) - \Lambda(t)| &\leq \frac{1}{Z_n(\theta) \times S_Y(\theta_-)} \sup_{0 \leq t \leq \theta} |Z_n(t) - S_Y(t_-)| \\
&+ \left(\frac{2}{S_Y(\theta)} - 1 \right) \sup_{0 \leq t \leq \theta} |N_n(t) - F_{1,Y}(t)|
\end{aligned}
$$

et de la même façon, nous pouvons prouver, que

$$
\begin{aligned}
\sup_{0 \leq t \leq \theta} |\Lambda_n(t_-) - \Lambda(t_-)| &\leq \frac{1}{Z_n(\theta) \times S_Y(\theta_-)} \sup_{0 \leq t \leq \theta} |Z_n(t) - S_Y(t_-)| \\
&+ \frac{1}{S_Y(\theta_-)} \sup_{0 \leq t \leq \theta} |N_n(t_-) - F_{1,Y}(t_-)| \\
&+ \left(\frac{1}{S_Y(\theta_-)} - 1 \right) \sup_{0 \leq t \leq \theta} |N_n(t) - F_{1,Y}(t)|,
\end{aligned}
$$

d'après (3.9), *il existe* $\alpha(\theta), \beta(\theta), \gamma(\theta) > 0$ *tels que*

$$\begin{aligned}\sup_{0\leq t\leq\theta}|S_n(t)-S_T(t)| &\leq \frac{\alpha(\theta)}{Y_n(\theta)}\sup_{0\leq t\leq\theta}|Z_n(t)-S_Y(t_-)|+\beta(\theta)\sup_{0\leq t\leq\theta}|N_n(t)-F_{1,Y}(t)| \\ &+ \gamma(\theta)\sup_{0\leq t\leq\theta}|N_n(t_-)-F_{1,Y}(t_-)|. \qquad (3.10)\end{aligned}$$

La relation

$$\sum_{n\geq 2} P\left(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|>5\sqrt{\frac{\log n}{n}}\right)<\infty,$$

la remarque la suivant et le fait que $Y_n(\theta) \xrightarrow{p.co.} S_Y(\theta^-) \neq 0$ *et que* $\sup_{0\leq t\leq\theta}|S_n(t)-S_T(t)| = \sup_{t\leq\theta}|S_n(t)-S_T(t)|$ *entraînent les résultats visés.*

3.2.2 Normalité asymptotique

L'EKM possède bien d'autre propriétés, dont sa normalité asymptotique qui a été montrée dans Breslow 1974, dont voici l'énoncé.

Théorème 3.2. *Si* $T^* < 1$ *avec* $H(T^*) < 1$ *et si* F_T *et* F_C *sont continues alors pour tout* $t \leq T^*$, *nous avons*

$$\sqrt{n}\left(\widehat{S}_n(t)-S_T(t)\right)\xrightarrow{\mathcal{L}} Z,$$

où Z *suit la loi normale centrée et de variance* $S_T^2(t)\int_0^t \frac{dS_{1,Y}(s)}{S_Y^2(s)}$.

3.3 Tests non paramétriques de comparaison de deux fonctions de survie pour des échantillons censurés

La comparaison de deux échantillons ou plus est un sujet plus important pour pour beaucoup d'études cliniques. Dans le cas de la censure à droite, la majorité des tests non-paramétriques sont basés sur des statistiques linéaires de rang étendues au cas censuré ; ils sont des généralisations de tests de rangs initialement proposés pour des données non-censurées. Ainsi, Mantel 1966 donne une extension de la statistique de Mantel-Haenszel pour données complètes. Ce test adapté est équivalent à la généralisation du test de Savage formalisé par Rao 1960 et au test du log-rang proposé par Peto 1972.

De même, des généralisations des tests de Wilcoxon et de Kruskal-Wallis (test de Wilcoxon pour $p+1$ échantillons) ont été proposés par Gehan 1965, Breslow 1970, Peto 1972 et Prentice 1978.

Notons S_1 et S_2 les fonctions de survie dans deux groupes A et B, tels que

• ; S_1 : échantillon de taille n_1 (x_i et x_i^+).

• ; S_1 : échantillon de taille n_1 (y_j et y_j^+).

les données suivies de + sont censurées à droite.

On souhaite tester

$$H_0 \ : \ S_1 = S_2.$$

contre

$$H_1 \ : \ S_1 > S_2.$$

ou

$$H_2 \ : \ S_1 < S_2.$$

ou

$$H_3 \ : \ S_1 \neq S_2.$$

Dans la suite, on va proposer un test non-paramétrique asymptotique qui marche en présence de données censurées à droites.

3.3.1 Tests Gehan

En 1965 Gehan a introduit un test non paramétrique qu'il est une généralisation du test de Wilcoxon-Mann-Whitney au cas des données censurées à droite.

Posons

$$U_{ij} = \begin{cases} 1, & \text{si } x_i > y_j \text{u} x_i^+ \geq y_j^+, \\ -1, & \text{si } x_i < y_j \text{u} x_i^+ \leq y_j^+, \\ 0, & \text{par ailleurs .} \end{cases}$$

La statistique utilisé pour cet test est

$$\mathcal{W} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij},$$

où la somme est sur tout n_1 et n_2. Notons que le calcul de $\mathcal{W}$ étant trop lourd pour n_1 et n_2 sont plus grands.
Mantel 1967 montre qu'il peut être calculé dans une alternative façon en attribuant un score à chaque observation en fonction de son classement relatif. Calcul de Gehan chaque observation dans l'échantillon A est comparée à chacune dans échantillon B. les deux échantillons sont combinés en un seul échantillon groupé d'observations, cela revient au même que de comparer chaque observation avec restant $n_1 + n_2 - 1$. Mantel 1967 trouve que

$$\mathcal{W} = \sum_{i=1}^{n_1} U_i$$

où la sommation est sur le U_i de l'échantillon A uniquement et $U_i = R_{1,i} - R_{2,i}$, avec $= R_{1,i}$ et $R_{2,i}$ sont calculés comme suit *bullet* Calcul de $R_{1,i}$

i. On note le rang des observations complètes de gauche à droites.
ii. On donne à une donnée censurée, le même rang que la donnée non censurée qui suit.
iii. Si les données complètes se répètent, on réduit leurs rang au plus petit.

On appelle $R_{1,i}$ le rang obtenu pour la i^{eme} observation.
bullet Calcul de $R_{2,i}$

i. On note le rang de toutes les observations de droites à gauche.

ii. Si les données complètes se répètent, on réduit leurs rang au plus petit.

iii. Réduire le rang des données censurées à 1. On appelle $R_{2,i}$ le rang obtenu pour la i^{eme} observation.

Notons que, sous l'hypothèse H_0

$$\mathcal{W} \sim \mathcal{N}(0, \sigma^2),$$

avec

$$\sigma^2 = Var(\mathcal{W}) = \frac{n_1 n2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{i=1}^{n_1+n_2} U_i^2.$$

Le test de Mantel 1967 est basé sur la statistique suivante

$$Z = \frac{\mathcal{W}}{\sigma^2},$$

avec

$$\mathcal{Z} \sim \mathcal{N}(0, 1).$$

Les régions de rejet sont

$$\begin{cases} Z > Z_\alpha, & \text{pour } H_1, \\ Z < -Z_\alpha, & \text{pour } H_2, \\ |Z| > Z_{\alpha/2}, & \text{pour } H_3, \end{cases}$$

où $P\left(Z > Z_\alpha / H_0\right) = \alpha$.

3.3.2 Tests log-rang

Soient $t_1 < t_2 < ... < t_N$ les différents temps de données complètes dans l' échantillon global obtenu par concaténation de tous les deux échantillons observés.

Nous notons, pour $i = \overline{1,n}$ et $j\overline{1,2}$

- d_{ij} : le nombre de données complètes dans le j^{eme} échantillon au temps t_i.
- n_{ij} : le nombre d'individus à risque dans le j^{eme} échantillon au temps t_i.
- $d_i = d_{i1} + d_{i2}$: le nombre de données complètes dans l'échantillon global au temps
- $n_i = n_{i1} + n_{i2}$: le nombre d'individus à risque dans l'échantillon global au temps

Pour chaque temps d'événement t_i, l'information peut être résumée sous forme de tableau suivant

Évenements à t_i	Groupe 1	Groupe 2	Toal
Nb réalisés à t_i	d_{i1}	d_{i2}	d_i
Nb non réalisés à t_i	$n_{i1} - d_{i1}$	$n_{i2} - d_{i2}$	$n_i - d_i$
Total juste avant t_i	n_{i1}	n_{i2}	n_i

On s'intéresse aux temps des données complètes, alors on remarque que n_{i1}, n_{i2} suivent une loi hypergéométrique avec les paramètres $(d_i,\ d_{i1},\ n_{i1})$ et $(d_i,\ d_{i2},\ n_{i2})$ avec des espérances à

l'instant t_i donnée par

$$\begin{cases} E_{i1} = d_i \frac{n_{i1}}{n_i} \\ E_{i2} = d_i \frac{n_{i2}}{n_i} \end{cases}$$

obtient par les formule

$$\begin{cases} V_{i1} = E_{i1} \left(\frac{n_i - d_i}{n_i}\right) \left(\frac{n_i - n_{i1}}{n_i}\right) \\ V_{i2} = E_{i2} \left(\frac{n_i - d_i}{n_i}\right) \left(\frac{n_i - n_{i2}}{n_i}\right) \end{cases}$$

Posons

$$E_1 = \sum_{i=1}^{n} E_{i1}, \qquad V_1 = \sum_{i=1}^{n} V_{i1},$$

et

$$E_2 = \sum_{i=1}^{n} E_{i2}, \qquad V_2 = \sum_{i=1}^{n} V_{i2},$$

alors, on peut montrer que sous H_0 les statistiques

$$\begin{cases} Z_1 = \frac{d_1 - E_1}{\sqrt{V_1}} \overset{TCL}{\sim} \mathcal{N}(0.1) \\ Z_2 = \frac{d_2 - E_2}{\sqrt{V_2}} \overset{TCL}{\sim} \mathcal{N}(0.1) \end{cases}$$

avec $d_l = \sum_{i=1}^{n} d_{il}$ pour $l = \overline{1,2}$.

Donc, la statistique du test qui combine les deux variables est donnée par

$$Z = Z_1^2 + Z_2^2 \rightsquigarrow\sim \chi_1^2,$$

où χ_1^2 est loi de khi deux à 1 degré de liberté.

Pour un test asymptotique de niveau α, la zone de rejet est telle que

$$R_\alpha = Z > c_\alpha,$$

où c_α est le quantile d'ordre $1 - \alpha$ de la loi χ_1^2.

3.4 Exercices

Exercice 1.

Dix patients atteints d'un cancer sont suivis jusqu'à la mort. Leurs temps de survie, en mois, sont

$$5,\ 6,\ 8,\ 10,\ 10,\ 10,\ 13,\ 13,\ 15,\ 19.$$

1. Estimer la fonction de survie du temps de survie à ce cancer.
2. Tracer la courbe de cet estimateur.

Exercice 2.

Sur dix patients atteints de cancer des bronches, on a observé les durées de survie suivantes, exprimées en mois.

$$7,\ 3+,\ 13,\ 10,\ 11+,\ 15+,\ 3,\ 2,\ 8+,\ 9,$$

les données suivies de + sont censurées à droite.

1. Calculons l'estimateur de la loi du temps de survie à cette maladie.
2. Tracer la courbe de cet estimateur.

Exercice 3.

Sur dix patients atteints de cancer des bronches, on a observé les durées de survie suivantes, exprimées en mois.

$$7,\ 3+,\ 13,\ 10,\ 10,\ 15,\ 3,\ 2,\ 8+,\ 9,$$

les données suivies de + sont censurées à droite.

1. Calculons l'estimateur de la loi du temps de survie à cette maladie.

2. Tracer la courbe de cet estimateur.

Exercice 4.

Donner une condition nécessaire et suffisante pour que l'estimateur de Kaplan-Meier s'annule.

Exercice 5.

Soit X_1, X_2, ... , X_n n v.a.r i.i.d. de même loi de X. Posons $T_n = \sup_{1 \leq i \leq n} X_i$. Montrer que

1. Si $T_X < +\infty$, alors $T_n \to T_X$ $(n \to \infty)$ p.s.

2. Si X et Y sont deux v.a.r. vérifiant $X < Y$ p.s alors $T_X < T_Y$ p.s.

3. Soient X et Y deux v.a.r. indépendantes, montrer que

$$T_{X \wedge Y} = T_X \wedge T_Y.$$

4. Soient X et Y deux v.a.r. indépendantes, montrer que

$$T_{X \vee Y} = T_X \vee T_Y.$$

Exercice 6. Répondre vrai ou faux avec explication.

A) Le but du test du Log-Rank est de voir si la survie est identique dans les deux groupes étudiés.

B) L'analyse de la survie est l'estimation de la probabilité de survenue d'un décés.

C) La probabilité $S(t_i)$ est la probabilité pour qu'un individu meure avant la date t_i.
D) Le test du Log-Rank correspond á un test du Chi 2 avec un seul degré de liberté.
E) La fonction de survie $S(t)$ est estimé en faisant le produit des survies instantanés sur tous les intervalles.

Exercice 7.

Soit le temps T (semaine) de rémission chez deux groupes de patients atteints de leucémie :
- Le premier groupe est soumis à un traitement
(24, 30, 42, 15+, 40+, 42+) (où $x+$ indique que x est censurée à droite).
- Le second groupe est soumis à un placebo
(10, 26, 28, 30, 41, 12+).
Peut-on conclure à l'efficacité du médicament ? (tester l'hypothèse de réaction identique).
1. Utiliser le test non paramétrique de Log-Rang.
1. Utiliser le test non paramétrique de Gehan.

Chapitre 4

Estimation des fonctions de densité et de régression pour des données censurées à droite

4.1 Estimation de la fonction de densité pour des données censurées à droite

4.1.1 Principe de l'estimation

Nous observons un échantillon de variables aléatoires i.i.d. $(Y_i = \min(T_i, C_i),\ \delta_i = 1_{\{T_i \leq C_i\}})_{1 \leq i \leq n}$ de même loi que $(Y = \min(T, C),\ \delta)$ où C est indépendante de T.

Nous supposons, dans la suite, que T admet une densité de probabilité f_T. En suivant le même raisonnement adopté par Rosenblatt

(1956) et Parzen (1962) dans le cas de données complètes, Foldes et al. (1981) ont proposé d'estimer f comme suit

$$f_n(x) = \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) dF_n(y),$$

où K est le noyau, (h_n) est la fenêtre et $F_n = 1 - S_n$ avec S_n étant l'estimateur de Kaplan-Meier de S_T.

4.1.2 Convergence presque sûre

Sous l'hypothèse de continuité de f au point x et sous des conditions assez faibles sur le noyau et la fenêtre, Foldes et al. (1981) ont montré la convergence presque sûre de $f_n(x)$ vers $f_T(x)$ dans le théorème suivant.

Théorème 4.1. *(voir Foldes et al. (1981) page 22)*
Si f_T est bornée et continue au point x, $S_C(T_{F_T}^-) < 1$ et K est une densité continue à droite et à variation bornée et si $h_n \to 0$, $h_n \left(\frac{\ln n}{n}\right)^{1/8} \to \infty$, nous avons

$$f_n(t) \xrightarrow{p.s.} f_T(t),$$

où $T_{F_U} = \sup\{t, \quad F_U(t) < 1\}$ est le point terminal d'une variable aléatoire U. Il est clair que si $t > T_{F_U}$ alors $S_U(t) = 0$.

4.1.3 Convergence presque complète

Dans cette section, nous allons donner la convergence presque complète ponctuelle de $f_n(x)$ vers $f_T(x)$, qui est plus forte que la presque sûre, en un point $x < T_{F_Y}$. Pour cela, nous avons besoins à les hypothèses suivantes.

H1 f_T est continue au point x.

H2 f_T est de classe $\mathcal{C}^2$ au voisinage de x.

H3 $\exists k, p, \varepsilon_0 \in \mathbb{R}_+^*, \forall y \in]x-\varepsilon_0, x+\varepsilon_0[, |f_T(x)-f_T(y)| \leq k|x-y|^p$.

H4 $h_n \to 0$ et $nh_n^2/\log n \to \infty$.

H5 K est une densité continue à droite, à variation bornée sur $\mathbb{R}$ et telle que :
$\exists M > 0, \forall u \in \mathbb{R}, |u| \geq M \Rightarrow K(u) = 0$ (i.e. K est à support compact).

H6 K est bornée.

H7 $\int uK(u)\,du = 0$ et $\int u^2K(u)\,du < \infty$.

Notons que les conditions sur le noyau et la fenêtre un peu plus fortes que les conditions de la convergence presque sure, ce qui n'est pas très contraignant.

Théorème 4.2. *Soit $x < T_{F_Y}$, Supposons que les conditions* (**H1**), (**H4**), (**H5**) *et* (**H6**) *sont satisfait, alors nous avons*

$$f_n(x) \xrightarrow{p.co.} f_T(x).$$

Preuve 4.1. *On considère la décomposition suivante*

$$\begin{aligned} |f_n(x) - f_T(x)| &= |f_n(x) - Ef_n(x) + Ef_n(x) - f_T(x)| \\ &\leq |f_n(x) - Ef_n(x)| + |Ef_n(x) - f_T(x)| \end{aligned} \quad (4.1)$$

le preuve de ce théorème revient à montrer les deux Lemmes ci-dessous, traitant chacun un des termes de la décomposition 4.1. Commençons par étudier le deuxième terme.

Lemme 4.1. *Sous* (**H4**) *et* (**H5**), *on a pour tout $\theta < T_{F_Y}$:*

$$\sup_{x \leq \theta} |f_n(x) - \mathbb{E}f_n(x)| = O_{p.co.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),$$

où $\mathbb{E}f_n(x) = \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) dF_T(y)$.

Preuve 4.2. *La démonstration de ce lemme s'inspire de celle du théorème 3.2. de Foldes et al. (1981).*
Soient $\theta < T_{F_Y}$ et $x \leq \theta$, nous avons :

$$|f_n(x) - \mathbb{E}f_n(x)| = \frac{1}{h_n}\left|\int K\left(\frac{x-y}{h_n}\right) d(F_n(y) - F_T(y))\right|$$

En posant $u = \frac{x-y}{h_n}$, nous obtenons :

$$|f_n(x) - \mathbb{E}f_n(x)| = \frac{1}{h_n}\left|\int K(u) d(\widetilde{F}_n(u) - \widetilde{F}(u))\right|$$

avec

$$\widetilde{F}_n(u) = F_n(x - uh_n) \text{ et } \widetilde{F}(u) = F_T(x - uh_n),$$

donc, par intégration par parties

$$\begin{aligned} |f_n(x) - \mathbb{E}f_n(x)| &= \frac{1}{h_n}\left|\int_{-M}^{M} K(u)d(\widetilde{F}_n(u) - \widetilde{F}(u))\right| \\ &\leq \frac{1}{h_n}\left|\int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u))dK(u)\right|. \quad (4.2) \end{aligned}$$

De plus, K étant à variation bornée et continue à droite, nous avons $K = K_1 - K_2$ où K_1 et K_2 sont deux fonctions croissantes et continues à droites, d'où :

$$\begin{aligned} |f_n(x) - \mathbb{E}f_n(x)| &\leq \frac{1}{h_n}\left|\int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u))dK_1(u)\right| \\ &\quad + \frac{1}{h_n}\left|\int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u))dK_2(u)\right| \\ &\leq \frac{\sup_{u>-M} |\widetilde{F}_n(u) - \widetilde{F}(u)|}{h_n}\left[\int_{-M}^{M} dK_1(u) + \int_{-M}^{M} dK_2(u)\right] \\ &\leq \frac{V_K}{h_n} \sup_{u>-M} |\widetilde{F}_n(u) - \widetilde{F}(u)| \\ &\leq \frac{V_K}{h_n} \sup_{u>-M} |F_n(x - uh_n) - F_T(x - uh_n)|, \end{aligned}$$

où V_K est la variation totale de K sur $\mathbb{R}$.
Soit $\theta^* \in]\theta, T_{F_Y}[$, puisque $h_n \to 0$, $\exists n_0 \in \mathbb{N}/\forall n \geq n_0 : Mh_n < \theta^* - \theta$, ce qui montre que $x - uh_n < \theta^*$ du fait que $u > -M$ et $x \leq \theta$. Par conséquent $\sup_{u>-M} |F_n(x - uh_n) - F_T(x - uh_n)| \leq$

$\sup_{t<\theta^*}|F_n(t) - F_T(t)|$.
Donc pour tout $x \leq \theta$ *nous avons :* $|f_n(x) - \mathbb{E}f_n(x)| \leq \frac{V_K}{h_n}\sup_{t<\theta^*}|F_n(t) - F_T(t)|$, *d'où :*

$$\sup_{x\leq\theta}|f_n(x) - \mathbb{E}f_n(x)| \leq \frac{V_K}{h_n}\sup_{t<\theta^*}|F_n(t) - F_T(t)| = O_{p.co.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right), \tag{4.3}$$

en vertu du théorème de Bochner.

Enfin, le deuxième terme de la décomposition (4.1), se contrôle par le lemme suivant.

Lemme 4.2. *Soit* $x < T_{F_Y}$, *sous* (**H1**),(**H4**),(**H5**) *et* (**H6**), *nous avons*

$$\mathbb{E}f_n(x) \underset{n\to\infty}{\longrightarrow} f_T(x),$$

Preuve 4.3. *En utilisant le changement de variable* $z = x - y$, *nous pouvons écrire :*

$$\begin{aligned}
\mathbb{E}f_n(x) &= \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) f_T(y)dy \\
&= \frac{1}{h_n}\int K\left(\frac{z}{h_n}\right) f_T(x-z)dz \\
&\overrightarrow{n\to\infty} \; f_T(x),
\end{aligned}$$

d'après le théorème de Bochner ($zK(z) \underset{|z|\to\infty}{\longrightarrow} 0$ *car* K *est à support compact).*

Théorème 4.3. *Soit $x < T_{F_Y}$,sous* (**H3**),(**H4**) *et* (**H5**), *nous obtenons*

$$f_n(x) - f_T(x) = O_{p.co.}\left(h_n^p + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

La démonstration de cette théorème est basée sue la décomposition (4.1) et le lemme 4.1 avec le lemme suivant.

Lemme 4.3. *Soit $x < T_{F_Y}$, sous* (**H3**),(**H4**) *et* (**H5**), *nous avons*

$$\mathbb{E}f_n(x) - f_T(x) = O(h_n^p).$$

Preuve 4.4. *En utilisant le changement de variable $u = \frac{x-y}{h_n}$, il s'ensuit que :*

$$\begin{aligned}
|\mathbb{E}f_n(x) - f_T(x)| &= \left|\frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) f_T(y)dy - f_T(x)\right| \\
&= \left|\int K(u) f_T(x - uh_n)du - f_T(x)\right| \\
&= \left|\int_{-M}^{M} K(u) f_T(x - uh_n)du - f_T(x)\int_{-M}^{M} K(u)du\right| \\
&= \left|\int_{-M}^{M} K(u)(f_T(x - uh_n) - f_T(x))du\right|,
\end{aligned} \tag{4.4}$$

et comme

$$h_n \to 0, \quad \exists n_1 \in \mathbb{N}/\forall n \geq n_1 : h_n < \frac{\varepsilon_0}{M},$$

donc pour tout $n \geq n_1$ nous avons

$$|x - uh_n - x| = |u|h_n < \varepsilon_0 \Rightarrow |f_T(x - uh_n) - f_T(x)| \leq kM^p h_n^p,$$

d'où

$$\begin{aligned} |\mathbb{E}f_n(x) - f_T(x)| &\leq kM^p h_n^p \int_{-M}^{M} K(u)du \\ &= kM^p h_n^p \\ &= O(h_n^p). \end{aligned}$$

Théorème 4.4. *Soit* $x < T_{F_Y}$, *sous* (**H2**), (**H4**), (**H5**) *et* (**H7**), *nous avons*

$$f_n(x) - f_T(x) = O_{p.co.}\left(h_n^2 + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Preuve 4.5. *Exercice.*

Notons que Földes A. et al. (1981) ont étudié la convergence presque sûre uniforme de f_n sous l'hypothèse de bornitude de la dérivée de f. et Boukeloua (2013) est donné la convergence presque complète uniforme sur un compact C sous l'hypothèse de continuité de f sur C.

4.2 Estimation de la fonction de régression pour des données censurées à droite

4.2.1 Principe de l'estimation

Nous observons un échantillon de variables aléatoires i.i.d. $(X_i, Z_i = \min(Y_i, C_i))_{1\leq i\leq n}$ de même loi que $(X, Z = \min(Y, C))$ où (X, Y)

à valeur dans $\mathbb{R}^d \times \mathbb{R}$.

Soit D_1 l'hypothèse comprenant les trois conditions suivantes

$D_{1,1}$ S_C continue et $T_{F_Y} < T_{F_C}$.

$D_{1,2}$ C et (X, Y) sont indépendants.

$D_{1,3}$ $T_{F_C} < \infty, \exists\ T$ avec $T < T_{F_Y}$ tq $Z_i \leq T$.

Sous l'hypothèse que C et (X, Y) sont indépendants, Nadaraya-Watson, construit simultanément un estimateur de la régression en 1964 (voir Nadaraya (1964) et Watson (1964)) en suivant l'idée de Carbonez et al. (1995). Soit

$$\begin{aligned} h : \mathbb{R}^d \times \mathbb{R} &\longrightarrow \mathbb{R} \\ (x, y) &\longmapsto h(x, y), \end{aligned}$$

une fonction mesurable et on considéré le problème de l'estimation de $E\left[h(X, Y)\right]$ à partir de l'èchantillon (X_i, Z_i, δ_i). Un estimateur sans biais de $E\left[h(X, Y)\right]$ est

$$\frac{1}{n} \sum_{i=1}^{n} \frac{h(X_i, Z_i)\delta_i}{S_C(Z_i)}.$$

Ceci est motivé par le fait que

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{h(X_i,Z_i)\delta_i}{S_C(Z_i)}\right] &= \frac{1}{n}\sum_{i=1}^{n}E\left[\frac{h(X_i,Z_i)\delta_i}{S_C(Z_i)}\right] \\ &= E\left[\frac{h(X,Z)\delta}{S_C(Z)}\right] \\ &= E\left[\frac{h(X,Y)\delta}{S_C(Y)}\right] \\ &= E\left[E\left(\frac{h(X,Y)\delta}{S_C(Y)}/(X,Y)\right)\right] \\ &= E\left[\frac{h(X,Y)}{S_C(Y)}E\left(\delta/(X,Y)\right)\right] \\ &= E\left(h(X,Y)\right), \end{aligned}$$

car

$$E\left(\delta/(X,Y)\right) = S_C(Y), \tag{4.5}$$

si les hypothèses $D_{1,1}$-$D_{1,3}$ sont satisfaites. En effet
Pour tout B dans $\sigma(X,Y)$ (tribu engendrée par le couple (X,Y)) il existe un borélien A tel que $B = (X,Y)^{-1}(A)$. L'indépendance de (X,Y) et C permet d'ecrire (2,2)

$$\begin{aligned} \int_B \delta d\mathbf{P} &= \int_B 1_{\{Y\leq C\}}d\mathbf{P} \\ &= \int_{A\times\mathbf{R}} 1_{\{y\leq u\}}d\mathbf{P}_{(X,Y,C)}(x,y,u) \\ &= \int_A\int_{\mathbf{R}} 1_{\{y\leq u\}}d\mathbf{P}_{(X,Y)}(x,y)\bigotimes d\mathbf{P}_C(u). \end{aligned}$$

Maintenant par le théorème de Fubini et la condition $D_{1,1}$, nous obtenons

$$\begin{aligned}\int_B \delta d\mathbf{P} &= \int_A \left(\int_{\mathbf{R}} 1_{\{u \geq y\}} d\mathbf{P}_C(u) \right) d\mathbf{P}_{(X,Y)}(x,y) \\ &= \int_A S_C(y-) d\mathbf{P}_{(X,Y)}(x,y) \\ &= \int_A S_C(y) d\mathbf{P}_{(X,Y)}(x,y) \\ &= \int_B S_C(Y) d\mathbf{P}.\end{aligned}$$

De plus, $S_C(Y)$ étant clairement mesurable par rapport à $\sigma(X,Y)$, le résultat en découle. Alors, on peut estimer r par

$$\tilde{r}(x) = \sum_{i=1}^{n} W_{i,n}(x) \frac{Z_i \delta_i}{S_C(Z_i)}, \tag{4.6}$$

avec $W_{n,i}(x)$ est donnée par

$$W_{n,i}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)}, \tag{4.7}$$

Mais en pratique la fontion de survie S_C des censures est souvent inconnue, donc il est impossible d'utiliser l'estimateur $\tilde{r}$ (4.6). Par conséquent, on remplace S_C par son estimateur de Kaplan et Meier (1958) S_n donné par

$$S_n(Z) = \begin{cases} \prod_{i=1}^{n} (1 - \frac{\delta_i}{n-i+1})^{1_{\{Z_i \leq t\}}} & \textbf{si} \quad t < M_n \\ \lim_{s \to M_n} S_n(s) & \textbf{si} \quad t \geq M_n \end{cases} \tag{4.8}$$

avec $M_n = \max(Z_i) = Z_n$ tq Z_i statistiquement ordonné $Z_{(1)} \leq Z_{(2)} \leq Z_{(3)} \leq \cdots Z_{(n)}$ et δ_i indicatrice de l'observation Z_i
D'après le théorème du Bitouzé et al. (1999) et sous la condition D_1, nous avons

$$\sup_{x \leq T} |S_n(t) - S_C(t)| \xrightarrow{p.co} 0. \tag{4.9}$$

Nous pouvons déduire de l'hypothèse $D_{1,3}$ que $S_C(T) > 0$. En vertu de (4.9), nous obtenons pour n suffisamment grand $S_n(T) > 0$ presque sure.
Donc, un estimateur calculable de $r(.)$ est obtenu par

$$\tilde{r}_n(x) = \sum_{i=1}^{n} W_{i,n}(x) \frac{Z_i \delta_i}{S_n(Z_i)}. \tag{4.10}$$

Remarquons que pour toute $1 \leq j \leq n$, $S_n(Z_j) = 0$ implique que $\delta_j = 0$.

4.2.2 Convergence presque complète

Dans cette section, nous montrons la convergence presque complète ponctuelle de l'estimateur à noyau de la régression en présence d'une censure à droite, en précisant la vitesse de convergence. Dans toute la suite, nous appellerons $\widehat{r}_n$ l'estimateur à noyau de la régression défini par (4.10) considéré pour une variable aléatoire X à valeurs dans $\mathbb{R}^d$, de densité f dont nous supposons l'existence.
Maintenant, nous introduisons les hypothèses de base permettant

de donner un théorème général sur la convergence presque complète ponctuelle de $\widehat{r}_n(x)$.

D_2 r et f sont "l" fois continument dérivable dans un voisinage de x.

D_3 $f(x) > 0$.

D_4 $h_n \longrightarrow 0$ et $\lim_{n\to+\infty} \frac{logn}{nh_n} = 0$.

D_5 K est borné, intégrable à support compact et $\int K(z)dz = 1$.

D_6 $\int t^j K(t)dt = 0$, $\forall j = \overline{1, l-1}$ et $0 < |\int t^l K(t)| < \infty$.

Théorème 4.5. *Sous les conditions D_1 et $D_2 - D_6$ nous avons*

$$\tilde{r}_n(x) - r(x) = O_{p.co}(h_n^l) + O_{p.co}\left(\sqrt{\frac{logn}{nh_n}}\right).$$

Preuve 4.6. *Posons*

$$\tilde{r}_n(x) = \sum_{i=1}^{n} W_{i,n}(x)\frac{\delta_i Z_i}{S_n(Z_i)} = \frac{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})\frac{\delta_i Z_i}{S_n(Z_i)}}{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} := \frac{\hat{g}_n(x)}{\hat{f}_n(x)}$$

et

$$\tilde{r}(x) = \sum_{i=1}^{n} W_{i,n}(x)\frac{\delta_i Z_i}{S_C(Z_i)} = \frac{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})\frac{\delta_i Z_i}{S_C(Z_i)}}{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} = \frac{\tilde{g}_n(x)}{\hat{f}_n(x)}.$$

où f_n et g_n sont des estimateurs de la densité f et de $g(x)$ avec $g(x) = r(x)f(x)$.
La démonstration du ce théorème est basé sur la décomposition suivante

$$\begin{aligned}
\tilde{r}_n(x) - r(x) &= \tilde{r}_n(x) - \tilde{r}(x) + \tilde{r}(x) - r(x) \\
&= \tilde{r}_n(x) - \tilde{r}(x) + \frac{\tilde{g}_n(x) - g(x)}{f_n(x)} + \frac{f(x) - f_n(x)}{f_n(x)} r(x) \\
&= \tilde{r}_n(x) - \tilde{r}(x) + \frac{1}{f_n(x)} \left[\tilde{g}_n(x) - E\tilde{g}_n(x) + E\tilde{g}_n(x) - g(x)\right] \\
&+ \frac{r(x)}{f_n(x)} \left[f(x) - Ef_n(x) + Ef_n(x) - f_n(x)\right]
\end{aligned}$$

et les lemmes suivantes.

Lemme 4.4. *Sous les conditions $D_{1,1}$, $D_{1,3}$ et D_4, nous obtenons*

$$\tilde{r}_n(x) - \tilde{r}(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Preuve 4.7. *Nous avons,*

$$\begin{aligned}
|\tilde{r}_n(x) - \tilde{r}(x)| &= \left| \sum_{i=1}^{n} W_{i,n}(x) \frac{Z_i \delta_i}{S_n(Z_i)} - \sum_{i=1}^{n} W_{i,n}(x) \frac{Z_i \delta_i}{S_C(Z_i)} \right| \\
&= \left| \sum_{i=1}^{n} W_{i,n}(x) Z_i \delta_i \left(\frac{1}{S_n(Z_i)} - \frac{1}{S_C(Z_i)} \right) \right| \\
&= \left| \sum_{i=1}^{n} W_{i,n}(x) Z_i \delta_i \left(\frac{S_C(Z_i) - S_n(Z_i)}{S_n(Z_i) S_C(Z_i)} \right) \right| \\
&\leq \frac{T}{S_C(T) S_n(T)} \sup_{t \leq T} |S_n(t) - S_C(t)|,
\end{aligned}$$

car $\sum_{i=1}^{n} W_{i,n}(x) = 1$. *D'aprés Bitouzé et al. (1999), nous avons*

$$\sup_{t \leq T} |S_n(t) - S_C(t)| = O_{p.co} \left(\sqrt{\frac{\log n}{n}} \right).$$

De plus, puisque $h_n \to 0$ *(hypothèse* D_4*), nous déduisons que*

$$|\tilde{r}_n(x) - \tilde{r}(x)| = O_{p.co} \left(\sqrt{\frac{\log n}{n h_n}} \right).$$

En effet,

$$\frac{\sqrt{\frac{\log n}{n}}}{\sqrt{\frac{\log n}{n h_n}}} = \sqrt{h_n} \to 0.$$

Lemme 4.5. *Sous les conditions* D_2, D_5, D_6 *et* $D_{1,2}$*, nous avons*

$$E\left[g_n(x)\right] - g(x) = O_{p.co}(h_n^l).$$

Preuve 4.8. *Nous avons*

$$\begin{aligned} E\left[\tilde{g}_n(x)\right] &= E\left(\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)}\right) \\ &= E\left(\frac{1}{h_n}K(\frac{x-X}{h_n})\frac{\delta Z}{S_C(Z)}\right) \\ &= \frac{1}{h_n}E\left(E\left(K(\frac{x-X}{h_n})\frac{\delta Z}{S_C(Z)}/X\right)\right) \\ &= \frac{1}{h_n}E\left(K(\frac{x-X}{h_n})E\left(\frac{\delta Z}{S_C(Z)}/X\right)\right) \end{aligned}$$

Or, D'autre part, la relation (4.5) nous permet d'écrire

$$\begin{aligned} E\left(\frac{\delta Z}{S_C(Z)}/X\right) &= E\left(E\left(\frac{\delta Y}{S_C(Y)}/(X,Y)\right)/X\right) \\ &= E\left(E\left(\frac{\delta Y}{S_C(Y)}/(X,Y)\right)/X\right) \\ &= E\left(\frac{Y}{S_C(Y)}E(\delta/(X,Y))/X\right) \\ &= E(Y/X) = r(X). \end{aligned}$$

Alors, posons $z = \frac{x-t}{h_n}$ *et puisque* $\int K(t)dt = 1$, *nous obtenons*

$$\begin{aligned} E\left[\tilde{g}_n(x)\right] - g(x) &= \frac{1}{h_n} E\left(K(\frac{x-X}{h_n})r(X)\right) - g(x) \\ &= \frac{1}{h_n}\int K(\frac{x-t}{h_n})r(t)f(t)dt - g(x)\int K(z)dz \\ &= \frac{1}{h_n}\int K(z)r(x-zh_n)h_n f(x-zh_n)dz - g(x)\int K(z)dz \\ &= \int K(z)\left(g(x-zh_n) - g(x)\right)dz. \end{aligned}$$

L'hypothèse D_2 *et le développement de Taylor de* g *au voisinage de* x *permettent d'écrire*

$$g(x-zh_n) = g(x) + \sum_{j=1}^{l}(-1)^j \frac{z^j h_n^j}{j!} + o(h_n^l)$$

Donc, d'après les hypothèses D_5 et D_6, nous avons

$$\begin{aligned}
|E\left[\tilde{g}_n(x)\right] - g(x)| &= \left| \int K(z) \left(\sum_{j=1}^{l} (-1)^j \frac{z^j h_n^j}{j!} g^{(j)}(x) + o(h_n^l) \right) dz \right| \\
&\leq \int K(z) \left| \sum_{j=1}^{l} \frac{z^j h_n^j}{j!} g^{(j)}(x) + o(h_n^l) \right| dz \\
&\leq \sum_{j=1}^{l} \frac{h_n^j}{j!} g^{(j)}(x) \int z^j K(z) dz + o(h_n^l) \int K(z) dz \\
&\leq \sum_{j=1}^{l-1} \frac{h_n^j}{j!} g^{(j)}(x) \int z^j K(z) dz + \frac{h_n^l}{l!} g^{(l)}(x) \int z^l K(z) dz + o(h_n^l) \\
&\leq \frac{h^l}{l!} g^{(l)}(x) \int K(z) z^l dz + o(h_n^l) \\
&= O(h_n^l) + o(h_n^l) = O_{p.co}(h_n^l)
\end{aligned}$$

Lemme 4.6. *Sous les conditions D_2, D_4, D_5 et $D_{1,3}$, nous obtenons*

$$\tilde{g}_n(x) - E\left[\tilde{g}_n(x)\right] = O_{p.co}\left(\sqrt{\frac{\log n}{n h_n}}\right)$$

Preuve 4.9. *Nous avons*

$$\begin{aligned}
\tilde{g}_n(x) - E\left[\tilde{g}_n(x)\right] &= \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)} \\
&- E\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\{K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)h_n} \\
&- E\left[K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)h_n}\right]\} \\
&:= \frac{1}{n}\sum_{i=1}^{n}\Gamma_i,
\end{aligned}$$

avec

$$\Gamma_i := Q_i - EQ_i \quad \texttt{et} \quad Q_i = K\left(\frac{x-X_i}{h_n}\right)\frac{\delta_i Z_i}{S_C(Z_i)h_n}.$$

Pour appliquer l'inégalité de Bernstein dans la proposition, il faut de montrer que

$$|\Gamma_1| \leq M_n, \quad E(\Gamma_1^2) \leq \sigma_n^2, \quad \frac{M_n}{\sigma_n^2} \leq C \quad \texttt{et} \quad \lim U_n = 0, \quad \texttt{avec} \quad U_n = \frac{\sigma_n^2}{n}\log n.$$

En vertu des hypothèses D_5 et $D_{1,3}$, nous obtenons

$$|\Gamma_1| \leq \frac{T}{S_C(T)h_n} = \frac{C}{h_n} := M_n.$$

Nous devons maintenant borner $E(\Gamma_1^2)$*, pour cela calculons* $E(Q_1^2)$*. D'après l'hypothèse* $D_{1,3}$*, nous avons*

$$\begin{aligned} E(Q_1^2) &= E\left(\frac{1}{h_n^2}K^2\left(\frac{x-X_1}{h_n}\right)\frac{\delta_1^2 Z_1^2}{G^2(Z_1)}\right) \\ &\leq \frac{T^2}{G^2(T)}E\left(\frac{1}{h_n^2}K^2\left(\frac{x-X_1}{h_n}\right)\right) \\ &\leq \frac{T^2}{G^2(T)}\int \frac{1}{h_n^2}K^2\left(\frac{x-t}{h_n}\right)f(t)dt, \end{aligned}$$

avec un changement de variable $z = \frac{x-t}{h_n}$ *et d'après l'hypothèse* D_5*, nous obtenons*

$$\begin{aligned} E(Q_1^2) &\leq \frac{T^2}{G^2(T)}\int \frac{1}{h_n^2}K^2(z)f(x-h_n z)h_n dz \\ &\leq \frac{T^2}{G^2(T)h_n}\int K^2(z)f(x-h_n z)dz \\ &\leq \frac{T^2}{G^2(T)h_n} = \frac{C'}{h_n} = \sigma_n^2. \end{aligned}$$

Donc, nous avons

$$E(\Gamma_1^2) = E\left((Q_1 - EQ_1)^2\right) = var(Q_1) \leq E(Q_1^2) \leq \sigma_n^2.$$

Maintenant, nous remarquons que

$$\frac{M_n}{\sigma_n^2} = \frac{\frac{C}{h_n}}{\frac{C'}{h_n}} \leq C$$

et d'après l'hypothèse D_4,

$$U_n = n^{-1}\sigma_n^2 \log n = \frac{C \log n}{nh_n} \longrightarrow 0.$$

Donc, il vient

$$\frac{1}{n}\sum_{i=1}^{n} \Gamma_i = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right),$$

d'où

$$\tilde{g}_n(x) - E\left[\tilde{g}_n(x)\right] = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Lemme 4.7. *Sous les conditions D_5 et D_6, nous avons*

$$f(x) - E\left[f_n(x)\right] = O_{p.co}(h_n^l)$$

Preuve 4.10. *Puisque $\int K(z)dz = 1$, nous avons*

$$\begin{aligned} f(x) - E\left[f_n(x)\right] &= f(x) - E\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right] \\ &= f(x) - E\left[\frac{1}{nh_n} nK\left(\frac{x - X}{h_n}\right)\right] \\ &= f(x) - \frac{1}{h_n} E\left[K\left(\frac{x - X}{h_n}\right)\right] \\ &= f(x)\int K(z)dz - \frac{1}{h_n}\int K\left(\frac{x - t}{h_n}\right) f(t)dt. \end{aligned}$$

Alors, posons $z = \frac{x-t}{h_n}$, *nous obtenons*

$$\begin{aligned} f(x) - E\left[f_n(x)\right] &= f(x)\int K(z)dz - \int K(z)f(x-h_nz)dz \\ &= \int K(z)\left[f(x) - f(x-h_nz)\right]dz. \end{aligned}$$

D'après D_2 *et le développement de Taylor de* f *au voisinage de* x *nous avons*

$$f(x - zh_n) = f(x) + \sum_{j=1}^{l}(-1)^{j)}\frac{z^j h_n^j}{j!}f^{(j)}(x) + o(h_n^l),$$

alors, nous trouvons

$$\begin{aligned} |f(x) - Ef_n(x) &\leq \sum_{j=1}^{l-1}\frac{h_n^j}{j!}f^{(j)}(x)\int z^j K(z)dz + \frac{h_n^l}{l!}f^{(l)}(x)\int z^l K(z) + o(h_n^l) \\ &\leq \frac{h^l}{l!}\int z^l K(z)f^{(l)}(x)dz + o(h_n^l) \\ &\leq O(h_n^l) + o(h_n^l) = O_{p.co}(h_n^l). \end{aligned}$$

Lemme 4.8. *Sous les conditions* $D_2 - D_6$, *nous obtenons*
i)

$$E\left[f_n(x)\right] - f_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

ii) $\exists \mu > 0$ *t.q :*

$$\sum_{l=1}^{\infty} P(\hat{f}_n(x) < \mu) < \infty$$

Preuve 4.11. *i) Nous avons*

$$\begin{aligned} f_n(x) - E\left[f_n(x)\right] &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right) - E\left(\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)\right)\right] \\ &:= \frac{1}{n}\sum_{i=1}^{n}\Lambda_i, \end{aligned}$$

avec

$$\Lambda_i := Q_i' - EQ_i' \quad \texttt{et} \quad Q_i' = \frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right).$$

Pour appliquer l'inégalité de Bernstein dans la proposition, il faut de montrer que

$$|\Lambda_1| \leq M_n, \quad E(\Lambda_1^2) \leq \sigma_n^2, \quad \frac{M_n}{\sigma_n^2} \leq C \quad \texttt{et} \quad \lim U_n = 0, \quad \texttt{avec} \quad U_n = \frac{\sigma_n^2}{n}\log n.$$

En vertu l'hypothèse D_5, nous obtenons

$$|\Lambda_1| \leq \frac{C}{h_n} := M_n.$$

Nous devons maintenant borner $E(\Lambda_1^2)$, pour cela calculons $E(Q_1'^2)$. Nous faisons un changement de variable $z = \frac{x-t}{h_n}$ et d'après l'hypothèse D_5, nous obtenons

$$\begin{aligned} E(Q_1'^2) &= \frac{1}{h_n^2}\int K^2\left(\frac{x-t}{h_n}\right) f(t)dt \\ &\leq \frac{1}{h_n}\int K^2(z) f(x-h_n z)dz \\ &\leq \frac{C'}{h_n} = \sigma_n^2. \end{aligned}$$

Donc, nous avons

$$E(\Lambda_1^2) = E\left((Q_1' - EQ_1)'^2\right) = var(Q_1') \leq E(Q_1'^2) \leq \sigma_n^2.$$

Maintenant, nous remarquons que

$$\frac{M_n}{\sigma_n^2} = \frac{\frac{C}{h_n}}{\frac{C'}{h_n}} \leq C$$

et d'après l'hypothèse D_4*,*

$$U_n = n^{-1}\sigma_n^2 \log n = \frac{C\log n}{nh_n} \longrightarrow 0.$$

Donc, nous trouvons

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda_i = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right),$$

d'où

$$\hat{f}_n(x) - E\left[\hat{f}_n(x)\right] = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

ii) Remarquons que $\hat{f}_n(x)$ *converge presque complètement vers* $f(x)$*, donc il suffit de poser* $\mu = \frac{f(x)}{2}$ *et de remarquer que*

$$P\left(|\hat{f}_n(x)| < \frac{f(x)}{2}\right) \leq P\left(|\hat{f}_n(x) - f(x)| > \frac{f(x)}{2}\right).$$

4.3 Exercices

Exercice 1.

Soit $x < T_{F_Y}$, sous (**H2**), (**H4**), (**H5**) et (**H7**), nous avons

$$f_n(x) - f_T(x) = O_{p.co.}\left(h_n^2 + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Exercice 2.

Considérons deux variables aléatoires (v.a.) continues T et X, T est la variable d'intérêt censurée à droite par la v.a. C et X est la covariable, où la variable explicative, pouvant prendre des valeurs dans $\mathbb{R}^d$, $d \geq 1$. C'est á dire que l'observation consiste en $(X, Y = \min(T, C), \delta)$ où $\delta = 1_{\{Y \leq C\}}$.

1. Supposons que C est indépendant de (T, X), montrer que

$$E\left[\frac{\delta\Phi(Y, X)}{1 - G(Y_-)}\right] = E\left[\Phi(T, X)\right],$$

où Φ et une fonction mesurable de $\mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ et $G(t) = P(C \leq t)$.

2. Sous l'hypothèse que T et C sont conditionnellement indépendants sachant X, déterminer la fonction $h(Y, X)$ telle que

$$E\left[\delta h(Y, X)\Phi(Y, X)\right] = E\left[\Phi(T, X)\right].$$

3. Pour estimer la fonction $F(t/x) = P(T \leq t/X = x)$, on introduit les fonctions suivantes $H_2(t/x) = P(Y \leq t/X = x)$, $H_1(t/x) = P(Y \leq t, \delta = 1/X = x)$, $H_0(t/x) = P(Y \leq t, \delta =$

$0/X = x)$ et $G(t/x) = P(C \leq t/X = x)$.

3.1. Exprimer les fonctions $H_i(t/x)$ $(i = 0, 1, 2)$ en fonction de F et G.

3.2. Notons par $\Lambda_Z(t/x)$ la fonction de hasard cumulé conditionnelle de Z sachant X. Exprimer $\Lambda_T(t/x)$ et $\Lambda_C(t/x)$ en fonction de H_0, H_1 et H_2.

3.3. Notons que l'estimateur de $g(t/x) = E(\varphi(Y)/X = x)$ est

$$\hat{g}(t/x) = \frac{\sum_{i=1}^{n} K\left(h^{-1}(x - X_i)\right) \varphi(Y_i)}{\sum_{i=1}^{n} K\left(h^{-1}(x - X_i)\right)},$$

où φ est une fonction Borélienne, K noyau et h tend vers 0.

Proposer des estimateurs pour H_0, H_1 et H_2. (Indication : Ecrire $H_i(t/x) = E(\varphi(Y)/X = x)$).

3.4. Déduire un estimateur de $F(t/x)$. (Indication : Ecrire $1 - F(t/x) = \Pi_{s \leq t}(1 - d\Lambda_T(s/x))$).

4. A partir de la première question, Proposer un estimateur de $E\left[|T - \alpha X|\right]$ où α est un réel.

Chapitre 5

Introduction à l'estimation non paramétrique pour des données doublement censurées

5.1 Estimation de la fonction de survie

Considérons trois variables aléatoires positives indépendantes T, L et R de fonctions de répartition respectives F_T, F_L et F_R et de fonctions de survie respectives S_T, S_L et S_R où T représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Dans le modèle I de Patilea et Rolin (2006), au lieu d'observer un échantillon de T on observe un échantillon

du couple (Z, A) où $Z = \max(\min(T, R), L)$ et

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

Ce modèle considéré la censure a droite et la censure à gauche comme deux phénomènes qui agissent indépendamment l'un de l'autre mais que l'un peut censurer l'autre. Patilea et Rolin (2006) ont proposé d'estimer S_T, fonction de survie de la variable d'intérêt T comme suit.

5.1.1 Principe de l'estimation

Considérons H la fonction de répartition de Z, elle peut s'écrire $\sum_{k=0}^{2} H^{(k)}(t)$ où

$$H^{(k)}(t) = P(Z \leq t, A = k), \quad \text{pour } k = 0, 1, 2.$$

En notant pour toute application R de $\mathbb{R}$ dans $\mathbb{R}, R(t_-)$la limite de R à gauche de t, lorsque cette limite existe, ces fonctions s'écrivent

$$\begin{aligned} H^{(0)}(t) &= \int_0^t F_L(u_-)\, S_R(u_-)\, dF_T(u), \\ H^{(1)}(t) &= \int_0^1 F_L(u_-)\, S_T(u) dF_n(u), \\ H^{(2)}(t) &= \int_0^t \{1 - S_T(u) S_n(u)\}\, dF_L(u), \end{aligned}$$

et c'est a partir de ces équations que l'estimateur est obtenu.
L'idée est de considérer dans un premier temps $Y = \min(X, R)$ et L dans un modèle de censure à gauche (c'est-à-dire que l'on considère une donnée complète si $A = 0$ ou $A = 1$ et censurée à gauche si $A = 2$), et d'estimer la fonction de répartition de Y, puis l'utiliser pour estimer la fonction de répartition de la variable d'intérêt T en considérant un modèle de censure à droite.
L'estimateur de la fonction de survie S_T ainsi obtenu, en remplaçant a la fin les fonctions $H^{(0)}, H^{(1)}$ et $H^{(2)}$ par leurs estimateurs empiriques, obtenus à partir d'un échantillon $(Z_i, A_i)_{1 \leq i \leq n}$, est donné par

$$S_{\mathrm{n}}\left(Z'_j\right) = 1 - F_{\mathrm{n}}\left(Z'_j\right) = \prod_{1 s i s j} \left\{1 - \frac{D_{01}}{U_{1-1} - N_{l-1}}\right\}$$

où $\left(Z'_j\right)_{1 \leq j \leq M}$ sont les valeurs distinctes des Z_i prises dans l'ordre croissant, et

$$D_{kj} = \sum_{1 \leq i \leq n} 1_{\left[Z_1 = X'_j, A_1 = k\right]}, \quad N_j = \sum_{1 \leq i \leq n} 1_{|Z_1 \leq X'_1|},$$

$$U_{j-1} = n \prod_{j \leq l s M} \left\{1 - \frac{D_{2l}}{N_l}\right\}$$

pour $0 \leq l \leq 2$ et $1 \leq j \leq M$
Soulignons le fait que si $L \equiv 0$ (pas de censure à gauche), $\hat{S}_n$ se réduit à l'estimateur de Kaplan-Mejer qui lui-même se réduit au complément à 1 de la fonction de répartition empirique si $R \equiv \infty$.

Patilea et Rolin (2006) ont introduit cet estimateur et montré sa convergence uniforme presque sûre et sa convergence en tant que processus vers un gaussien sous des conditions sa convergence en tant que processus vers un gaussien sous des conditions d'identifiabilité du modèle. $\hat{S}_{\mathrm{n}}$ intervenant plus loin au dénominateur dans les expressions d'estimateurs de la fonction de régression, le résultat suivant donne une condition nécessaire et suffisante pour qu'il s'annule.
Dans un souci de clarté, nous notons dans la suite de ce chapitre D_{kj} par $D_{k,j}$.

Lemme 5.1. *i) Une condition nécessaire et suffisante pour que* $S_n(Z_{kj}) = 0$ *pour la première fois et reste nul est :* $D_{0,k_0} \neq 0, D_{1,k_a} = 0 \text{et} \forall j > k_0, D_{0,j} = D_{1,j} = 0$ *si* $k_n \neq M$.
ii) $\hat{S}_n$ *s'annule pour la première fois en* Z'_M *si et seulement si* $D_{0,M} \neq 0$ *et* $D_{1,M} = 0$.

Preuve 5.1. *1. Commençons par montrer que pour tout* $k, 0 \leq k \leq M-1$, *nous avons*

$$U_k \geq N_{k-}$$

En effet

$$\begin{aligned} U_k =& \pi \left(\frac{N_{k+1} - D_{2,k+1)}}{N_{k+1}} \right) \left(\frac{N_{k+2} - D_{2,(k+2)}}{N_{k+2}} \right) \cdots \left(\frac{N_M - D_{2M}}{N_M} \right) \\ =& \left(\frac{N_k + D_{0,(k+1)} + D_{1,(k+1)}}{N_{k+1}} \right) \cdots \left(\frac{N_{M-2} + D_{0,(M-1)} + D_{1,(M-1)}}{N_{M-1}} \right) \\ & \times \left(N_{M-1} + D_{0,M} + D_{1,M} \right) \\ \geq & \frac{N_k}{N_{k+1}} \times \cdots \times \frac{N_{M-2}}{N_{M-1}} \times N_{M-1} = N_{k-} \end{aligned}$$

Remarquons que s'il existe j tel que $j > k$ avec $D_{1,j} \neq 0$ ou $D_{0,j} \neq 0$ alors $U_k > N_k$.

2 Soit k_0 le premier indice k tel que $D_{0,k} = U_{k-1} - N_{k-1}$ (le premier k pour lequel $\hat{S}_{\mathrm{n}}(Z_k) = 0$). Nous avons alors :

$$D_{0,k0} \neq 0 \text{ et } D_{0,k0} = U_{k0-1} - N_{k0-1}.$$

Par ailleurs

$$U_{ka-1} = n \left(1 - \frac{D_{2,ka}}{N_{ka}} \right) \cdots \left(1 - \frac{D_{2,M}}{N_M} \right) = \left(1 - \frac{D_{2,k0}}{N_{k0}} \right) U_{k00}$$

D'apres (L.) et (1.3), il vient

$$\begin{aligned} D_{0,k_0} + N_{kj} - 1 &= \left(\frac{N_{k0} - D_{2,k_0}}{N_{kk_1}} \right) U_{k_0} \\ &= \left(\frac{N_{k0} - 1 + D_{0,k_0} + D_{1,k_0}}{N_{kj}} \right) U_{k0}; \end{aligned}$$

et en vertu de ([.1]), nous devons avoir $D_{1,kq} = 0$, donc $U_{k_0=N_k}$ alors $D_{0,k_a} \neq 0, D_{1,k_0} = 0$ et $\forall j > k_0, D_{1,j} = D_{0,j} = 0$, (au-delà de k_0, ce qui montre que la condition énonóke est nécessaire .

Montrons que la condition nécessaire est aussi suffisante. Supposons que $D_{1,k_0} = 0, D_{0,k_0} \neq 0$ *et* $\forall j > k_0, D_{1,j} = D_{0,j} = 0$*, et montrons que* $S_n\left(Z'_{ka}\right) = 0$*. Nous avons*

$$\begin{aligned} U_{k_0-1} &= n\left(\frac{N_{k+}-1+D_{0,k_0}}{N_{k_2}}\right)\left(\frac{N_{k_0}}{N_{k_0+1}}\right)\cdots\left(\frac{N_M-1}{N_M}\right) \\ &= N_{k_2-1} + D_{0_1k_21}; \end{aligned}$$

avec $D_{0,k_0} \neq 0$*, autrement dit* $\hat{S}_n\left(Z'_{k_0}\right) = 0$*.*

5.2 Estimation de la fonction de régression

5.2.1 Principe de l'estimation

Lorsque Y est soumise à une censure mixte, nous estimons pour toute fonction $h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, \mathrm{Eh}(X, Y)$ par

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1_{\{A_i=0\}} Z_i}{\hat{S}_n\left(Z_i\right)\hat{F}_n\left(Z_i\right)},$$

où $\hat{S}_n$ et $\hat{F}_n$ sont respectivement les estimateurs de S_R et F_L dont nous rappelons les expressions à la section 5.1 Ceci est motivé par le fait que.

$$E\left(\frac{1_{\{A=0\}}h(X,Z)}{S_R(Z)F_L(Z)}\right) = E\left(E\left(\frac{1_{\{A_i=0\}}h(X,Y)}{S_R(Z)F_L(Z)}\right)/(X,Y)\right)$$
$$= E\left(\frac{h(X,Y)}{S_R(Y)F_L(Y)}E\left(1_{\{A_i=0\}}/(X,Y)\right)\right)$$
$$= E(h(X,Y))$$

car

$$E\left(1_{\{A_1=0\}}/X,Y\right) = S_R(Y)F_L(Y),$$

si les hypothèses $H_{1,1}, H_{1,2}, H_{1,4}$ ci-dessous sont satisfaites. En effet

Pour tout B dans $\sigma(X,Y)$ (tribu engendrée par le couple (X,Y)) il existe un borélien C tel que $B = (X,Y)^{-1}(C)$. L'indépendance de (X,Y) et (L,R) permet d'ecrire

$$\int_B (1_{A=0})\, dP = \int_B (1_{L<Y<R})\, dP$$
$$= \int_{C\times\mathbb{R}_+^2} (1_{l<y<r})\, dP_{(X,Y,L,R)}$$
$$= \int_{C\times\mathbb{R}_+^2} (1_{l<y<r})\, dP_{(X,Y)} \otimes dP_{(L,R)}.$$

Maintenant par le théorème de Fubini et l'indépendance de R et L, nous obtenons

$$\begin{aligned}\int_B (1_{A-0})\, dP &= \int_C \left(\int_{\mathbb{R}_+^2} (1_{l<y<r})\, dP_{(L,R)} \right) dP_{(X,Y)} \\ &= \int_C \left(\int_{\mathbb{R}+} (1_{l<y})\, dP_L \times \int_{\mathbb{R}+} (1_{y<r})\, dP_R \right) dP_{(X,Y)} \\ &= \int_C (F_L(y) S_R(y))\, dP_{(X,Y)} \\ &= \int_B F_L\left((Y_1)\right) S_R\left(Y_1\right) dP\end{aligned}$$

car F est continue. De plus, $F_L(Y)S_R(Y)$ étant clairement mesurable par rapport à $\sigma(X,Y)$, le résultat en découle.

La suite de ce chapitre est vouée à l'étude des estimateurs à poids que nous définirons pour Y dans ce cadre de censure.

Notons pour toute variable aléatoire U, F_U (resp. S_U) sa fonction de répartition (resp. survie, $S_U = 1 - F_U$). $T_U = \sup\{t : F_U(t) < 1\}$ et $I_U = \inf\{t : F_U(t) \neq 0\}$ dénotent les points terminaux du support de U.

Soit H_1 l'hypothèse comprenant les cinq conditions suivantes

$-H_{1,1} : Y, R$ et L sont indépendantes,

$-H_{1,2} : (R, L)$ et (X, Y) sont indépendantes,

$-H_{1,3} : \exists T < T_R$ et $I > I_L$ tels que

$\forall n \in \mathbb{N}, \forall i (1 \leq i \leq n), \quad A_i = 0 \Rightarrow I \leq Z_i \leq T$ a.s.

$-H_{1,4} : F_L$ est continue sur $]0, +\infty[$,

$-H_{1,5} : T_R \vee T_L < \infty$.

Nous aurons aussi besoin de l'hypothèse d'identifiabilité suivante $-H_2 : I_Y \leq I_L < I_R$ et $T_R \leq T_Y$.

Remarquons que puisque $I_L < Z_i < T_R$ dès que $A_i = 0$, l'hy pothèse $H_{1,3}$ semble ne pas être trop restrictive.

Notons $\{W_j, 1 \leq j \leq M\}$ les valeurs distinctes de $\{Z_i, 1 \leq i \leq n\}$, rangées par ordre croissant. Posons $D_{kj} = \sum_{i-1}^{n} 1_{\{Z_1 - W_j, A_1 - k\}}$ et $N_j = \sum_{i=1}^{n} 1_{\{Z_i \leq W_j\}}$. L'estimateur produit-limite de S_R donné dans Patilea et Rolin (2006) est

$$\hat{S}_n(W_j) = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{D_{1l}}{U_{l-1} - N_{l-1}} \right\} \text{ où } U_{j-1} = n \prod_{j \leq l \leq M} \left\{ 1 - \frac{D_{2l}}{N_l} \right\}$$

Patilea et Rolin (2006) ont montré que sous les hypothèses H_2 et $H_{1,1}$

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}^+} \left| \hat{S}_n(t) - S_R(t) \right| = 0. \quad \text{p.s.}$$

Par inversion du temps, nous obtenons l'estimateur inversé de celui de Kaplan-Meier, qui est un estimateur de F_L (censure à gauche) et il est donné par

$$\hat{F}_n(W_j) = \prod_{j \in l \leq M} \left\{ 1 - \frac{1_{\{A_l - 2\}}}{l} \right\}.$$

En adaptant le théorème, de type Glivenko-Cantelli, de Stute et Wang (1993), nous obtenons

$$\lim_{n \to \infty} \sup_{I_L < t} \left| \hat{F}_n(t) - F_L(t) \right| = 0 \quad \text{p.s.},$$

si F_L est continue ce qui est supposé dans l'hypothèse $H_{1,4}$. Nous pouvons déduire de l'hypothèse $H_{1,3}$ que

$$S_R(T) > 0 \text{ et } F_L(I) > 0.$$

En vertu de (3.5), (3.7) et (3.8, nous obtenons pour n suffisamment grand

$$\hat{S}_n(T) > 0 \text{ et } \hat{F}_n(I) > 0 \text{ p.s.}$$

En vertu de (3.2), nous proposons comme estimateurs poids de $r(x)$

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{(A_i - 0)} \frac{Z_i}{\hat{S}_n(Z_i)\, \hat{F}_n(Z_i)} \left(\frac{0}{0} := 0 \right).$$

En appliquant le lemme 1.1 à l'estimateur de S_{RU} on voit que k_0 est le plus petit entier naturel k tel que $\hat{S}_n(Z_k) = 0$ ssi

$$D_{1k_0} \neq 0, D_{0k_0} = 0 \text{ et } \forall j > k_0 D_{1j} = D_{0j} = 0,$$

ce qui implique que $S_n(Z_i) \neq 0$ dès que $A_i = 0$. Nous pouvons aussi remarquer que $\hat{F}_n(Z_i) \neq 0$ dans l'expression de $r_n(x)$ donnée en (3.10) dès que $A_i = 0$.

5.2.2 Convergence presque complète ponctuelle

Avant d'énoncer le théorème qui contient les résultats de la convergence presque complète (p.co), nous avons besoin de quelques hypothèses.

h_1 : Les fonctions de répartition des variables Y, R et L sont continues, $I_Y \leq I_L < I_R$ et $T_R < T_Y$.

$h_{2,1}$: r et f sont ℓ fois continument dérivables dans un voisinage de x.

$h_{2,2} : f(x) > 0$,

$h_{2,3} : \lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} \frac{nh_n}{\log n} = +\infty$.

$h_{2,4}$: K est bornée, intégrable, à support compact et $\int K(t)dt = 1$.

$h_{2,5} : \int t^j K(t)dt = 0, \forall j = 1, \ldots, \ell - 1$ et $0 < \left|\int t^\ell K(t)\right| < \infty$.

Soit S un sous-ensemble compact de $\mathbb{R}$.

$h'_{2,1}$: r et f sont ℓ fois continûment dérivables autour de S

$h'_{2,2} : \exists\theta > 0$, inf $_{x\in S} f(x) > \theta$,

$h'_{2,6} : \exists\beta > 0, \exists C < \infty, \forall x \in S, \forall y \in S : |K(x) - K(y)| \leq C|x - y|^\beta$

Nous terminons la liste des hypothèses par les conditions introduites au section 5.2.1 en vue de traiter notre type de censure.

$H_{1,2}$: (R, L) et (X, Y) sont indépendants.

$H_{1,3}$: $\exists T < T_R$ et $I > I_L$ tels que $\forall n \in \mathbb{N}, \forall i(1 \leq i \leq n), A_i = 0 \Rightarrow I \leq Z_i \leq T$ p.s.

Le théorème suivant donne le taux de convergence presque complète ponctuelle de r_n.

Théorème 5.1. *Sous les hypothèses $h_1, h_{2,1} - h_{2,5}, H_{1,2}$ et $H_{1,3}$,*

nous avons

$$r_n(x) - r(x) = O\left(h_n^{\ell}\right) + O_{a.\infty}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Preuve 5.2. *Posons*

$$\begin{aligned}\hat{r}_n(x) &= \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i - 0\}} \frac{Z_i}{S_R\left(Z_i\right) F_L\left(Z_i\right)} \\ &= \frac{\frac{1}{nh_m}\sum_{i-1}^{n} K\left(\frac{x-X_n}{h_n}\right) \frac{1_{i,A_1=01} Z_i}{S_R(Z_i) F_L(Z_t)}}{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_1}{h_n}\right)} \\ &:= \frac{\hat{r}_{\mathrm{n},N}(x)}{f_n(x)},\end{aligned}$$

où

$$f_{\mathrm{n}}(x) = \frac{1}{nh_{\mathrm{n}}} \sum_{\mathrm{i}=1}^{n} K\left(\frac{x - X_i}{h_{\mathrm{n}}}\right),$$

est l'estimateur à noyau de f, puis effectuons la décomposition suivante

$$r_n(x) - r(x) = r_{\mathrm{n}}(x) - \hat{r}_{\mathrm{n}}(x) + \frac{\hat{r}_{n,N} - g(x)}{f_{\mathrm{n}}(x)} + \frac{f(x) - f_n(x)}{f_n(x)} r(x),$$

où $g = rf$. La démonstration du théorème est alors une conséquence directe des lemmes suivants.

Lemme 5.2. *Sous les hypothèses $h_1, h_{2,3}$ et $H_{1,3}$ nous avons*

$$r_n(x) - \hat{r}_n(x) = O_{p.ox.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Preuve 5.3. *Les définitions de* $r_n(x)$*, de* $\hat{r}_n(x)$ *et l' hypothèse* $H_{1,3}$ *nouspermettent d'écrire*

$$
\begin{aligned}
&\mid r_n(x) - \hat{F}_n(x) \mid \\
&\qquad \leq \left| \sum_{i=1}^{n} W_{n,i}(x) 1_{(A_t - 0)} \frac{Z_i}{\hat{S}_{\mathrm{n}}(Z_i)\, \hat{F}_n(Z_i)} - \sum_{i=1}^{n} W_{n,i}(x) 1_{[A_t - 0)} \frac{Z_i}{S_R(Z_i)\, F_L(Z_i)} \right| \\
&\leq T \sum_{i=1}^{n} W_{n,i}(x) 1_{[A_i - 0\}} \\
&\qquad \times \left| \frac{F_L(Z_i)\, S_R(Z_i) - \hat{F}_n(Z_i)\, S_R(Z_i) + \hat{F}_{\mathrm{n}}(Z_i)\, S_R(Z_i) - \hat{F}_n(Z_i)\, \hat{S}_n(Z_i)}{\hat{F}_{\mathrm{n}}(I) \hat{S}_n(T) F_L(I) S_R(T)} \right| \\
&\leq T \frac{\sup_{t > L_L} \left| \hat{F}_n(t) - F_L(t) \right| + \sup_{t \in \mathbb{R}} \left| \hat{S}_n(t) - S_R(t) \right|}{\hat{F}_n(I) \hat{S}_{\mathrm{n}}(T) F_L(I) S_R(T)} \sum_{i=1}^{n} W_{n,i}(x).
\end{aligned}
$$

En tenant compte de h_1*, nous pouvons appliquer la relation 2.16 et le théorème 1 dans Bitouzé et al. (1999). De plus* $\sum_{i=1}^{n} W_{n,i}(x) = 1$*, nous pouvons donc conclure que*

$$
r_n(x) - \hat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{n}} \right).
$$

De plus, puisque $\lim_{n \to \infty} h_n = 0$ *(hypothèse* $h_{2,3}$ *), nous déduisons que*

$$
r_n(x) - \hat{r}_n(x) = O_{p.\infty}\left(\sqrt{\frac{\log n}{n h_n}} \right).
$$

Lemme 5.3. *Sous les hypothèses* $h_{2,1}, h_{2,4}, h_{2,5}$ *et* $H_{1,2}$ *nous obtenons*

$$
E\hat{r}_{n,N}(x) - g(x) = O\left(h_{\mathrm{n}}^{\ell}\right)
$$

Preuve 5.4. *Puisque* (X_i, Z_i, A_i) *sont i.i.d., nous avons*

$$\begin{aligned} &E\hat{r}_{N,n}(x) - g(x) \\ &\qquad = E\left[\frac{1}{h_n}K\left(\frac{x-X_1}{h_n}\right)E\left(\frac{1_{[A_1-0]}Y_1}{F_L\left(Y_1\right)S_R\left(Y_1\right)}/X_1\right) - g(x)\right] \\ &\qquad = \int \frac{1}{h_n}K\left(\frac{x-u}{h_n}\right)E\left\{\frac{1_{[A_1-0\}}Y_1}{F_L\left(Y_1\right)S_R\left(Y_1\right)}/X_1 = u\right\}f(u)du - g(x). \end{aligned}$$

L'utilisation de la relation (3.3) permet d'écrive

$$\begin{aligned} E\left[\frac{1_{[A_1-0\}}Y_1}{F_L\left(Y_1\right)S_R\left(Y_1\right)}/X_1\right] &= E\left[\frac{Y_1}{F_L\left(Y_1\right)S_R\left(Y_1\right)}E\left(1_{\{A_1-0\}}/\left(X_1,Y_1\right)\right)/X_1\right] \\ &= r\left(X_1\right). \end{aligned}$$

Donc, et puisque $\int K(t)dt = 1$, *nous obtenons*

$$\begin{aligned} E\hat{r}_{N,n}(x) - g(x) &= \int \frac{1}{h_n}K\left(\frac{x-u}{h_n}\right)g(u)du - g(x) \\ &= \int \left(g\left(x - zh_n\right) - g(x)\right)K(z)dz. \end{aligned}$$

Les hypothèses $h_{2,1}, h_{2,4}$ *et* $h_{2,5}$ *et le développement de Taylor au voisinage de* x *permettent d'écrire*

$$E\hat{r}_{N,\mathrm{n}}(x) - g(x) = (-1)^{\ell}h_n^{\ell}\int z^{\ell}K(z)\frac{g^{\ell}(x)dz}{\ell!} + o\left(h^{\ell}\right).$$

Lemme 5.4. *Sous les hypothèses* $h_{2,1}, h_{2,3}, h_{2,4}$ *et* $H_{1,3}$ *nous arons*

$$E\hat{r}_{n,N}(x) - \hat{r}_{n,N}(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Preuve 5.5. *Nous appliquons l'inégalité de Bemstein (corollaire A.1). Pour cela, posons*

$$\Gamma_{\mathrm{i}} = \frac{1_{(A_t-0)} Z_i}{h_{\mathrm{n}}} \frac{K\left(\frac{x-X_i}{h_n}\right)}{S_R\left(Z_i\right) F_L\left(Z_i\right)} \text{ et } U_{\mathrm{i}} = \Gamma_{\mathrm{i}} - E\Gamma_i.$$

Il est clair que $\hat{r}_{N,n}(x) - E\hat{r}_{N,n}(x) = \frac{1}{n}\sum_{i=1}^{n} U_i$. *En vertu des hy pothèses* $h_{2,4}$ *et* $H_{1,3}$, *nous obtenons*

$$|\Gamma_i| \leq \frac{T}{h_{\mathrm{n}} S_R(T) F_L(I)} := M.$$

Nous devons maintenant borner $\sigma^2 = EU_1^2$, *pour cela calculons* $E\Gamma_1^2$.

$$\begin{aligned} ET_1^2 &= \frac{1}{h_n^2} E\left\{ E \frac{1_{[A_1-0\}} Z_1^2}{S_R^2\left(Z_1\right) F_L^2\left(Z_1\right)} K^2\left(\frac{x-X_1}{h_n}\right) / X_1 \right\} \\ &= \frac{1}{h_n} \int \phi\left(x - y h_n\right) f\left(x - y h_n\right) K^2(y) dy, \end{aligned}$$

où $\phi(u) = E\left[\frac{1_{\{A_1=01} Z_1^2}{S S_2(Z_1) F L(Z_1)} / X_1 = u\right]$. *En raison des hypothèses* $h_{2,1}, h_{2,4}$ *et* $H_{1,3}$ *nous concluons qu' il existe une constante* C *telle que* $EU_1^2 \leq \frac{C}{h_n}$, *donc pour* $\varepsilon < \frac{\sigma^2}{\Pi\pi}$, *ilvient*

$$\mathbb{P}\left[|\hat{r}_{N,n}(x) - E\hat{r}_{N,n}(x)| > \varepsilon\right] \leq 2e^{-\frac{nx^2 k_n}{4C}}.$$

En vertu de l'hypothèse $h_{2,3}$, *nous pouvons poser* $\varepsilon = \varepsilon_0 \sqrt{\frac{\log n}{\pi h_m}}$. *Finalement pour* ε_0 *suffisamment grand, le terme de droite de* (4.4) *est le terme général d'une série convergente.*

Lemme 5.5. *Sous les hypothèses $h_{2,1} - h_{2,5}$, nous avons*

$$Ef_n(x) - f(x) = O\left(h_n^{\ell}\right), \quad Ef_n(x) - \hat{f}_n(x) = O_{p-\infty}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

$$et\ \exists \eta > 0, \sum_{n=1}^{\infty} P\left(\hat{f}_n(x) < \eta\right) < \infty.$$

Preuve 5.6. *La preuve des deux premières relations se fait de la même manière que celles des lemmes* 4.2 *et* 4.3 *en utilisant la constante 1 au lieu de la variable* $\frac{1_{\{A_1=0\}}Z_i}{S_n(Z_i)F_L(Z_i)}$. *Il reste à prouver la dernière relation. Remarquons que* $\hat{f}_n(x)$ *converge presque complètement vers* $f(x)$, *donc il suffit de poser* $\eta = \frac{f(x)}{2}$ *et de remarquer que*

$$P\left(\left|\hat{f}_n(x)\right| < \frac{f(x)}{2}\right) \leq P\left(\left|\hat{f}_n(x) - f(x)\right| > \frac{f(x)}{2}\right).$$

5.3 Exercices

Exercice 1.

Etant donné une covariable aléatoire X à valeurs dans $\mathbb{R}^d$, une variable réponse Y positive censurée mixte par deux variables R et L positives et nous disposons seulement d'observations du triplet (X, Z, A) où $Z = \max(\min(Y, R), L)$ et

$$A = 0, \quad si \quad L < Y < R,$$

$$A = 1, \quad si \quad L < R \leq Y,$$

$$A = 2, \quad si \quad \min(Y, R) \leq L.$$

1. Sous les hypothèses H_1, H_2 et H_3 ci-dessous, montrer que

$$E\left[\frac{1_{A=0}h(X,Z)}{S_R(Z)F_L(Z)}\right] = E\left[h(X,Y)\right],$$

où S_R est la fonction de survie de R, F_L est la fonction de répartition de L et h est une fonction mesurable de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$.

2. Proposer un estimateur sans biais de $E\left[\Phi(X,Y)\right]$.

H_1 : Y ,R et L sont indépendantes,

H_2 : (R, L) et (X, Y) sont indépendantes,

H_3 : F_L est continue sur $]0, +\infty[$.

Exercice 2.

Montrer que l'estimateur de Patilea et Rolin généralise l'estimateur de Kaplan-Meier.

Exercice 3.

Trouver une condition nécessaire et suffisante pour que l'estimateur de Patilea et Rolin s'annule.

Bibliographie

Bitouzé, D., Laurent, B., and Massart, P. (1999). A dvoretzky-kiefer-wolfowitz type inequality for the Kaplan-Meier estimator. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*. 35, 735–763.

Boukeloua, M (2012/2013) Etude des estimateurs de la fonction de répartition et de la densité dans un modèle de censure. *Université Frères Mentouri Constantine*. (Mémoires de Master).

Carbonez, A., L. Gyorfi, and E. C. Vander Meulen. (1995). Partitioning estimates of a regression function under random censoring. *Statistics and Decision*. 13, 21–7.

Foldes A., Rejto L. and Winter B.B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data II : the estimation of density and failure rate. *Period. Math. Hungar*. 12, 15–29.

Guerras, Z. and Boulamnekher, A. (2017/2018) Méthodes para-

métriques pour l'analyse de survie. *Université Frères Mentouri Constantine*. (Mémoires de Master).

Herbert Leiderma, P., Babu, B., Kagia, J., Kraemer, H. C. and Leiderman, G. F. (1973). African infant precocity and some social influences during the first year. *Nature*, 242, 247-249. *Computational Statistics & Data Analysis*, 69, 154–172.

Kaplan, E. M., and P. Meier. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 53(282) :457–81. doi :10.2307/2281868.

Kebabi, K., and Messaci, F., (2012) Rate of the almost complete convergence of a kernel regression estimate with twice censored data. *Statistics and Probability Letters*, 82(11), 1908-1913.

Leiderman, G. F. (1973). African infant precocity and some social influences during the first year. *Nature*, 242, 247–249.

Nadaraya, E. A. (1964). On estimating regression. *Theor. Probab. Appl.*. 9. 157–159.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhya A* , 26, 359–372.

Patilea, V and Rolin, J. M. (2006). Product-limit estimators of the survival function with twice censored data. *Ann. Statist.* 34(2), 925–938.

Parzen E. (1962). On estimating of a probability density and mode. *Annals of Mathematicla Statistics*, 33, 1065–1076. *Computational Statistics & Data Analysis*, 69, 154–172.

Rosenblatt M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832–837.

Shorack R. G. and Wellner W. J. (1986) . *Empirical processes with applications to statistics. John Wiley Sons.*