MINISTERE DE L'ENSEIGNEMENT SUPERIEUR,
DE LA RECHERCHE SCIENTIFIQUE ET DE
L'INNOVATION
****
UNIVERSITE NAZI BONI
****
UNITE DE FORMATION ET DE RECHERCHE EN
SCIENCES ET TECHNIQUES
******
DEPARTEMENT DE MATHEMATIQUES ET INFORMATIQUE

MEMOIRE

pour obtenir le diplôme de
Master de Mathématiques Appliquées de l'Université Nazi BONI
Option: **Modélisations et Calculs Scientifiques**

Soutenu publiquement le 16 Novembre 2019 par:

OUEDRAOGO Yacouba

Analyse mathématique de la sensibilité de l'intervalle de confiance de Kaplan Meier par rapport à la médiane de suivi

Spécialité: **Statistique Appliquée**

Directeur de mémoire : **Dr Boureima SANGARÉ**
Co-directeur du mémoire : **Dr Serge M. A. SOMDA**

Membres du Jury :

M. Sado TRAORÉ,	Professeur titulaire	UNB (Président)
M. Boureima SANGARÉ,	Maître de conférences	UNB (Directeur)
M. Serge M. A. SOMDA,	Maître assistant	UNB (Co-directeur)

Dédicace

■ *À mon père et à ma mère.*
À mes frères Oumarou,
Mamadou.
À ma grande sœur chérie ■

Remerciements

Je remercie DIEU, le très Miséricordieux de m'avoir donné la force, la santé et le courage pour terminer ce mémoire de fin de cycle. La réalisation de ce document a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner ma reconnaissance.

Je remercie tout d'abord, mon directeur de mémoire, le Docteur Boureima SANGARÉ, pour la confiance qu'il m'a accordé, pour ses qualités pédagogiques et scientifiques. Merci pour tous vos conseils et vos remarques tout au long de ce mémoire. J'ai été honoré de travailler sous votre direction. Vous aviez dédié beaucoup de votre temps à discuter avec moi à propos de mon avenir et je vous en suis très reconnaissant. Trouvez ici l'expression d'une grande reconnaissance pour tout ce que vous avez fait pour moi. J'espère seulement un jour vous montrer ma reconnaissance.

Merci à mon Co-directeur de mémoire, le Docteur Serge M. A. SOMDA, pour la qualité de votre encadrement et votre bienveillance. Ce mémoire n'aurait pas pu voir le jour sans la grande confiance que vous m'avez accordé. Merci pour le grand intérêt que vous avez porté à ce document. Je crois que les multiples discussions avec vous m'ont permis de voir la recherche autrement, et m'ont donné un véritable goût pour la recherche. J'ai été initié à un domaine (bio-statistique) plein d'intérêt. Je garde espoir, que ce n'est que le début d'une véritable aventure scientifique.

Je remercie le Professeur Sado TRAORÉ, pour avoir accepté présider cette soutenance.

Merci au Dr Satafa SANOGO pour les lectures et les critiques qui ont permis d'améliorer le document.

Aux Enseignants de l'UFR/ST, notamment : Théodore Marie Yves TAPSOBA, Joseph BAYARA, Idrissa KABORÉ, Jean De Dieu ZABSONRÉ, Aboudramane GUIRO, Adama OUEDRAOGO, Abdoulaye SERÉ, Ismaël NYANKINI, Jean Louis ZERBO, Herman SORÉ, Télesphore TIENDREBÉOGO, je dis merci pour les enseignements et les conseils qui m'ont été d'un grand interêt.

Merci à tous mes promotionnaires de Master pour les bons moments passés ensembles et les soutiens mutuels durant les périodes déprimantes de ce cursus Universitaire.

Eric DABONÉ, je tiens à te remercier pour tes multiples conseils et idées que tu as consenti à la création de ce document.

Mr Ibrahima DIALLO, trouvez ici l'expression d'une grande reconnaissance, votre ri-

gueur et gentillesse est à désirer. Merci pour tout.

Merci à Son Noël pour la correction du document.

La vie au Centre Muraz (Centre de calcul) n'aurait pas été aussi agréable sans la présence des étudiants stagiaires et de l'équipe Appui Méthodologique, pour la collaboration lors de mon stage. Nous avons partagé nos difficultés autour d'un café ! Ballo Cheick, Bouda Luc, Gomina Honorine, Kafando Moussa, Kiemdé Christelle, Nikiema Zoulkarnayni, Paré Toudala, Traoré Issouf, Yara Mimbouré, Mme Florence Kangoyé, Ouattara Filomène : merci pour tous ces moments de détente, pour les répétitions, les formations au Centre de Calcul.

Enfin, j'arrive à ma famille, merci maman, papa pour votre amour, vos bénédictions quotidiennes et vos multiples sacrifices que vous faites pour vos enfants. Oumarou, on pourra tout dire, mais ta gentillesse est sans limite, elle m'a permis d'être à ce niveau. Grande sœur Fati, merci pour ton amour inconditionnel, je t'aime. Mamadou, quand je crois avoir perdu tout espoir, il me suffit de t'appeler, et ma vie devient plein d'espoir. J'ai de la chance de vous avoir comme ma famille.

L'existence est faite de pages que l'on tourne. Chacune de ces pages est un chapitre du livre de l'histoire de notre vie. Parfois le destin, ou le hasard de la vie, fait que l'on doit quitter des personnes chères. Des personnelles essentielles, des personnes qui nous ont beaucoup appris. Ce message d'au revoir est l'expression de mon émotion profonde. Adieu ***John Emmanuel Marie SAWADOGO****, mon cher ami.*

Résumé

L'estimateur de Kaplan-Meier de la fonction de survie est un indicateur très largement utilisé dans les études de durée de vie, notamment en recherche clinique. La précision de cet estimateur pose cependant problème en présence de censure lourde. La médiane de suivi est souvent utilisée pour juger de la qualité du suivi. Toutefois, il n'existe pas de règle de décision claire pour apprécier la qualité des estimations produites en fonction de cette médiane. L'objectif de cette étude est d'évaluer la qualité de l'estimation du taux de survie à une date t selon les différentes valeurs de la médiane de suivi afin de proposer une règle de décision objective. Une analyse de la sensibilité de l'intervalle de confiance de Kaplan-Meier en fonction de la médiane de suivi a été effectuée. La couverture de l'intervalle de confiance et sa précision étaient les critères de jugement. Une simulation par Monte Carlo a été réalisée à l'aide du logiciel R. Notre étude a permis d'observer le fort impact de la médiane de suivi sur la qualité de l'estimateur de Kaplan-Meier. Pour les faibles valeurs de la médiane, le taux de couverture de l'intervalle de confiance observé est très bas, il croît avec la médiane jusqu'à atteindre un plateau qui correspond au niveau de confiance voulu. Le plateau correspond au niveau à partir duquel l'estimateur est de qualité. Après plusieurs réplications, nous avons trouvé que pour estimer la survie après un temps t dans une cohorte, il fallait un suivi médian d'au moins deux-tiers du temps. Cette étude donne des éléments factuels et objectifs pour évaluer la qualité des suivis des cohortes, chose très répandue dans la recherche en santé.

Mots clés : Médiane de suivi, censure à droite, Kaplan-Meier, intervalle de confiance, Simulation par Monte Carlo, Analyse de sensibilité.

Abstract

Kaplan-Meier's estimate of the survival function is a very widely used indicator in life-time studies, especially in clinical research. The accuracy of this estimator, however, is problematic in the presence of heavy censorship. Median tracking is often used to judge the quality of follow-up. However, there is no clear rule of decision to assess the quality of estimates produced based on this median. The objective of this study is to evaluate the quality of the estimate of the survival rate at a date t according to the different values of the median of follow-up in order to propose an objective decision rule. A sensitivity analysis of the Kaplan-Meier confidence interval as a function of the median follow-up was performed. The coverage of the confidence interval and its accuracy were the judgment criteria. A Monte Carlo simulation was performed using R software. Our study has observed the strong impact of the median of follow-up on the quality of the Kaplan-Meier estimator. For low values of the median (heavy censorship), the coverage rate of the observed confidence interval is very low. He is growing with the median until reaching a plateau that corresponds to the desired level of confidence. The plateau corresponds to the level from which the estimator is of quality. After several replications, we found that to estimate survival after a time t in a cohort, a median follow-up of at least two-thirds of the time was required. This study provides factual and objective elements for assessing the quality of cohort follow-ups, a very common topic in health research.

Keywords: Median follow-up, right censoring, Kaplan-Meier, interval of confidence, Simulation, Sensitivity analysis.

Table des matières

Liste des figures

Chapitre

1 Introduction générale

■ *Si vous n'échouez pas de temps à autre, c'est que vous ne faites rien de très innovant* ■

Woody Allen

Sommaire

1.1 Contexte et justification du sujet de mémoire

Les études de cohorte consistent à observer la survenue d'évènement d'intérêt dans le temps au sein d'une population définie. Ces études sont beaucoup utilisées pour leurs évidences en santé et leur niveau de preuve élevé ; mais elles sont aussi utilisées dans d'autres disciplines autre que la santé (par exemple : la sociologie, la finance, en physique,..). Dans les sciences médicales par exemple, elles consistent à recruter des sujets indemnes d'une pathologie d'intérêt et de les suivre au cours du temps pour identifier au niveau individuel la survenue de cette pathologie. En épidémiologie, la survie des patients est utilisée dans le cadre de la surveillance des maladies au niveau d'une population, mais aussi en recherche clinique pour l'évaluation des stratégies thérapeutiques ainsi que pour l'identification de facteurs pronostiques et la quantification de leurs effets. Des modèles peuvent être construits pour essayer de mieux comprendre le développement de certaines maladies ou d'évaluer l'efficacité d'un traitement face à une maladie. C'est pourquoi, les institutions de santé exigent généralement un programme d'enregistrement et de suivi des cas de maladie, pour qu'une stratégie de lutte soit approuvée. Les études de cohortes ont certes beaucoup d'avantages, cependant, elles ont des limites et leur mise en œuvre n'est pas sans difficultés diverses. En effet, pendant les phases

d'observations des participants, toutes les dépenses de santé sont à la charge des commanditaires de l'étude ; ce qui rend ces études très coûteuses sans oublier les perdus de vue et les évènements intermittents dont la qualité du suivi y dépend fortement. De ce fait, la nécessité de mettre en place un critère d'arrêt précoce d'une étude ou des critères pour juger la qualité des estimations du taux de survie produite dans le cas des études de longue durée est plus que nécessaire, au vu des coûts énormes que nécessitent ces études, mais aussi pour la prise de décision concernant l'utilisation de nouvelles interventions. Il est vrai que l'estimateur de la fonction de survie le plus utilisé est l'estimateur de Kaplan-Meier, cependant, plusieurs auteurs ont discuté de la faiblesse de l'estimateur de la variance de Kaplan-Meier pour les données fortement censurées. Des auteurs recommandent dans ce sens, en présentant des résultats d'analyse de survie, de donner des indications sur le suivi en précisant la médiane de suivi [31]. La médiane de suivi est très souvent utilisée comme indicateur de la qualité du suivi. Plus elle est élevée, plus le suivi sera considéré comme bon et plus les valeurs présentées seront fiables. Cependant la littérature ne présente pas de règle de décisions précises sur quelle valeur seuil de la médiane considérée.

1.2 Etat de l'art de l'analyse de survie

L'analyse de survie est une méthode permettant de synthétiser des probabilités de survenue d'un événement chez les sujets ayant en commun un événement d'origine, en tenant compte du délai écoulé entre ces deux événements [3]. L'analyse de survie est très souvent utilisée dans de nombreux domaines allant des sciences de la santé, les sciences biologiques, l'industrie, les assurances, l'astronomie, la sociologie, etc. L'objet de cette méthode est généralement une cohorte d'individus. Ils sont suivis jusqu'à l'apparition de l'événement d'intérêt. Cependant, il peut arriver que certains individus soient suivis sans que jamais l'observateur n'enregistre cet événement d'intérêt. On parle dans ce cas de censure ou d'individus censurés. Pendant longtemps, les auteurs ont proposé de simplement retirer ces individus pour lesquels l'information était incomplète, limitant l'analyse aux individus ayant complété l'étude. Cette approche introduit cependant un biais important. Des approches ont donc été proposées pour utiliser à bon escient l'information partielle disponible chez les sujets censurés.

En 1912, une première méthode fut introduite par Böhmer [6], la méthode actuarielle. Elle fait le bilan des occurrences de survenue de l'événement étudié à intervalles fixes en tenant compte des données censurées. En 1958, Edward Kaplan et Paul Meier, ont proposé un estimateur qui prend en compte toute l'information collectée dont les durées de survie des individus censurés avant leur perte de vue [16]. La fonction de survie de cet estimateur est une courbe en escalier décroissante qui indique le taux de survie dans le temps. La médiane de survie est obtenue à partir de cette courbe, c'est la date à laquelle le taux survie est de 50%. L'article décrivant l'estimateur de Kaplan Meier (KM)

[16] est le plus cité dans la littérature scientifique. C'est la méthode la plus courante utilisée dans l'analyse des données de survie et lorsqu'il s'agit de faire une comparaison entre deux groupes de participants [10]. Cela dénote l'importance de cette approche pour l'application des statistiques. Cependant, il existe des questions ouvertes sur la précision de celle-ci. La première variance de cet estimateur proposée par Kaplan et Meier à partir de la formule de Greenwood [13] découle d'une approximation de Taylor. Des intervalles de confiance ont été proposés à partir de cette formule.

Une approche simple consiste à proposer un intervalle de confiance identique à celle d'une espérance d'une distribution normale. Cette approche est confortée en théorie par la convergence en loi de la moyenne empirique de toute distribution vers la loi normale centrée et réduite (théorème de la limite centrale). Cependant, elle pose problème parce qu'en l'appliquant, on retrouve souvent des bornes négatives ou supérieures à l'unité, ce qui n'est pas convenable pour une probabilité. En 1990, Borgan et Liestol [7] ont proposé une approche identique en utilisant une transformation par le logarithme de l'indicateur. Meeker et Escobar [20], eux, proposaient en 1998, la transformation par la fonction logit. D'autres méthodes se fondant sur les transformations de la variance de Greenwood sont disponibles dans la littérature [29]. La plus utilisée est la transformation recommandée par Kalbfleish et Prentice [15]. Celle-ci propose une transformation $log(-log)$.

D'autres approches de calcul de la précision de l'estimateur de Kaplan-Meier sont également proposées dans la littérature [29]. Cutler et Ederer [9], en 1958, ont proposé de considérer le taux de survie comme une distribution binomiale en adaptant à chaque fois la taille de l'échantillon. Ils parlent de "taille réelle de l'échantillon (*the effective sample size*)". Peto et al, en 1977, [22] ont proposé une méthode alternative en utilisant une autre taille d'échantillon différente de celle de Cutler et Ederer. L'intervalle de confiance de Peto n'est généralement pas approprié puisque les bornes inférieures et supérieures peuvent être respectivement inférieures à 0 et supérieures à 1. En 1978, Rothman [25] propose un intervalle de confiance assez complexe basé sur la taille réelle de l'échantillon de Cutler et la variance de Greenwood pour résoudre le problème d'asymétrie de l'intervalle de confiance basé sur la variance de Greenwood, mais cet intervalle de confiance manque de précision lorsque le nombre d'individus à risque devient faible [25]. Enfin la littérature [8, 11, 30] propose des approches alternatives, notamment en utilisant des lois dites exactes. Une large évaluation de ces différentes méthodes a été proposée par Sorgho [29]. La qualité d'un intervalle de confiance de type Greenwood dépend fortement du taux de perdus de vue. En effet, si le rythme des censures est lent, nous serons très proches de la distribution binomiale et les estimations seront de qualité. Par contre, si la censure est lourde et très rapide, le dénominateur de la variance se retrouve très vite faussé, ce qui pourrait biaiser l'estimation de la variance. Il est alors courant de voir les chercheurs évaluer la qualité de statistiques produites par le rythme de perdus de vue. Un indicateur classique de ce rythme de perdus de vue est la médiane de suivi. La médiane de suivi est différente de la médiane de survie décrite plus haut. Il s'agit de la

date à laquelle 50% des sujets suivis sont censurés (*Méthode de Kaplan-Meier Inversé*) [26]. Dans la littérature, le suivi médian décrit la maturité des données, la stabilité de la courbe de Kaplan-Meier et la qualité du suivi. Cependant, certains auteurs ne partagent pas le même avis [27]. En 2002, Pocock et al. [24] réclamaient que le suivi médian soit cité dans les présentations cliniques de sorte à informer les lecteurs de l'étendue du suivi. En 2007 plusieurs auteurs [31] ont insisté sur l'utilité de donner en présence de résultats d'analyse de survie la médiane de suivi, pour exprimer la qualité des études mener.

1.3 Présentation du travail

1.3.1 Le travail scientifique

La médiane de suivi devient de plus en plus utilisée comme critère de jugement de la qualité de l'estimation de Kaplan-Meier. Cependant, on ne trouve pas, dans la littérature, des règles de décision pour évaluer cette qualité. Dans ce sujet de recherche, nous proposons une approche basée sur la simulation numérique et l'analyse de sensibilité pour mesurer la précision de l'estimateur de Kaplan Meier à la médiane de suivi. Il s'agit de trouver une relation entre la qualité de l'intervalle de confiance et la médiane de suivi. Des recommandations claires et objectives serons faites. Pour atteindre cet objectif, une analyse de sensibilité de l'intervalle de confiance en fonction de la médiane de suivi sera faite en utilisant les techniques de simulation par Monte Carlo. En d'autres termes, nous allons :

- étudier la longueur (la précision) de l'intervalle de confiance par rapport aux différentes valeurs de la médiane de suivi,
- étudier le taux de couverture (la stabilité) de l'intervalle de confiance par rapport aux différentes valeurs de la médiane de suivi,
- étudier la dispersion du taux de suivi autour d'un point, en fonction des différentes valeurs de la médiane de suivi.

1.3.2 Le contenu du mémoire

Pour une meilleure compréhension, le document sera structuré de la manière suivante.

Après ce chapitre introductif dans lequel nous avons présenté le contexte général et les objectifs de notre étude, nous ferons un tour d'horizon des définitions et concepts de base liés à la notion d'analyse de survie dans le chapitre 2. Nous exposons par ailleurs, les notions préliminaires d'analyse de survie que nous utiliserons par la suite. Nous définissons en outre différents concepts de l'analyse de survie, parmi lesquels, la notion de censure, les principales fonctions et distributions en analyse de survie. Une historique de l'estimateur de Kaplan-Meier et de la médiane de suivi ont été proposés afin de

mieux comprendre le choix et l'importance des estimateurs que nous avons utilisé dans nos travaux.

Pour ce qui concerne le chapitre 3, il marque le début de notre méthode de simulation tout en se basant sur des modèles de Monte Carlo ainsi qu'à leurs analyses mathématiques. Ce chapitre présente une brève description de la méthode de simulation de Monte Carlo, notre stratégie de simulation ainsi que les critères de comparaison que nous avons adopté au cours de cette étude.

Le chapitre 4 est consacré aux résultats de la simulation numérique avec l'introduction des schémas numériques et à l'interprétation des différents résultats numériques obtenus, suivit d'une discussion et de quelques recommandations. Nous proposons par exemple notre propre critère de jugement de la qualité de l'estimation du taux de survie à une date t en se servant de la médiane de suivi.

Enfin, le dernier chapitre sera consacré à la conclusion et aux perspectives de recherche.

Chapitre

2 Introduction à l'analyse de survie

Sommaire

L'analyse de survie s'intéresse au délai d'apparition d'un événement au cours du temps. Il constitue un domaine de la statistique qui s'intéresse à mesurer le temps jusqu'à un événement particulier (souvent appelé temps d'échec, ou temps de survie). Les applications de telles analyses sont multiples, nous pouvons citer comme exemples de temps d'échec : la durée de fonctionnement de pièces avant une défaillance en fiabilité industrielle, la durée de grèves ou de périodes de non-emploi en économie, la durée de vie de patients lors d'essais cliniques. Ce type d'analyse est particulièrement utile dans la recherche biomédicale. Des modèles peuvent être construits pour essayer de mieux comprendre le développement de certaines maladies où d'évaluer l'efficacité de divers traitements ou la résistance du patient face à une maladie.

La première méthode d'analyse de survie, la méthode actuarielle, est apparue en 1912 [6]. Elle est utilisée dans le domaine médical pour la première fois en 1950 [5]. En 1958, Kaplan et Meier [16] présentent d'importants résultats concernant l'estimation non paramétrique de la fonction de survie, ils étudient l'espérance, la variance et les propriétés asymptotiques. Mantel (1966) [18], a étudié la statistique du log-rank pour comparer deux distributions de survie.

2.1 Définitions des concepts de base

Pour mener à bien une analyse de survie d'une population, il est nécessaire de connaître les définitions suivantes :

- *Cohorte* : ensemble de sujets inclus dans une étude au même moment, et suivis dans des conditions standardisées pendant une durée prédéfinie.

- *Événement d'intérêt* : événement auquel on s'intéresse au cours de l'étude. Par exemple : décès lié à un AVC (Accident Vasculaire Cérébral), complication, rechute, disparition de symptômes. On utilisera l'analyse de survie dès qu'il y aura une notion de durée jusqu'à la survenue de l'évènement d'intérêt.

- *Durée de survie* : délai entre la date d'origine et la date de survenue de l'événement.

- *Date d'origine* : c'est la date correspondante au point de départ de la surveillance. Elle peut être différente pour chaque sujet selon les modalités d'inclusion du sujet. Dans certains cas, la date d'origine peut être antérieure à l'inclusion dans l'étude. On parle alors de cohorte historique.

- *Date de point* : c'est la date choisie pour faire le bilan.

- *Date des dernières nouvelles* : c'est la date la plus récente à laquelle on a recueilli des informations sur le patient, notamment la survenue ou non de l'événement d'intérêt.

- *Un perdu de vue* : un sujet est perdu de vue lorsque sa surveillance est interrompue avant la date de point et que l'événement d'intérêt ne s'est pas produit.

- *Temps de recul* : délai entre la date d'origine et la date de point, c'est-à-dire le délai maximum potentiel de suivi pour un sujet. Les reculs minimum et maximum d'une série de sujets définissent donc l'ancienneté de cette série.

- *Temps de participation* : durée de surveillance pour chaque sujet utilisée dans l'estimation de la survie.

2.2 Censure et Troncature

2.2.1 Notion de censure

En analyse de survie, la censure est une condition dans laquelle la valeur d'une mesure ou d'une observation n'est que partiellement connue. En d'autres termes tous les individus qui n'ont pas connu d'*événement* au cours d'une période donnée sont considérés comme des observations censurées. Plusieurs types de censure existent et nous nous limiterons, dans le cadre de ce mémoire, à la censure dite *censure à droite*. Cependant, nous donnerons une définition de la censure à gauche et de la censure par intervalle. Nous reprenons quelques définitions tels que données par Philippe Saint Pierre [23]

2.2.1.1 Censure à droite et censure à gauche

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

- *Censure de Type I (censure fixe)*
  Soit C une valeur fixée, au lieu d'observer les variables $T_1, ..., T_n$ qui nous intéressent, on observe T_i uniquement lorsque $T_i \leq C$, sinon on sait uniquement que $T_i > C$: On utilise la notation suivante : $X_i = T_i \wedge C = \min_{1 \leq i \leq n} (T_i, C)$
- *Censure de type II (censure en attente)*
  Elle est présente quand on décide d'observer les durées de survie des n individus jusqu'à ce que k d'entre eux aient subi l'événement et d'arrêter l'étude à ce moment là.
- *Censure de Type III (censure aléatoire)*
  Soit C_i une variable aléatoire positive ou nulle représentant le temps de censure pour l'individu i et T_i sa durée de survie. La durée T_i est dite censurée à droite si $C_i < Ti$. Le temps d'observation de l'individu i est alors $X_i = T_i \wedge C_i = \min_{1 \leq i \leq n} (T_i, C_i)$ et son statut est donné par $\delta_i := 1_{(Ti \leqslant Ci)}$ qui vaut 1 si l'individu i a connu l'évènement et 0 s'il est censuré.
  La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par :
  - ✓ la perte de vue : le patient quitte l'étude en cours et on ne le revoit plus. Ce sont des patients perdus de vue.
  - ✓ l'arrêt ou le changement du traitement : les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.

✓ la fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients exclu-vivants. Les perdus de vue (et les exclusions) et les exclu-vivants correspondent à des observations censurées mais les deux mécanismes sont de natures différentes (la censure peut être informative chez les perdus de vue).

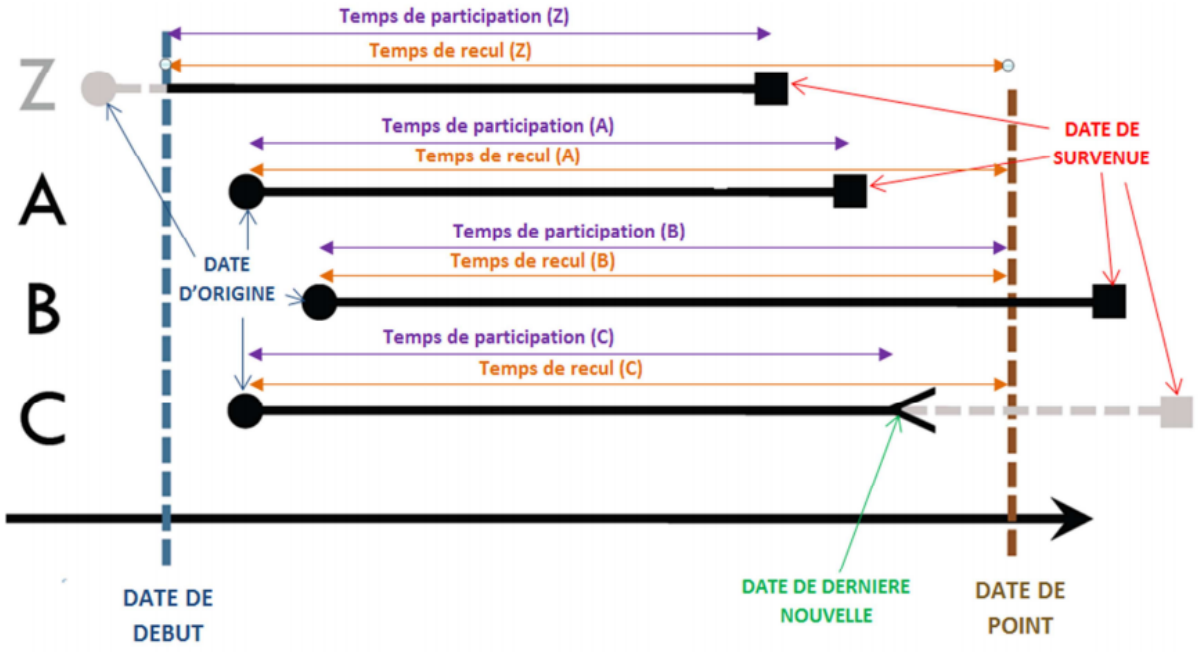


FIGURE 2.1 – Exemple de censure aléatoire

⇒ Le patient A est suivi entièrement durant la période d'étude.

⇒ Le patient B est encore vivant après la date de point. Son décès surviendra peut être plus tard, mais hors de la période d'étude.

⇒ Le patient C n'est pas suivi durant toute la période d'étude. Il était vivant jusqu'au moment où on l'a perdu de vue (*date de dernière nouvelle*).

⇒ Le patient Z, sa maladie est apparue avant la date du début de l'étude (*cohorte historique*).

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires $(X, \delta) : X = T \vee C = \max(T, C)\,, \delta = 1_{T \geq C}$. Comme pour la censure à droite, on suppose que la censure C est indépendante de T.

2.2.1.2 Censure par intervalle

Un individu est censuré par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues.

Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'événement s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à de la censure à droite.

2.2.2 Notion de troncature

Une observation est dite tronquée si elle est observée conditionnellement à un événement. Seuls les individus qui n'ont pas présenté l'événement étudié à la date d'inclusion dans l'étude sont retenus pour celle-ci. On parle généralement de troncature à gauche. Autrement dit, un individu qui avait déjà connu l'événement étudié ne fait plus partie de cette étude. En d'autres termes, une variable aléatoire y est dit être tronqué à gauche si, pour une valeur de seuil c, la valeur exacte de y est connu pour tous les cas $y > c$, mais inconnu pour tous les cas $y \leq c$. De même, la troncature à droite signifie que la valeur exacte de y est connu dans les cas où $y < c$, mais inconnu quand $y \geq c$.

Remarque 2.2.1. *La phénomène de troncature est très différent de la censure :*

- *dans le cas de la troncature, on perd complètement l'information sur les observations en dehors de la plage d'observation : les systèmes ne sont pas observables, donc même pas répertoriés. La troncature élimine de l'étude une partie des systèmes.*
- *dans le cas de la censure, on a connaissance du fait qu'il existe une information, mais on ne connaît pas sa valeur précise, simplement le fait qu'elle excède un seuil ; dans le cas de la troncature on ne dispose pas de cette information.*

2.3 Principales fonctions en analyse de survie

Supposons que la durée de survie T soit une variable aléatoire positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes.

2.3.1 Fonction de survie, fonction de répartition et fonction densité de probabilité

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

$$S(t) = P(T > t),\ t \geq 0. \tag{2.1}$$

C'est une fonction monotone, décroissante et continue vérifiant :

$$S(0) = 1 \; et \; \lim_{t\to\infty} S(t) = 0.$$

La fonction de répartition F fait correspondre à un temps t, la probabilité de décéder avant le temps t :

$$F(t) = P(T \leq t) = 1 - S(t). \tag{2.2}$$

La densité de probabilité f, est une fonction $f(f \geq 0)$ qui vérifie :

$$\forall t \geq 0, \; F(t) = \int_0^t f(u)du \tag{2.3}$$

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

$$f(t) = F'(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{2.4}$$

Elle représente la probabilité instantanée de décéder dans un petit intervalle de temps après t. En tout point t de continuité de f, on a F dérivable en t et

$$f(t) = F'(t) = -S'(t).$$

2.3.2 Risque instantané λ et le taux de risque cumulé h

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

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t} \frac{P(t < T < t + \Delta t)}{P(T > t)} \\ &= \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt} \ln(S(t)). \end{aligned}$$

Elle représente la probabilité de décéder à la date t. *Le taux de risque cumulé* est l'intégrale du risque instantané :

$$h(t) = \int_0^t \lambda(u)du = -\ln(S(t)). \tag{2.5}$$

La fonction de survie s'exprime donc en fonction du taux de mortalité cumulé (ou du taux instantané) par la relation suivante (d'après la fonction 2.5) :

$$S(t) = \exp(-h(t)) = \exp\left(-\int_0^t \lambda(u)du\right). \tag{2.6}$$

2.4 Quantité associée à la distribution de survie

2.4.1 Moyenne et Variance de la durée de survie

En raison de censure et de troncature, la plupart des études de survie ne permettent pas d'obtenir une bonne estimation du temps de la survie moyenne, raison pour laquelle cet estimateur est presque ignoré en analyse de survie au profit de la médiane de survie qui lui constitue un très bon indicateur. Le temps moyen de survie est naturellement définit par

$$E(X) = \int_0^\infty tf(t)dt \tag{2.7}$$

La variance de la durée de survie $V(X)$ est définie par :

$$V(X) = \int_0^\infty t^2 f(t)dt - (E(X))^2 \tag{2.8}$$

En remplaçant $f(t)$ par $-S^{'}(t)$ et en faisant une intégration par partie, on obtient :

$$V(X) = \left[-t^2 S(t)\right]_0^{+\infty} + 2\int_0^\infty tS(t)dt - (E(X))^2 \tag{2.9}$$

$$V(X) = 2\int_0^\infty tS(t)dt - (E(X))^2 \tag{2.10}$$

Ainsi, on peut déduire l'espérance et la variance à partir de n'importe laquelle des fonctions F, S, f, h, λ.

2.4.2 Quantiles de la durée de survie

Définition 2.4.1. *Soit $p \in]0,1[$. Le quantile associé à p, noté t_p est une fonction q donnée par*

$$t_p \equiv q(p) = \inf\{t : F(t) \geq p\} \tag{2.11}$$

$$= \inf\{t : S(t) \leq 1-p\} \tag{2.12}$$

Quand F est strictement croissante et continue alors elle est bijective et $\forall p \in]0,1[$, on a :

$$t_p \equiv q(p) = F^{-1}(p), \tag{2.13}$$

$$= S^{-1}(1-p) \tag{2.14}$$

Pour p fixé, le quantile t_p est le temps auquel une proportion p de la population à disparu.

La médiane de la durée de survie est le temps t pour lequel la probabilité de survivre est égale à 0.5. Dans le cas d'une fonction en escalier, généralement la courbe de Kaplan-Meier, il est difficile de trouver la date t exacte pour laquelle la probabilité de ne pas présenter l'événement soit égale à $\frac{1}{2}$. La médiane est donc définie par,

$$M = \inf\{t : S(t) = 0.5,\ (t \geq 0)\}$$

Cependant, il est possible d'obtenir un intervalle de confiance du temps médian avec un risque α.

2.5 Estimateur de Kaplan-Meier

Une analyse de survie a pour principe d'estimer la probabilité de survie à différents intervalles de temps. On appelle fonction de survie S_i à un instant t_i donné, la probabilité d'être vivant à cet instant

$$S(t_i) =: \frac{Nombre\ de\ survivant\ \text{à}\ l'instant\ t_i}{Nombre\ d'individu\ suivi} \tag{2.15}$$

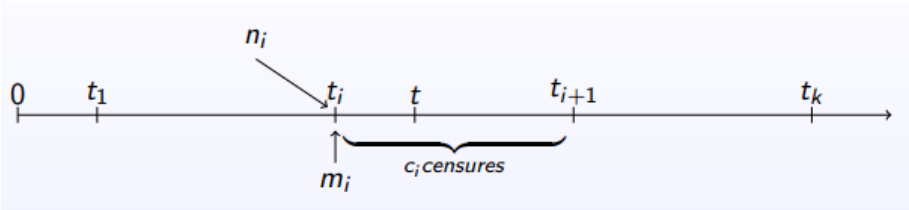
Cet indicateur serait aisé à calculer si on connaissait le statut vivant/décédé de tous les sujets au moment de chaque mesure. En pratique les données sont le plus souvent incomplètes. La qualité des données est le problème le plus important dans l'analyse des données, dans lequel on peut inclure la pertinence des informations collectées [4].
Kaplan et Meier ont proposé en 1958 une méthode non paramétrique d'estimation de la probabilité de survie après une durée t en prenant en compte toute l'information disponible même chez les individus censurés.

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

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= P(T \geq t | T \geq t-1) P(T \geq t-1) \\ &= P(T \geq t | T \geq t-1) \times \cdots \times P(T \geq 1 | T \geq 0) P(T \geq 0) \\ &= Q_t \times Q_{t-1} \times \cdots \times Q_1 \times 1 \end{aligned} \tag{2.16}$$

avec $Q_t = P(T \geq t | T \geq t-1)$

Soit $X_i = \min((T_i, C_i), \delta_i)_{1\leq i\leq n}$ un échantillon de durée censurée à droite. Notons $(t_i)_{1\leq i\leq k}$ les k temps distincts de réalisations d'événements observés ordonnés.
$t_1 < t_2 < \cdots < t_k$.



- n_i nombre de sujets à risque en t_i (juste avant t_i),
- m_i nombre de sujets décédés entre t_i et t_{i+1},
- c_i nombre de sujets censurés entre t_i et t_{i+1},
- q_{t_i} : probabilité d'être en vie en t_{i+1} sachant qu'on était en vie en t_i. C'est la proportion de survivants en t_i, c'est-à-dire $\dfrac{n_i - m_i}{n_i}$.

Le nombre de sujets à risque en t_{i+1} est donné par

$$n_{i+1} = n_i - m_i - c_i,$$

c'est-à-dire, la soustraction du nombre de sujets présents en t_i par le nombre de sujets décédés avant t_i et du nombre de sujets censurés entre t_i et t_{i+1}. L'estimateur de Kaplan-Meier est donné par la formule suivante :

$$\hat{S}_{KM}(t) = \prod_{i:t_i\leq t} \hat{q}_{t_i} = \prod_{i:t_i\leq t} \left(1 - \frac{m_i}{n_i}\right) \tag{2.17}$$

Pour $t \leq t_1$, on a par convention $\hat{S}_{KM}(t) = 1$. $\hat{S}_{KM}(t)$ est une fonction en escalier avec des sauts aux temps de décès observés. Elle n'atteint 0 que si le plus grand délai observé de l'échantillon correspond à un événement.

2.6 Précision et limites de l'estimateur de Kaplan-Meier

2.6.1 Précision de l'estimateur

2.6.1.1 Formule de Greenwood

Pour apprécier la précision de l'estimation de $S(t)$, il est utile d'estimer la variance de l'estimateur. S'il n'y a pas d'observations censurées, alors la formule (2.17) se réduit à la fonction de distribution empirique (2.15).

La variance de l'estimateur de Kaplan-Meier est estimée par la formule de Greenwood :

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{m_j}{n_j(n_j - m_j)} \tag{2.18}$$

En l'absence de censure ni de troncature (donnée complète) (2.18) devient

$$\hat{\sigma}^2(t) = \hat{S}(t)\frac{(1 - \hat{S}(t))}{n}, \quad avec \quad \hat{S}(t) = \hat{S}_{KM}(t)$$

l'estimateur standard de la variance binomiale.

$$IC(\alpha) = \left[\hat{S}(t) \pm Z_{1-\alpha/2}\hat{\sigma}^2_G(t)\right]$$

En effet (2.18) est obtenu à partir de *la méthode delta* qui assure que pour une fonction f et une variable aléatoire X on a :

$$f(X) = f(c) + f'(c)(X - c). \tag{2.19}$$

Ceci conduit donc à :

$$\begin{aligned} E(f(X)) &= f(c) + f'(c)E(X - c) && (2.20)\\ &= f(c) + f'(c)(E(X) - c). && (2.21)\\ &et \\ V(f(X)) &= [f'(c)]^2 V(X - c) && (2.22)\\ &= [f'(c)]^2 V(X). && (2.23) \end{aligned}$$

Ainsi, si Y est une variable aléatoire suivant la loi normale de moyenne μ et de variance σ^2, alors $f(Y)$ est une distribution normale de moyenne $f(\mu)$ et de variance $[f'(\mu)]^2.V(Y) = [f'(\mu)]^2.\sigma^2$. On obtient facilement ce résultat en prenant $c = \mu$. Par conséquent, en prenant $c = \mu$, on a :

- $f(Y) = \log(Y)$, on obtient que $E[f(Y)] = \log(\mu)$ et $V(f(Y)) = (\frac{1}{\mu})^2.\sigma^2$;

- $f(Y) = \exp(Y)$, on obtient que $E[f(Y)] = \exp(\mu)$ et $V(f(Y)) = (\exp(\mu))^2.\sigma^2$.

Soit X_i la variable aléatoire qui associe le nombre d'événements à la date t_i. La probabilité d'avoir m_i événements à la date t_i est donnée par $\frac{m_i}{n_i}$, donc $X_i \sim \mathcal{B}(n_i, \frac{m_i}{n_i})$.
Posons $\lambda_i = \frac{m_i}{n_i}$, on a :

$$Var(X_i) = n_i\lambda_i(1 - \lambda_i) \tag{2.24}$$

En passant au logarithme, l'équation (2.17) devient :

$$\log(\hat{S}(t)) = \sum_{i:t_i \leq t} \log\left(1 - \frac{X_i}{n_i}\right) \tag{2.25}$$

$$Var(\log(\hat{S}(t))) = Var\left(\sum_{t_i \leq t} \log(1 - \frac{X_i}{n_i})\right) \tag{2.26}$$

$$= \sum_{i:t_i \leq t} Var(\log(1 - \frac{X_i}{n_i})) \tag{2.27}$$

$$= \sum_{i:t_i \leq t} \left(\frac{1}{1 - \frac{X_i}{n_i}}\right)^2 \times Var(\frac{X_i}{n_i}) \tag{2.28}$$

$$= \sum_{i:t_i \leq t} \left(\frac{1}{1 - \frac{m_i}{n_i}}\right)^2 \times \frac{Var(X_i)}{n_i^2} \tag{2.29}$$

$$= \sum_{i:t_i \leq t} \left(\frac{1}{1 - \lambda_i}\right)^2 \times \frac{\lambda_i(1 - \lambda_i)}{n_i} \quad d'après\ (2.24) \tag{2.30}$$

$$= \sum_{i:t_i \leq t} \frac{\lambda_i}{n_i(1 - \lambda_i)} \tag{2.31}$$

$$= \sum_{i:t_i \leq t} \frac{m_i}{n_i(n_i - m_i)} \tag{2.32}$$

Notons $\hat{S}(t) = \exp(Y(t)), \implies Y(t) = \log(\hat{S}(t))$. En appliquant (2.23), on obtient

$$Var(\hat{S}(t)) = Var\left[\exp(Y(t))\right], \tag{2.33}$$

$$= \left[\exp(Y(t))\right]^2 \times Var(Y(t)), \tag{2.34}$$

$$= \left[\hat{S}(t)\right]^2 \times Var(Y(t)), \tag{2.35}$$

$$= \left[\hat{S}(t)\right]^2 \times Var(\log(\hat{S}(t)) \tag{2.36}$$

$$= \left[\hat{S}(t)\right]^2 \times \sum_{i:t_i \leq t} \frac{m_i}{n_i(n_i - m_i)} \tag{2.37}$$

$$\boxed{Var(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \times \sum_{i:t_i \leq t} \frac{m_i}{n_i(n_i - m_i)}}$$

En absence de censure ni de troncature (données complètes), on a $n_{i+1} = n_i - m_i$ et l'estimateur de Kaplan-Meier se simplifie :

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{m_i}{n_i}\right), \tag{2.38}$$

$$= \prod_{i:t_i \leq t} \frac{n_i - m_i}{n_i} = \prod_{i:t_i \leq t} \frac{n_{i+1}}{n_i}, \tag{2.39}$$

$$= \frac{n_t}{n_1}, \tag{2.40}$$

avec n_1 le nombre d'individus au début de l'étude. La variance est donc estimée comme suit :

$$Var(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \times \sum_{i:t_i \leq t} \frac{m_i}{n_i(n_i - m_i)}, \tag{2.41}$$

$$= \left[\frac{n_i}{n_1}\right]^2 \times \sum_{i:t_i \leq t} \frac{n_i - n_{i+1}}{n_i \times n_{i+1}} \tag{2.42}$$

$$= \left[\frac{n_i}{n_1}\right]^2 \times \sum_{i:t_i \leq t} \left(\frac{1}{n_{i+1}} - \frac{1}{n_i}\right), \tag{2.43}$$

$$= \left[\frac{n_i}{n_1}\right]^2 \times \left[\left(\frac{1}{n_2} - \frac{1}{n_1}\right) + \left(\frac{1}{n_3} - \frac{1}{n_2}\right) + \dots\right], \tag{2.44}$$

$$= \left[\frac{n_i}{n_1}\right]^2 \times \left(\frac{1}{n_i} - \frac{1}{n_1}\right), \tag{2.45}$$

$$= \left[\frac{n_i}{n_1} \times \frac{1}{n_1}\right] \times \left(1 - \frac{n_i}{n_1}\right), \tag{2.46}$$

$$= \frac{\hat{S}(t)\left(1 - \hat{S}(t)\right)}{n_1}. \tag{2.47}$$

Ainsi,

$$\boxed{\hat{\sigma}^2 = \hat{S}(t)\frac{(1 - \hat{S}(t))}{n}},$$

avec n l'effectif initial.

2.6.1.2 IC de la transformation $\log - \log$

Kalbfleish et Prentice (2002) [14, 15] ont proposé un intervalle de confiance du taux de survie $S(t)$ basé sur une transformation logarithmique $\log - (\log)$ qui est beaucoup utilisé en pratique et par beaucoup de logiciels (R, $STATA$, $Python$). L'intervalle de confiance asymétrique est donné par :

$$\exp\left(-\exp(C_+(t))\right) < S(t) < \exp\left(-\exp(C_-(t))\right) \tag{2.48}$$

où

$$C_\pm(t) = \log\Big(-\log(\hat{S}(t))\Big) \pm \sqrt{\hat{V}(t)}$$

avec

$$\hat{V}(t) = \frac{1}{(\log(\hat{S}(t))^2} \sum_{t_j \leq t} \frac{m_j}{n_j(n_j - m_j)}$$

Finalement, en considérant $g(t)$ donnée par $g(t) = \log(-\log(t))$

$$\hat{\sigma}_g(t) = \frac{\hat{\sigma}_G(t)}{\hat{S}(t)\log[\hat{S}(t)]}$$

Au niveau de confiance de (1-α), les bornes de l'intervalle de confiance sont données par :

$$IC(\alpha) = [\hat{S}(t) \times \exp(\pm Z_{1-\frac{\alpha}{2}}\hat{\sigma}_g(t))] \tag{2.49}$$

En effet la variance log-(log) de Kalbfleish, est obtenue en prenant $g(t) = \log(Y(t))$, avec $Y(t) = -\log(\hat{S}(t))$. En appliquant (2.23) à $g(t) = \log(Y(t))$, on obtient :

$$\begin{aligned}
Var(g(t)) &\approx \left(\frac{1}{Y(t)}\right)^2 \times Var(Y(t)) && (2.50)\\
&\approx \left(\frac{1}{Y(t)}\right)^2 \times Var(-\log(\hat{S}(t))) && (2.51)\\
&\approx \left(\frac{1}{Y(t)}\right)^2 \sum_{i=1}^{n} \frac{m_i}{n_i(n_i - m_i)} && (2.52)\\
&\approx \left(\frac{1}{\log(\hat{S}(t))}\right)^2 \sum_{i=1}^{n} \frac{m_i}{n_i(n_i - m_i)} && (2.53)
\end{aligned}$$

Au niveau de confiance de $(1-\alpha)$, les bornes de l'intervalle de confiance sont données par :

$$\left[g(t) \pm Z_{\alpha/2}\sqrt{\hat{V}}\right], \ avec\ \hat{V} = Var(g(t)) \tag{2.54}$$

On obtient

$$\left[\log(-\log\hat{S}(t)) \pm Z_{\alpha/2}\sqrt{\hat{V}}\right]. \tag{2.55}$$

Prenons

$$\begin{aligned}
C_+ &= \log(-\log\hat{S}(t)) + Z_{\alpha/2}\sqrt{\hat{V}} && (2.56)\\
C_- &= \log(-\log\hat{S}(t)) - Z_{\alpha/2}\sqrt{\hat{V}}. && (2.57)
\end{aligned}$$

Ce qui donne l'intervalle de confiance asymétrique définit par :

$$\exp\left[-\exp(C_+(t))\right] < \hat{S}(t) < \exp\left[-\exp(C_+(t))\right] \tag{2.58}$$

2.6.2 Limite de l'approche de Greenwood

L'approche d'estimation de la précision de l'estimateur de Kaplan-Meier par la variance de Greenwood est la plus utilisée dans la littérature. Cependant, elle fait l'objet de nombreuses critiques, notamment pour sa faiblesse dans le cas de censures importantes. D'autres approches sont proposées dans la littérature pour une meilleure estimation de la précision. Culter et Enderer [9] ont utilisé les tables de mortalité pour expliquer l'effet du suivi partiel sur la qualité du suivie. Sur une cohorte de 126 patients au total obtenue sur une période de 5 ans, ils ont réussi à montrer que la précision du taux de survie dépendait fortement de la qualité du suivi, et que la prise en compte des patients entrés après et avant l'étude réduisait considérablement l'erreur type sur l'estimation du taux de survie.
Lors d'une étude de suivie de cohorte, un nombre important de patients peuvent être retirés vivants avant la fin de l'étude, ce qui pourraît entacher la qualité de l'estimation de Kaplan-Meier. Conscient que la fiabilité du résultat dépend de la taille de l'échantillon, c'est-à-dire, du nombre de cas observés, Cutler et Ederer [9] ont introduit le concept "*effective sample size*" : la taille réelle de l'échantillon.
Le concept est introduit par la formule (2.59) ci-dessous permettant de calculer l'erreur type d'un taux de survie p, lorsque tous les cas ont été observés jusqu'à la mort ou à la fin de l'étude. En effet, lorsqu'il n'y a pas de pertes ni de retraits dans une analyse de table de mortalité, la probabilité cumulée de survie suit une distribution binomiale. On a

$$\hat{s} = \sqrt{\frac{p(1-p)}{I}} \tag{2.59}$$

Ainsi, si dans une étude, on aimerait avoir à la fin de l'étude un taux de survie p avec une erreur s, sans perte de suivi durant l'étude, la taille initiale de la cohorte sera donnée par la formule suivante :

$$I = \frac{p(1-p)}{\hat{s}^2}$$

Dans leur exemple, Cutler et Ederer ont montré qu'une analyse de la table de mortalité de 126 personnes équivalait à un suivi de 68 personnes pendant 5 ans, c'est-à-dire que la taille "effective" de l'échantillon serait de 68.
Kenneth J. Rothman [25], lui, propose aussi d'utiliser la formule de Greenwood comme indicateur des informations pour obtenir la taille effective I de l'échantillon comme ses prédécesseurs Cutler et Ederer. En utilisant la probabilité de survie cumulée p et la taille effective I de l'échantillon, il obtient une limite de confiance plus précise en supposant que $I\hat{p}$ suit une distribution binomiale. L'avantage de cette méthode est qu'elle donne un intervalle de confiance symétrique autour de $\hat{S}(t)$.

$$p^* = \frac{I}{I+z^2}\left|\hat{p} + \frac{z^2}{2I} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{I} + \frac{z^2}{4I^2}}\right| \tag{2.60}$$

L'intervalle de confiance de Rothman conduit à des valeurs de bornes asymétriques par rapport à la valeur de $S(t)$, donnant de bonnes précisions pour les valeurs de $S(t)$ plus petites, contrairement à celle de Greenwood qui n'arrive pas à donner une bonne précision des valeurs de $\hat{S}(t)$ plus petites. Peto et al. [22] ont proposé une variance cette fois, en utilisant une autre taille de l'échantillon notée

$$n'_j = \frac{n_j}{\hat{S}(t)}$$

avec $t_j \leq t \leq t_{j+1}$ et n_j le nombre d'individus à risque à la date j. La variance de Peto et al. [22] est donnée par :

$$\hat{\sigma}_p^2(t) = (\hat{S}(t))^2 \times \frac{1 - \hat{S}(t)}{n'_j}$$

et l'intervalle de confiance de Peto par :

$$[\hat{S}(t) - Z_{\alpha/2} \times \hat{\sigma_p}(t); \hat{S}(t) + Z_{\alpha/2} \times \hat{\sigma_p}(t)]$$

Cette méthode n'est généralement pas appropriée puisque les bornes inférieures et supérieures peuvent être respectivement inférieures à 0 et supérieures à 1.

2.7 La médiane de suivi

2.7.1 Définition

La censure est la principale cause des difficultés rencontrées en analyse de survie. La médiane de suivi, selon l'approche de Kaplan-Meier inversée [26] est la date à laquelle le taux de valeurs censurées est à 50%. Pour tout individu de la cohorte, en plus de T la variable aléatoire représentant le délai avant l'événement, on considère C la variable aléatoire représentant le temps de participation de l'individu à l'étude. Si $T > C$ alors l'individu quittera l'étude avant qu'on observe sa date de l'événement et sera par conséquent censuré à droite. La médiane de suivi telle que définie est alors obtenue par :

$$M_e = \min\{t \mid P(C > t) = 0.5\} \quad (2.61)$$

$$= \min\{t \mid S^*(t) = 0.5\} \quad (2.62)$$

avec $S^*(t)$ le taux de censure à la date t.

2.7.2 Méthodes d'estimation du suivi

Dans sa revue de littérature Schemper [26] a énuméré six méthodes qui ont été utilisées pour quantifier la médiane de suivi. On a les notations suivantes :

- t_{1i} est la date d'entrée de l'individus i ;
- t_{2i} est la date finale enregistrée ;
- t_3 est la date de fin d'étude ;
- $s_i = 1$ si la date de décès est connue, sinon $s_i = 0$ si la date est censurée à la date t_{2i}

On obtient alors :

1. *Une méthode basée sur le temps d'observation* $(T_{ob_i} = t_{2i} - t_{1i})$; $Me_{ob} = Me_{T_{ob}}$.
   Pour chaque patient on observe le temps d'entrée et la date de décès de l'individu. Mais cette méthode à des limites car le temps d'observation dépend considérablement de la vitesse à laquelle les décès se produisent. Donc pour un même suivi sur deux études différentes, le temps d'observation sera beaucoup plus faible pour l'étude où les décès sont plus fréquemment observés.
2. *Une méthode basée sur le temps de censure* $T_{ce_i} = (t_{2i} - t_{1i} \; si \; s_i = 0)$; $Me_{ce} = Me_{T_{ce}}$.
   Ici, la date de décès n'est pas connue. Cette méthode en dit moins sur la qualité du suivi car elle ne concerne que les périodes censurées.
3. *Une méthode basée sur le temps avant la fin de l'étude* $(T_{f_i} = t_3 - t_{1i})$; $Me_f = Me_{T_f}$. Pour chaque individu, on calcule le temps entre la date de fin d'étude et la date d'entrée sans tenir compte que le temps soit censuré ou pas. Elle peut donc surestimer le suivi réel.
4. *Une méthode basée sur le temps Connu.*
   Dans cette méthode on calcule le temps de la façon suivante :

$$T_{co_i} = \begin{cases} t_{2i} - t_{1i} & \text{si } s_i = 0 \\ t_3 - t_{1i} & \text{si } s_i = 1 \end{cases}$$

   $Me_{co} = Me_{T_{co}}$. Ici les individus décédés avant la fin de l'étude sont considérés non censurés jusqu'à la date de point. Cette méthode de suivi a tendance à surestimer le suivi potentiel des patients présentant une perte croissante de suivi, comme cela a été montré dans [26]. Néanmoins, elle est très efficace s'il y a un risque de perte de suivi élevé.
5. *Méthode de Kaplan-Meier inversé*
   Elle se calcule de la même manière que la méthode de Kaplan-Meier [16]. Le décès $(s_i = 1)$ est censuré par un temps d'observation réel mais reste inconnu pour un individu. Dans le cas où le temps est censuré $(s_i = 0)$ on obtient un temps final.

6. *Suivi selon la distribution de Korn*
   L'approche de Korn [17] estime la probabilité qu'un individu i soit suivi avant un instant t donné, t strictement positive. Cette fonction est donnée par la probabilité suivante :
   $$P(t) = P(T > t|E > t)P(E > t)$$
   avec $P(E > t)$ la proportion des sujets vérifiant $T_{f_i} > t$. T est le temps jusqu'à la perte du suivi de l'individu. Quant à $P(t) = P(T > t|E > t)$ elle se détermine de l'estimateur de Kaplan-Meier du suivi potentiel ("reverse Kaplan-Meier"). En absence de risque de perte de suivi, le suivi de Korn présente un très meilleur résultat de suivi [26].

Pour étudier la qualité du suivi médian estimée par ces méthodes, Schemper et al [26] ont mené une simulation de Monte Carlo. 5000 temps de suivi échantillonés à partir d'une distribution exponentielle ont été utilisés et censurés à partir de la procédure Gehan et Thomas [12] pour modéliser un essai clinique. Le temps écoulé jusqu'à la perte du suivi suit également une distribution exponentielle avec un taux de risque spécifié. À l'issu de cette étude Schemper et al. estiment que le suivi médian, les quantiles du suivi potentiel estimés de Kaplan-Meier inversé sont de bons indicateurs pour quantifier le suivi médian. L'avantage est qu'ils peuvent facilement être obtenus avec un logiciel standard.

Chapitre

3 Méthodologie

Sommaire

3.1 Simulation de variables aléatoires

3.1.1 Simulation de la loi uniforme

Générer une suite de variables aléatoires $(U_n)_{n\geq 1}$ indépendantes et suivant la loi uniforme sur $[0,1]$ est impossible sur un ordinateur : d'abord parce que les nombres entre 0 et 1 sont en fait de la forme $k/2^p$ avec $k \in \{0,...,2^p-1\}$; ensuite parce que vérifier qu'une suite $(U_n)_{n\geq 1}$ est bien indépendante et identiquement distribué $(i.i.d.)$ est une question très délicate. En pratique, les logiciels se basent sur une suite dite pseudo-aléatoire qui a la forme $X_{n+1} = f(X_n)$ avec X_n à valeurs dans un ensemble fini, typiquement de taille plus grande que $2p$. L'objectif est alors d'en déduire une suite (U_n) qui ressemble autant que possible à une suite $i.i.d.$ uniforme sur $\{0,...,2p-1\}$. Comme f est déterministe, un tel générateur est périodique. Il existe de nombreuses méthodes pour générer de telles suites. Sous R (comme sous Python et Matlab) via la commande $runif$, on utilise :

$$X_{n+1} = AX_n \ (mod\ 2) \tag{3.1}$$

où X_n est un vecteur de 19937 bits et A une matrice bien choisie. La variable pseudo-aléatoire (U_n) correspond alors aux 32 derniers bits de X_n. Proposé par Matsumoto

et Nishimura en 1998 [19], ce générateur a une période de longueur $219937 - 1$. Il est possible de changer de générateur sous R, il suffit de le spécifier avant l'appel à runif via la fonction RNGkind.

Les fonctions *random* des principaux langages de programmation ou logiciels sont bâties sur des algorithmes arithmétiques dont le plus simple correspond au générateur congruentiel linéaire. Il s'agit de générer une suite de nombres $(X_n)_{n\geq 1}$ vérifiant une relation de récurrence

$$X_{n+1} = aX_n + c \ mod \ M \tag{3.2}$$

et d'en déduire une suite $(U_n)_{n\geq 1}$ à valeurs dans $[0, 1[$ en prenant $U_n = X_n/M$. Par exemple la fonction rand de Scilab utilise (3.2) avec $M = 231,\ a = 843\ 314\ 861$ et $c = 453816693$. La suite (U_n) ainsi construite est complètement déterministe, car périodique. Cependant, sa période est tellement grande qu'on peut en pratique la considérer comme une suite aléatoire.

3.1.2 Simulation d'une loi quelconque

3.1.2.1 Méthode d'inversion de la fonction de répartition

Soit X une variable aléatoire réelle de fonction de répartition F, définie par

$$F(x) := P(X < x).$$

F est une fonction croissante, continue à droite. L'ensemble de ses discontinuités est au plus dénombrable et F a pour limite 0 en -∞ et 1 en $+\infty$. Dans le cas particulier où F est continue et strictement croissante sur tout $\mathbb{R}$, elle réalise une bijection de $\mathbb{R}$ sur $]0, 1[$ et admet donc un inverse

$$F^{-1} :]0, 1[\to \mathbb{R}$$

Si U est une variable aléatoire de loi uniforme sur $]0, 1[$, alors

$$Y := F^{-1}(U)$$

a même loi que X.
En calculant la fonction de répartition de Y, on obtient :

$$\forall x \in \mathbb{R}, P(Y \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x) \tag{3.3}$$

Cette propriété permet donc, à partir d'un générateur aléatoire fournissant des réalisations d'une variable aléatoire uniforme, de simuler une variable aléatoire de même loi que X, en admettant que l'on sache calculer F^{-1}. Ceci nous conduit à poser la définition suivante.

Définition 3.1.1. *(Inverse généralisée) Si F est la fonction de répartition d'une variable aléatoire, son inverse généralisé (ou fonction quantile) est définie par :*

$$\forall u \in]0,1[,\ F^{-1}(u) := \inf\{x \in \mathbb{R};\ F(x) \geq u\}. \tag{3.4}$$

Remarque 3.1.1. *L'équation $F(x) = u$ peut avoir soit une solution unique, soit aucune solution, soit une infinité de solutions. La détermination graphique de $F^{-1}(u)$ est également possible.*

Théorème 3.1.1. *(Méthode d'inversion) Soient X une variable aléatoire réelle de fonction de répartition F et U une variable aléatoire de loi uniforme sur $]0,1[$. Alors X et $F^{-1}(U)$ ont même loi.*

Preuve. *F^{-1} est une fonction croissante donc borélienne, ceci dit elle est mesurable. F étant strictement croissante, la preuve se résume à l'équivalence*

$$\forall u \in [0,1],\ \forall x \in \mathbb{R},\ F^{-1}(u) \leq x \Leftrightarrow u \leq F(x), \tag{3.5}$$

avec F^{-1} définit par (3.4).

$$\textit{Soit} \quad A_u := \{t \in \mathbb{R};\ F(t) \geq u\}$$

d'où $F^{-1}(u) = \inf A_u$ par continuité de F.
Montrons l'implication :

$$F^{-1}(u) \leq x \Rightarrow u \leq F(x).$$

Soit $F^{-1}(u) \leq x$. Alors pour $n \geq 1$, $F^{-1}(u) \leq x + \frac{1}{n}$ et comme $F^{-1}(u) = \inf A_u$, il existe un $t_n \in A_u$ tel que $F^{-1}(u) \leq t_n \leq x + \frac{1}{n}$.
Par définition de A_u, on a $F(t_n) \geq u$ donc par croissance de F on obtient

$$u \leq F(t_n) \leq F\left(x + \frac{1}{n}\right).$$

De la continuité de F au point x on en déduit en faisant tendre n vers l'infini que $u \leq F(x)$. $x \in \mathbb{R}$ et $u \in]0,1[$ étant quelconque, l'implication directe est vérifiée.
L'implication réciproque est immédiate. En effet, si $u \leq F(x)$, cela signifie que $x \in A_u$. Or $F^{-1}(u) = \inf A_u$, donc $F^{-1}(u) \leq x$

Par conséquent si F est explicite, on calculera F^{-1} et pour générer un échantillon $X_1, X_2, \ldots, X_n$ de variables aléatoires indépendantes et de même loi de fonction de répartition F, on générera un échantillon $U_1, U_2, \ldots; U_n$ de variables de loi uniforme sur $[0;1]$ et on posera $X_i = F^{-1}(U_i)$.

Exemple 3.1. *Application à la loi exponentielle (Méthode d'inversion)*

On rappelle que X suit une loi exponentielle de paramètre λ ($\lambda > 0$) si pour tout $t \in \mathbb{R}^+$,

$$P(X > t) = \exp(-\lambda t)$$

Nous noterons cette loi $\mathcal{E}(\lambda)$. Si F est la fonction de répartition de X, nous avons alors $F(t) = 1 - e^{-\lambda t}$ (pour $t \geq 0$) et

$$F^{-1}(x) = -\frac{\ln(1-x)}{\lambda}$$

Si $U \sim \mathcal{U}(]0,1[)$, nous avons, d'après le résultat ci-dessus :

$$-\frac{\ln(1-U)}{\lambda} \sim \mathcal{E}$$

Comme $1 - U$ a même loi que U, on a :

$$-\frac{\ln(U)}{\lambda} \sim \mathcal{E}$$

3.1.2.2 Méthode de rejet

On veut simuler une variable aléatoire X de densité f, mais f est trop compliquée pour que ceci puisse se faire directement. On dispose par contre d'une densité auxiliaire g telle que :

1. on sait simuler Y de densité g ;
2. il existe une constante c telle que, pour tout y, $f(y) \leq cg(y)$;
3. pour tout y, on sait calculer le rapport $f(y)/(cg(y))$.

Ceci étant acquis, la proposition suivante montre comment simuler suivant la densité f.

Proposition 3.1.1. *Soient f et g deux densités de probabilité sur $\mathbb{R}$ vérifiant la relation suivante :*

$$\exists c > 0 \text{ tel que } \forall x \in \mathbb{R},\ f(x) \leq cg(x) \quad \text{avec} \quad g(x) \neq 0.$$

Soit X une variable aléatoire réelle, X ayant g pour densité de probabilité et Y une variable aléatoire indépendante de X suivant la loi uniforme standard $\mathcal{U}[0,1]$. La loi conditionnelle de X sachant $Y \leq \dfrac{f(X)}{cg(X)}$ a pour densité f.

Preuve. *Soit $q(x) = f(x)/cg(x)$, f et g étant les densités de probabilité. On a : $0 \leq q(x) \leq 1$. Calculons $P(Y \leq q(x))$ en restant dans le cas strictement continu. Par définition on a :*

$$P(Y \leq q(x)) = \int_{x=-\infty}^{x=+\infty} \int_{y=0}^{y=q(x)} h(x,y)dxdy$$

avec $h(x, y)$ la densité de probabilité du couple (X, Y). Comme X et Y sont deux variables indépendantes, on a :

$$h(x, y) = 1 \times g(x) = g(x)$$

donc

$$\begin{aligned} P(Y \leq q(x)) &= \int_{x=-\infty}^{x=+\infty} \int_{y=0}^{y=q(x)} g(x)dxdy && (3.6)\\ &= \int_{x=-\infty}^{x=+\infty} g(x) \left(\int_{y=0}^{y=q(x)} dy \right) dx && (3.7)\\ &= \int_{x=-\infty}^{x=+\infty} g(x)q(x)dx && (3.8)\\ &= \int_{x=-\infty}^{x=+\infty} \frac{f(x)}{c} dx && (3.9)\\ &= \frac{1}{c} \left[F(x) \right]_{-\infty}^{+\infty} && (3.10)\\ &= \frac{1}{c} && (3.11) \end{aligned}$$

Soit B un sous-ensemble de $\mathcal{R}$. On a par définition :

$$P(X \in B | Y \leq q(x)) = \frac{P(X \in B \text{ et } Y \leq q(X))}{P(Y \leq q(X))},$$

d'après le résultat précédent on a :

$$\begin{aligned} P(X \in B | Y \leq q(x)) &= cP(X \in B \text{ et } Y \leq q(X)) && (3.12)\\ &= c \int_{x \in B} g(x) \left(\int_{y=0}^{y=q(x)} dy \right) dx && (3.13)\\ &= c \int_{x \in B} g(x)q(x)dx \quad avec\ q(x) = f(x)/cg(x) && (3.14)\\ &= c \int_{x \in B} \frac{f(x)}{c} dx = \int_{x \in B} f(x)dx. && (3.15) \end{aligned}$$

Ce qui montre que f est la densité de probabilité de la loi conditionnelle de X sachant

$$Y \leq \frac{f(x)}{cg(x)}.$$

3.1.2.3 Méthode de transformation

Soit Y et T deux variables réelles telles que $Y = T(X)$ où T est une transformation connue. Soient $x_1, x_2, \ldots, x_n$, n valeurs de X_n déjà simulées.
La méthode de composition permet d'avoir n valeurs simulées $y_1, y_2, \ldots, y_n$ de Y en

appliquant la transformation

$$y_i = T(x_i),\ \forall i \in [\![0\ ;\ n]\!]$$

Remarque 3.1.2. *La méthode d'inversion est un cas particulier de la méthode de transformation ou X suit $\mathcal{U}(0,1)$.*

Exemple 3.2. *La loi log-normale. On dit que Y suit la loi log-normale de paramètre m et σ^2 si*

$$Y = \exp(X) \text{ ou } X \text{ suit } \mathcal{N}(m,\sigma).$$

Si on veut simuler n valeurs $y_1, y_2, \ldots, y_n$ de Y, la procédure suivante peut être utilisée :

- *simuler n valeurs $x_1, x_2, \ldots, x_n$ de X suivant $\mathcal{N}(m,\sigma)$,*
- *en déduire n valeurs simulées $y_1, y_2, \ldots, y_n$ de Y en posant :*

$$y_i = \exp(x_i),\ \forall i = 1 \text{ à } n$$

3.1.2.4 Méthode de composition

Supposons que la densité de la variable X que l'on cherche à simuler puisse s'écrire sous la forme :

$$f(x) = \sum_i^r p_i f_i(x)$$

où $p_1, \ldots, p_r$ représentent une distribution discrète ($0 < p_i < 1$ et $\sum p_i = 1$) et f_i sont des densités définies sur $[a_i, b_i]$. Alors X peut être simulé à partir de l'algorithme suivant :

- *Simuler $I \sim p_1, ..., p_r$,*
- *Simuler X suivant $f_I(x)$, $\quad x \in [a_i, b_i]$,*
- *retourner X.*

3.2 La Méthode de Monte Carlo

3.2.1 Description de la méthode de Monte Carlo

Le terme méthode de Monte Carlo désigne toute méthode visant à calculer une valeur numérique en utilisant des procédés aléatoires, c'est-à-dire des techniques probabilistes (ou plus généralement stochastiques).
Pour calculer une quantité I, la première étape est de la mettre sous forme d'une espérance $I = \mathbb{E}(X)$ avec X une variable aléatoire. Si on sait simuler des variables $X_1, X_2, \ldots$ indépendantes et identiquement distribuées, alors nous pouvons approcher X par :

$$X \approx \frac{X_1 + X_2 + \dots X_N}{N} \tag{3.16}$$

avec N "grand", sous réserve d'application de la loi des grands nombres. C'est ce type d'approximation que l'on appelle la *méthode de Monte Carlo*

Sa convergence est donnée par la loi forte des grands nombres :

Théorème 3.2.1. *(Loi forte des grands nombres) Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires i.i.d., à valeurs dans $R^d (d \in \mathbb{N}^*)$. On suppose que $\mathbb{E}(|X_1|) < +\infty$. Alors*

$$\frac{X_1 + X_2 + \dots X_N}{N} \xrightarrow[N\to\infty]{p.s} E(x) \tag{3.17}$$

Dans le cas où la variable est continue, cette espérance mathématique s'exprime sous la forme d'une intégrale.
Une application classique des méthodes Monte-Carlo est le calcul des quantités du type

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx,$$

où $\varphi : R^d \to \mathbb{R}$ est une fonction donnée et X un vecteur aléatoire de densité f suivant laquelle on sait simuler. Dans ce contexte, l'estimateur Monte-Carlo de base est défini par

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i),$$

où les X_i sont générées de façon *i.i.d.* selon f. La quantité $\hat{I}_n$ est appelée l'approximation de I par la méthode de Monte Carlo.
L'estimateur $\hat{I}_n$ est sans biais, c'est-à-dire que $\mathbb{E}[I_n] = I$. Il converge d'après la loi forte des grands nombres (3.2.1). Plus n devient grand, plus l'estimation se rapproche de la vraie valeur I

3.2.2 Simulation de la méthode de Monte Carlo dans le cas pratique

Dans le cas pratique, on utilise l'algorithme suivant, pour l'approximation de Monte Carlo :

- générer en utilisant un générateur de bonne qualité n variables aléatoires (valeurs simulées d'une variable aléatoire suivant la loi uniforme sur [0,1]) : $u_1, u_2, \dots, u_n$.
- choisir une densité f définie sur le support Δ (et facile à simuler par exemple par la méthode d'inversion) pour simuler à partir des u_i, n valeurs d'une variable X, soit $x_1, x_2, \dots x_n$,
- déterminer les quantités $\varphi(x_1), \varphi(x_2) \dots, \varphi(x_n)$,
- calculer la moyenne arithmétique de ces quantités soit $\hat{I}$.

3.2.3 La précision de l'estimateur de Monte-carlo

Par définition l'on a :

$$V(\hat{I}) = V\left(\frac{\sum_{i=1}^{n} g(X_i)}{n}\right),$$

avec g un générateur de variables aléatoires. Comme les $g(X_i)$ sont $i.i.d$, il s'en suit que

$$V(I) = \frac{V(g(X))}{n} = \frac{E(g^2(X)) - [E(g(X))]^2}{n} \tag{3.18}$$

$$= \frac{E(g^2(X)) - I^2}{n}. \tag{3.19}$$

On peut approximer I^2 par $\hat{I}^2$. Quant à $E(g^2(X))$, on peut utiliser la méthode de Monte Carlo pour l'approximer à

$$\hat{J} = \frac{\sum_{i=1}^{n} g^2(x_i)}{n}.$$

Les x_i sont les valeurs simulées de X par f. Une approximation de $V(\hat{I})$ est donnée par

$$\widehat{V(\hat{I})} = \frac{1}{n}(\hat{J} - \hat{I}^2)$$

3.3 Notre stratégie de simulation et critère de comparaison

Les simulations permettent de prédire le comportement du sujet étudié sans avoir à passer par la construction de prototypes où la réalisation d'essais réels, coûteux et/ou difficiles à mettre en place ; ce qui est un avantage essentiel en matière de coûts de production. La simulation numérique consiste à simuler sur ordinateur des phénomènes physiques en utilisant des logiciels élaborés sur la base de modèles théoriques mathématiques.

Les techniques de simulation Monte Carlo sont utilisées pour simuler des systèmes déterministes avec des paramètres stochastiques. La technique s'appuie sur l'échantillonnage des distributions des quantités incertaines. Elles sont aujourd'hui utilisées pour simuler des phénomènes physiques complexes dans plusieurs domaines scientifiques et appliqués : Radioactivité, physique des hautes énergies, économétrie, logistique, ...

3.3.1 Stratégies de simulation

3.3.1.1 Hypothèses

Nous considérons une cohorte d'individus présentant un risque constant λ d'apparition d'événement. Dans ce sens, la distribution des délais suit une loi exponentielle.

Nous considérons ensuite une censure non informative, c'est-à-dire que la censure suit une distribution uniforme indépendante des délais des événements.

Le scénario de la cohorte est inspiré des données discutées par Somda et al. [28]. L'étude de référence concerne des patients souffrant de tumeurs des cellules germinales non séminomateuses au stade métastatique. Au total 246 patients étaient observés pour des stades de bons pronostics selon les critères IGCCCG (International Germ Cell Consensus Classification Group). Le suivi médian était de 48,6 mois avec un intervalle de confiance à 95% IC= [47,2 ;53,7]). 12 patients étaient décédés avant 5 ans avec un taux de survie globale à 5 ans estimé à 95% à IC= [90,7 ;97,3]. En plus de ces résultats de référence, nous considérons la même cohorte avec un pronostic moyen (survie à 5 ans de 50%) et avec un faible pronostic (survie à 5 ans de 10%).

Pour trouver la relation entre le suivi médian et la couverture de l'intervalle de confiance, nous avons considéré plusieurs valeurs de suivi médian entre 0 et 170 mois par pas de un mois. Pour un niveau de confiance $1-\alpha = 95\%$, pour atteindre la précision de 6 points de pourcentage nous avons trouvé que 10 000 réplications à chaque étape étaient suffisantes [2].

3.3.1.2 Paramètres

Le temps T de survie est généré suivant une distribution exponentielle de paramètre λ. Le taux de survie après un temps t est égal à

$$S(t) = p(T > t) = \exp(-\lambda t) \tag{3.20}$$

Le risque instantané d'apparition de l'événement λ est obtenu à partir du taux de survie s à la date $t = 5$ ans soit 60 mois. On obtient :

$$\lambda = -\frac{\log(s)}{t} \tag{3.21}$$

Les dates de censures C sont supposées aléatoires, indépendantes des dates de réalisations des événements T et générées selon une distribution uniforme sur $[0; d]$ ou d est la date de point. La densité de probabilité de la variable C est :

$$f(t) = \begin{cases} \frac{1}{d} & \text{si } 0 \leq t \leq d \\ 0 & \text{sinon} \end{cases} \tag{3.22}$$

Soit M_e la médiane de suivi, alors M_e est la date pour laquelle le taux de censure est

de $\frac{1}{2}$, c'est-à-dire $P(C \geq M_e) = \frac{1}{2}$. Comme C suit une loi uniforme, on obtient :

$$\begin{aligned} P(C \geq t) &= 1 - P(C < t) = 1 - \int_{-\infty}^{t} f(x)dx = 1 - \int_{-\infty}^{t} \frac{1}{d}dx \\ &= 1 - \frac{t}{d} \end{aligned}$$

donc

$$P(C \geq M_e) = \frac{1}{2} = 1 - \frac{M_e}{d} \Rightarrow d = 2.M_e \tag{3.23}$$

$$\boxed{d = 2 \times M_e}$$

3.3.2 Critères de comparaison

Pour les critères de comparaisons, nous avons observé deux éléments essentiels : *la stabilité* (ou le taux de couverture) et *la précision de l'intervalle de confiance* (ou la prudence). L'estimateur de Kaplan-Meier et l'intervalle de confiance ont été calculés pour chaque échantillon. Pour chacune des valeurs du suivi médian considérés, $N = 10\ 000$ échantillons de taille $n = 246$ ont été générés aléatoirement.

1. *La couverture* est obtenue en comparant le taux de survie réel s_t aux bornes des intervalles de confiance de niveau $1 - \alpha$ de $\hat{s}_t$. Pour chaque itération de la médiane, la couverture représente le pourcentage de fois où l'intervalle de confiance de l'estimateur couvre la vraie valeur du paramètre.

   Soient $X_1, X_2, ..., X_i$ des variables aléatoires i.i.d tel que :

$$X_i = \begin{cases} 1 & \text{si } s_t \text{ est dans l'intervalle} \\ 0 & \text{si } s_t \text{ n'est pas dans l'intervalle} \end{cases} \tag{3.24}$$

   Alors

$$X_i \sim \mathfrak{B}(1, p) \tag{3.25}$$

   On pose :

$$P(X_i = 1) = 1 - \alpha \text{ et } P(X_i = 0) = \alpha \tag{3.26}$$

   Pour N itérations indépendantes, posons

$$X = \sum_{i=1}^{N} X_i \tag{3.27}$$

   on obtient :

$$X \sim \mathfrak{B}(N, 1 - \alpha)$$

La couverture est donnée par :

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \Longrightarrow E(\bar{X}) = \frac{1}{N} \times E\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N} \times \sum_{i=1}^{N} E(X_i) = 1 - \alpha \tag{3.28}$$

et la variance de $\bar{X}$ est donnée par :

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{N} X_i}{N} X_i\right) = \frac{1}{N^2} \times Var\left(\sum_{i=1}^{N} X_i\right) = \frac{\alpha(1-\alpha)}{N} \tag{3.29}$$

donc

$$Z = \frac{\sqrt{N}\left[\bar{X} - (1-\alpha)\right]}{\sqrt{\alpha(1-\alpha)}} \sim \mathcal{N}(0,1)$$

Pour N assez grand, par exemple, $N \geq 30$ et $N(1-\alpha) \geq 5$ posons :

$$E(X_i) = \mu, \quad Var(X_i) = \sigma^2 \text{ donc}$$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

D'après le théorème de la limite centrale :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$$

Il est possible alors de proposer un nouvel intervalle de confiance de $\bar{X}$ de niveau $1 - \alpha'$, donné par

$$\left[\bar{X} - z_{\alpha'/2}\frac{\sigma}{\sqrt{N}}; \bar{X} + z_{\alpha'/2}\frac{\sigma}{\sqrt{N}}\right] \tag{3.30}$$

2. La *longueur de l'intervalle de confiance à* $(1-\alpha)$. Pour chaque simulation, la précision est obtenue en faisant la différence entre la borne supérieure et la borne inférieure pour chaque apparition de l'événement. Les intervalles de confiance de plus petites amplitudes impliquent des estimations plus précises. En choisissant $L(t)$ comme la longueur moyenne de l'intervalle de confiance à la date t, on obtient :

$$L(t) = \frac{1}{N}.\sum_{i=1}^{N}\left(B_j^{sup}(t) - B_j^{inf}(t)\right) \tag{3.31}$$

Avec $B_j^{sup}(t)$ et $B_j^{inf}(t)$ respectivement les bornes supérieures et inférieures de l'intervalle de confiance à la j^{eme} itération. Ce calcul permet d'observer la précision des différents intervalles.

Chapitre

4 Résultats et Discussion

Sommaire

4.1 Résultats de la simulation

Les 10 000 réplications ont été éffectuées pour les trois scénarios ($s = 95\%, s = 50\%, s = 10\%$) pour chacune des valeurs de la médiane de suivi de 0 à 170 mois.

Les figures (4.1), (4.2) et (4.3) donnent les distributions des taux de survie estimés à 60 mois pour des médianes de suivi de 15, 38, 60 et 120. Lorsque la médiane de suivi est de 15 mois, nous avons une grande dispersion du taux de survie allant de 0 à 1. La médiane de cette box-plot est de 0.98. Le taux de survie réel est largement surestimé. Pour la médiane de 38 mois, la dispersion est moins présente, avec une médiane de 0.95 pour cette box-plot. Les valeurs extrêmes sont moins importantes. Lorsque la médiane de suivie est de 60 et 120 mois, nous observons une dispersion moins forte autour de la médiane obtenue de 0.95. Plus la médiane de suivi devient grande, plus la dispersion autour du taux de survie réel diminue (le taux de survie estimé se rapproche du taux de survie réel).

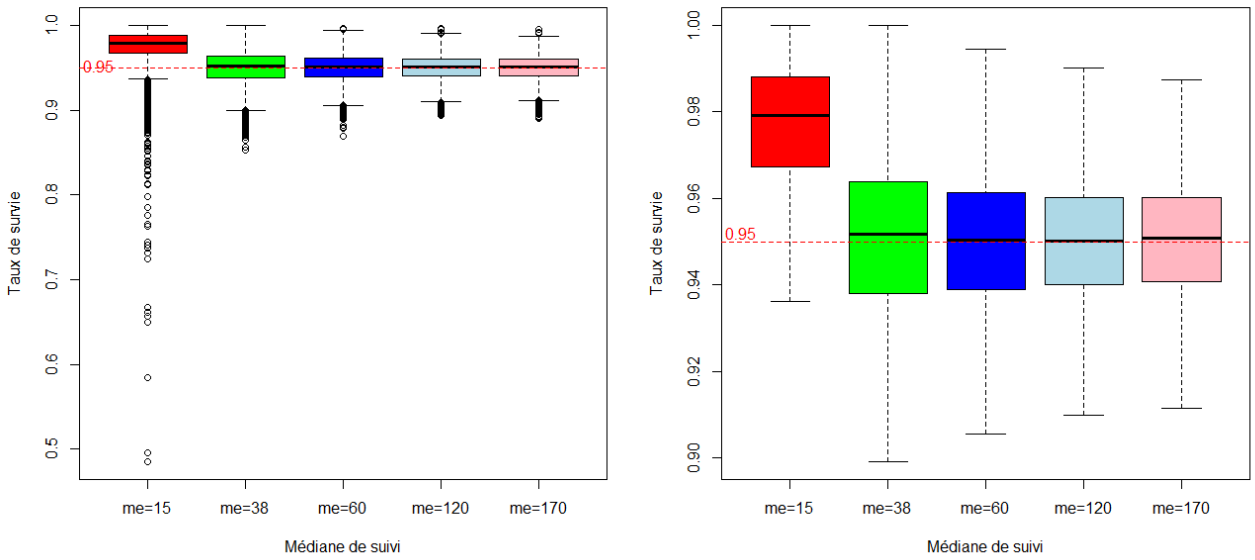


Figure 4.1 – Taux de survie en fonction de la médiane de suivi ($t = 60; s = 95\%$)

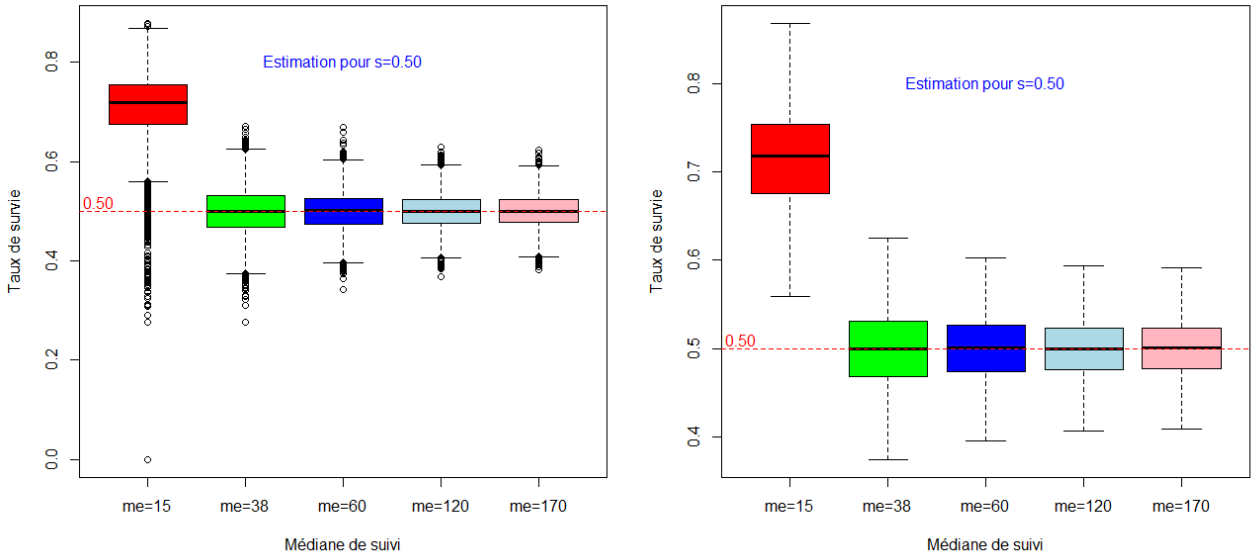


FIGURE 4.2 – Taux de survie en fonction de la médiane de suivi ($t = 60; s = 50\%$)

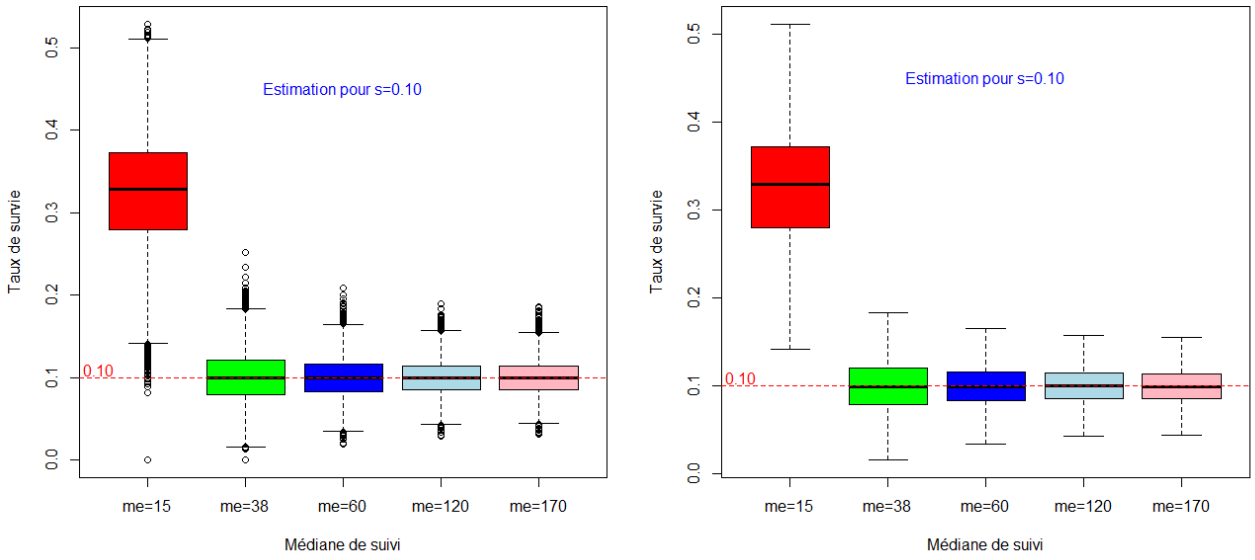


FIGURE 4.3 – Taux de survie en fonction de la médiane de suivi ($t = 60; s = 10\%$)

La figure (4.4) montre le taux de couverture en fonction de la médiane de suivi. La courbe est croissante pour les médianes inférieures à 38. Elle présente un plateau lorsque la médiane de suivi devient supérieure ou égale 38 avec une très bonne couverture de la vraie valeur. Ainsi, l'estimation de l'intervalle de confiance est anti-conservatrice pour les valeurs de la médiane de suivi inférieure à 38 et très conservatrice pour les suivis médians supérieurs ou égale à 38. La qualité de la couverture évolue dans le même sens que la médiane de suivi.

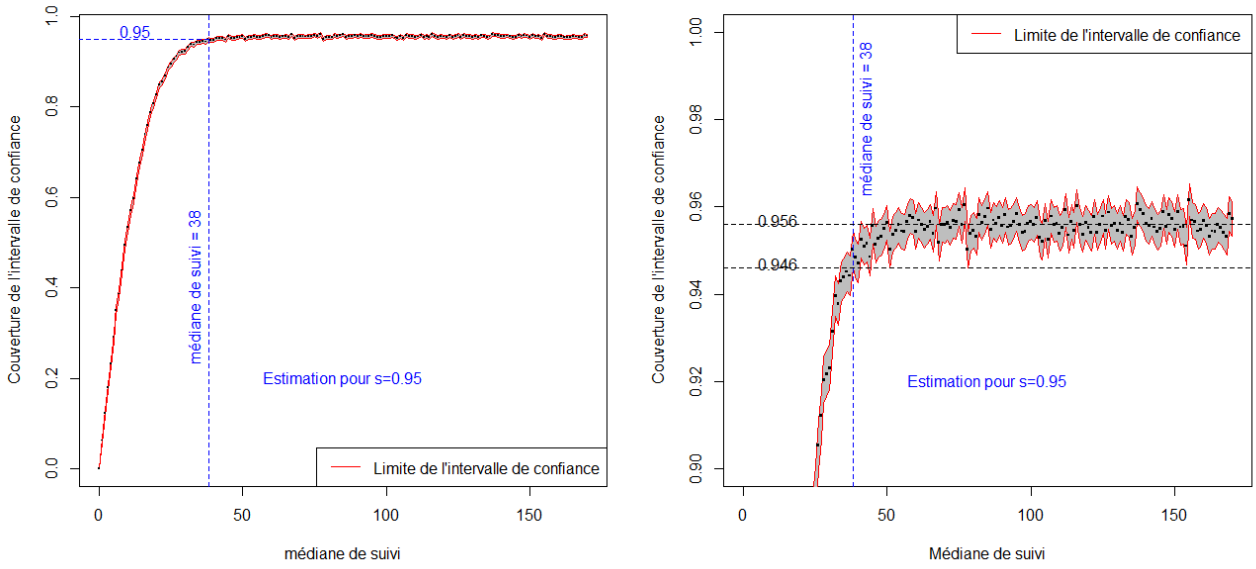


FIGURE 4.4 – Taux de couverture en fonction de la médiane de suivi ($t = 60$ mois, $s = 95\%$)

Pour le cas de mauvais pronostic (s=0.50 ou s=0.10), on obtient respectivement la figure (4.5). Les résultats sont identiques. La méthode est conservatrice dès que la médiane de suivi est supérieure à 34. Un vrai plateau s'observe, chaque taux de couverture observé se situe de 94,4% à 95,6%.

La figure (4.6)donne la précision, c'est-à-dire, la longueur moyenne des intervalles de confiance du taux de survie à la date t en fonction de la médiane de suivi. On observe une courbe strictement croissante de 0 à 31 mois et décroissante de 31 mois à T convergeant vers 0.06. Cependant on obtient une bonne précision à partir de 33 mois de suivi médian

La figure (4.7) montre l'évolution du taux de survie moyen en fonction de la médiane de suivi. La courbe du taux de survie en fonction de la médiane de suivi montre pour les petites valeurs de la médiane de suivi, un taux de survie très supérieur au taux de survie normale de 95%, c'est la partie décroissante de la courbe. À partir de la médiane de suivi de 30, nous obtenons un taux de survie moyen presque constant (on observe un plateau) autour de 95%, c'est-à-dire, pour les valeurs de suivi médian supérieur à $\frac{t}{2}$ (soit $0.5t$). En reprenant la simulation avec des taux de survie relativement faibles, c'est-à-dire dans les cas de mauvais pronostic ($s = 5\%,\ s = 10\%\ s = 50\%$), on arrive à la même conclusion : Pour les valeurs du suivi médian supérieur à $\frac{t}{2}$, on obtient une survie moyenne très proche de la vraie valeur.

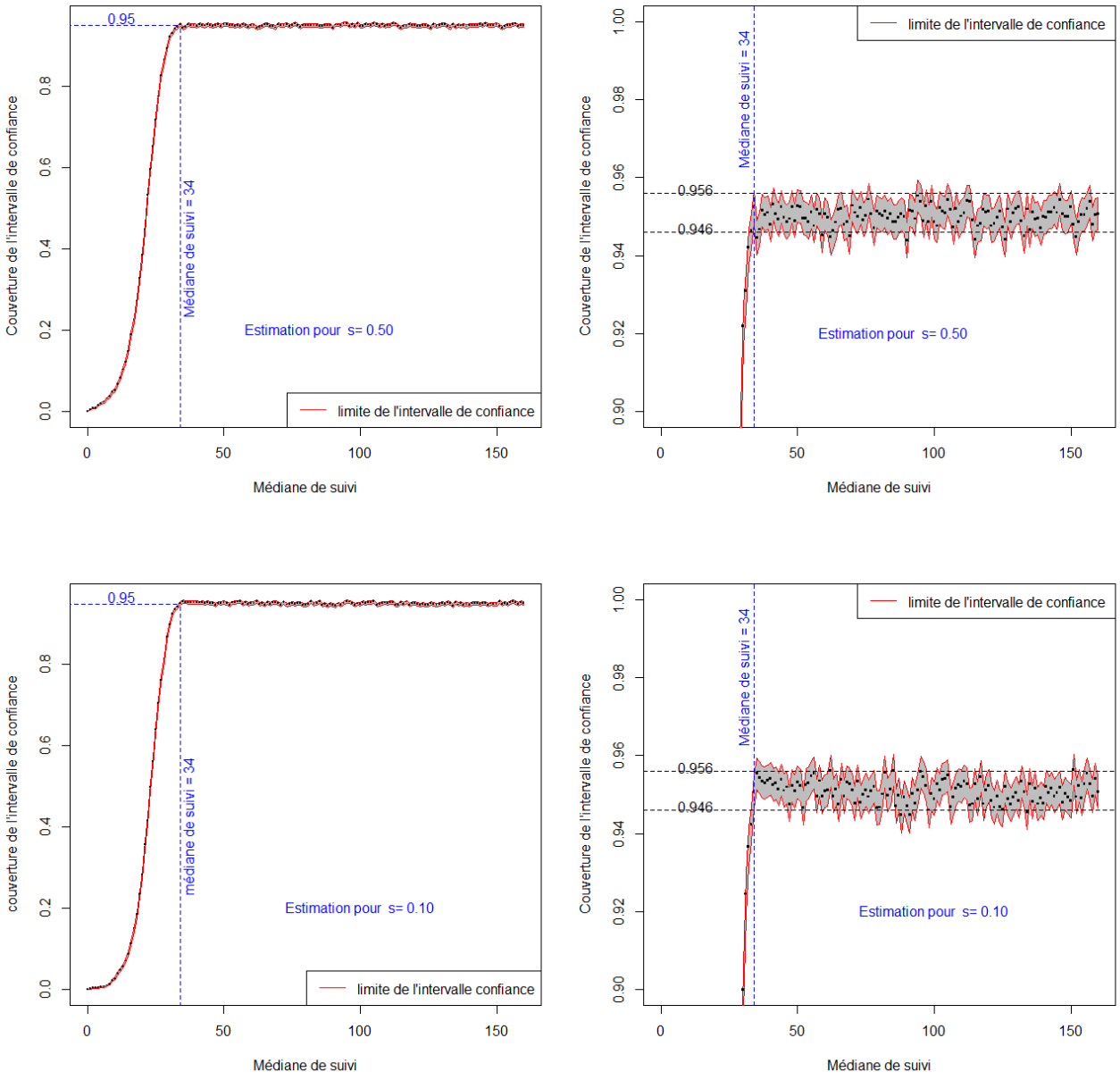


FIGURE 4.5 – Taux de couverture en fonction de la médiane de suivi

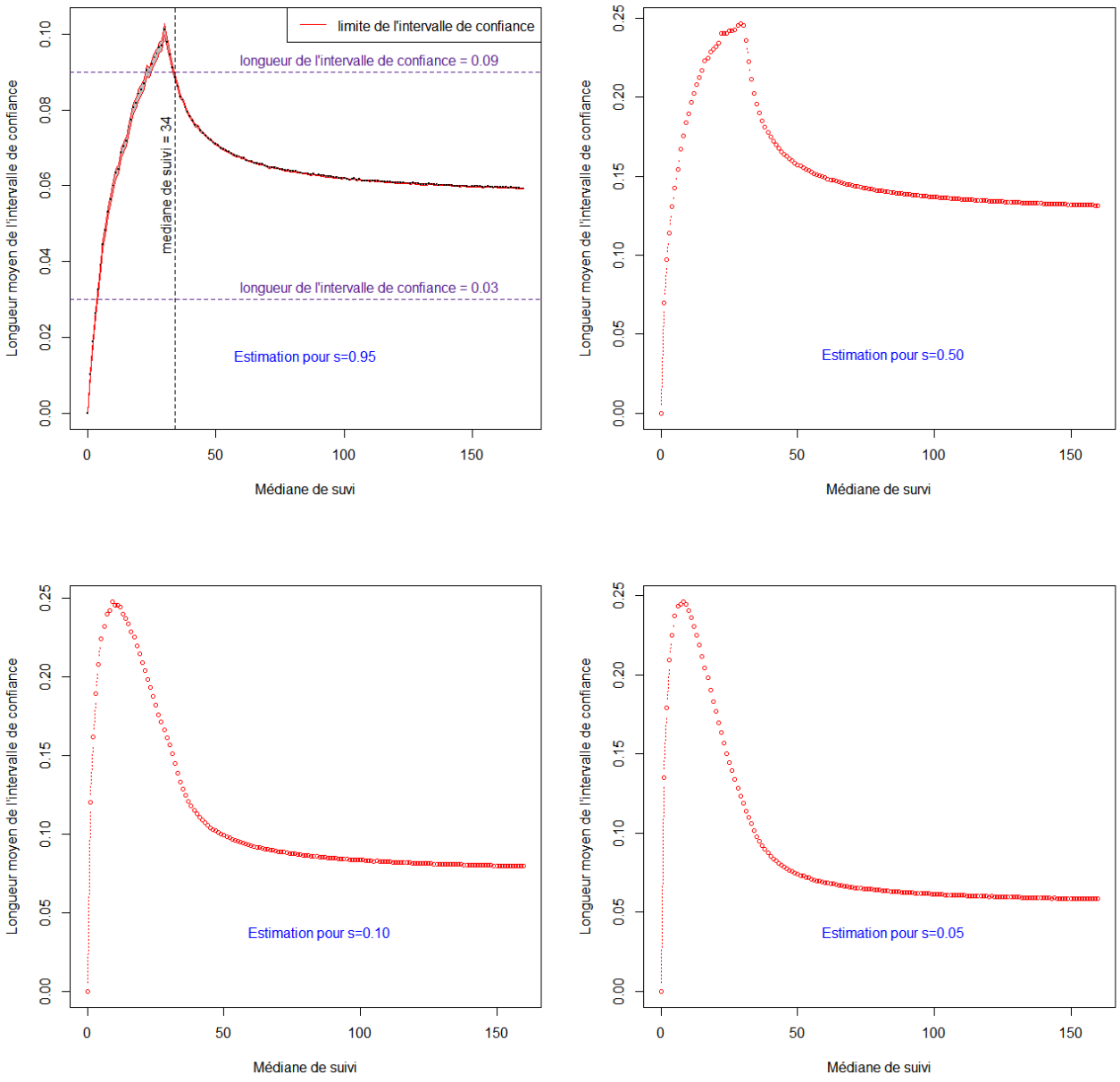


FIGURE 4.6 – Longueur de l'intervalle en fonction de la médiane de suivi à $t = 60$ mois

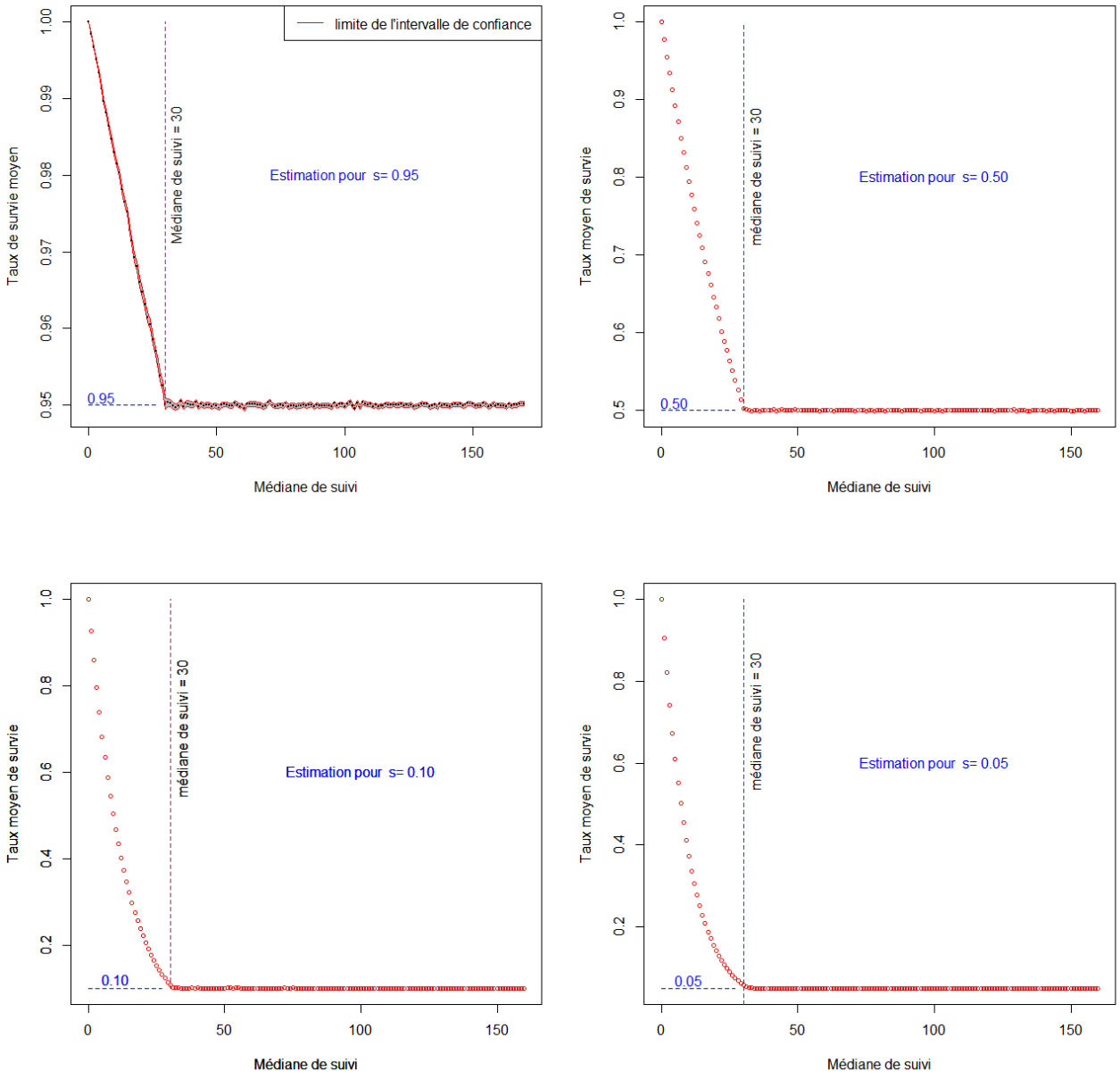


FIGURE 4.7 – Taux de survie en fonction de la médiane de suivi à $t = 60$ mois

4.2 Interprétation des résultats

L'intervalle de confiance de Kaplan-Meier offre une très bonne couverture quand il y a peu de censure. En effet à partir d'un certain niveau de la médiane de suivi la couverture est constante autour du niveau de confiance $1-\alpha$. De plus la dispersion est très faible ce qui dénote une bonne précision de l'estimateur. Pour les faibles valeurs de la médiane de suivi, c'est-à-dire quand la censure est importante, la couverture est insuffisante. Pour un intervalle de confiance calculé à 95%, on se rend compte que la probabilité de contenir la vraie valeur du paramètre est inférieure à 95%. L'estimation est anti-conservatrice. Toutefois, plus la censure diminue, c'est-à-dire plus la médiane de suivi augmente, plus l'estimation s'améliore.

Notre simulation montre une mauvaise couverture pour les valeurs de la médiane inférieures à 38 et une bonne couverture pour les valeurs supérieures ou égale à 38. Le taux de survie étant à $t = 60$ mois, cette simulation montre une bonne couverture pour les valeurs supérieures à $\frac{38}{60}t$ soit $0.633t$. Lorsque le taux de survie à $t = 5$ ans était de 50%, nous avons observé que l'estimation était efficace à partir du seuil de $0.55t$. De même pour le taux survie de 10% nous avons obtenu un seuil de $0.56t$.

Pour ce qui concerne la longueur de l'intervalle de confiance, celle-ci nous donne une information sur la précision de l'estimateur. Nous observons que pour ce critère les estimations convergent vers la valeur de référence au fur et à mesure que la médiane de suivi augmente. Il est cependant difficile de définir une valeur seuil à partir de laquelle la précision serait acceptable.

4.3 Discussion

À la fin de cette étude nous pouvons dégager les résultats suivants :

- La couverture et la précision de l'intervalle de confiance de l'estimateur de Kaplan-Meier est très sensible à la médiane de suivi ;
- L'intervalle de confiance de Kaplan-Meier présente une bonne couverture et une précision adéquate à partir de certaines valeurs seuils de la médiane de suivi ;
- Les valeurs seuils sont atteintes à partir de deux tiers (63%) de la date de l'estimation.

L'analyse de sensibilité a été réalisée en utilisant des techniques de simulation par Monte Carlo. Il s'agit de méthodes fortement recommandées dans la littérature [21]. Pour des raisons de validité, un exemple réel a été considéré [28]. Ce scénario a été légèrement modifié pour en produire trois autres. Les échantillons ont été tirés par génération de nombre pseudo aléatoires selon des algorithmes disponibles dans les logiciels. Pour évaluer chaque scénario 10.000 réplications étaient effectuées. Ce nombre a été trouvé suffisant pour apprécier la qualité du taux de couverture. Ainsi des seuils de confiance

ont pu être définis lorsque la couverture était bonne ou mauvaise. Enfin notre analyse de sensibilité a porté sur 170 points différents ce qui a pour objet de proposer des décisions précises. Finalement, les observations étaient faites sur 246 individus $\times 10.000$ réplications $\times$ 170 points, ce qui fait un grand nombre d'informations.

Une des principales limite que nous relevons dans notre étude de sensibilité est que nous avons considéré une seule date d'estimation à savoir 60 mois. Ceci est dû au fait que la problématique de la médiane de suivi se rencontre plus fréquemment pour les longues durées de vie (5 ans au plus).

Dans les cohortes simulées, la censure était considérée non informative. Cela signifie que le retrait d'un individu du suivi avant l'événement d'intérêt était indépendant de la probabilité de cet événement et surtout était indépendant du temps d'observation. Cette hypothèse bien que pratique est peu réaliste. En effet, plus les participants durent dans l'étude plus ils risqueront de la quitter avant son terme. Le risque de censure est donc généralement lié au temps d'observation.

L'intervalle de confiance de l'estimateur de Kaplan-Meier a été calculé en utilisant la variance de Greenwood avec une transformation $\log -(\log)$ [15, 14]. La variance de Greenwood est connue pour son manque de puissance en cas de censure importante. D'autres méthodes existent dans la littérature pour le calcul de l'intervalle de confiance, notamment en utilisant des distributions exactes. Sorgho [29] a évalué 11 méthodes différentes d'estimation de l'intervalle de confiance et proposé des alternatives dans les cas où la méthode log-(log) perd sa puissance.

En 1958, Cutler et Ederer [9] avaient montré à partir des tables de mortalité que la prise en compte des patients entrés après et avant l'étude réduisaient considérablement l'erreur type sur l'estimation du taux de survie. Cette étude est en conformité avec notre étude. En augmentant la taille des échantillons au fur et à mesure, on obtiendra une médiane de suivi beaucoup plus grande, ce qui réduit considérablement l'erreur type sur l'estimation du taux de survie.

Dans la littérature, la médiane de suivi est utilisée pour juger la qualité du suivi [27], à partir de cette étude nous pensons que le suivi médian serait un très bon critère de jugement de la précision de l'intervalle de confiance. Elle pourrait servir à juger la nécessité de poursuive une étude dans le cas ou le suivi médian est atteint très tôt ou encore nous ramener à revoir la taille de notre échantillon (augmenter la taille de l'échantillon) si on désire poursuivre l'étude comme le proposait Cutler et Ederer [9]. En effet, il serait moins utile de poursuive une étude si, on est pas à mesure de donner une bonne précision du taux de survie obtenu.

En 2002, Pocock et al [24], avait proposé que l'on représente les courbes de survie tant que la proportion de patients à risque est de l'ordre de 10% à 20% de l'effectif initial. Cependant, une taille supérieure à 20% de l'effectif initial ne garantirait pas la qualité de la précision du taux de survie à un instant t. Il serait nécessaire dans le cas de censure lourde, d'avoir un suivi médian supérieur à deux tiers de t.

Cette étude permet également de répondre à certains auteurs, tel que Shuster [27] qui ont refusé l'idée que le suivi médian puisse apporter de l'importance à l'estimation du taux de survie et aussi de la qualité au suivi. En effet, le suivi médian exprime d'abord un taux lourd de valeurs censurées, dont l'impact sur la qualité de l'estimateur de Kaplan-Meier est considérable [1]. De plus la médiane de suivi apporte un intervalle de validité pour la précision de la qualité de la variance.

En recherche clinique, notamment en analyse de survie, l'objectif est souvent de mesurer le taux de survie à une certaine date t. Nous proposons à partir de cette étude un seuil en deça duquel la qualité de la précision de l'intervalle diminue. Si, pour une étude l'objectif est de mesurer le taux de survie à une date T, il serait nécessaire que le suivi médian soit au moins supérieur à deux tiers de la date de l'estimation. Cette étude apporte une règle de décision claire et objective pour évaluer la qualité de l'estimation d'un taux de survie.

Chapitre

5 Conclusion générale et perspectives de recherche

Sommaire

5.1 Bilan du travail

Au terme de cette étude nous avons montré que la médiane de suivi peut être considérée comme un critère d'arrêt précoce d'une étude. Cette étude servira de critère pour la validation de l'estimation du taux de survie à un instant t dans les études de survie de longue durée. En effet, au vu des frais qu'engendrent les études de cohorte, nos résultats permettront de juger de la nécessité de poursuivre une étude. Ainsi, nous mettons en place un intervalle de temps, dans lequel la qualité de l'intervalle de confiance est acceptable.

Cette qualité de l'intervalle de confiance du taux de survie à une date t dépend de la médiane de suivi. Plus la médiane de suivi est grande, meilleure est la qualité de l'intervalle de confiance. Plus précisément, au delà d'une médiane de suivi de $\frac{2}{3} \times t$ nous avons une très bonne qualité en terme de couverture et de la longueur de l'intervalle de confiance du taux de survie à la date t.

5.2 Perspectives de recherche

Après avoir mis en évidence l'importance du suivi médian sur la qualité de l'estimation, nous comptons poursuivre nos recherches sur la prise en compte de ce facteur dans la conception des études. Notamment, nous développerons une méthode de calcul taille optimale de l'échantillon en tenant compte du suivi médian. Dans cette étude nous avons également pu remarquer que la méthode de Kaplan-Meier, paraissait moins efficace en présence de censure lourde, nous tenterons de trouver une relation analytique entre le suivi médian et la qualité de l'estimateur de Kaplan-Meier. Enfin, nous allons reprendre

la même étude pour les autres méthodes de calcul de l'intervalle de confiance afin de proposer des régles de décisions équivalentes.

Bibliographie

[1] K. A. Adeleke. A simulation study on kaplan meier non-parametric survival methods. *Journal of the Nigerian Mathematical Society*, 31, 243-254, 2012. 45

[2] A. Agresti. Categorical data analysis. 2013. 33

[3] T. Ancelle. Statistique épidémiologie. *3 ème Editions*. Maloine, 2011. 2

[4] T. Ancelle. Statistique épidémiologie. *4 ème Editions*. Maloine, 2017. 14

[5] J. Berkson and R. P. Gage. Calculation of survival rates for cancer. In *Proceedings of the staff meetings Mayo Clinic*, volume 25, page 270, 1950. 8

[6] P. E. Böhmer. Theorie der unabhängigen wahrscheinlichkeiten. In *Rapports Memoires et Procès verbaux de Septième Congrès International d'actuaires Amsterdam*, volume 2, pages 327–343, 1912. 2, 8

[7] Ø. Borgan and K. Liestøl. A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*, pages 35–41, 1990. 3

[8] C. B. Borkowf. A simple hybrid variance estimator for the Kaplan–Meier survival function. *Statistics in medicine*, 24, 6, 827-851, 2005. 3

[9] S. J. Cutler and F. Ederer. Maximum utilization of the life table method in analyzing survival. *Journal of chronic diseases*, 8, 6, 699-712, 1958. 3, 20, 44

[10] İ. Etikan, S. Abubakar, and R. Alkassim. The Kaplan-Meier estimate in survival analysis. *Biom Biostatistics Int J.*, 5, 2, 00128, 2017. 3

[11] M. P. Fay. Two-sided exact tests and matching confidence intervals for discrete data. *R journal*, 2, 1, 53-58, 2010. 3

[12] E. A. Gehan and D. G. Thomas. The performance of some two-sample tests in small samples with and without censoring. *Biometrika*, 56, 1, 127-132, 1969. 23

[13] M. Greenwood. The errors of sampling of the survivorship tables. *Reports on public health and medical subjects*, 33, 1926. 3

[14] J. D. Kalbfleisch. The statistical analysis of failure time data. Technical report, 2002. 18, 44

[15] J. D Kalbfleisch and R. L Prentice. The statistical analysis of failure time data. John Wiley & Sons. 3, 18, 44

[16] L. Edward Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53, 282, 457-481, 1958. 2, 3, 8, 22

[17] E. L. Korn. Censoring distributions as a measure of follow-up in survival analysis. *Statistics in medicine*, 5, 3, 255-260, 1986. 23

[18] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50, 163-170, 1966. 8

[19] M. Matsumoto and T. Nishimura. Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8, 1, 3-30, 1998. 26

[20] W. Q. Meeker and L. A Escobar. Statistical methods for reliability data. John Wiley & Sons. 2014. 3

[21] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American statistical association*, 44, 247, 335-341, 1949. 43

[22] R. Peto, M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. G. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient II. analysis and examples. *British journal of cancer*, 35, 1, 1, 1977. 3, 21

[23] S. Philippe. Introduction à l'analyse des durées de survie. http ://www.lsta.upmc.fr/psp/Cours-Survie-1.pdf, 2015. 9

[24] S. J. Pocock, T. C. Clayton, and D. G Altman. Survival plots of time-to-event outcomes in clinical trials : good practice and pitfalls. *The Lancet*, 359, 9318, 1686-1689, 2002. 4, 44

[25] K. J Rothman. Estimation of confidence limits for the cumulative probability of survival in life table analysis. *Journal of chronic diseases*, 31, 8, 557-560, 1978. 3, 20

[26] M. Schemper and T. L. Smith. A note on quantifying follow-up in studies of failure time. *Controlled clinical trials*, 17, 4, 343-346, 1996. 4, 21, 22, 23

[27] J. J. Shuster. Median follow-up in clinical trials. *Journal of Clinical Oncology*, 9, 1, 191-192, 1991. 4, 44, 45

[28] S. M. A. Somda, S. Culine, C. Chevreau, K. Fizazi, E. Leconte, A. Kramar, and T. Filleron. A statistical approach to determine the optimal duration of post-treatment follow-up : Application to metastatic nonseminomatous germ cell tumors. *Clinical genitourinary cancer*, 15, 2, 230-236, 2017. 33, 43

[29] B. Sorgho, M. A. S. Somda, and D. Barro. Estimation de l'intervalle de la fonction de survie de Kaplan- Meier. Mémoire de master. *Université de Ouagadougou*, 2018. 3, 44

[30] R. L. Strawderman, M. I. Parzen, and M. T. Wells. Accurate confidence limits for quantiles under random censoring. *Biometrics*, pages 1399–1415, 1997. 3

[31] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke. The strengthening the reporting of observational studies in epidemiology (STROBE) statement : guidelines for reporting observational studies. *Annals of internal medicine*, 147, 8, 573-577, 2007. 2, 4