UNIVERSITÉ DU QUÉBEC À MONTRÉAL

INVESTIGATION SUR LES CALCULS DE TAUX D'INCIDENCE À L'AIDE D'APPROCHES ISSUES DE L'ANALYSE DE L'HISTORIQUE DES ÉVÉNEMENTS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

LUCK MATUNGULU LUKUSA

FÉVRIER 2022

REMERCIEMENTS

Le travail de mémoire est rarement l'œuvre d'une seule personne. Le présent n'échappe pas à cette tradition. Pour cette raison, je tiens tout d'abord à exprimer ma profonde gratitude à ma directrice de recherche, Professeure Juli Atherton, pour son soutien sans faille, pour ses précieux conseils et pour la confiance qu'elle m'a témoigné en acceptant la direction de mon mémoire. Ses encouragements constants, ses observations critiques, sa grande disponibilité, sa patience et ses magnifiques qualités humaines m'ont énormément aidé dans la réalisation de ce travail. J'apprécie grandement le temps qu'elle a passé à corriger aussi bien les mathématiques que la présentation générale. Cela a énormément contribué à une meilleure version de mon mémoire. Je lui dis sincèrement merci.

J'ai conscience que mon parcours au sein du cursus de maîtrise au Département de mathématiques à l'UQAM n'aurait pas abouti aux résultats escomptés sans le soutien remarquable de nos distingué(e)s professeur(e)s. Pour cette raison, ils méritent un profond sentiment de gratitude de notre part. Je pense particulièrement à madame Sorana Froda, aux messieurs René Ferland, Fabrice Larribe, François Watier et Karim Oualkacha.

Un grand merci à madame Gisèle Legault, analyste de l'informatique au laboratoire des cycles supérieurs en mathématiques et à monsieur Jean Rancourt, bibliothécaire à la Bibliothèque des sciences, respectivement pour leurs assistance et soutien informatique et bibliographique. Je tiens à remercier également madame Isabella Couture, secrétaire et assistante à la gestion des études au Département de mathématiques, pour sa patience et ses qualités humaines.

Mes remerciements vont aussi à mes collègues du Département de mathématiques, avec qui j'ai passé des moments agréables. Je pense particulièrement à Patrick Tian, Julien St-Pierre et Amadou Diogo Barry, merci pour vos soutiens et vos collaborations sans faille. Je dis également merci à Ismaila Ba pour sa sympathie et son sens de l'humour.

J'exprime ma profonde reconnaissance à ma mère, mes sœurs et frères qui m'ont toujours encouragé malgré les milliers de kilomètres qui nous séparent. Je pense en particulier à mon frère Pierre Lutumba wa Lukusa pour son soutien moral qui m'a donné la force d'aller jusqu'au bout. Je pense également à mes frères Papy Joseph Musangu Lukusa et Dieudonné Kalonji Lukusa disparus pendant mon parcours au sein du cursus de maîtrise. Ils nous ont quittés très tôt pendant que la famille avait encore besoin d'eux. Je n'aurai plus jamais la chance de fêter avec eux la fin de ma maîtrise. Qu'ils trouvent auprès de Dieu un repos éternel et paisible !

Mes pensées vont également à Rady Fahmy, qui, malgré sa détermination, a perdu le grand combat contre le cancer. Quel que soit sa demeure actuelle, je lui murmure mes sincères remerciements pour m'avoir accompagné dans ma lutte de retrouver la santé de mes yeux malgré ses propres problèmes de santé. Son grand cœur d'aider, sa gentillesse et son humanisme nous manqueront à jamais. Que l'éternel Dieu lui donne un repos paisible auprès de lui !

Un très grand merci à mon amour, Florine Totokani, pour son soutien inestimable, son amour inconditionnel, ses encouragements sincères et sa patience. Avec elle, nous dédions ce mémoire à nos filles, Florida Lukusa et Floriana Lukusa ainsi qu'à mes parents, David Lukusa Lutumba Matungulu et Anastasie Mujinga Mutombo.

À toutes et à tous, merci !

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Le taux d'incidence est l'une des mesures les plus couramment utilisées en épidémiologie bien que sa modélisation laisse parfois quelques lacunes et voir des confusions d'interprétation. Dans la littérature épidémiologique, le taux d'incidence est présenté comme le nombre d'événements sur les personnes-temps à risque. L'objectif de notre travail est de présenter quelques autres approches mathématiques pour modéliser et estimer le taux d'incidence, et de lever certaines équivoques sur les hypothèses d'application. Premièrement, nous montrons que le taux moyen défini par Selvin est le taux d'incidence. L'expression de Selvin s'applique aux fonctions de risque non constantes. Nous considérons plusieurs façons d'estimer l'expression de Selvin. Nous examinons également plusieurs façons d'approximer les personnes-temps au dénominateur pour les données de type recensement. Par la suite, nous simplifions le modèle à la fonction de risque constante et nous montrons que le risque dans ce cas est égal au taux d'incidence. Deuxièmement, nous considérons les modèles de Markov homogènes continus à plusieurs états. Dans ces modèles, avec la fonction de risque constante, l'estimateur du maximum de vraisemblance (EMV) pour le risque est l'estimateur du maximum de vraisemblance du taux d'incidence. Nous concluons ce travail par des idées pour d'autres extensions susceptibles d'être investiguées dans les travaux futurs.

Mots-clés : Épidémiologie, fonction de risque constante, taux d'incidence, analyse de survie, modèles multi-états, chaîne de Markov homogène continue, analyse de l'historique des événements.

INTRODUCTION

Il existe de nombreuses mesures d'occurrence de décès et de maladie (ci-après dénommés événements) en épidémiologie. Celle qui est couramment utilisée, mais souvent mal comprise est le taux d'incidence (TI) défini comme suit :

$$TI = \frac{\#\text{ d'événements}}{\text{personnes-temps}}. \tag{1}$$

Le nombre (#) d'événements est calculé sur une période bien spécifiée et les personnes-temps représentent le temps total passé par tous les individus *à risque*[1] au cours de la même période. Le calcul à l'expression (1) est souvent cité comme étant un bon moyen pour gérer les pertes de vue qui surviennent lorsqu'un individu quitte une population ou un échantillon avant qu'un événement ne soit observé. Lorsque l'intervalle de temps étudié devient infinitésimal, le taux d'incidence (TI) se rapproche de la *fonction de risque* dans l'analyse de survie.

Dans cette introduction, nous commençons par expliquer la motivation de ce mémoire et nous donnons une brève revue de la littérature. Nous finissons cette introduction par un aperçu de la structure du document.

1. Il s'agit de la somme des durées cumulées sur l'ensemble de la population à l'étude et sur l'ensemble de la durée de suivi, pendant laquelle les individus sont susceptibles d'être enregistrés comme de nouveaux cas.

0.1 Motivation et revue de la littérature

Malheureusement, le taux d'incidence est souvent mal compris en épidémiologie. Certaines raisons pour cela sont déjà présentes dans la littérature épidémiologique :

1. Difficultés d'interprétation lorsque les taux d'incidence sont souvent confondus avec les probabilités. Voir la citation suivante de (Vandenbroucke, 1985) :

   > Part of the blame for this situation - as (Morgenstern *et al.*, 1980) call it in the introduction of their paper - might be with the elementary introductions in biostatistics and epidemiology which might have been kept too elementary. When browsing through such introductions to statistics in medicine or through elementary textbooks of epidemiology, one usually finds that the rate of mortality is treated as a peculiar kind of probability or as if it were a case fatality rate, with very little awareness of the underlying theoretical assumptions.

2. Difficultés d'interprétation causées par la dépendance numérique du choix de l'unité de temps dans les personnes-temps. Voir ci-dessous la citation de (Vandenbroucke, 2003) :

   > The author criticized publications on death rates in hospitals by Farr and Nightingale in the 1860s. The author rehashed the 130-year-old accusation that Farr and Nightingale had used a "wrong" rate to overstate their political message about differences in death rates between hospitals. Farr and Nightingale had published death rates up to "90 per 100 per year" for particular hospitals for which they envisaged reforms. Interested parties at the time thought that these rates were impossible and thus wrong. Still, the explanation is straight forward : in hospital several persons in succession occupy a single bed over a year, which yields 1 person-year of observation. Several persons might have died in that bed, which might even lead to death rates larger than unity. A simple solution to restore "intuitive credibility" is to calculate death rates per person-month or person-week : a death rate of 90 per 100 patient-years becomes 1.73 per 100 patient-weeks. The latter number would not as quickly have drawn accusations of "political spin", although it

> remains exactly the same incidence. Like the 19th century critics of Farr and Nightingale, the 1996 author concluded that these high mortality rates were used because of Nightingale's political agenda.

3. Un manque de rigueur dans la littérature épidémiologique. Voir la citation de l'éditeur suivant (Elandt-Johnson, 1975) :

> Editor's note.—We suspect that most epidemiologists occasionally or regularly misuse one or another of the terms so carefully defined by Dr. Elandt-Johnson. We share her pessimism regarding the likelihood of changing this situation.

Étant donné que nous sommes maintenant en 2021, de nombreuses références citées ci-dessus sont anciennes. Des manuels plus modernes tels que (Jewell, 2003) et (Rothman *et al.*, 2008) donnent de bien meilleures présentations de concepts tels que les taux d'incidence, les taux de prévalence et l'incidence cumulée. Souvent, des diagrammes et exemples simples tels que ceux présentés dans la section 1.1.2 sont utilisés pour illustrer les calculs des taux d'incidence à partir de l'expression (1). Malheureusement, la discussion qui suit les présentations de ces exemples est très intuitive et manque de rigueur. Par conséquent, une certaine incertitude, des confusions et une mauvaise utilisation des taux d'incidence demeurent comme problème. Les points importants à aborder sont les suivants :

4. Préciser si le taux d'incidence est calculé pour une population ou un échantillon. Si l'expression (1) est calculée pour une population, le résultat est un paramètre de la population. En revanche, si l'expression (1) est calculée pour un échantillon, le résultat est une estimation et il faudrait dans ce cas, considérer les propriétés de l'estimateur, comme l'estimation de sa variance. Cette distinction est rarement discutée ; les exceptions sont (Rosner, 2010) et (Walker, 1991) qui tiennent compte de l'estimation de la variance.

5. Le plus souvent, il n'y a pas des discussions concernant la population sous-jacente. En particulier, qu'il s'agisse d'une population ouverte (avec des entrées ou des sorties) ou d'une population fermée (population fixe), le concept reste mal compris. Voir (Vandenbroucke et Pearce, 2012) qui indiquent :

   > « ... the concept of dynamic populations as the basis for incidence rate calculations remains often inadequately understood. »

6. L'estimation du nombre des personnes-années (ou plus généralement de l'unité de temps) au dénominateur du taux d'incidence. Le livre de (Elandt-Johnson et Johnson, 1999) fait la distinction entre les données de *type recensement* où le nombre de personnes exposées au risque doit être estimé et les données de *type expérimental* où le nombre de personnes exposées au risque peut être déterminé directement. Par exemple, une femme qui a subi une hystérectomie n'est pas à risque de cancer de l'ovaire ; cependant, bien que le nombre de femmes dans une population puisse être disponible, le nombre de femmes ayant subi une hystérectomie pourrait ne pas être disponible.

   > When the date of death is recorded in a specific area over a specific period of time, estimates of the number exposed to risk are usually based on census data. For convenience, we use the term census-type data generally to describe data in which the numbers exposed to risk are estimated indirectly. When records are available from which the numbers exposed to risk can be ascertained directly we, again for convenience, use the term experimental-type data. Sometimes these terms may not appear to be very relevant to the data actually under consideration. Their function is to remind ourselves what type of data we are considering. (Elandt-Johnson et Johnson, 1999).

7. Concernant les données de type recensement, lorsque les personnes-temps doivent être estimés et, surtout dans les cas où la fonction de survie est loin d'être linéaire ou quand ils ne tiennent pas compte de la structure sous-jacente de la population, cela peut conduire à sous-estimer considérablement

les personnes-temps à risque. Cette mauvaise estimation peut être atténuée par un calcul sur un petit intervalle. Dans les cas où la population augmente ou diminue presque linéairement, les approximations sont plus précises.

8. Spécifier si l'événement peut se produire une seule fois pour chaque individu (comme un décès) ou plusieurs fois par individu (comme une maladie). Lorsque l'événement peut se produire plusieurs fois par personne, une attention particulière doit être accordée au concept *d'être à risque* de vivre l'événement et son implication lors du calcul des personnes-temps au dénominateur. Par exemple, les hommes et les femmes qui ont subi une hystérectomie ne sont plus à risque de développer un cancer de l'ovaire. Autoriser plusieurs événements par personne provoque également une confusion quant à l'intradépendance entre les événements d'un même individu. Voir (Windeler et Lange, 1995).

9. Bien qu'il n'est pas nécessaire de préciser l'échelle de temps (âge ou instant du calendrier) pour laquelle le taux d'incidence est calculé en appliquant l'expression (1), faire une telle distinction est nécessaire afin de calculer l'estimateur dans le contexte des données. En se référant à l'exemple 3 du chapitre 1, on voit qu'il n'est pas nécessaire de connaitre l'échelle de temps pour calculer le taux d'incidence. Par contre, connaitre l'échelle de temps signifie que l'on sait que les taux d'incidence ont été calculés par intervalle d'âge. Dans ce contexte, on peut voir que certains intervalles d'âge peuvent avoir des taux plus élevés que d'autres intervalles. Cette précision est également nécessaire à tout développement théorique concernant les taux d'incidence. Aussi, comme nous le discutons dans la section 1.1.1, connaître l'échelle de temps permet de déterminer si la population est ouverte (dynamique) ou non.

Dans ce mémoire, en utilisant des techniques d'analyse de survie (Lawless, 2003), (Klein et Moeschberger, 2003) et (Kalbfleisch et Prentice, 2002), d'analyse de l'historique des événements (Aalen *et al.*, 2008) et de la modélisation multi-états (Beyersmann *et al.*, 2012) et (Cook et Lawless, 2018), nous abordons les points 1 à 9 ci-dessus.

La spécification de modèles oblige à examiner attentivement toute hypothèse concernant la population ou l'échantillonnage implicite dans le calcul des taux d'incidence ainsi que toute hypothèse utilisée pour approximer les personnes-temps au dénominateur. Dans les cas où le taux d'incidence est calculé pour un échantillon, une approche de modélisation permet d'examiner plus attentivement l'estimation de la variance de l'estimateur. Dans les manuels actuels d'épidémiologie, il existe une tendance à présenter les taux d'incidence de manière informelle (voir (Rothman *et al.*, 2008) et (Jewell, 2003)). Certains auteurs (voir (Clayton et Hills, 1993), (Selvin, 2004) et (Selvin, 2008)) proposent une approche plus mathématique. Ici, nous étudions le taux d'incidence de l'expression (1) à l'aide des modèles d'analyse de survie et d'analyse de l'historique des événements. Parmi les manuels antérieurs traitant du taux d'incidence, les écrits de (Selvin, 2004) et de (Selvin, 2008) nous semblent les plus pertinents. Ces deux travaux de Selvin seront le point de départ pour le chapitre 2 de ce mémoire.

Nous empruntons deux chemins principaux. Le premier suit l'approche de (Selvin, 2004) et de (Selvin, 2008) pour des fonctions de risque qui ne sont pas nécessairement constantes. La deuxième approche considère des situations de modélisation plus complexes et utilise des modèles simples de Markov homogènes (et donc stationnaires) à états finis (également appelés modèles multi-états). Dans ces modèles, la fonction de risque est constante. À notre connaissance, ces détails n'apparaissent pas dans la littérature épidémiologique concernant les taux d'incidence.

Il convient de noter que le livre de (Rothman *et al.*, 2008) évoque la modélisation multi-états dans son texte, mais cette modélisation n'y est pas développée explicitement. L'application des modèles multi-états au taux d'incidence rappelle les travaux effectués par (Coeurjolly *et al.*, 2012) qui utilisent un modèle multi-états (modèle de Markov homogène continu) pour modéliser le risque attribuable.

0.2 Aperçu du mémoire

Ce mémoire est composé de ce chapitre introductif et de quatre autres chapitres. Le chapitre 1 présente des exemples spécifiques de la littérature épidémiologique qui motivent ce travail. Il fournit de nombreuses références aux concepts d'analyse de survie et aux modèles multi-états simples nécessaires au développement de notre travail. Le chapitre 2 se concentre sur l'approche de Selvin et sur la façon d'estimer les personnes-temps à risque. Le chapitre 3 restreint ensuite la fonction de risque à une constante et utilise des modèles simples issus de l'analyse de survie et des modèles multi-états pour examiner les exemples épidémiologiques donnés au chapitre 1. Enfin, des réflexions et des idées pour de futures investigations sont fournies dans la conclusion. Concernant la conclusion, il sera utile de revenir sur les observations faites ci-dessus à propos des points 1 à 9. Certains éléments importants que nous considérons dans le mémoire sont présentés ci-dessous dans les points 10 à 16.

10. Les points 3 et 9 sont liés, car s'il y avait eu plus de rigueur et que la modélisation des taux d'incidence avait été présentée, alors les différentes échelles de temps auraient dû être introduites et mises en évidence dans la littérature épidémiologique.
11. Lorsque la fonction de risque dans l'analyse de survie est constante, elle est égale au taux d'incidence.

12. Les populations fermées à gauche peuvent être modélisées avec une censure à droite. Les populations ouvertes peuvent être modélisées en utilisant une censure à droite et une troncature à gauche.

13. Lorsque les données de type expérimental introduites au point 6 sont disponibles, nous concevons qu'il y a probablement suffisamment de détails dans les données pour modéliser directement la fonction de risque ; il y a ainsi moins d'intérêt pour le calcul du taux d'incidence qui est davantage une mesure sommaire.

14. Nous concluons que les taux d'incidence sont plus pratiques pour les données de type recensement (voir le point 6). Dans cette situation, on ne peut pas modéliser la fonction de risque et les taux d'incidence deviennent plus intéressants. L'estimation de personnes-temps au dénominateur de l'expression (1) (voir le point 7) est essentielle.

15. Dans la litérature, peu d'attention est accordée à la possibilité d'avoir plusieurs événements par sujet (voir le point 8). Le fait de placer le taux d'incidence dans un cadre de modèles multi-états clarifie le concept de temps à risque. Pour les présentations voulant éviter beaucoup de détails mathématiques, il est intéressant de ne présenter que le schéma du modèle multi-états et de relier le taux d'incidence estimé à une intensité de transition constante pour entrer dans un état particulier.

16. Bien qu'il s'agisse en grande partie d'un exercice pédagogique, le modèle présenté doit être utile et explicite afin de mettre en évidence les hypothèses utilisées. Par exemple, dans la littérature épidémiologique, la censure (et/ou la troncature) *non informative* et *non indépendante* ne sont (à notre connaissance) jamais abordées dans les calculs des taux d'incidence. Comme nous le verrons, les deux sont implicites dans les modèles de ce mémoire.

CHAPITRE I

MATÉRIEL PRÉLIMINAIRE

Dans ce chapitre, nous présentons les notions d'épidémiologie, d'analyse de survie et de modélisation multi-états auxquelles nous faisons référence dans les chapitres futurs. Tout au long de la présentation, nous nous référons aux points 1 à 9 discutés dans la section 0.1 de l'introduction. À chaque fois, nous rappellons également la formule du taux d'incidence présentée à l'expression (1).

1.1 Épidémiologie

Dans cette section, nous élaborons deux sujets (points 4 et 5 de la section 0.1) mentionnés dans l'introduction, à savoir, le type de population ou d'échantillon et une présentation des exemples simples couramment trouvés dans les manuels d'épidémiologie. Un troisième sujet avec des implications importantes pour chacun des éléments ci-dessus est l'axe du temps (point 9 de la section 0.1). La discussion des populations et des échantillons dans la section 1.1.1 nous aidera à comprendre l'approximation des personnes-temps au dénominateur de l'expression (1) et sera nécessaire lorsqu'on essaie de construire un cadre théorique. Un objectif principal de ce mémoire est d'utiliser les exemples de la section 1.1.2 comme point de départ et de les étendre dans un contexte théorique pour mieux comprendre à la fois le concept de taux d'incidence tel qu'il est actuellement utilisé en épidémiologie et

les hypothèses implicites dans le calcul du taux d'incidence.

1.1.1 Types de population

Plusieurs de ces définitions peuvent être trouvées dans le texte de (Rothman *et al.*, 2008). Ces définitions s'appliquent à la fois aux populations et aux échantillons.

- Population fermée ou fixe : n'ajoute aucun nouveau membre au fil du temps et ne perd des membres que lorsqu'un événement d'intérêt survient (par exemple, un décès).

- Population fermée à gauche : toute population ou cohorte qui n'ajoute pas de nouveaux membres, mais peut en perdre avant que l'événement d'intérêt survienne (par exemple, perte de vue avant le décès). Dans une *cohorte*, l'appartenance est fixée, généralement par un événement déterminant. Par exemple, une cohorte de naissance est la cohorte définie en partie par le fait d'être né(e) à un moment donné (par exemple, toutes les personnes nées au Québec en 2010 constituent la cohorte de naissance québécoise pour 2010).

- Population ouverte : c'est une population ou cohorte qui peut ajouter et perdre des membres au fil du temps. Une population ouverte est également appelée *population dynamique*. Par exemple, des membres peuvent arriver en raison de la naissance et de l'immigration et partir en raison de la migration et du décès. Une population *stationnaire* ou une population en *état stable* est une population où le nombre de personnes entrant dans la population est équilibré par le nombre de personnes quittant la population au cours d'une période quelconque dans les limites des niveaux pertinents de certaines covariables comme l'âge et le sexe. Par exemple, de nombreuses sous-populations définies par des covariables telles que le sexe et le groupe sanguin peuvent être considérées à l'état d'équilibre que pendant de courtes périodes.

(Rothman *et al.*, 2008) précise que l'ouverture ou la fermeture d'une population ou d'un échantillon dépend de l'échelle de temps utilisée. Par exemple, toutes les personnes qui ont déjà consommé une drogue particulière constitueraient une population fermée si le temps était mesuré depuis le début de leur consommation de la drogue. Par contre, ces personnes constitueraient cependant une population ouverte dans le temps calendaire, car de nouveaux utilisateurs pourraient s'accumuler sur une période de temps. Ce point est souvent omis dans les discussions ailleurs.

1.1.2 Exemples utilisés pour illustrer le calcul du taux d'incidence

Dans cette sous-section, nous présentons des exemples utilisés dans les présentations standards de l'épidémiologie. Ces types d'exemple sont un point de départ de ce mémoire où l'objectif est d'utiliser des techniques d'analyse de survie et d'analyse de l'historique des événements, à savoir les modèles multi-états, pour approfondir les cadres théoriques possibles qui peuvent être développés pour les calculs de taux d'incidence.

La discussion suivante est basée sur les ressources populaires en épidémiologie entre autres : (Clayton et Hills, 1993), (Rothman *et al.*, 2008), (Dicker *et al.*, 2012), (Jewell, 2003) et (Breslow et Day, 1980). Il ne s'agit pas d'une liste exhaustive, mais simplement d'un échantillon représentatif de la manière dont les taux d'incidence sont présentés en épidémiologie et de fournir un point de départ pour la discussion dans ce mémoire. Le livre de (Breslow et Day, 1980) est moins récent, mais n'en est pas moins un livre classique.

Tout au long de la discussion, nous insistons sur le type de population ou d'échantillon et l'échelle de temps utilisée. Ces points importants sont souvent omis dans les présentations simples des taux d'incidence.

Les exceptions sont (Rothman *et al.*, 2008) (comme discuté dans la section 1.1.1) et (Jewell, 2003) qui répertorient les quatre échelles de temps suivantes : 1) l'âge ; 2) le temps écoulés depuis l'exposition à un facteur de risque spécifique ; 3) le temps du calendrier et 4) le temps écoulé depuis le diagnostic. À l'exception de l'échelle de temps en 3), toutes les échelles de temps susmentionnées peuvent être modélisées intuitivement à l'aide de l'analyse de survie. Habituellement, la première étape consiste à aligner tous les temps de survie sur un axe horizontal commun avec zéro représentant soit l'âge zéro, soit le temps d'exposition ou le temps de diagnostic. Le temps de survie de l'événement d'intérêt est le temps positif qui part de l'origine à l'événement. Lors de la modélisation partant de l'analyse de survie, il est plus facile d'imaginer une cohorte puisque la taille de la population ou de l'échantillon est fixe.

Dans (Clayton et Hills, 1993), (Rothman *et al.*, 2008), (Jewell, 2003) et (Breslow et Day, 1980), l'analyse de survie est mentionnée et évoquée à des degrés divers. (Clayton et Hills, 1993) ne mentionne pas le taux d'incidence et aucun de (Rothman *et al.*, 2008), (Jewell, 2003) et (Breslow et Day, 1980) n'entre en détail sur l'analyse de survie et les taux d'incidence que nous présentons dans ce mémoire. Nous reportons la discussion de (Selvin, 2004) et (Selvin, 2008) au chapitre 2.

Trois schémas sont inclus dans (Rothman *et al.*, 2008) pour illustrer les calculs de taux d'incidence. Un pour montrer différents modèles de mortalité, un deuxième pour illustrer l'hypothèse d'état stable et un troisième (voir l'exemple 1 ci-dessous) pour montrer une population fermée (point 5 de la section 0.1) suivie pendant une certaine période.

Exemple 1.

La figure 1 tirée de la figure 3-4 de (Rothman *et al.*, 2008) illustre un échantillon ou une population fermée suivie pendant 19 ans.

Bien que les unités de temps soient données sur l'axe horizontal, l'échelle de temps n'est pas définie concrètement comme l'âge. Le taux d'incidence est de 5 cas divisé par $2 + 2(4) + 8 + 14 + 4(19) = 108$ personnes-années ; soit 0.0462963 (personnes-années)$^{-1}$.

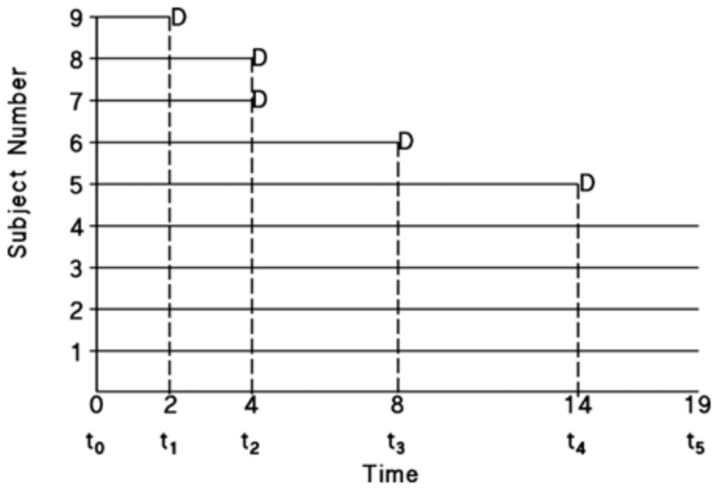


Figure 3-4 • Example of a small closed population with end of follow-up at 19 years.

Figure 1.1 Schéma (tiré de la figure 3-4 de (Rothman *et al.*, 2008)) illustrant une population fermée de 9 sujets suivis pendant 19 ans.

En plus de ce que nous avons énuméré ci-dessus, (Jewell, 2003) aborde les événements multiples possibles (voir le point 8 de la section 0.1) et leur effet sur le calcul des personnes-années. (Rothman *et al.*, 2008) mentionne également la possibilité d'événements multiples, mais il restreint ensuite la discussion aux premières occurrences. Dans l'exemple 2 ci-dessous, nous présentons le premier exemple de (Jewell, 2003) qui illustre très bien le contraste dans les calculs des personnes-années lorsqu'il y a présence d'événements multiples contre lorsqu'il n'y a aucun événement multiple.

Exemple 2.

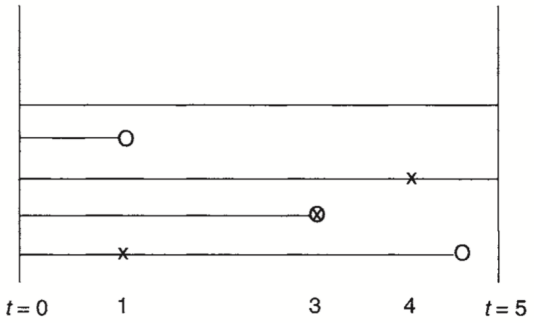


Figure 2.2 *Schematic illustrating calculation of an incidence proportion, point prevalence, and incidence rate. Population of 5; the symbols represent: O, death; x, incident case of disease. Here, lines represent time alive.*

Figure 1.2 Schéma (tiré de la figure 2.2 de (Jewell, 2003)) illustrant un échantillon ou une population fermée à gauche de 5 sujets. Les symboles dans ce schéma représentent : O, la mort et x, cas incident de la maladie. Les lignes horizontales représentent le temps de survie.

En particulier, l'exemple de la figure 2 (tiré de la figure 2.2 de (Jewell, 2003)) souligne que le calcul des personnes-années au dénominateur dépend du concept d'*être à risque*. D'une part, si les personnes se rétablissent immédiatement et redeviennent de nouveau à risque, le taux d'incidence est égal à 0.16 (soit 3 cas divisé par 5+1+5+3+4.5 = 18.5 personnes-années). D'autre part, si les personnes ne sont plus jamais à risque (par exemple, elles développent une immunité) ou si elles sont à risque pour un événement différent, il y a moins des personnes-années au dénominateur et le taux d'incidence est plutôt égal à 0.21 (soit 3 cas divisé par $5 + 1 + 4 + 3 + 1 = 14$ personnes-années).

L'échelle de temps n'est pas explicitement définie, cependant, puisque tous les temps de survie sont alignés à $t = 0$, l'implication dans (Jewell, 2003) est que l'axe x est soit l'âge, soit le temps après l'exposition ou le temps après le diagnostic.

Puisque dans chaque cas, il y a un événement déterminant à $t = 0$ (être né, être exposé ou être diagnostiqué), les individus représentés dans la figure 2 forment une cohorte.

(Breslow et Day, 1980) fournit l'exemple 3 ci-dessous où l'échelle de temps est l'âge (voir le point 9 de la section 0.1), mais la population ou l'échantillon est dynamique (voir le point 5 de la section 0.1) ce qui implique que les individus entrent et sortent à des âges différents.

Exemple 3.

Le taux d'incidence calculé concerne le diagnostic de cancer. Cet exemple est tiré de la figure 2.1 de (Breslow et Day, 1980) et est répliqué dans notre figure 3 ci-dessous. Notez que les gens entrent dans l'étude à des âges différents.

(Breslow et Day, 1980) ont un certain nombre de discussions intéressantes. Premièrement, ils soulignent qu'un intervalle de temps court pour le calcul du taux d'incidence conduit à un dénominateur plus stable. Ils disent que cela est dû à la dynamique changeante de la population en raison des naissances, des décès et des migrations (voir le point 7 de la section 0.1). Bien qu'elle ne soit pas signalée explicitement par (Breslow et Day, 1980), l'observation ci-dessus joue un rôle important dans l'estimation des personnes-temps au dénominateur et cela n'est pas un problème pour l'estimateur présenté à l'expression (1) lui-même. Une deuxième raison invoquée pour avoir un intervalle de temps court est qu'un intervalle de temps plus long passerait à côté de tout changement rapide du taux d'incidence instantané. À noter que cette seconde raison touche à la différence entre le taux d'incidence instantané (essentiellement la fonction de risque) et le taux moyen sur la période définie pour le calcul de l'expression (1). Le tableau 2.1 de (Breslow et Day, 1980) fournit des comparaisons des taux d'incidence calculés avec la vraie

Fig. 2.1 Schematic illustration of age-specific incidence rates. (D = diagnosis of cancer; W = withdrawn, disease free.)

	Period 1	Period 2	Period 3
Number of events	2	4	3
Total observation time	59.7	93.5	43.6
rate	0.034	0.043	0.069

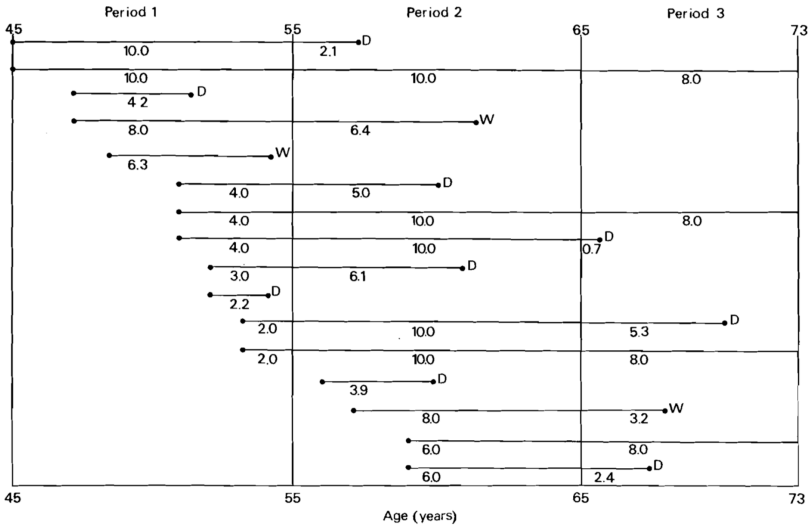


Figure 1.3 Schéma (tiré de la figure 2.1 de (Breslow et Day, 1980)) illustrant une population dynamique dont l'échelle de temps est l'âge.

valeur des personnes-jours et en utilisant également une approximation médiane des personnes-jours. Le tableau 2.2 dans (Breslow et Day, 1980) présente les taux d'incidence calculés en utilisant une approximation différente pour les personnes-années et explique également comment une approximation plus précise pourrait être utilisée.

(Breslow et Day, 1980) et (Boyle et Parkin, 1996) discutent des taux d'incidence du cancer. Les points pratiques suivants se dégagent. Le premier concerne le point 2 dans la section 0.1. Naturellement, il est souhaitable de ne pas avoir des taux d'incidence ni trop grands ni trop petits ; par conséquent, comme le cancer est une maladie assez rare, les unités de temps sont souvent pour 100 000 personnes-années lors de l'estimation des taux d'incidence du cancer.

Le deuxième point pratique est qu'on a généralement le nombre de cas de cancer pour une population donnée, mais en raison de problèmes financiers et logistiques, il n'y a pas de suivi détaillé de la population elle-même. La conséquence est que les personnes-années à risque ne peuvent pas être calculées exactement et doivent être estimées à partir des données de recensement (d'où le terme de données de type recensement). De plus, pour certaines maladies, telles que le cancer de l'ovaire, seules certaines personnes sont à risque, à savoir les femmes qui n'ont pas subi d'hystérectomie. Cette identification des femmes dans une population n'est pas toujours disponible. Dans la pratique, nous aurons comme résultat, l'inclusion de certaines personnes au dénominateur qui ne sont pas réellement à risque. Voir le point 6 de la section 0.1 qui mentionne les données de *type recensement* par rapport aux données de *type expérimental*.

(Dicker *et al.*, 2012) ont quatre exemples de taux d'incidence et peut-être les plus variés. Tous leurs exemples utilisent l'échelle de temps du calendrier. Sur les quatre, dans tous sauf un, les personnes-années doivent être approximées et dans le seul exemple où les personnes-années ne sont pas approximées, elles sont fournies. Dans l'exemple 4 ci-dessous, nous montrons l'exemple venant de la figure 3.1 de (Dicker *et al.*, 2012) reproduit dans la figure 4 de ce mémoire.

Exemple 4.

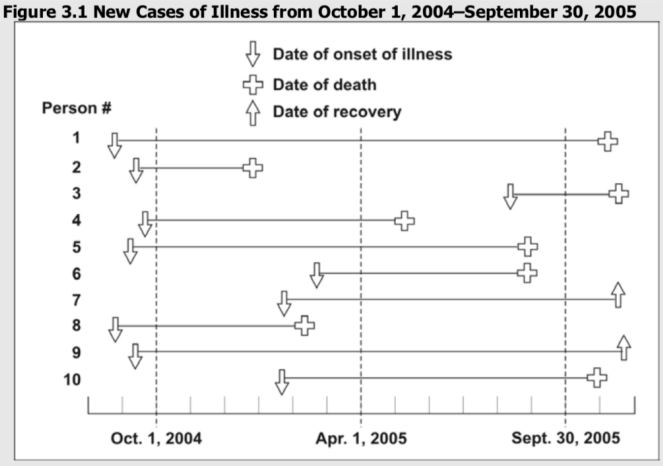
La figure 4 (tirée de la figure 3.1 de (Dicker *et al.*, 2012)) a une échelle de temps du calendrier. La question demande d'estimer le taux d'incidence de la maladie. Le lecteur est censé estimer les personnes-années à partir du graphique en utilisant une approximation médiane au 1er avril 2005. On suppose qu'il y a une population ou un échantillon de 20 personnes au total. Les temps de maladie pour dix des 20 personnes qui sont tombées malades sont illustrés. Les dix autres (non représentés) sont supposés être sains pendant tout l'intervalle d'étude.

Étant donné que deux personnes sont décédées avant le point médian, soit le 1er avril 2005 et que l'intervalle d'étude est d'un an, il y a 18 personnes-années dans l'approximation du point médian du dénominateur. Au cours de l'intervalle d'un an du 1er octobre 2004 au 30 septembre 2005, il y a eu quatre nouveaux cas de maladie (représentés par des flèches vers le bas). Par conséquent, le taux d'incidence de la maladie est estimé à 4/18 (personnes-années)$^{-1}$. Un point subtil qui n'est pas abordé dans l'exemple est le fait qu'une fois que les gens sont malades, on suppose généralement qu'ils ne peuvent pas redevenir malades avant d'avoir récupéré. Les personnes-années dans le calcul du taux d'incidence à l'expression (1) sont des années à risque.

EXAMPLES: Incidence versus Prevalence

Figure 3.1 represents 10 new cases of illness over about 15 months in a population of 20 persons. Each horizontal line represents one person. The down arrow indicates the date of onset of illness. The solid line represents the duration of illness. The up arrow and the cross represent the date of recovery and date of death, respectively.

Figure 3.1 New Cases of Illness from October 1, 2004–September 30, 2005



Example A: *Calculate the incidence rate from October 1, 2004, to September 30, 2005, using the midpoint population (population alive on April 1, 2005) as the denominator. Express the rate per 100 population.*

Incidence rate numerator	= number of new cases between October 1 and September 30 = 4 (the other 6 all had onsets before October 1, and are not included)
Incidence rate denominator	= April 1 population = 18 (persons 2 and 8 died before April 1)
Incidence rate	= (4 / 18) x 100 = 22 new cases per 100 population

Figure 1.4 Exemple de calcul du taux d'incidence à partir de la figure 3.1 de (Dicker *et al.*, 2012) où l'axe des abscisses représente l'instant du calendrier.

Dans l'exemple 5 ci-dessous, nous montrons un exemple tiré de (Dicker *et al.*, 2012) qui n'a pas de figure, mais fournit suffisamment d'informations pour calculer un taux d'incidence. Notez qu'il n'y a pas de discussion quant à savoir si la population est dynamique ou non (voir le point 5 de la section 0.1) ; cependant, les personnes-années sont estimées à partir de la population en milieu d'année (voir le point 7 de la section 0.1). Ce type d'estimations sera discuté dans le chapitre 2.

Exemple 5.

Un deuxième exemple tiré de (Dicker *et al.*, 2012).

EXAMPLES: Calculating Incidence Rates (Continued)

Example C: *In 2003, 44,232 new cases of acquired immunodeficiency syndrome (AIDS) were reported in the United States.[5] The estimated mid-year population of the U.S. in 2003 was approximately 290,809,777.[6] Calculate the incidence rate of AIDS in 2003.*

Numerator	= 44,232 new cases of AIDS
Denominator	= 290,809,777 estimated mid-year population
10n	= 100,000
Incidence rate	= (44,232 / 290,809,777) x 100,000
	= 15.21 new cases of AIDS per 100,000 population

Figure 1.5 Un exemple de calcul du taux d'incidence tiré de (Dicker *et al.*, 2012). L'axe des abscisses représente l'instant du calendrier. Aucune figure n'est fournie pour cet exemple.

Nous avons choisi de présenter de nombreux exemples avec leurs figures originales. À l'exception de l'exemple 4, toutes les figures des exemples donnés ci-dessus illustrent les temps de survie ou les durées de la maladie d'intérêt par des lignes horizontales.

Dans chaque exemple, nous avons présenté une copie exacte d'une figure de la littérature épidémiologique. Cela a été fait délibérément pour que le lecteur sache exactement quelles informations ont été fournies et quelles hypothèses sont généralement formulées, le cas échéant.

Une grande contribution de ce mémoire est de considérer différents modèles théoriques pour chaque situation tout en mettant l'accent sur les hypothèses de chaque modèle. L'objectif de ce mémoire est de considérer différents paramètres pour l'estimateur du taux d'incidence tout en utilisant des techniques d'analyse de survie et d'analyse de l'historique des événements.

1.2 Analyse de l'historique des événements

Dans cette section, nous introduisons brièvement les sujets de l'analyse de survie et des modèles multi-états tout en mettant particulièrement l'accent sur le matériel nécessaire dans la suite de notre travail. (Klein et Moeschberger, 2003), (Lawless, 2003), (Kalbfleisch et Prentice, 2002) sont de très bonnes références pour l'analyse de survie et (Aalen *et al.*, 2008) et (Cook et Lawless, 2018) sont de bonnes références pour les modèles multi-états. Lorsque cela s'avère nécessaire, nous renvoyons le lecteur aux démonstrations dans les références mentionnées au début de cette section 1.2 et en appendice plutôt que de les reproduire ici dans le corps de notre travail.

1.2.1 Analyse de survie

L'analyse de survie est une discipline statistique qui analyse des données en considérant des valeurs positives. Le nom d'analyse de survie survient parce que souvent ces valeurs positives représentent les *durées de vie* complètes d'individus ou *temps de survie* d'individus atteints d'une maladie particulière. En général, on note une variable aléatoire positive T pour représenter les durées de vie. Ici, on suppose que les durées de vie des sujets sont considérées comme identiquement distribuées et indépendantes (iid) ; donc chaque sujet i a une durée de vie représentée par T_i.

Dans de nombreuses présentations, l'indicateur i est supprimé et, le cas échéant, nous faisons de même ici.

Nous commençons l'introduction par un échantillon ou une population fermée. Pour simplifier notre présentation, nous supposons que T est une variable aléatoire continue. On note la densité de la variable aléatoire T par $f(t)$. Deux autres fonctions très utiles dans l'analyse de survie sont la fonction de survie $S(t)$ et la fonction de risque $h(t)$. La fonction de survie est simplement la probabilité qu'une durée de vie soit supérieure à une valeur t :

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

De toute évidence, la fonction de survie est égale à un moins la fonction de répartition $F(t) = P(T \leq t)$. En effet, lorsqu'ils considèrent une période de temps $[0, T]$ en l'absence de censure et de troncature, les épidémiologistes se réfèrent à l'estimation empirique $\hat{F}(T) = \frac{\#\text{ de cas}}{\#\text{ total}}$ comme incidence cumulative et utilisent souvent la notation $I(t)$. En revenant à l'exemple 1 nous voyons que l'incidence cumulée à t_2 est $\hat{F}(t_2) = P(T \leq t_2) = 3/9 = 1/3$.

La fonction de risque joue un rôle central à la fois dans l'analyse de survie et dans ce mémoire. Elle est définie comme :

$$h(t) = \lim_{dt \to 0} \frac{P(T \in [t, t + dt) | T \geq t)}{dt}. \tag{1.1}$$

La fonction de risque est également connue sous le nom de taux d'incidence instantané. Il est facile de voir lorsque l'on considère une population fermée à gauche et qu'on laisse tomber la limite $dt \to 0$ tout en considérant dt petit que la fonction de risque peut s'écrire comme :

$$h(t) \approx \frac{P(T \in [t, t + dt) | T \geq t)}{dt}.$$

Nous interprétons le numérateur $P(T \in [t, t+dt)|T \geq t)$ par

$$\hat{P}(T \in [t, t+dt)|T \geq t) = \frac{\#\text{ d'événements dans l'intervalle } [t, t+dt)}{\#\text{ de personnes à risque dans l'intervalle } [t, \infty)}.$$

D'où,

$$\hat{h}(t) = \frac{\#\text{ d'événements dans l'intervalle } [t, t+dt)}{[\#\text{ de personnes à risque dans l'intervalle } [t, t+dt)] \times dt}$$

puisque le nombre (#) de personnes à risque dans l'intervalle $[t, \infty)$ est le même que le nombre (#) de personnes à risque dans l'intervalle $[t, t+dt)$ pour une population fermée à gauche.

Notons que le dénominateur approxime (surestime légèrement) les personnes-années dans l'intervalle $[t, t+dt)$, ce qui fait que $\hat{h}(t)$ est approximativement égal au taux d'incidence lorsque l'intervalle $[t, t+dt)$ ou dt est petit.

Lorsque le temps de survie T est continu, on montre facilement que $h(t) = \frac{f(t)}{S(t)}$ ce qui conduit à $S(t) = \exp(-\int_0^t h(u)du)$. La distribution exponentielle $f(t) = \lambda \exp(-\lambda t)$ est la seule distribution continue pour T qui conduit à une fonction de risque constante, $h(t) = \lambda$.

1.2.2 Données incomplètes et échantillonnage biaisé

Un point fort de l'analyse de survie est sa capacité à gérer certains types de données incomplètes et de biais de sélection. Le type de données incomplètes le plus courant est celui des données censurées à droite. Cela se produit, par exemple, lorsque des individus sont suivis longitudinalement et deviennent *perdus de vue*. La censure à droite intervient également à la fin d'une étude ou d'une période d'observation.

Les individus ayant des temps de survie plus longs ont plus de chance d'être perdus de vue et, par conséquent, ignorer les données des individus censurés à droite entraîne une sous-estimation du temps de survie et conduit à un biais de sélection appelé *troncature à droite* (où l'on sous-estime les temps de survie). Par conséquent, pour éviter ce biais de sélection, l'analyse de survie comprend des méthodes pour analyser les temps de survie incomplets censurés à droite ainsi que les temps de survie complètement observés.

La figure 1 dans l'exemple 1 fournit un exemple simple de censure à droite où les sujets 1 à 4 ont vécu plus longtemps que le temps d'étude. Mathématiquement, la censure à droite peut être représentée par une variable aléatoire de censure C (encore une fois, nous supprimons l'indice du sujet i). Ainsi, en laissant T représenter le vrai temps de survie, la donnée réellement observée est le temps de survie observé $X = \min(T, C)$ avec un indicateur $\delta = I(T \geq C)$. Cette notation est souvent utilisée pour le type de censure à droite le plus simple, appelé *censure aléatoire* où la variable aléatoire T est statistiquement indépendante de la variable aléatoire C.

Un autre type de biais de sélection est appelé *troncature à gauche* ou *entrée retardée* (*delayed entry* en anglais). Contrairement à la troncature à droite où il y a une tendance à manquer les temps de survie plus longs, la troncature à gauche est un biais de sélection où certains individus qui ont tendance à avoir les temps de survie plus courts sont omis. La figure 3 dans l'exemple 3 a de nombreux sujets avec une entrée différée. Pour représenter mathématiquement la troncature à gauche, nous introduisons une variable aléatoire de troncature à gauche Y indiquant la différence entre l'origine de temps de survie et le temps de survie auquel l'observation a commencé. Dans l'exemple 3, chaque sujet a un temps de troncature à gauche égal à l'âge auquel il est entré dans la période d'observation.

En l'absence de censure à droite, le temps de survie T_i pour chaque sujet i est observé conditionnellement à l'événement $\{T_i > Y_i\}$. En présence d'une censure à droite, on suppose généralement qu'un sujet ne peut pas être censuré tant qu'il n'est pas en observation. Pour cette raison, C et Y ne sont pas indépendants puisque $P(C > Y) = 1$. Sous censure à droite, le temps de survie observé $X_i = \min(T_i, C_i)$ pour chaque sujet i est observé conditionnellement à l'événement $\{X_i > Y_i\}$.

Pour la censure et la troncature ci-dessus, il est courant de supposer qu'elles sont *non informatives* et *indépendantes*. (Kalbfleisch et Prentice, 2002) ont une bonne discussion de ces concepts et (Lawless, 2003) discute l'hypothèse d'*indépendance* en relation avec la troncature à gauche. En ce qui concerne l'hypothèse d'*indépendance* pour la censure, il est important de comprendre qu'il s'agit d'une hypothèse plus faible que celle de la censure aléatoire où T et C sont statistiquement indépendants. Nous utilisons ces hypothèses tout au long du mémoire. Par souci de concision, nous renvoyons le lecteur aux références ci-dessus et nous ne les présentons pas explicitement dans ce travail, cependant elles sont implicites dans chaque vraisemblance que nous présentons.

En ce qui concerne les vraisemblances en général, la manipulation la plus simple en statistique se fait lorsque nous avons n variables aléatoires indépendantes et identiquement distribuées (iid) et des valeurs t_i complètement observées. La vraisemblance (1.2) ci-dessous est exprimée en termes d'une densité f qui décrit la distribution de n variables aléatoires T_i (iid) dans la population :

$$L \propto \prod_{i=1}^{n} f(t_i). \tag{1.2}$$

Dans la suite de ce mémoire, nous considérons les différentes vraisemblances qui surviennent dans l'analyse de survie en présence d'une censure à droite, un type de données incomplètes, et d'une troncature à gauche, un type de biais de sélection.

Au chapitre 3, nous considérons les différentes situations de données motivées par les exemples de la section 1.1.2 avec l'hypothèse d'une fonction de risque constante. Pour chaque situation d'analyse de survie ci-dessous, nous exprimons à la fois la vraisemblance en termes d'une densité générale f et également en termes de la densité exponentielle $f(t) = \lambda \exp(-\lambda t)$ pour n individus iid.

1.2.3 Vraisemblances dans l'analyse de survie : cas de non-censure à droite et de non-troncature à gauche

En analyse de survie, on peut observer des temps sans censure ni troncature. Dans ce cas, la vraisemblance des données complètes (1.2) est la vraisemblance appropriée à utiliser. Pour les données distribuées de façon exponentielle, nous avons :

$$\begin{aligned} L &\propto \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) \\ &= \lambda^n \exp(-\lambda \sum_{i=1}^{n} t_i). \end{aligned} \tag{1.3}$$

1.2.4 Vraisemblances dans l'analyse de survie : cas de censure à droite et de non-troncature à gauche

Supposons un échantillon de taille n avec des temps observés $X_i = \min(T_i, C_i)$. Nous utilisons l'indicateur $\delta_i = I(T_i \leq C_i)$ pour indiquer la censure à droite. Les lettres minuscules x_i sont utilisées pour représenter les temps observés réalisés.

En supposant que nous avons une censure indépendante et non informative, on a :

$$L \propto \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}. \tag{1.4}$$

Les valeurs qui sont observées exactement contribuent à la vraisemblance par f et les valeurs censurées à droite contribuent par S. Sur la base d'une loi exponentielle, la vraisemblance (1.4) devient :

$$\begin{aligned} L &\propto \prod_{i=1}^{n} (\lambda \exp(-\lambda x_i))^{\delta_i} (\exp(-\lambda x_i))^{1-\delta_i} \\ &= \lambda^{\sum_{i=1}^{n} \delta_i} \exp(-\lambda \sum_{i=1}^{n} x_i). \end{aligned} \tag{1.5}$$

1.2.5 Vraisemblances dans l'analyse de survie : cas de censure à droite et de troncature à gauche

Comme mentionné ci-dessus, dans le cas de la troncature à gauche et de la censure à droite, nous conditionnons sur l'événement $\{X_i > Y_i\}$ impliquant que le temps de survie est supérieur au temps de troncature. En supposant la censure et la troncature non informatives ainsi qu'en supposant l'indépendance, nous avons la vraisemblance conditionnelle aux n événements $\{X_i > Y_i\}$ et les temps de troncature réalisés y_i :

$$L \propto \prod_{i=1}^{n} \left(\frac{f(x_i)}{S(y_i)} \right)^{\delta_i} \left(\frac{S(x_i)}{S(y_i)} \right)^{1-\delta_i}. \tag{1.6}$$

En termes d'une loi exponentielle, nous avons les expressions de la densité $f(x_i|X_i > Y_i, y_i) = \frac{\lambda \exp(-\lambda x_i)}{\exp(-\lambda y_i)}$ et celle de la fonction de survie $S(x_i|X_i > Y_i, y_i) = \frac{\exp(-\lambda x_i)}{\exp(-\lambda y_i)}$.

D'où, l'expression (1.6) devient :

$$\begin{aligned} L &= \prod_{i=1}^{n} f(x_i|X_i > Y_i, y_i)^{\delta_i} S(x_i|X_i > Y_i, y_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda^{\delta_i} \left(\frac{\exp(-\lambda x_i)}{\exp(-\lambda y_i)} \right) \\ &= \prod_{i=1}^{n} \lambda^{\delta_i} \left(\exp(-\lambda(x_i - y_i)) \right) \\ &= \lambda^{\sum_{i=1}^{n} \delta_i} \left(\exp(-\lambda \sum_{i=1}^{n} (x_i - y_i)) \right). \end{aligned} \tag{1.7}$$

En résumé, les données complètes, les données censurées à droite et les données tronquées à gauche sont les situations de données qui relèvent de l'analyse de survie que nous considérons dans ce mémoire.

Certains exemples de la section 1.1.2 justifient des modèles plus complexes et nous présentons donc quelques techniques pertinentes de l'analyse de l'historique des événements dans la section ci-dessous. Il est important de noter que le domaine de l'analyse de survie est contenu dans celui l'analyse de l'historique des événements, car il traite des modèles multi-états les plus simples ayant juste deux états, vivant et mort, où mort est un état absorbant.

1.3 Deux modèles multi-états et un modèle d'événements récurrents

Nous présentons brièvement deux modèles multi-états simples et un modèle d'événements récurrents sur lesquels nous revenons au chapitre 3 lorsque nous considérons des modèles pour traiter les exemples de la section 1.1.2. Les modèles multi-états sont basés sur des chaînes de Markov en temps continu avec un petit espace d'état.

Nous considérons ici les plus simples de ces modèles, à savoir ceux qui sont homogènes dans le temps. Chaque modèle multi-états est généralement présenté avec une figure simple pour indiquer les états qui peuvent être occupés et les transitions possibles entre ces états. Les bonnes références pour les chaînes de Markov en temps continu sont (Grimmett et Stirzaker, 2001) et (Cox et Miller, 1965). Les deux modèles multi-états que nous considérons ici sont le modèle du *processus de renouvellement alterné* et le modèle *maladie-décès* (*illness-death* en anglais). Voir les figures 1.6 et 1.7.

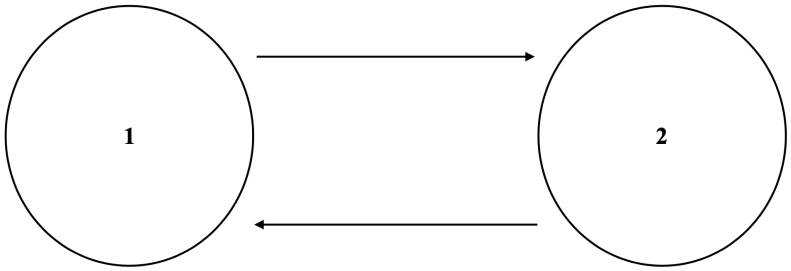


Figure 1.6 Schéma d'un modèle du processus de renouvellement alterné. Les cercles 1 et 2 représentent respectivement l'état 1 (état sain) et l'état 2 (état malade) qu'un sujet peut occuper.

Dans la section 1.3.3, nous présentons également un modèle simple d'événements récurrents basé sur une loi de Poisson.

Soit $X(t)$ l'état du modèle au temps t, et j et h deux états dans l'espace d'état $S = \{1, \ldots, N\}$. De plus, soit $P_{hj}(s,t) = P(X(t) = j | X(s) = h, H_s-)$ la probabilité de transition de h au temps s vers j au temps t ($s < t$), conditionnellement à l'historique juste avant le temps s, H_s-. Puisque nous avons un modèle de Markov, $P_{hj}(s,t) = P(X(t) = j | X(s) = h)$; c'est-à-dire la probabilité de transition ne dépend que du dernier état visité et non de l'historique H_s-.

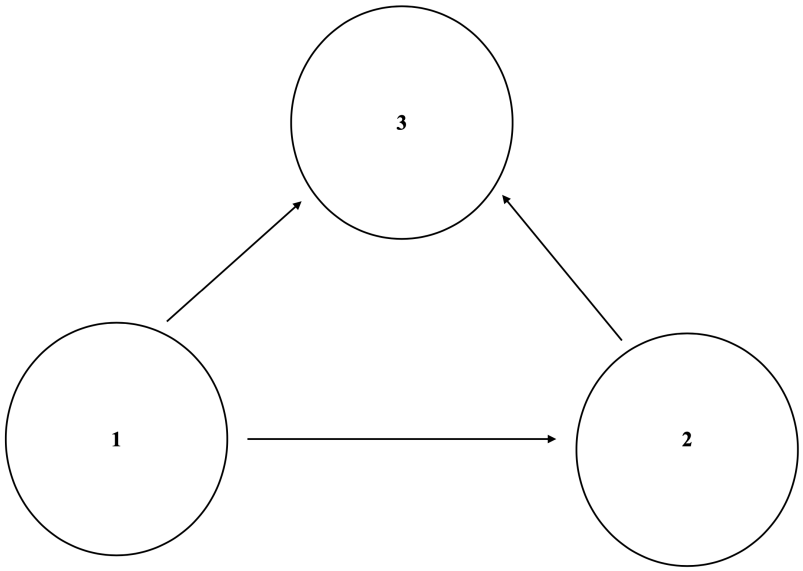


Figure 1.7 Schéma d'un modèle *maladie-décès* (*illness-death* en anglais). Les cercles 1, 2 et 3 représentent respectivement l'état 1 (état sain), l'état 2 (état malade) et l'état 3 (état décès) qu'un sujet peut occuper.

L'intensité de transition pour une transition $h \to j$ est :

$$\lambda_{hj}(t) = \lim_{dt \to 0} \frac{P_{hj}(t, t+dt)}{dt}. \tag{1.8}$$

Pour un modèle homogène dans le temps (stationnaire), on a $P_{hj}(s, t^*) = P_{hj}(0, t)$ où $t = t^* - s$ et les intensités de transition λ_{hj} sont constantes.

Dans notre présentation du processus de renouvellement alterné et du modèle *maladie-décès*, nous voyons que la fonction de risque de l'analyse de survie et l'intensité de transition, bien qu'elles sont non équivalentes, peuvent coïncider pour certaines transitions dans certains modèles multi-états.

Les intensités de transition constituent la matrice de *taux de transition* également connue sous le nom de matrice *génératrice* $A(t)$. Dans le cas de temps homogènes, puisque les intensités sont constantes, $A(t) = A$.

Les éléments de A sont notés (a_{ij}) pour la i^e ligne et la j^e colonne. La somme $\sum_{i=1}^{N} a_{ij} = 0$. Tous les éléments non diagonaux sont des intensités de transition qui sont égales ou supérieures à 0 et les éléments diagonaux sont inférieurs ou égaux à zéro. En résumé, pour tout état i, $\sum_{j\neq i, j=1}^{N} a_{ij} = -a_{ii}$.

Pour le modèle homogène dans le temps, les *équations directes* (*forward equations* en anglais) se simplifient en

$$\frac{dP_{ij}(0,t)}{dt} = \sum_{k=1}^{N} P_{ik}(0,t)a_{kj}. \tag{1.9}$$

À partir de ces équations différentielles, on peut trouver les probabilités de transition $P_{ij}(0,t)$.

Notez qu'il existe également des *équations inverses* (*backward equations* en anglais). Par souci de concision, nous ne présentons ici que les équations directes puisque ce sont les équations que nous utilisons pour présenter le modèle *maladie-décès* dans la suite de ce mémoire. Voir (Karlin et Taylor, 1981) pour plus d'informations et une dérivation des équations inverses.

Notons que puisque nous sommes dans le contexte du modèle de Markov homogène à temps continu, les distributions des temps de séjour (temps passés dans chaque état) dans les vraisemblances sont exponentielles.

1.3.1 Vraisemblances : Processus de renouvellement alterné

Pour la chaîne de Markov homogène à temps continu, les temps de séjour dans chaque état sont distribués de manière exponentielle. Ce modèle est très simple à écrire puisqu'étant dans l'état 1, la seule option est d'aller à l'état 2 et quand on est dans l'état 2, la seule option est d'aller à l'état 1.

Pour chaque individu i, nous notons la durée passée dans l'état 1 (c'est le temps de séjour passé dans l'état 1) comme x_{ij} et la durée passée dans l'état 2 (c'est le temps de séjour passé dans l'état 2) comme y_{ik} où $j = 1, 2, \ldots$ et $k = 1, 2, \ldots$.

Les distributions pour les temps complets dans chaque état sont :

$$x_{ij} \sim \exp(\lambda_{12})$$

et

$$y_{ik} \sim \exp(\lambda_{21}).$$

Notons que λ_{12} est l'intensité de transition de l'état $1 \to 2$ et λ_{21} est l'intensité de transition de l'état $2 \to 1$.

Soit τ_i le nombre total de transitions effectuées par la personne i dans l'intervalle d'observation de longueur T. Nous considérons des situations où tous les sujets commencent à l'état 1 (par exemple *état sain*), donc si τ_i est impair alors la séquence du temps se termine par x_i et si τ_i est pair alors la séquence du temps se termine par y_i. Lorsque tous les individus commencent à l'état 1 au temps $t = 0$, la vraisemblance est :

$$L \propto \prod_{i=1}^{n} \left[\lambda_{12}^{\lceil \frac{\tau_i}{2} \rceil} \left\{ \prod_{j=1}^{\lceil \frac{\tau_i}{2} \rceil + I(\tau_i = \text{pair})} \exp(-\lambda_{12} x_{ij}) \right\} \lambda_{21}^{\lceil \frac{\tau_i}{2} \rceil - I(\tau_i = \text{impair})} \left\{ \prod_{k=1}^{\lceil \frac{\tau_i}{2} \rceil} \exp(-\lambda_{21} y_{ik}) \right\} \right]. \tag{1.10}$$

Le symbole $\lceil \frac{\tau_i}{2} \rceil$ est la partie entière supérieure du nombre réel $\frac{\tau_i}{2}$ (positif ou nul).

Lorsque τ_i représente le nombre d'événements (le nombre de fois où la personne i est passée de l'*état sain* vers l'*état malade*) ou encore lorsque τ_i est impair, la dernière observation pour l'individu i est y_{ik} pour une certaine valeur de k et est *coupée* ou censurée à droite à la fin de l'intervalle d'observation de longueur T.

Lorsque τ_i est pair (ou lorsque τ_i représente le nombre de fois où la personne i est passée de l'*état malade* vers l'*état sain*), la dernière observation est x_{ij} pour une certaine valeur de j et est *coupée* ou censurée à droite à la fin de l'intervalle d'observation de longueur T. Dans la vraisemblance (1.10) ci-dessus, nous utilisons les notations x_{ij} et y_{ik} pour les deux temps d'observation, à savoir, les temps complets et les temps censurés à droite. Dans l'analyse de survie classique, nous appelons ces temps comme étant les temps observés. La censure à droite est traitée comme étant indépendante et non informative.

Bien que cela ne soit pas nécessaire dans le développement de la vraisemblance ci-dessus, par souci d'exhaustivité, nous présentons les solutions des *équations directes* pour le processus de renouvellement alterné dans l'appendice A.

1.3.2 Vraisemblances : Modèle *maladie-décès*

Contrairement au processus de renouvellement alterné, une fois qu'un sujet quitte l'état 1, il ne peut pas y revenir et une fois qu'un sujet quitte l'état 2, il ne peut pas y revenir également. S'ils sont suivis assez longtemps, les sujets finissent par se retrouver dans l'état d'absorption 3. Par conséquent, dans le modèle *maladie-décès*, nous pouvons utiliser les *équations directes* afin de développer la vraisemblance pour les temps de séjour. On note x_i, y_i et z_i les temps de séjour dans les états 1, 2 et 3 respectivement. Nous utiliserons également les mêmes notations pour les temps de séjour censurés à droite.

Soient λ_{12}, λ_{13} et λ_{23} respectivement les intensités de transition de $1 \to 2$, de $1 \to 3$ et de $2 \to 3$. La matrice génératrice du modèle *maladie-décès* est :

$$A = \begin{bmatrix} -(\lambda_{12}+\lambda_{13}) & \lambda_{12} & \lambda_{13} \\ 0 & -\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 \end{bmatrix}. \tag{1.11}$$

Les équations directes sont :

$$\begin{aligned}
\frac{dP_{11}(0,t)}{dt} &= -(\lambda_{12}+\lambda_{13})P_{11}(0,t) \\
\frac{dP_{12}(0,t)}{dt} &= -\lambda_{23}P_{12}(0,t) + \lambda_{12}P_{11}(0,t) \\
\frac{dP_{13}(0,t)}{dt} &= \lambda_{13}P_{11}(0,t) \\
\frac{dP_{22}(0,t)}{dt} &= -\lambda_{23}P_{22}(0,t) \\
\frac{dP_{23}(0,t)}{dt} &= \lambda_{23}P_{22}(0,t) \\
\frac{dP_{33}(0,t)}{dt} &= 0.
\end{aligned}$$

En utilisant les conditions initiales $P_{11}(0,0) = 1$, $P_{22}(0,0) = 1$, $P_{33}(0,0) = 1$, $P_{12}(0,0) = 0$, $P_{13}(0,0) = 0$ et $P_{23}(0,0) = 0$, nous avons les solutions :

$$\begin{aligned}
P_{11}(0,t) &= \exp(-(\lambda_{12}+\lambda_{13})t) \\
P_{12}(0,t) &= \frac{\lambda_{12}}{\lambda_{12}+\lambda_{13}-\lambda_{23}}(\exp(-\lambda_{23}t) - \exp(-(\lambda_{12}+\lambda_{13})t)) \\
P_{13}(0,t) &= \frac{\lambda_{13}}{(\lambda_{12}+\lambda_{13})}(1 - \exp(-(\lambda_{12}+\lambda_{13})t)) \\
P_{22}(0,t) &= \exp(-\lambda_{23}t) \\
P_{23}(0,t) &= 1 - \exp(-\lambda_{23}t) \\
P_{33}(0,t) &= 1.
\end{aligned}$$

Notons que $P_{21}(0,t) = 0$, $P_{31}(0,t) = 0$ et $P_{32}(0,t) = 0$ pour toutes les valeurs de t étant donné que dans le modèle *maladie-décès*, contrairement au modèle de renouvellement alterné, on ne peut pas revenir à un état après l'avoir quitté. Les équations directes sont utiles pour former la vraisemblance des temps de séjour observés.

Soit T la durée fixée de la période d'observation et rappelons que x_i est le temps de séjour éventuellement censuré dans l'état 1 pour le sujet i et y_i est le temps de séjour éventuellement censuré dans l'état 2 pour le sujet i. Nous avons les contributions possibles suivantes à la vraisemblance.

— Si le sujet i reste dans l'état 1 jusqu'à la fin de l'intervalle d'observation, la contribution est :

$$\begin{aligned} & P_{11}(0, x_i) \\ & = \exp(-(\lambda_{12} + \lambda_{13})x_i). \end{aligned}$$

Ici, nous avons une observation censurée à droite et $x_i = T$.

— Si le sujet i passe à l'état 2 et reste à l'état 2 jusqu'à la fin de l'intervalle d'observation, la contribution est :

$$\begin{aligned} & P_{11}(0, x_i)\lambda_{12}P_{22}(0, y_i) \\ & = \exp(-(\lambda_{12} + \lambda_{13})x_i)\lambda_{12}\exp(-\lambda_{23}y_i). \end{aligned}$$

Ici, nous avons une observation complète x_i suivie d'une observation censurée à droite y_i et $y_i = T - x_i$.

— Si i passe à l'état 2 puis à l'état 3, la contribution est [1] :

$$\begin{aligned} & P_{11}(0, x_i)\lambda_{12}P_{22}(0, y_i)\lambda_{23} \\ & = \exp(-(\lambda_{12} + \lambda_{13})x_i)\lambda_{12} \exp(-\lambda_{23}y_i)\lambda_{23}. \end{aligned}$$

— Si i passe directement à l'état 3, la contribution à la vraisemblance est :

$$\begin{aligned} & P_{11}(0, x_i)\lambda_{13} \\ & = \exp(-(\lambda_{12} + \lambda_{13})x_i)\lambda_{13}. \end{aligned}$$

Compte tenu des exemples épidémiologiques que nous avons présentés à la section 1.1.2, nous nous intéressons à λ_{12} et nous exprimons donc la vraisemblance en fonction de ce paramètre qui décrit la première transition et nous introduisons la notation suivante pour chaque sujet i :

$$\delta_{i12} = \begin{cases} 1, \text{ si l'individu } i \text{ passe directement à l'état 2} \\ 0, \text{ sinon} \end{cases}$$

$$\delta_{i13} = \begin{cases} 1, \text{ si l'individu } i \text{ passe directement à l'état 3} \\ 0, \text{ sinon} \end{cases}$$

$$\delta_{i11} = \begin{cases} 1, \text{ si l'individu } i \text{ reste dans l'état 1} \\ 0, \text{ sinon.} \end{cases}$$

---

1. Si nous traitons le temps de séjour x_i pour l'état 1 comme un temps de survie, nous voyons que la fonction de risque est égale à $\lambda_{12} + \lambda_{13}$. En revanche, traiter y_i comme un temps de survie montre que sa fonction de risque est égale à l'intensité λ_{23}.

Bien entendu, un seul de δ_{i12}, δ_{i13} ou δ_{i11} peut être égal à 1 et donc, pour tout individu i, $\delta_{i12} + \delta_{i13} + \delta_{i11} = 1$. Notez que les individus i passant de l'état 2 à l'état 3 ont $\delta_{i12} = 0$ et $\delta_{i13} = 0$.

Puisque nous n'avons besoin de résoudre que λ_{12} dans notre exemple, par souci de simplicité, nous n'incluons explicitement dans la vraisemblance ci-dessous que les termes qui contiennent λ_{12}. [2]

$$\begin{aligned} L &\propto \prod_{i=1}^{n} (P_{11}(0, x_i))^{\delta_{i11}} (P_{11}(0, x_i)\lambda_{12})^{\delta_{i12}} \\ &\quad \times (\text{termes qui ne dépendent pas de } \lambda_{12}) \\ &= \prod_{i=1}^{n} \exp(-(\lambda_{12} + \lambda_{13})x_i)\lambda_{12}^{\delta_{i12}} \\ &\quad \times (\text{termes qui ne dépendent pas de } \lambda_{12}) \\ &= \exp\left(-(\lambda_{12} + \lambda_{13}) \sum_{i=1}^{n} x_i\right) \lambda_{12}^{\sum_{i=1}^{n} \delta_{i12}} \\ &\quad \times (\text{termes qui ne dépendent pas de } \lambda_{12}). \end{aligned} \tag{1.12}$$

1.3.3 Vraisemblances : Événements récurrents

Dans la situation où l'on redevient à risque dès la fin de l'événement, la modélisation d'événements récurrents est appropriée (Cook et Lawless, 2007). Avec ce type de modélisation, le nombre d'événements vécus par une personne est proportionnel à la durée de son suivi. Soit λ le taux du processus de Poisson et $\tilde{T}_i$ le temps de suivi de la personne i. Laissons K_i représenter le nombre d'événements pour la

2. Notez qu'en incluant les paramètres λ_{13} et λ_{23} dans la vraisemblance, on peut facilement la développer afin de trouver les estimateurs du maximum de vraisemblance (EMV) de λ_{13} et λ_{23}.

personne i, alors :

$$P(K_i = k_i) = \frac{(\lambda \tilde{T}_i)^{k_i} \exp(-\lambda \tilde{T}_i)}{k_i!}.$$

Soit x_i le temps de survie observé pour l'individu i et y_i le temps de troncature pour l'individu i ; nous pouvons écrire la vraisemblance de k_i pour les n iid observations par :

$$P(K_i = k_i) = \prod_{i=1}^{n} \frac{(\lambda(x_i - y_i))^{k_i} \exp(-\lambda(x_i - y_i))}{k_i!} \tag{1.13}$$

où k_i est le nombre d'événements pour l'individu i.

1.4 Estimation dans l'analyse de survie

Dans la section 1.2.1 nous avons mentionné l'estimation empirique $\hat{F}(t) = \frac{\#\text{ de } t_i \leq t}{\#\text{ total}}$. Nous avons également l'estimation empirique pour la fonction de survie $\hat{S}(t) = \frac{\#\text{ de } t_i > t}{\#\text{ total}}$. Ces estimateurs sont essentiellement non paramétriques [3] et représentent respectivement les estimations pour la fonction de répartition et la fonction de survie pour le cas où les temps de survie sont complètement observés.

De manière analogue, nous avons les estimateurs non paramétriques de Kaplan-Meier [4] pour le cas d'une censure à droite indépendante et non informative. Cet estimateur a été initialement présenté dans l'article (Kaplan et Meier, 1958) et est également présenté dans n'importe quel livre sur l'analyse de survie, voir par exemple (Kalbfleisch et Prentice, 2002), (Klein et Moeschberger, 2003) ou (Lawless, 2003).

---

3. Ces estimateurs sont des estimateurs discrets qui accordent un poids au temps d'événements. Strictement parlant, s'il n'y a pas d'observations liées, nous avons n paramètres (ou poids) à estimer.

4. Ces estimateurs sont des estimateurs discrets qui mettent la masse de probabilité aux k temps d'événements observés. S'il y a k événements non censurés et aucun lien, alors nous avons $k < n$ paramètres à estimer.

Nous utilisons l'estimateur de Kaplan-Meier au chapitre 2 de ce mémoire. Pour dériver l'estimateur de Kaplan-Meier, on attribue un poids aux temps d'événements observés. Pour tout événement qui a été censuré à droite, nous répartissons le poids vers la droite (aux temps ultérieurs de l'événement). L'estimateur de Kaplan-Meier peut être dérivé en optimisant la vraisemblance (1.4). Voir l'appendice C.2. Ci-dessous, nous présentons l'expression de l'estimateur de Kaplan-Meier.

$$\hat{S}_{KM}(t) = \begin{cases} 1, & t < t_{(1)} \\ \prod_{t_{(j)} \leq t}(1 - \hat{h}_j), & t \geq t_{(1)} \end{cases} \tag{1.14}$$

où $\hat{h}_j = \frac{d_j}{n_j} = \frac{\#\text{ d'événements à } t_{(j)}}{\#\text{ de personnes à risque à } t_{(j)-}}$ et $t_{(j)}$ est la j^e statistique d'ordre.

1.5 Résumé

Dans les chapitres 2 et 3 nous reviendrons sur les exemples présentés de la section 1.1.2.

— Lorsque le taux d'incidence est calculé sur l'intervalle $(0, t_5)$, l'exemple 1 dans la section 1.1.2 représente une population fermée. Si l'intervalle s'étend au-delà de t_5, on peut traiter la fin de l'observation en utilisant la censure à droite. Puisque l'échantillon est fermé à partir de $t = 0$, il n'y a pas de troncature à gauche. De plus, comme les personnes-années peuvent être calculées, les données sont de type expérimental. Cet exemple (exemple 1 de la section 1.1.2) est revisité dans les chapitres 2 et 3 :

  — dans la section 2.1.3 du chapitre 2, nous revisitons cet exemple en utilisant l'expression de Selvin dans l'exemple 1 ;

— dans la section 3.2.1 du chapitre 3, nous reprenons cet exemple dans l'exemple 1 en traitant la fin de l'intervalle d'observation avec une censure à droite et en utilisant une fonction de risque constante.

— L'exemple 2 dans la section 1.1.2 représente une cohorte fermée à gauche ($n = 5$). Il n'y a pas de troncature à gauche puisque tous les sujets sont suivis à partir du temps $t = 0$. L'exemple présente des données de type expérimental. Il y a deux possibilités :

1. *Immunité ou maladie chronique* : on ne peut pas retomber malade tout de suite. Cet exemple est traité deux fois. Une fois dans l'exemple 2 de la section 3.2.2 en traitant la mort comme une censure à droite par rapport à l'événement maladie. Deuxièmement, dans l'exemple 4 de la section 3.3.1 en utilisant un schéma d'état du modèle *maladie-décès*.
2. *Pas d'immunité* : une personne est de nouveau à risque de tomber malade ou de mourir. La fin de l'étude peut être représentée par une censure à droite. Cette situation peut être modélisée à l'aide d'un modèle d'événements récurrents. Voir l'exemple 5 dans la section 3.3.2.

— Le type de données dans l'exemple 3 de la section 1.1.2 provient généralement d'une étude de cohorte avec le calendrier comme échelle de temps. Par la suite, les données sont remises pour une échelle de temps d'âge qui introduit une troncature à gauche (l'âge de chaque personne entrée dans l'étude correspond à son temps de troncature). Le temps de survie est l'âge au moment du diagnostic du cancer. Le taux d'incidence recherché est l'incidence du cancer (sur différentes périodes de 5 ans). Notez que, bien que l'échantillon soit fermé à gauche en temps du calendrier, il est ouvert en fonction de l'âge. Nous analysons cet exemple en utilisant la troncature à gauche et la censure à droite dans l'exemple 3 de la section 3.2.3.

— Dans l'exemple 4 de la section 1.1.2, 20 individus ($n = 20$) ont été suivis pendant un an. Il y a eu un total de 10 événements, mais 6 d'entre eux se sont produits avant le début de la période de référence (le 1^{er} octobre 2004). Les lignes horizontales individuelles sont trompeuses, car le temps à risque est avant l'événement ($\Downarrow$). Toutes les récupérations ($\Uparrow$) sont survenues après l'année d'intérêt (le 30 septembre 2005). Les données sont de type recensement et les personnes-temps doivent être approximées. Nous revisitons cet exemple dans la section 3.3.3 afin de montrer pourquoi il est important d'avoir un meilleur cadre théorique lors de la discussion sur le taux d'incidence.

— L'exemple 5 de la section 1.1.2 présente à nouveau des données de type recensement. Contrairement aux exemples précédents qui impliquaient un petit nombre d'individus, nous avons ici un exemple basé sur une population. Le nombre de décès dans la population américaine est fourni et le nombre de personnes-années à risque doit être estimé. Notez qu'il n'y a pas de discussion à savoir si la population est fermée ou ouverte. Cet exemple est repris dans l'exemple 2 de la section 2.3.2.

Dans le chapitre 3, nous fournissons également des dérivations pour les intensités de transition constantes en utilisant le processus de renouvellement alterné. Ici, au lieu de supposer que la personne risque immédiatement de retomber malade, nous accordons une période pendant laquelle la personne est malade puis de nouveau en bonne santé. Les taux d'incidence (intensités de transition constantes) peuvent être estimés pour chaque état du modèle. Nous abordons d'abord le cas où tous les individus commencent dans l'état sain au temps $t = 0$ (voir figure 3.1) puis, nous continuons avec le cas d'une cohorte dont l'échelle de temps est l'instant du calendrier (voir figure 3.2). Nous discutons du modèle de renouvellement alterné dans la conclusion.

CHAPITRE II

L'ESTIMATEUR DE SELVIN ET L'APPROXIMATION DES PERSONNES-TEMPS

Dans ce chapitre, nous utilisons les exemples 1 et 5 présentés dans la section 1.1.2. Nous examinons l'estimateur de Selvin dans la section 2.1. Ensuite, dans la section 2.3, nous nous référons à l'estimateur de Selvin pour calculer les approximations des personnes-temps au dénominateur. Notez que, contrairement au chapitre 3, nous ne supposons pas ici une fonction de risque constante.

2.1 L'estimateur de Selvin

Dans son livre (Selvin, 2008), Selvin introduit une expression du taux d'incidence en fonction de la fonction de survie. Il prend la forme d'un taux moyen et est donné par :

$$\text{taux moyen} = \frac{S(t) - S(t+dt)}{\int_t^{t+dt} S(u)du}. \tag{2.1}$$

Bien que cela ne soit pas explicitement indiqué dans (Selvin, 2008), il est supposé que l'expression de Selvin (2.1) s'applique à une population fermée, comme dans l'exemple 1. Il n'y a aucune restriction sur la distribution de temps de survie.

Il est intéressant de noter que (Elandt-Johnson et Johnson, 1999) introduisent également un taux moyen pour une fonction continue qui prend la même forme que l'expression (2.1). De plus, notez que puisque la fonction de survie $S(t)$ est un paramètre de la population, l'expression de Selvin (2.1) peut être considérée comme un paramètre de cette population. Notez également qu'il n'y a pas de contraintes sur la distribution S. Par conséquent, bien que nous considérerons brièvement le cas d'un risque constant à la section 2.1.2, la fonction de risque n'a pas besoin d'être constante pour utiliser l'expression de Selvin.

Lorsque l'intervalle de temps est court, on peut montrer que la fonction de risque est approximativement égale à l'expression (2.1) :

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{-dS(t)/dt}{S(t)} \\ &= -\lim_{dt\to 0} \frac{S(t+dt)-S(t)}{S(t)dt} \end{aligned} \tag{2.2}$$

qui pour dt petit devient

$$\begin{aligned} &\approx \frac{S(t)-S(t+dt)}{S(t)dt} \\ &\approx \frac{S(t)-S(t+dt)}{\int_t^{t+dt} S(u)du} \\ &= \text{taux moyen.} \end{aligned} \tag{2.3}$$

Par conséquent, on peut voir pourquoi la fonction de risque est parfois appelée taux d'incidence instantané.

2.1.1 Interprétation de l'expression de Selvin

Soit N_0 représentant la taille de la population ou de l'échantillon au début de l'intervalle d'étude. En multipliant N_0 par le numérateur et le dénominateur de l'expression (2.1), on a :

$$\text{taux moyen} = \frac{N_0(S(t) - S(t+dt))}{N_0(\int_t^{t+dt} S(u)du)}. \tag{2.4}$$

On commence par considérer le numérateur de l'expression (2.4) et on montre qu'il représente un certain nombre de cas (événements). Par la suite, nous interpréterons le dénominateur pour montrer qu'il représente les personnes-temps. Tout au long de la présentation, il est clair qu'une hypothèse de population fermée est requise.

Concernant le numérateur, on a :

$$N_0 \times (S(t) - S(t+dt)) = N_0 \times P(T \in (t, t+dt]). \tag{2.5}$$

Le nombre (#) de personnes N_0 fois la probabilité de mourir dans l'intervalle de temps $(t, t+dt]$ correspond au nombre (#) attendu d'événements survenus dans l'intervalle $(t, t+dt]$.

Pour le dénominateur, la figure 2.1 ci-dessous est utile dans cette interprétation. Au lieu de diviser l'aire de l'intégrale $N_0 \times (\int_t^{t+dt} S(u)du)$ en termes de barres verticales comme c'est le cas pour l'intégration de Riemann, nous divisons l'aire en barres horizontales. En considérant la barre du haut, nous voyons que sa hauteur $N_0 \times (S(t) - S(t^*))$ équivaut au nombre attendu de personnes qui meurent dans l'intervalle $(t, t^*]$ et sa longueur $(t^* - t)$ est équivalente à la durée de l'intervalle. Par conséquent, l'aire $N_0 \times (S(t) - S(t^*)) \times (t^* - t)$ est une approximation de personnes-temps pour les personnes décédées dans l'intervalle $(t, t^*]$.

En mettant tout ensemble, nous voyons que l'intégrale $N_0 \times (\int_t^{t+dt} S(u)du)$ fournit une approximation de personnes-temps de l'échantillon ou de la population pour la période d'intérêt. Avec cette interprétation de l'expression (2.1), on voit son équivalence avec l'estimateur du taux d'incidence à l'expression (1).

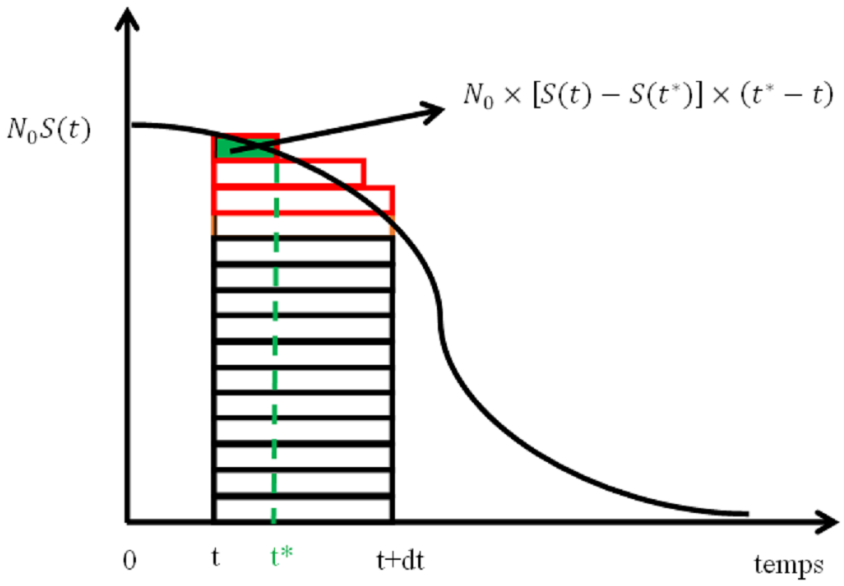


Figure 2.1 Schéma d'une courbe de survie $S(t)$. Diviser les zones en dessous de la courbe en barres horizontales permet d'interpréter les différentes zones en termes de personnes-temps.

2.1.2 Fonction de risque constante

Lorsque les temps de survie suivent une loi exponentielle, la fonction de risque est constante. Si les temps de survie sont paramétrés par λ, on a $S(t) = \exp(-\lambda t)$ et $h(t) = \lambda$.

Dans l'expression (2.1),

$$\begin{aligned}
\text{taux moyen} &= \frac{S(t) - S(t+dt)}{\int_t^{t+dt} S(u)du} \\
&= \frac{(\exp(-\lambda t) - \exp(-\lambda(t+dt)))}{\int_t^{t+dt} \exp(-\lambda u)du} \\
&= \frac{(\exp(-\lambda t) - \exp(-\lambda(t+dt)))}{\frac{(\exp(-\lambda t+dt) - \exp(-\lambda(t)))}{-\lambda}} \\
&= \lambda.
\end{aligned} \tag{2.6}$$

Par conséquent, lorsque la fonction de risque est constante, l'expression de Selvin est égale à la fonction de risque.

Nous reviendrons à la situation d'un risque constant au chapitre 3. Pour le reste du chapitre, nous explorons le cas où la fonction de risque n'est pas constante.

2.1.3 Fonction de risque non constante

Dans cette section, nous considérons l'estimation de l'expression de Selvin (2.1) à partir de données lorsqu'on n'émet pas l'hypothèse d'un risque constant.

Dans un échantillon fermé ($i = 1, \ldots, n$), les personnes ne sont perdues de vue dans l'échantillon qu'en faisant l'expérience de l'événement pendant la période d'intérêt. Dans cette situation, on peut estimer la fonction de survie en utilisant la fonction de survie empirique :

$$\hat{S}(t) = P(T > t) = \frac{\#\text{ de } t_i > t}{n}. \tag{2.7}$$

Puisque la fonction de survie empirique $\hat{S}$ est une fonction échelonnée, il est facile de montrer que l'estimateur par injection ou « plug-in » du

$$\text{taux moyen} = \frac{\hat{S}(t) - \hat{S}(t+dt)}{\int_t^{t+dt} \hat{S}(u)du} \tag{2.8}$$

conduit à l'expression (1). Cette situation est reflétée par l'exemple 1 au chapitre 1. Voir l'exemple 1 ci-dessous et voir l'appendice C pour un exemple plus général.

Exemple 1. En se référant à l'exemple 1 de la sous-section 1.1.2, nous avons $t = 0$, $dt = 19$ et $t + dt = 19$. La fonction de survie empirique est :

$$\hat{S}(t) = \begin{cases} 1, \text{ pour } 0 \leq t < 2 \\ 8/9, \text{ pour } 2 \leq t < 4 \\ 6/9, \text{ pour } 4 \leq t < 8 \\ 5/9, \text{ pour } 8 \leq t < 14 \\ 4/9, \text{ pour } 14 \leq t < 19. \end{cases} \tag{2.9}$$

En se référant à l'équation (2.1), nous avons :

$$\begin{aligned} \hat{S}(t) - \hat{S}(t+dt) &= \hat{S}(0) - \hat{S}(19) \\ &= 9/9 - 4/9 \\ &= 5/9 \end{aligned}$$

pour le numérateur et

$$\begin{aligned}\int_t^{t+dt} \hat{S}(u)du &= \int_0^{19} \hat{S}(u)du \\ &= 1(2) + 8/9(4-2) + 6/9(8-4) + 5/9(14-8) + 4/9(19-14) \\ &= 108/9\end{aligned}$$

pour le dénominateur.

L'estimateur du taux d'incidence par « plug-in » à l'équation (2.8) vaut 5/108 qui équivaut au résultat obtenu en utilisant l'expression (1) dans l'exemple 1 de la sous-section 1.1.2.

2.2 L'estimateur de Selvin pour les populations ou échantillons fermés à gauche

Une population fermée à gauche peut être modélisée en utilisant la censure à droite. Inspiré par l'exemple 1 ci-dessus, on pourrait d'abord se demander si la réplication de l'approche de l'estimateur par « plug-in » ci-dessus en remplaçant l'estimateur empirique par l'estimateur de Kaplan-Meier pourrait fournir le même estimateur qu'à l'expression (1) dans le cas de censure à droite. L'exemple simple présenté à l'appendice C.2 montre que cette approche ne reproduira l'expression du taux d'incidence de l'équation (1) que dans certaines circonstances. Spécifiquement :

— la figure C.2 avec censure à droite dans l'intervalle d'observation nous conduit à ce que le taux d'incidence ne soit pas reproduit ;

— la figure C.2 sans censure à droite dans l'intervalle d'observation nous conduit à reproduire le taux d'incidence.

Nous discutons plus en détail de cette situation dans l'appendice C.

2.3 Approximation des personnes-temps au dénominateur

Tel que mentionné au point 6 de la section 0.1, le livre de (Elandt-Johnson et Johnson, 1999) fait la distinction entre les données de type recensement où le nombre de personnes exposées au risque au dénominateur doit être estimé et les données de type expérimental où le nombre de personnes exposées au risque ne doit pas être estimé. Il est clair que dans de nombreuses situations de données réelles, le nombre de cas serait enregistré ; cependant, soit les informations nécessaires pour calculer les personnes-temps pourraient ne pas être disponibles, soit même si les informations étaient disponibles, il serait difficile de récupérer les informations pour calculer les personnes-temps.

Étant donné que l'estimation de personnes-temps est une composante essentielle de l'estimation du taux d'incidence, nous considérons certaines approches couramment utilisées pour estimer les personnes-temps au dénominateur de l'expression (1).

2.3.1 Population fermée (Selvin)

Dans le livre de (Selvin, 2008), après avoir introduit l'expression (2.1) qui s'applique clairement aux populations fermées, Selvin explique comment estimer le dénominateur. Il approxime l'aire dans l'intervalle $(t, t + dt]$ et sous la courbe de survie en divisant l'aire en un triangle et un rectangle. La figure 2.1 représente le triangle en rouge et le rectangle en noir.

Simplement, l'approximation est :

— rectangle : $dt \times N_0 \times S(t + dt)$;

— triangle : $\frac{1}{2} \times dt \times N_0 \times [S(t) - S(t + dt)]$.

Bien sûr, l'approche ci-dessus (diviser la zone en un triangle et un rectangle) fonctionnerait pour estimer la zone sous une courbe qui augmente ou diminue presque linéairement sur l'intervalle d'intérêt. Si la courbe s'écarte de ce comportement de manière significative, l'approximation triangle-rectangle pour la zone sera inexacte.

2.3.2 Populations et échantillons dynamiques

Le chapitre 1 de (Selvin, 2004) suit celui de (Selvin, 2008) sauf que ce dernier introduit un taux moyen en termes d'une fonction $y(t)$ qui donne le nombre de personnes en vie au temps t :

$$\text{taux moyen} = \frac{y(t + dt) - y(t)}{\int_t^{t+dt} y(u)du}. \tag{2.10}$$

Dans les deux ouvrages de Selvin précités, il y a très peu de discussions concernant le type de population. Immédiatement, on peut voir que si la population diminue, le numérateur (représentant le nombre de cas) sera négatif. Même si l'on ajoute des signes de valeur absolue, c'est-à-dire

$$\text{taux moyen} = \frac{|y(t + dt) - y(t)|}{\int_t^{t+dt} y(u)du},$$

on constate que des problèmes subsistent. Considérons les personnes qui immigrent dans la population ; cela augmentera la valeur $y(t+dt)$ et donc la valeur $y(t+dt)-y(t)$ augmentera également, mais pas parce que le nombre de cas a augmenté.

Nous suggérons qu'une expression plus appropriée serait

$$\text{taux moyen} = \frac{\#\text{ de cas}}{\int_t^{t+dt} y(u)du}, \tag{2.11}$$

où $y(t)$ est défini comme la taille de la population à risque pour l'événement d'intérêt.

Habituellement, le nombre (#) de cas est facilement déterminé à partir des certificats de décès ou des dossiers médicaux. Par contre, les personnes-temps à risque sont beaucoup plus difficiles à déterminer. Elles peuvent cependant être modélisés par $\int_t^{t+dt} y(u)du$, en utilisant un argument similaire à celui présenté dans la section 2.1.1.

Pour approximer $\int_t^{t+dt} y(u)du$ nous devons estimer l'aire sous la courbe $y(t)$. Souvent, une simple approximation rectangulaire égale à $y(\text{au point-médian}) \times (t, t + dt]$ est utilisée. Cette approximation fonctionne assez bien, surtout si y ne varie pas beaucoup tout au long de l'intervalle d'étude. Une solution telle que discutée dans (Breslow et Day, 1980) est de prendre des intervalles de temps plus petits.

Nous revenons à l'exemple 5 de la sous-section 1.1.2 qui utilise l'approximation du point-médian. Bien que la possibilité d'immigration ou d'émigration ne soit pas abordée dans l'exemple, elle doit être comprise puisque nous parlons de la population des États-Unis.

Exemple 2. Ici nous avons :

$$\begin{aligned} y(\text{au point-médian } (t, t + dt]) &= 290809777 \\ \text{intervalle } (t, t + dt] &= 1 \text{ année.} \end{aligned} \tag{2.12}$$

Par conséquent, le taux d'incidence est de :

$$\begin{aligned} TI &= \frac{44\ 232}{290\ 809\ 777} \\ &= 0.0001520994\ \frac{\text{cas}}{\text{personnes-années}}. \end{aligned}$$

En multipliant le numérateur et le dénominateur par 100 000 personnes-années, on obtient 15.2 cas par 100 000 personnes-années. Notez que la présentation dans (Dicker *et al.*, 2012) ne fournit pas clairement les unités de calcul ; voir le point 2 de la section 0.1.

CHAPITRE III

FONCTION DE RISQUE CONSTANTE ET SITUATIONS DE DONNÉES COMPLEXES

Dans ce chapitre, nous considérons les exemples épidémiologiques de la sous-section 1.1.2 pour des fonctions de risque et d'intensité constantes. Nous nous limitons à des situations de données expérimentales et utilisons des techniques d'analyse de l'historique des événements présentées dans la section 1.2.

3.1 Fonction de risque constante

Avant de considérer les exemples de la sous-section 1.1.2 et le taux d'incidence, nous considérons le cas d'un échantillon de n temps de survie *iid* complètement observés.

Comme indiqué à l'appendice B, sans utiliser aucune technique d'analyse de l'historique des événements, on peut écrire la vraisemblance pour n temps de survie t_i *iid* qui suivent une distribution exponentielle et trouver l'estimateur du maximum de vraisemblance (EMV) pour le paramètre λ :

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}. \tag{3.1}$$

Nous voyons que pour ce cas très simple d'un risque constant et de n temps

de survie *iid*, l'EMV pour le paramètre λ prend la forme de l'expression (1), c'est-à-dire $\left(\frac{\#\text{ d'événements}}{\text{personnes-temps}}\right)$. Les t_i dans l'expression (3.1) sont les temps de survie complètement observés et non censurés.

Dans la suite, nous revenons aux exemples présentés à la sous-section 1.1.2 et nous utilisons les techniques de la section 1.2 pour estimer λ (la fonction de risque constante). Au chapitre 2, nous avons démontré que la fonction de risque constante est égale au taux d'incidence (voir l'expression (3.1)).

Nous verrons que bon nombre des situations de données complexes présentées à la sous-section 1.1.2 peuvent être modélisées en utilisant la censure à droite et la troncature à gauche. Dans les exemples pratiques, il est important d'être clair sur l'échelle de temps que l'on utilise. L'exemple 3 de la section 1.1.2 illustre bien ce point. En outre, les situations comportant plusieurs événements peuvent être modélisées à l'aide de modèles multi-états plus complexes et de modèles à événements récurrents. L'exemple 2 de la sous-section 1.1.2, celui présentant le cas où il y a absence d'immunité après infection est un de cette situation.

3.2 Modélisation à l'aide de l'analyse de survie

Ci-dessous, nous traitons les exemples 1, 2 (pour le cas avec immunité) et 3 en utilisant les techniques présentées à la sous-section 1.2.1.

3.2.1 Échantillon fermé

L'exemple 1 de la sous-section 1.1.2 est un échantillon fermé observé sur une période fixe. La période fixe est utilisée dans le calcul du taux d'incidence et la censure à droite peut être utilisée pour modéliser les temps observés pour les sujets qui sont encore en vie à la fin de l'intervalle d'étude.

Par conséquent, nous utilisons une censure à droite indépendante pour estimer λ. Supposons un échantillon de taille n avec des temps observés $X_i = \min(T_i, C_i)$. Nous utilisons un indicateur $\delta_i = I(T_i \leq C_i)$ pour indiquer la censure à droite. Les temps de survie observés sont représentés par des lettres minuscules x_i.

Revenant à la vraisemblance (1.5), en prenant le log, on a :

$$\sum_{i=1}^{n} \delta_i \log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

et en maximisant, on obtient :

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}. \tag{3.2}$$

Notez que l'expression (3.2) est précisément l'estimation dans l'équation (1). Notez également la similitude avec l'expression (3.1) ; $\sum_{i=1}^{n} x_i$ et $\sum_{i=1}^{n} t_i$ sont les personnes-temps observées dans l'échantillon, cependant $\sum_{i=1}^{n} x_i$ tient compte de la censure à droite due à la fin de l'intervalle fixe d'observation.

Dans l'exemple ci-dessous, nous considérons la situation des données de l'exemple 1 de la sous-section 1.1.2.

Exemple 1.

En considérant l'exemple 1 de la sous-section 1.1.2, le terme $\sum_{i=1}^{n} \delta_i$ donne le nombre d'événements observés, soit 5 et le terme $\sum_{i=1}^{n} x_i$ représente les personnes-temps à risque, soit 108. Ainsi, l'estimation $\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}$ est égale à 5/108.

3.2.2 Échantillon ou population fermée à gauche

Dans une population fermée à gauche, des individus peuvent être perdus de vue avant d'avoir l'événement d'intérêt. L'exemple 2 de la sous-section 1.1.2 où les sujets ne peuvent avoir l'événement qu'une seule fois est l'un de cas. Ici, si un sujet meurt avant d'avoir eu la maladie, il est censuré à droite par rapport à l'événement d'avoir la maladie.

En ce qui concerne la modélisation, nous sommes dans la situation de la section 3.2.1. Auparavant, C_i représentait la censure pour le sujet i due à la fin de l'intervalle d'étude uniquement, alors que maintenant C_i représente la perte de vue causée par la mort avant de vivre l'événement d'intérêt plus la censure à droite causée par la fin de l'intervalle d'étude.

Ci-dessous, nous revisitons l'exemple 2 de la sous-section 1.1.2.

Exemple 2.

En étiquetant les sujets 1 à 5 de haut en bas, les données observées sont : $(x_1, \delta_1) = (5, 0)$, $(x_2, \delta_2) = (1, 0)$, $(x_3, \delta_3) = (4, 1)$, $(x_4, \delta_4) = (3, 1)$ [1], $(x_5, \delta_5) = (1, 1)$. Il s'ensuit que $\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i} = \frac{3}{5+1+4+3+1} = \frac{3}{14} = 0.21$.

3.2.3 Troncature à gauche avec censure à droite

Bien que l'on puisse modéliser mathématiquement la troncature à gauche sans censure à droite, cela se produit rarement dans la pratique. L'exemple 3 de la sous-section 1.1.2 lorsqu'il est modélisé avec comme échelle de temps l'âge est un exemple où la troncature à gauche et la censure à droite se produisent ensemble.

---

1. Notez la définition de δ_i ; donc $\delta_i = 1$ quand $T_i = C_i$.

La situation de l'exemple 3 de la sous-section 1.1.2 se produit généralement lorsqu'une cohorte est prise avec un échelle de temps du calendrier. Ici, il s'agit d'un échantillon d'individus qui sont tous collectés au même temps du calendrier et suivis dans le temps. Plus tard, lorsque les données sont rééchelonnées en termes d'âge, nous constatons que, bien que tous les sujets aient été collectés au même temps du calendrier, ils avaient des âges différents au moment de la collecte des données. Ces différents âges se traduisent par des données tronquées. La censure à droite dans ces exemples se produit en raison du retrait d'un individu dans l'étude et de la fin de l'intervalle d'étude ou de la collecte de données.

Ainsi, l'exemple 3 de la sous-section 1.1.2 représente une population dynamique sur une échelle d'âge. Les gens entrent en observation à un certain âge y_i (le temps de troncature), puis ils se retirent ou subissent l'événement qui est un cancer à l'âge x_i (le temps observé).

Toujours en utilisant le risque constant, nous pouvons modéliser cet exemple 3 en utilisant la troncature à gauche pour l'âge à l'entrée et la censure à droite pour ceux qui se retirent et pour la fin de l'observation. La vraisemblance pour cela a été développée à l'expression (1.7).

Encore une fois, en prenant le log de l'expression (1.7) on a :

$$\sum_{i=1}^{n} \delta_i \log(\lambda) - \sum_{i=1}^{n} \lambda(x_i - y_i)$$

et en maximisant, on obtient :

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} (x_i - y_i)}. \tag{3.3}$$

Comme on peut le voir à l'expression (3.3), $(x_i - y_i)$ est le nombre de personnes-temps d'observation pour l'individu i et $\sum_{i=1}^{n}(x_i - y_i)$ est le nombre de personnes-temps en observation pour tous les n individus. L'expression (3.3), représentant l'EMV pour le risque constant lorsqu'il y a troncature à gauche et censure à droite, équivaut exactement à l'expression (1) du taux d'incidence.

Exemple 3.

L'exemple 3 de la sous-section 1.1.2 peut être traité avec une troncature à gauche. Ici, l'échelle de temps est l'âge et différentes périodes d'observation sont utilisées pour calculer l'incidence. Pour illustrer, nous utilisons la période de 55 à 65. La notation pour chaque individu i est :

$$\begin{aligned} y_i = &\text{ temps de troncature observé} \\ x_i = &\text{ durée de vie observée} \\ \delta_i = &\text{ indicateur de la censure} \end{aligned}$$

Les données pour la période d'âge de 55 ans à 65 ans sont présentées dans le tableau 3.1 ci-dessous.

Le petit code R présenté à l'appendice D montre les calculs de l'estimateur du paramètre constant λ pour les données ci-dessous. Nous trouvons que

$$\begin{aligned} \hat{\lambda} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n}(x_i - y_i)} \\ &= \frac{4}{93.5} = 0.043, \end{aligned}$$

soit 4 300 cancers par 100 000 personnes-années.

Tableau 3.1 Les données de (Breslow et Day, 1980).

y_i	x_i	δ_i
55	57.1	1
55	65	0
55	61.4	0
55	60	1
55	65	0
55	65	0
55	61.1	1
55	65	0
55	65	0
56	59.9	1
57	65	0
59	65	0
59	65	0

3.3 Modélisation à l'aide de techniques issues des modèles multi-états et d'événements récurrents

Dans cette section, nous considérons des modèles plus complexes issus de l'analyse de l'historique des événements pour calculer les taux d'incidence. Rappelons que dans ce chapitre nous nous concentrons sur des situations de données complexes et les simplifions en prenant un risque constant et des fonctions d'intensité de transition constantes. Cela signifie que tous nos modèles tombent dans la situation de la modélisation de Markov homogène à temps continu.

En utilisant différentes hypothèses de modélisation, l'exemple 2 de la sous-section 1.1.2 est un bon exemple pour illustrer certains des modèles ci-dessous.

3.3.1 Modèle *maladie-décès*

Ici, nous revenons à l'exemple 2 de la sous-section 1.1.2 où les individus sont immunisés après l'infection. Auparavant, nous considérions la mort comme une censure à droite des cas de maladie et modélisons la fonction de risque constante. Nous présentons maintenant un modèle plus détaillé (modèle *maladie-décès*) qui représente mieux la situation et nous voyons que le taux d'incidence à l'expression (1) est équivalent à notre estimation de l'intensité de transition constante entre l'état sain (1) et l'état malade (2).

En se référant au modèle *maladie-décès* et à la vraisemblance (1.12), nous voyons que le paramètre λ_{12} représente le taux d'incidence de la maladie. Notez que dans l'exemple 2 de la sous-section 1.1.2 et dans le développement de notre modèle *maladie-décès*, chaque sujet commence l'observation dans l'état 1 (l'état sain) au temps $t = 0$.

En prenant le log de la vraisemblance (1.12) et les dérivées partielles par rapport à λ_{12}, nous maximisons (1.12), pour trouver le taux d'incidence de la maladie. C'est-à-dire que nous trouvons l'estimation pour le paramètre constantλ_{12}.

Le log de la vraisemblance (1.12) est :

$$\sum_{i=1}^{n} -(\lambda_{12} + \lambda_{13})x_i + \sum_{i=1}^{n} \delta_{i12} \log \lambda_{12}. \tag{3.4}$$

Si nous fixons la dérivée partielle par rapport à λ_{12} égale à zéro, nous trouvons $\hat{\lambda}_{12} = \frac{\sum_{i=1}^{n} \delta_{i12}}{\sum_{i=1}^{n} x_i}$ [2]. Rappelons que x_i est le délai (temps passé) à l'état 1 (sain) pour l'individu i.

---

2. Si on avait laissé le terme λ_{13}, on trouverait vraisemblablement $\hat{\lambda}_{13} = \frac{\sum_{i=1}^{n} \delta_{i13}}{\sum_{i=1}^{n} x_i}$

Donc $\sum_{i=1}^{n} x_i$ représente toutes les personnes-temps pour les n individus passés dans l'état 1. Par conséquent, l'estimation de l'intensité de transition constante λ_{12} du modèle *maladie-décès* est égal à l'expression (1).

En revenant à l'exemple (2) de la sous-section 1.1.2 où les patients sont de nouveau à risque, nous utilisons maintenant le modèle *maladie-décès*.

Exemple 4.

Bien que nous puissions collecter toutes les données pour le modèle *maladie-décès*, on constate que les données nécessaires pour cet exercice sont exactement les données que nous avons collectées à l'exemple 2 de la sous-section 3.2.2. En supposant que les temps de séjour observés dans les états 1, 2 et 3 pour l'individu i soient respectivement x_i, y_i et z_i, nous avons :

— $(\delta_{111}, \delta_{112}, \delta_{112}) = (1, 0, 0)$ et $(x_1, y_1, z_1) = (5, 0, 0)$.

— $(\delta_{211}, \delta_{212}, \delta_{212}) = (0, 0, 1)$ et $(x_2, y_2, z_2) = (1, ?, ?)$ [3]

— $(\delta_{311}, \delta_{312}, \delta_{312}) = (0, 1, 0)$ et $(x_3, y_3, z_3) = (4, 1, 0)$.

— $(\delta_{411}, \delta_{412}, \delta_{412}) = (0, 1, 0)$ et $(x_4, y_4, z_4) = (3, 0, 2)$.

— $(\delta_{511}, \delta_{512}, \delta_{512}) = (0, 1, 0)$ et $(x_5, y_5, z_5) = (1, 3.5, 0.5)$.

Nous estimons l'intensité de transition constante comme :

$$\begin{aligned}\hat{\lambda}_{12} &= \frac{\sum_{i=1}^{n} \delta_{i12}}{\sum_{i=1}^{n} x_i} \\ &= \frac{3}{5+1+4+3+1} = \frac{3}{14} \\ &= 0.21,\end{aligned}$$

3. Nous ne pouvons pas déduire les temps y_2 et z_2 du diagramme.

soit 21 000 cas par 100 000 personnes-années, ce qui équivaut à l'expression (1). On voit que le calcul conduit au même résultat qu'à l'exemple 2 de la sous-section 3.2.2.

3.3.2 Événements récurrents (répétés)

La deuxième hypothèse mentionnée dans l'exemple 2 de la sous-section 1.1.2 est que les personnes sont immédiatement de nouveau à risque après avoir contracté la maladie. Cela correspond à un modèle d'événements récurrents (répétés). Le plus simple est basé sur un modèle de Poisson. La vraisemblance pour ce modèle est donnée à l'expression (1.13).

En prenant le log de cette vraisemblance, on a :

$$\sum_{i=1}^{n}\{k_i \log(\lambda(x_i - y_i)) - \lambda(x_i - y_i) - \log k_i!\}. \tag{3.5}$$

En effectuant la dérivée partielle par rapport à λ et en l'égalant à zéro, nous trouvons l'EMV de λ :

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n}(x_i - y_i)}. \tag{3.6}$$

Puisque k_i est le nombre d'événements observés pour l'individu i, $\sum_{i=1}^{n} k_i$ est donc le nombre total d'événements observés pour tous les n individus. Encore une fois, x_i est le temps observé et y_i est le temps de troncature. Par conséquent, $(x_i - y_i)$ est le nombre de personnes-temps d'observation pour la personne i et $\sum_{i=1}^{n}(x_i - y_i)$ est le nombre de personnes-temps pour les n individus. Ce modèle conduit également au taux d'incidence défini à l'équation (1).

Ci-dessous, nous revenons à l'exemple 2 de la sous-section 1.1.2. Rappelons dans les dérivations ci-dessus que x_i est le temps observé jusqu'à la mort ou la fin de l'intervalle d'étude pour l'individu i et y_i est le temps de troncature pour l'individu i (tous les temps de troncature pour l'exemple 2 sont nuls).

Exemple 5.

En revenant à l'exemple 2 de la sous-section 1.1.2, nous avons :

— $(x_1, y_1) = (5, 0); k_1 = 0$

— $(x_2, y_2) = (1, 0); k_2 = 0$

— $(x_3, y_3) = (5, 0); k_3 = 1$

— $(x_4, y_4) = (3, 0); k_4 = 1$

— $(x_5, y_5) = (4.5, 0); k_5 = 1$

En estimant nous obtenons :

$$\begin{aligned}\hat{\lambda} &= \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n}(x_i - y_i)} \\ &= \frac{1+1+1}{5+1+5+3+4.5} = \frac{3}{18.5} \\ &= 0.16,\end{aligned}$$

soit 16 000 cas par 100 000 personnes-années.

3.3.3 Processus de renouvellement alterné

Enfin, nous présentons une situation motivée par l'exemple 2 de la section 1.1.2. Imaginez qu'après l'infection, il y ait une période aléatoire de récupération. Nous présentons ci-dessous cette situation où chaque personne commence au temps $t = 0$ dans un état sain. Voir la figure 3.1.

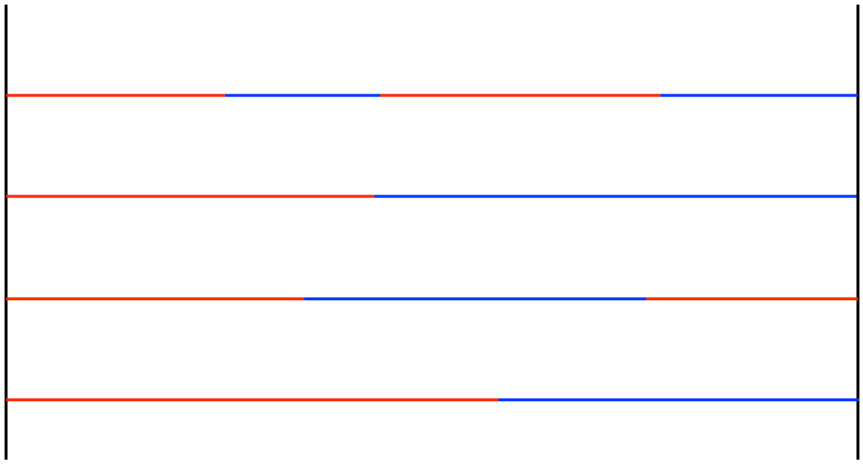


Figure 3.1 Schéma de $n = 4$ sujets alternant entre les états sain (rouge) et malade (bleu). Tous les sujets commencent à l'état sain (1) au temps $t = 0$ (disons le temps du calendrier au diagnostic).

Un tel modèle est un processus de renouvellement alterné. Le nombre de transitions d'un état à un autre que nous observons pour l'individu i est τ_i. Nous rappelons que le dernier temps de séjour sera censuré à droite à la fin de la période d'observation.

Pour estimer l'incidence de la maladie (passage de l'état *sains* vers l'état *malades*), λ_{12}, nous avons pour chaque sujet i les situations suivantes :

— τ_i est impair alors $\frac{\tau_i}{2} = \lceil \frac{\tau_i}{2} \rceil$ est le nombre de transitions vers l'état 2, malade. Par exemple, si $\tau_i = 1$ alors $\lceil \frac{\tau_i}{2} \rceil = 1$, il y a eu une transition vers l'état malade et si $\tau_i = 3$ alors $\lceil \frac{\tau_i}{2} \rceil = 2$, donc il y a eu deux transitions vers l'état malade ;

— τ_i est pair alors $\frac{\tau_i}{2} = \lceil \frac{\tau_i}{2} \rceil$ est le nombre de transitions vers l'état 2, malade. Par exemple, si $\tau_i = 2$ alors $\lceil \frac{\tau_i}{2} \rceil = 1$, il y a eu une transition vers l'état malade et si $\tau_i = 4$ alors $\lceil \frac{\tau_i}{2} \rceil = 2$, donc il y a eu deux transitions vers l'état malade.

Pour estimer l'incidence du retour dans l'état sain (passage de l'état malade vers l'état sain), λ_{21}, nous avons pour chaque sujet i les situations suivantes :

— τ_i est impair alors $\frac{\tau_i}{2} = \lfloor \frac{\tau_i}{2} \rfloor$ est le nombre de transitions vers l'état 1, sain. Par exemple, si $\tau_i = 1$ alors $\lfloor \frac{\tau_i}{2} \rfloor = 0$, il n'y a eu aucune transition vers l'état sain et si $\tau_i = 3$ alors $\lfloor \frac{\tau_i}{2} \rfloor = 1$, donc il y a eu une transition vers l'état sain ;

— τ_i est pair alors $\frac{\tau_i}{2} = \lfloor \frac{\tau_i}{2} \rfloor$ est le nombre de transitions vers l'état 1, sain. Par exemple, si $\tau_i = 2$ alors $\lfloor \frac{\tau_i}{2} \rfloor = 1$, il y a eu une transition vers l'état sain et si $\tau_i = 4$ alors $\lfloor \frac{\tau_i}{2} \rfloor = 2$, donc il y a eu deux transitions vers l'état sain.

En prenant le log de la vraisemblance (1.10) pour tous les n sujets, on a :

$$\sum_{i=1}^{n} \left[(\lceil \frac{\tau_i}{2} \rceil) \log \lambda_{12} - \sum_{j=1}^{\lceil \frac{\tau_i}{2} \rceil + I(\tau_i = \text{pair})} \lambda_{12} x_{ij} \right.$$
$$\left. + (\lceil \frac{\tau_i}{2} \rceil - I(\tau_i = \text{impair})) \log \lambda_{21} - \sum_{k=1}^{\lceil \frac{\tau_i}{2} \rceil} \lambda_{21} y_{ik} \right].$$

En effectuant les dérivées partielles par rapport à λ_{12} et λ_{21} on trouve :

$$\hat{\lambda}_{12} = \frac{\sum_{i=1}^{n} (\lceil \frac{\tau_i}{2} \rceil)}{\sum_{i=1}^{n} \sum_{j=1}^{\lceil \frac{\tau_i}{2} \rceil + I(\tau_i = \text{pair})} x_{ij}}, \tag{3.7}$$

$$\hat{\lambda}_{21} = \frac{\sum_{i=1}^{n} (\lceil \frac{\tau_i}{2} \rceil - I(\tau_i = \text{impair}))}{\sum_{i=1}^{n} \sum_{k=1}^{\lceil \frac{\tau_i}{2} \rceil} y_{ik}}. \tag{3.8}$$

À noter que dans chaque cas on revient à la définition du taux d'incidence de l'expression (1). L'expression (3.7) estime l'incidence de la maladie, car le temps à risque est tout le temps passé dans l'état 1 (l'état sain). Un sujet ne peut devenir malade que s'il était en bonne santé.

Le dénominateur $\sum_{i=1}^{n} \sum_{j=1}^{\lceil \frac{\tau_i}{2} \rceil + I(\tau_i = \text{pair})} x_{ij}$ est l'ensemble des personnes-temps passées dans l'état 1 (état sain) par les n individus. Le numérateur $\sum_{i=1}^{n} (\lceil \frac{\tau_i}{2} \rceil)$ est le nombre de transitions de l'état sain vers l'état malade observées pour les n individus. De même, l'expression (3.8) estime l'incidence de redevenir en bonne santé (dans l'état sain), car le temps à risque est tout le temps passé dans l'état 2 (état malade). Un sujet ne peut passer à l'état sain que s'il était malade. Le dénominateur $\sum_{i=1}^{n} \sum_{k=1}^{\lceil \frac{\tau_i}{2} \rceil} y_{ik}$ est l'ensemble des personnes-temps passées dans l'état 2 (état malade) par les n individus. Le numérateur $\sum_{i=1}^{n} (\lceil \frac{\tau_i}{2} \rceil - I(\tau_i = \text{impair}))$ est le nombre de transitions de l'état malade vers l'état sain observées pour tous les n individus.

Bien que nous ne présentons pas la dérivation, on peut étendre le travail ci-dessus à la situation où une cohorte est prise au temps du calendrier (voir la figure 3.2 ci-dessous). Ici, comme dans la sous-section 1.2.5, il y a un temps de troncation associé au premier séjour dans l'état 1 pour chaque individu. Le temps du calendrier auquel les données ont été collectées est notre ligne de base. Le premier temps de séjour n'est que partiellement observé. Le temps d'observation avant le temps du calendrier constitue la troncature à gauche. Il existe différentes approches pour traiter les temps de séjour tronqués. Souvent, s'il y a beaucoup de temps observés dans l'intervalle d'étude, les temps de séjour tronqués sont abandonnés. Certains auteurs supposent la stationnarité lorsqu'ils traitent la troncature. De nombreux auteurs conditionnent sur le temps de troncature (voir par exemple (Cook et Lawless, 2018)).

Les calculs, quand on conditionne sur les temps de troncature (développés, mais non présentés dans ce travail), donnent des estimations pour les intensités de transition constantes de la même forme qu'à l'expression (1).

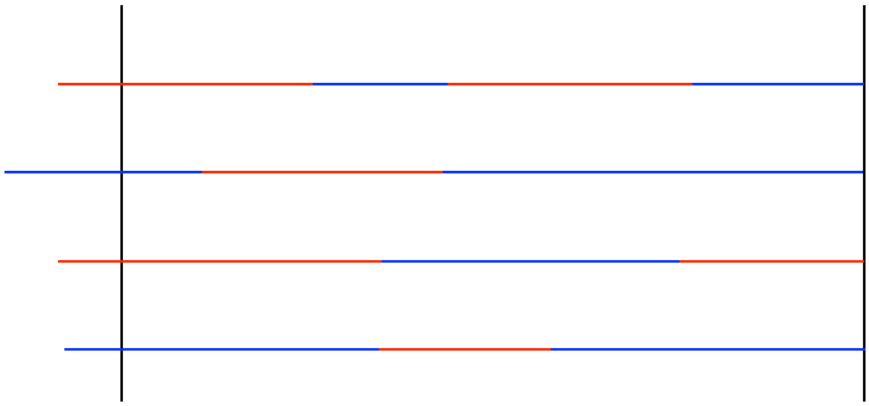


Figure 3.2 Schéma de $n = 4$ sujets alternant entre les états sain (rouge) et malade (bleu). Il s'agit d'une cohorte dont l'échelle de temps est l'instant du calendrier et où le premier temps de séjour dans l'état 1 pour chaque individu est précédé d'un temps de troncation.

CONCLUSION

Dans cette conclusion, nous résumons brièvement chaque chapitre et discutons des extensions possibles pour un travail futur. Avant de commencer notre discussion, revenons d'abord à l'exemple 4 de la sous-section 1.1.2 du chapitre 1. Cet exemple 4 tiré de (Dicker *et al.*, 2012) est présenté de manière confuse par ces auteurs. Le cadre théorique présenté dans ce mémoire peut aider à clarifier cet exemple. Le fait que le modèle montre à la fois le début de la maladie et le début du rétablissement suggère un modèle de renouvellement alterné où il y a une période de rétablissement avant qu'un individu puisse retomber malade (voir la section 3.3.3 pour plus de détails). Le cas où les individus passent de l'état malade à l'état absorbant (décès) sans avoir une période de rétablissement suggère un modèle *maladie-décès*. Ces points ne sont pas abordés dans l'exemple 4 de (Dicker *et al.*, 2012). De plus, dans cet exemple, on peut se demander si les sujets qui ont été malades, mais pas encore guéris (comme le sujet 1) devraient être inclus lors de l'estimation des personnes-temps au dénominateur.

Réflexions sur le chapitre 2

L'expression de Selvin fournit une approche générale où le risque n'a pas besoin d'être constant. Il existe un certain nombre d'extensions intéressantes possibles au travail que nous avons présenté.

Section 2.2 :

Dans cette section, nous avons appliqué l'idée d'estimateur par « plug-in » de la section 2.2 au cas de la censure à droite. Naïvement, nous avons utilisé l'estimateur de Kaplan-Meir pour la fonction de survie dans l'expression de Selvin (2.1). Nous avons vu que lorsqu'il y avait une censure dans l'intervalle d'observation, l'estimateur par « plug-in » n'estimait pas le taux d'incidence donné à l'expression (1).

Une deuxième tentative consisterait à redéduire l'estimateur de Selvin pour le cas avec censure à droite. Soit S la fonction de survie pour les temps de survie et soit G la fonction de survie pour les temps de censure. En utilisant les temps observés $X_i = \min(T_i, C_i)$, on peut calculer les personnes-temps à risque dans l'intervalle $(t, t + dt]$ en termes de la variable aléatoire X_i comme $\int_t^{t+dt} S(u)G(u)du$. Il serait donc intéressant d'étudier l'estimateur :

$$\frac{S(t) - S(t + dt)}{\int_t^{t+dt} S(u)G(u)du}. \tag{3.9}$$

Premièrement, nous voudrions vérifier l'interprétation comme nous l'avons fait dans la sous-section 2.1.1 et deuxièmement, nous voudrions étudier un estimateur par « plug-in » comme cela a été fait dans la sous-section 2.1.3. On utiliserait les estimateurs empiriques pour S et G. En cas de succès, on pourrait étendre cette approche au cas avec troncature à gauche seule et par la suite au cas avec troncature à gauche et censure à droite.

Comme nous l'avions précisé au point 15 de la section 0.2 de l'introduction, dans le cadre des modèles multi-états (modélisation de Markov homogène continu), le taux d'incidence estimé peut-être relié à une intensité de transition constante.

Il serait intéressant d'essayer une approche similaire de l'estimateur de Selvin au taux de transitions non constants pour certains modèles multi-états où les fonctions de survie de certains temps de séjour sont estimées par des estimateurs de type produit-limite.

Section 2.3 :

Ici, nous avons considéré quelques approches couramment utilisées pour estimer les personnes-temps au dénominateur lorsqu'on est en possession de données de type recensement.

Il serait intéressant d'estimer plus finement les personnes-temps lorsque les populations sont dynamiques ou ouvertes. Des paramètres estimant la dynamique des populations existent sur de nombreux sites gouvernementaux tels que :

`https://www.statcan.gc.ca/eng/subjects-start/population_and_demography`

ce qui pourrait être utile pour estimer les personnes-années à risque. Des références en démographie peuvent être utiles dans (Keyfitx et Caswell, 2005). Un facteur limitatif serait bien sûr l'exactitude de ces estimateurs qui devrait également être étudiée.

Réflexions sur le chapitre 3

Dans ce chapitre, nous avons estimé une fonction de risque constante qui est égale au taux d'incidence pour une variété de situations. Pour chaque exemple et chaque modèle que nous avons utilisé, l'estimation du risque constant était toujours égale à l'expression (1), soit le nombre de cas par personnes-temps à risque.

Les vraisemblances que nous avons utilisées ont été présentées dans la section 1.2 avec l'hypothèse que la fonction de risque était constante (Markov homogène).

Une deuxième hypothèse était que toutes les censures à droite et troncatures à gauche étaient indépendantes. Il serait intéressant d'utiliser des méthodes modernes telles que la probabilité inverse des poids de censure (en anglais :*Inverse Probability of Censoring Weights*) pour étudier les situations où la censure n'est pas indépendante. On pourrait faire des simulations pour voir quel effet la censure et la troncature non indépendantes pourraient avoir sur la fonction de risque constante et, par conséquent, le taux d'incidence et comparer les résultats à ceux obtenus à partir de l'expression (1).

À noter que les exemples que nous avons présentés à la sous-section 1.1.2 sont très typiques des exemples présentés dans la littérature épidémiologique et aucun d'entre eux ne mentionne la censure ou la troncature ni l'hypothèse d'indépendance.

Extensions plus générales

Les premières extensions plus générales que nous considérons concernent l'estimation de la variance. L'estimation de la variance pour le taux d'incidence repose souvent sur une approximation normale asymptotique et/ou sur la loi de Poisson (voir (Rosner, 2010) et (Newman, 2001)). Ci-dessous, nous nous concentrons sur la présentation de Poisson (souvent en utilisant une loi normale asymptotique).

Un argument simple (voir par exemple la section 10.1.3 dans (Newman, 2001)) justifiant la loi de Poisson est le suivant : lorsque les événements sont rares et qu'il n'y a pas beaucoup de censure, on peut prendre les personnes-temps cumulés comme constantes et égales au nombre de personnes multiplié par la longueur de l'intervalle d'étude. En réalité, les personnes-temps sont plus petites qu'elles devraient être, mais avec des événements rares et peu de censure, elles ne devraient pas être beaucoup plus petites.

Ceci justifie alors le fait de considérer le nombre d'événements comme suivant une loi de Poisson de paramètre $(nT\lambda)$ où n est le nombre d'individus, T est la longueur de l'intervalle et λ est le taux d'incidence. Soit d le nombre d'événements,

$$d \sim \text{Poisson}(nT\lambda).$$

Nous pouvons noter qu'en prenant les personnes-temps comme constantes et égales à nT, la vraisemblance est proportionnelle à la vraisemblance de la loi exponentielle,

$$L \propto \exp(-\lambda nT)\lambda^d.$$

L'estimateur du maximum de vraisemblance est $\hat{\lambda} = \frac{d}{nT}$, et étant donné que le dénominateur est considéré comme constant, la variance est estimée comme $\text{Var}[\hat{\lambda}] = \frac{d}{n^2T^2}$. Les intervalles de confiance sont souvent construits en conséquence soit avec une approximation normale, soit avec un intervalle de confiance exact utilisant une loi de Poisson (voir (Rosner, 2010)). Par exemple, avec l'approximation normale on obtient :

$$\frac{d}{nT} \pm z_{\alpha/2}\frac{\sqrt{d}}{nT}, \tag{3.10}$$

où le niveau de confiance $C = 1-\alpha$ et $z_{\alpha/2}$ est le quantile $1-\alpha/2$ de la loi normale standard. Souvent, au lieu d'utiliser nT, on utilise les personnes-années observées pendant le suivi.

Ci-dessous, en s'inspirant des travaux non publiés de (Atherton *et al.*, 2018) [4], nous introduisons une présentation plus formelle de l'approximation de Poisson. Considérons le taux d'incidence dans l'intervalle $[0, t)$ et une cohorte (échantillon fermé) commençant au temps $t = 0$. Pour l'instant, nous autorisons toute distribution F pour la fonction de survie. Notons par $N(t)$ le processus de comptage qui compte le nombre d'événements du temps 0 à t. On voit facilement que pour un temps fixe T le processus de comptage suit la loi binomiale avec une probabilité de succès égale à $F(T)$:

$$P[N(T-) = k] = \frac{n!}{k!(n-k)!}[F(T)]^k[S(T)]^{n-k}. \tag{3.11}$$

Ensuite, avec l'approximation bien connue de la loi binomiale par la loi de Poisson, en laissant $\tilde{\lambda} = nF(T)$, nous avons :

$$P[N(T-) = k] \approx \frac{\tilde{\lambda}^k \exp(-\tilde{\lambda})}{k!} \tag{3.12}$$

quand $n \to \infty$ et $p \to 0$.

Ci-dessus, nous avons montré l'approximation de Poisson qui fonctionne pour toute distribution du temps de survie tant que $F(T)$ est petite pour la longueur de l'intervalle de temps T.

Si nous restreignons $F(t) \approx \exp(-\lambda t)$ à être exponentielle, alors :

$$\tilde{\lambda} = np = nF(T) = n[1 - \exp(-\lambda T)].$$

---

4. Atherton, J., Ferland, R. et Lefebvre, G. (2018). Using Order Statistics to Link the Hazard Function and the Intensity Process in the Presence of Right Censoring, Left Truncation and Covariates. *Submitted to Statistical Science.*

En utilisant une approximation de Taylor pour un petit λT, nous avons $\tilde{\lambda} \approx \lambda nT$. C'est la relation bien connue entre le paramètre dans la loi de Poisson et le paramètre dans la loi exponentielle.

(Kalbfleisch et Prentice, 2002) dérivent les intervalles de confiance de Wald et de score pour le paramètre des temps de survie à loi exponentielle en présence d'observations indépendantes censurées à droite. Peut-être, sans surprise, les deux intervalles de confiance sont égaux à l'expression (3.10) avec nT remplacé par les personnes-temps observées.

Nous pensons qu'il y a deux raisons principales pour lesquelles peu d'études ont été faites concernant l'estimation de la variance du taux d'incidence :

— dans le cas de données de type expérimental, il y a très probablement suffisamment de détails pour modéliser la fonction de risque fournissant beaucoup plus d'informations que la mesure récapitulative du taux d'incidence. Dans de tels cas, la modélisation de la fonction de risque serait préférée et il y aurait peu d'intérêt à calculer le taux d'incidence et son intervalle de confiance ;

— dans le cas de données de type recensement, les données ne sont pas assez fines pour modéliser la fonction de risque et le taux d'incidence ne peut être qu'approximé. Lors de l'estimation de la variance, nous estimons, au mieux, la variance pour un estimateur qui ne peut être qu'approximé.

Malgré les observations ci-dessus, nous pensons qu'il est important d'étudier de manière plus approfondie le comportement de la variance pour le taux d'incidence dans la littérature. À notre connaissance, cela n'a pas encore été fait.

Nous concluons en listant quelques extensions générales. Beaucoup de ces idées pourraient être étudiées à l'aide de simulations. Nous suggérons d'utiliser des fonctions de risque non constantes et d'étudier les situations où la loi de Poisson ne peut être vérifiée. Avec des simulations, on pourrait calculer la couverture réelle des intervalles de confiance et peut-être développer des règles empiriques d'application de l'approximation de Poisson couramment en usage et présentée ci-dessus.

— L'hypothèse de Poisson a été critiquée dans le passé (Windeler et Lange, 1995). Il serait intéressant de revenir à la cohorte fermée et à l'expression (3.11) pour les cas où $F(T)$ est plus grande et n est plus petit. Ici, les événements ne seraient pas rares et la variabilité aléatoire des personnes-temps jouerait un rôle plus important.

— *Pour l'estimation de la variance sans censure (une approche paramétrique)* : on pourrait considérer différentes distributions paramétriques dans l'expression (2.1) de Selvin puis essayer la méthode du bootstrap paramétrique et non paramétrique pour l'estimation de la variance.

— *Pour l'estimation de la variance sans censure (une approche non paramétrique)* : on pourrait appliquer le bootstrap non paramétrique (échantillonnage avec remise) pour accéder à la variabilité de notre estimateur par « plug-in » basé sur l'expression (2.1) de Selvin.

Les deux dernières idées ci-dessus peuvent être étendues aux cas avec censure à droite et troncature à gauche une fois que l'on a trouvé des estimateurs appropriés pour le taux d'incidence basés sur l'expression (2.1) de Selvin.

APPENDICE A

PROCESSUS DE RENOUVELLEMENT ALTERNÉ

Nous pouvons utiliser les équations directes pour résoudre $P_{11}(0,t)$, $P_{12}(0,t)$,$P_{21}(0,t)$ et $P_{22}(0,t)$ avec les conditions initiales $P_{11}(0,0)=1$, $P_{12}(0,0)=0$, $P_{21}(0,0)=0$ et $P_{22}(0,0)=1$. Les équations directes sont :

$$\begin{aligned}
\frac{dP_{11}(0,t)}{dt} &= \lambda_{21}P_{12}(0,t) - \lambda_{12}P_{11}(0,t) \\
\frac{dP_{12}(0,t)}{dt} &= \lambda_{12}P_{11}(0,t) - \lambda_{21}P_{12}(0,t) \\
\frac{dP_{21}(0,t)}{dt} &= -\lambda_{12}P_{21}(0,t) + \lambda_{21}P_{22}(0,t) \\
\frac{dP_{22}(0,t)}{dt} &= \lambda_{12}P_{21}(0,t) - \lambda_{21}P_{22}(0,t).
\end{aligned} \tag{A.1}$$

En additionnant les deux premières équations ainsi que les deux dernières équations, nous obtenons :

$\frac{dP_{11}(0,t)}{dt} + \frac{dP_{12}(0,t)}{dt} = 0$ et $\frac{dP_{21}(0,t)}{dt} + \frac{dP_{22}(0,t)}{dt} = 0$.

En intégrant et en utilisant les conditions initiales, nous trouvons :

$$P_{12}(0,t) = 1 - P_{11}(0,t) \tag{A.2}$$

et

$$P_{21}(0,t) = 1 - P_{22}(0,t). \tag{A.3}$$

En substituant les équations (A.2) et (A.3) dans les équations directes, nous obtenons deux équations différentielles chacune en termes de $P_{11}(0,t)$ et $P_{22}(0,t)$ uniquement. Cela conduit aux solutions :

$$\begin{aligned} P_{11}(0,t) &= \frac{\lambda_{12}}{\lambda_{12}+\lambda_{21}} \exp(-(\lambda_{12}+\lambda_{21})t) + \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} \\ P_{22}(0,t) &= \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} \exp(-(\lambda_{12}+\lambda_{21})t) + \frac{\lambda_{12}}{\lambda_{12}+\lambda_{21}}. \end{aligned} \tag{A.4}$$

Pour trouver $P_{12}(0,t)$ et $P_{21}(0,t)$, on revient aux équations (A.2) et (A.3).

Dans les équations ci-dessus, t est le temps écoulé depuis le début du processus. La vraisemblance, cependant, doit être exprimée en termes des temps de séjour.

APPENDICE B

EMV POUR UNE DISTRIBUTION EXPONENTIELLE

Ci-dessous, nous présentons rapidement la solution de l'EMV pour des temps de survie complètement observés. Référons-nous à la section 2.1.2 :

$$L = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i)$$

$$\ell = \log(L) = n \log \lambda - \sum_{i=1}^{n} \lambda t_i$$

$$\frac{d\ell}{d\lambda} = 0$$

$$\frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

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

APPENDICE C

EXTENSIONS DE L'ESTIMATEUR DU TAUX D'INCIDENCE PAR PLUG-IN

Dans cet appendice, nous présentons différents schémas pour illustrer les estimateurs du taux d'incidence par injection ou « plug-in » en utilisant deux estimateurs de la fonction de survie (la fonction de survie empirique et la fonction de survie de Kaplan-Meier) dans l'expression (2.1) de Selvin. Pour les schémas C.1, C.2 et C.3, chaque ligne horizontale représente la durée de survie d'un sujet jusqu'à l'événement ou à la censure. Nous considérons que les sujets étaient sains au début de l'étude. Le rond (O) sur les lignes horizontales symbolise une perte de vue du sujet (censure) et la croix (X) symbolise l'événement (maladie). L'âge zéro (0) est l'âge du sujet lorsqu'il entre dans l'étude. Nous commençons par présenter le cas simple de non-censure (figure C.1) et finissons par les cas censurés (figures C.2 et C.3).

C.1 L'estimateur non paramétrique de la fonction de survie empirique : cas de non-censure

Dans cette section nous présentons un exemple fictif de données non censurées où tous les sujets ont connu l'événement. Nous utilisons l'estimateur de la fonction de survie empirique dans l'expression (2.1) de Selvin afin d'estimer le taux d'incidence.

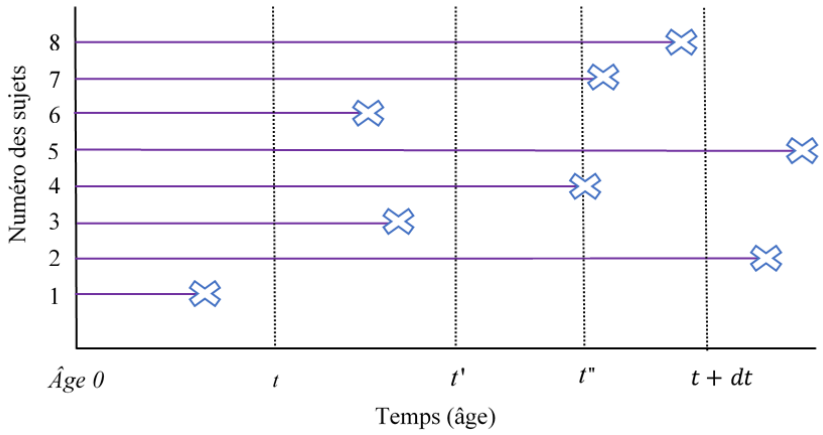


Figure C.1 Exemple fictif d'une population fermée sans censure de 8 sujets.

Soit n_t le nombre de personnes à risque au temps t. Pour cet exemple, l'estimateur du taux d'incidence par injection ou « plug-in » est alors calculé comme suit :

$$\begin{aligned}\frac{\hat{S}(t)-\hat{S}(t+dt)}{\int_t^{t+dt}\hat{S}(u)du} &= \frac{\frac{n_t}{n}-\frac{n_{t+dt}}{n}}{\frac{n_t}{n}(t'-t)+\frac{n_{t'}}{n}(t''-t')+\frac{n_{t''}}{n}(t+dt-t'')} \\ &= \frac{n_t-n_{t+dt}}{n_t(t'-t)+n_{t'}(t''-t')+n_{t''}(t+dt-t'')}\end{aligned} \tag{C.1}$$

ce qui équivaut à l'expression (1) du taux d'incidence.

C.2 Estimateur du taux d'incidence en utilisant la fonction de survie de Kaplan-Meier (KM)

Dans cette section nous présentons un autre exemple fictif avec des données censurées à droite. Nous utilisons l'estimateur de la fonction de survie de Kaplan-Meier dans l'expression (2.1) de Selvin afin d'estimer le taux d'incidence.

Nous remarquons à la figure C.2 que le premier temps d'événement est observé au temps t' et que le nombre de personnes à risque au temps t' est de $n_{t'} = n-1$, celui au temps t'' est de $n_{t''} = n-3$ et celui au temps $t+dt$ est de $n_{t+dt} = n-4$.

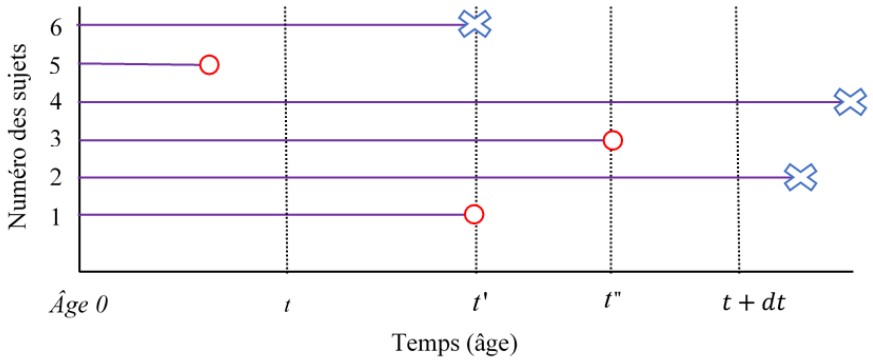


Figure C.2 Exemple fictif d'une population fermée censurée à droite de 6 sujets.

La survie pour cet exemple est obtenue comme ci-dessous :

$$\begin{aligned}
&\hat{S}_{KM}(t) = 1 \\
&\hat{S}_{KM}(t') = \left(1 - \frac{1}{n-1}\right) \\
&\hat{S}_{KM}(t'') = \hat{S}_{KM}(t') = \left(1 - \frac{1}{n-1}\right) \\
&\hat{S}_{KM}(t+dt) = \hat{S}_{KM}(t'') = \left(1 - \frac{1}{n-1}\right).
\end{aligned} \tag{C.2}$$

L'estimateur du taux d'incidence par « plug-in » est obtenu, après simplification pour $n = 6$ et $dt = (t' - t) + (t'' - t') + (t + dt - t'')$, comme suit :

$$\begin{aligned}
&\frac{\hat{S}_{KM}(t) - \hat{S}_{KM}(t+dt)}{\int_t^{t+dt} \hat{S}_{KM}(u)du} \\
&= \frac{1 - (1 - \frac{1}{n-1})}{1 \times (t' - t) + (1 - \frac{1}{n-1}) \times (t'' - t) + (1 - \frac{1}{n-1}) \times (t + dt - t'')} \\
&= \frac{1}{5(t' - t) + 4(t'' - t') + 4(t + dt - t'')}.
\end{aligned} \tag{C.3}$$

L'estimateur du taux d'incidence tel que donné en (C.3) n'est pas équivalent à l'estimateur du taux d'incidence à l'équation (1). Une explication serait que

l'estimateur de la fonction de survie de KM est une fonction en escalier qui, en cas de censure, ne descend pas. En effet, pour notre exemple, $\hat{S}_{KM}(t+dt) = \hat{S}_{KM}(t'') = \hat{S}_{KM}(t')$, c'est-à-dire, le nombre de personnes à risque au temps t' reste le même qu'au temps t''.

Maintenant, essayons de modifier l'exemple de la figure C.2. Supposons que les sujets 1 et 3 ayant été censurés dans l'exemple précédent à la figure C.2 ont connu l'événement et les sujets 2 et 4 sont toujours censurés par la fin d'étude (c'est-à-dire qu'ils ont connu l'événement en dehors de l'intervalle d'étude). Que pourrait-il se passer dans cette nouvelle situation ? Graphiquement, nous pouvons présenter ce schéma modifié comme ci-dessous.

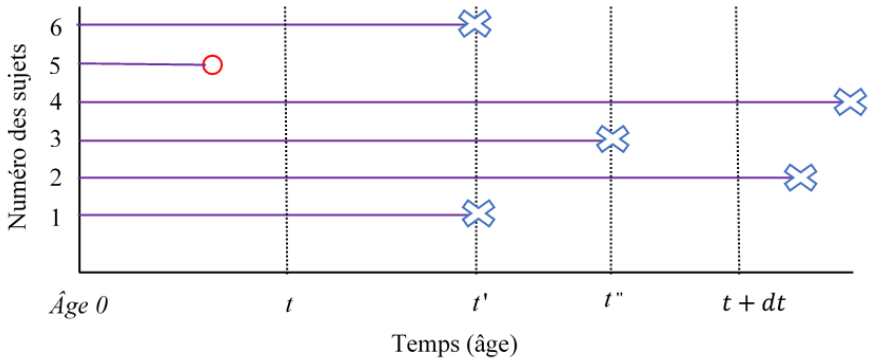


Figure C.3 Exemple modifié (population fermée censurée à droite de 6 sujets).

Dans ce nouveau cadre, nous avons :

$$\begin{aligned}
\hat{S}_{KM}(t) &= 1 \\
\hat{S}_{KM}(t') &= (1 - \frac{2}{n-1}) \\
\hat{S}_{KM}(t'') &= (1 - \frac{2}{n-1})(1 - \frac{1}{n-3}) \\
&= \frac{(n-1)(n-3) - (n-1) - 2(n-3) + 2}{(n-1)(n-3)} \\
\hat{S}_{KM}(t+dt) &= \hat{S}_{KM}(t'') = (*).
\end{aligned} \tag{C.4}$$

Alors, l'estimation du taux d'incidence par « plug-in », est obtenu, après simplification pour $n = 6$, $dt = (t'-t)+(t''-t')+(t+dt-t'')$ et $(t''-t) = (t'-t)+(t''-t')$, comme suit :

$$\begin{aligned}
&\frac{\hat{S}_{KM}(t) - \hat{S}_{KM}(t+dt)}{\int_t^{t+dt} \hat{S}_{KM}(u)du} \\
&= \frac{1 - \frac{(n-1)(n-3)+(n-1)+2(n-3)-2}{(n-1)(n-3)}}{1 \times (t'-t) + (1-\frac{2}{n-1}) \times (t''-t') + (*) \times (t+dt-t'')} \\
&= \frac{3}{5(t'-t) + 3(t''-t') + 2(t+dt-t'')}.
\end{aligned} \tag{C.5}$$

Cette fois-ci, après les modifications faites à l'exemple de la figure C.2, l'expression (C.5) est équivalente au taux d'incidence à l'équation (1).

APPENDICE D

CODE R : CALCULS DU TAUX D'INCIDENCE

Ce code R reprent les calculs du taux d'incidence présentés à l'exemple 3 de la sous-section 3.2.3.

```
x <- c(57.1,65,61.4, 60,65,65,61.1,65,65,59.9,65,65,65)
y <- c(55,55,55,55,55,55,55,55,55,56,57,59,59)
delta <- c(1,0,0,1,0,0,1,0,0,1,0,0,0)
Lambda <- sum(delta)/(sum(x-y))
Lambda
```

RÉFÉRENCES

Aalen, O. O., Borgan, O. et Gjessing, H. K. (2008). *Survival and Event History Analysis : A process Point of View*. Statistics for Biology and Health. Springer Science & Business Media.

Beyersmann, J., Schumacher, M. et Allignol, A. (2012). *Competing Risks and Multistate Models with R*. Springer Science & Business Media.

Boyle, P. et Parkin, D. M. (1996). Méthodes statistiques pour les registres. In O. M. Jensen, D. M. Parkin, R. MacLeannan, C. S. Muir, et R. G. Skeet (dir.). *Enregistrement des Cancers, Principes et Methodes ; IARC Publications Scientifiques*, numéro 95 chapitre 9, p. 135. Centre International de Recherches sur le Cancer (OMS).

Breslow, N. E. et Day, N. E. (1980). *Statistical Methods in Cancer Research VOLUME 1 - The Analysis of Case-Control Studies*. WHO-IARC Scientific Publications No. 32. International Agency for Research on Cancer, Lyon, France.

Clayton, D. et Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press.

Coeurjolly, J. F., Nguile-Makao, M., Timsit, J. F. et Liquet, B. (2012). Attributable risk estimation for adjusted disability multistate models : Application to nosocomial infections. *Biomedical Journal, 54(5)*, 600–616.

Cook, R. J. et Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media, New York, NY.

Cook, R. J. et Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. Chapman and Hall/CRC.

Cox et Miller. (1965). *The Theory of Stochastic Processes*. Chapman and Hall/CRC.

Dicker, R. C., Coronado, F., Koo, D. et Parrish, R. G. (2012). *Principles of Epidemiology in Public Health Practice : An Introduction to Applied Epidemiology and Biostatistics*.

Elandt-Johnson, R. C. (1975). Reviews and commentary - definitions of rates : Some remarks on their use and misuse. *American Journal of Epidemiology, 102(4)*, 267–271.

Elandt-Johnson, R. C. et Johnson, N. L. (1999). *Survival Models and Data Analysis*. John Wiley & Sons, Inc.

Grimmett, G. et Stirzaker, D. (2001). *Probability and Random Processes (Third Edition)*. Oxford University Press Inc., New York.

Jewell, N. P. (2003). *Statistics for Epidemiology*. Chapman and Hall/CRC.

Kalbfleisch, J. et Prentice, R. (2002). *The Statistical Analysis of Failure Time Data (Second Edition)*. John Wiley & Sons, Inc.

Kaplan, E. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53(282)*, 457–481.

Karlin, S. et Taylor, H. M. (1981). *A Second Course in Stochastic Process (First Edition)*. Academic Press.

Keyfitx, N. et Caswell, H. (2005). *Applied Mathematical Demography*. Springer Science & Business Media.

Klein, J. et Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Inc.

Morgenstern, H., Kleinbaum, D. et Kupper, L. L. (1980). Measures of disease incidence used in epidemiologic research. *International Journal of Epidemiology, 9(1)*, 97–104.

Newman, S. (2001). *Biostatistical Methods in Epidemiology*. John Wiley & Sons, Inc.

Rosner, B. (2010). *Fundamentals of Biostatistics (Seventh Edition)*. Belmont : Pacific Grove. Brooks/Cole, Boston, USA.

Rothman, K. J., Greenland, S. et Lash, T. L. (2008). *Modern Epidemiology*. Chapman and Hall/CRC.

Selvin, S. (2004). *Statistical Analysis of Epidemiologic Data*. Oxford Scholarship.

Selvin, S. (2008). *Survival Analysis for Epidemiologic and Medical Research*. Cambridge University Press.

Vandenbroucke, J. P. (1985). Reviews and commentary : On the rediscovery of a distinction. *American Journal of Epidemiology, 121(5)*, 627–628.

Vandenbroucke, J. P. (2003). Continuing controversies over "risks and rates" - more than a century after william farr's "on prognosis". *Soz - Praventivmed, 48*, 216–218.

Vandenbroucke, J. P. et Pearce, N. (2012). Education corner : Incidence rates in dynamic populations. *International Journal of Epidemiology, 41*, 1472–1479.

Walker, A. M. (1991). *Observation and Inference*. Epidemiology Resources Inc.

Windeler, J. et Lange, S. (1995). Events per person year - a dubious concept. *BMJ : British Medical Journal, 310(6977)*, 454–456.