Etude des délais de survenue des effets indésirables médicamenteux à partir des cas notifiés en pharmacovigilance : problème de l'estimation d'une distribution en présence de données tronquées à droite

Fanny Leroy

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

UNIVERSITÉ PARIS-SUD

UFR DE MÉDECINE

ÉCOLE DOCTORALE DE SANTÉ PUBLIQUE ED420

Année : 2014

Numéro attribué par la bibliothèque

THÈSE

en vue d'obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS-SUD

en Santé Publique

Spécialité : **BIOSTATISTIQUE**

Présentée et soutenue publiquement le 18 Mars 2014 par :

Fanny LEROY

Etude des délais de survenue des effets indésirables médicamenteux à partir des cas notifiés en pharmacovigilance : problème de l'estimation d'une distribution en présence de données tronquées à droite

Directeurs de thèse : Mme. Pascale TUBERT-BITTER et M. Jean-Yves DAUXOIS

Composition du jury :

Mme. Catherine QUANTIN, Pr.	Présidente
Mme. Shulamith GROSS, Pr.	Rapporteur
M. Michal ABRAHAMOWICZ, Pr.	Rapporteur
M. Bernard BÉGAUD, Pr.	Examinateur
M. Philippe BROËT, Pr.	Examinateur
M. Jean-Yves DAUXOIS, Pr.	Directeur de thèse
Mme. Pascale TUBERT-BITTER	Directrice de thèse

Remerciements

En premier lieu, je tiens à remercier chaleureusement mes directeurs de thèse, Pascale Tubert-Bitter et Jean-Yves Dauxois, pour la confiance qu'ils m'ont accordée, leur gentillesse, leurs encouragements et leur disponibilité tout au long de ces quatre années passées ensemble. Merci Pascale pour ta bienveillance, ton soutien permanent et ton optimisme à toute épreuve. Merci Jean-Yves pour toutes ces heures passées dans le train, l'avion, le métro ou encore le taxi, encadrer une thèse à distance n'est pas chose simple. Je te remercie pour tous ces efforts consentis, pour ta persévérance et ton investissement permanent.

Je remercie le centre régional de pharmacovigilance de Bordeaux, et tout particulièrement Hélène Théophile et Françoise Haramburu, de nous avoir fourni les données qui ont motivé ce travail et qui ont permis d'illustrer les méthodes développées tout au long de cette thèse. Un grand merci à elles pour leur aide ô combien précieuse dans la bonne exploitation de ces données et l'interprétation des résultats.

Je tiens à remercier les membres du jury qui m'ont fait l'honneur d'évaluer ce travail. Merci à Catherine Quantin d'avoir accepté de présider le jury ; merci à mes rapporteurs, Shulamith Gross et Michal Abrahamowicz, pour le temps qu'ils ont consacré à la lecture minutieuse du manuscrit, pour leurs commentaires précieux et constructifs. Merci enfin à mes examinateurs, Bernard Bégaud et Philippe Broët, d'avoir accepté d'évaluer mon travail.

Je remercie l'ensemble de l'équipe Biostatistique du CESP ; merci à Thierry Moreau, ex-directeur de l'U780, de m'avoir si bien accueillie à mon arrivée ; à Ghislaine Breton et

Edith Lesieux-Potier pour leur patience et leur persévérance dans toutes les démarches administratives ; à Aurélien Latouche pour ses conseils avisés ; à Sylvie Escolano pour sa spontanéité et sa bonne humeur communicative et merci à Ismaïl Ahmed pour ses lectures attentives et ses remarques pertinentes. Un grand merci également à Rachel Nadif et à l'ensemble de l'équipe Epidémiologie Respiratoire et Environnementale du CESP pour leur soutien et leur bonne humeur au quotidien.

Je remercie l'Université Paris-Sud qui a financé les trois premières années de ce travail de thèse grâce à une allocation de l'École Doctorale 420, la Fondation ARC pour la recherche sur le cancer qui m'a accordée une bourse de quatrième année (DOC20121206 119) et l'équipe Biostatistique qui a pris part au financement des derniers mois de cette thèse ainsi qu'aux congrès et séminaires auxquels j'ai participé tout au long de ces quatre ans.

Je remercie l'Université Paris Descartes de m'avoir donné l'opportunité d'enseigner au sein de l'UFR Mathématiques et Informatique. Ce fut une expérience très enrichissante, tant sur le plan professionnnel que personnel.

Mes prochains remerciements s'adressent à tous mes collègues du CESP que j'ai eu le plaisir de rencontrer et de côtoyer tout au long de ces quatre années, et tout particulièrement à Dorota, Helena, Juliette, Alexia, Margarita, Elsa, Ismaïl, Mohammed, Orianne, Margaux, Annabelle, Marta, Emilie, Zhen, Sylwester, Jonathan et Yves.

Un grand merci à tous mes amis qui ont suivi, de près ou de loin, toutes les péripéties survenues au cours de ces quatre ans. Merci à Eléonore, Joffrey, Mouna et David pour leur oreille attentive et leur optimisme sans faille.

Merci à ma famille et à la famille de Nicolas pour leur soutien et leurs encouragements au quotidien. Merci Maman pour tous ces déjeuners du jeudi midi, toutes ces pâtes, pizzas et sashimis engloutis pour la bonne cause. Enfin, j'écris ces derniers mots pour Nicolas, qui n'a jamais douté et qui est capable aujourd'hui de faire un résumé de ma thèse presque aussi bien que moi : MERCI.

Résumé

Ce travail de thèse porte sur l'estimation paramétrique du maximum de vraisemblance pour des données de survie tronquées à droite, lorsque les délais de troncature sont considérés déterministes. Il a été motivé par le problème de la modélisation des délais de survenue des effets indésirables médicamenteux à partir des bases de données de pharmacovigilance, constituées des cas notifiés. Les distributions exponentielle, de Weibull et log-logistique ont été explorées.

Parfois le caractère tronqué à droite des données est ignoré et un estimateur naïf est utilisé à la place de l'estimateur pertinent. Une première étude de simulations a montré que, bien que ces deux estimateurs - naïf et basé sur la troncature à droite - puissent être positivement biaisés, le biais de l'estimateur basé sur la troncature est bien moindre que celui de l'estimateur naïf et il en va de même pour l'erreur quadratique moyenne. De plus, le biais et l'erreur quadratique moyenne de l'estimateur basé sur la troncature à droite diminuent nettement avec l'augmentation de la taille d'échantillon, ce qui n'est pas le cas de l'estimateur naïf.

Les propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance ont été étudiées. Sous certaines conditions, suffisantes, cet estimateur est consistant et asymptotiquement normal. La matrice de covariance asymptotique a été détaillée. Quand le délai de survenue est modélisé par la loi exponentielle, une condition d'existence de l'estimation du maximum de vraisemblance, assurant ces conditions suffisantes, a été obtenue. Pour les deux autres lois, une condition d'existence de l'estimation du maximum de vraisemblance a été conjecturée.

A partir des propriétés asymptotiques de cet estimateur paramétrique, les intervalles de confiance de type Wald et de la vraisemblance profilée ont été calculés. Une seconde étude de simulations a montré que la couverture des intervalles de confiance de type Wald pouvait être bien moindre que le niveau attendu en raison du biais de l'estimateur du paramètre de la distribution, d'un écart à la normalité et d'un biais de l'estimateur de la variance asymptotique. Dans ces cas-là, la couverture des intervalles de la vraisemblance profilée est meilleure.

Quelques procédures d'adéquation adaptées aux données tronquées à droite ont été présentées. On distingue des procédures graphiques et des tests d'adéquation. Ces procédures permettent de vérifier l'adéquation des données aux différents modèles envisagés.

Enfin, un jeu de données réelles constitué de 64 cas de lymphomes consécutifs à un traitement anti TNF-α issus de la base de pharmacovigilance française a été analysé, illustrant ainsi l'intérêt des méthodes développées.

Bien que ces travaux aient été menés dans le cadre de la pharmacovigilance, les développements théoriques et les résultats des simulations peuvent être utilisés pour toute analyse rétrospective réalisée à partir d'un registre de cas, où les données sur un délai de survenue sont aussi tronquées à droite.

Abstract

This work investigates the parametric maximum likelihood estimation for right-truncated survival data when the truncation times are considered deterministic. It was motivated by the modeling problem of the adverse drug reactions time-to-onset from spontaneous reporting databases. The families of the exponential, Weibull and log-logistic distributions were explored.

Sometimes, right-truncation features of spontaneous reports are not taken into account and a naïve estimator is used instead of the truncation-based estimator. Even if the naïve and truncation-based estimators may be positively biased, a first simulation study showed that the bias of the truncation-based estimator is always smaller than the naïve one and this is also true for the mean squared error. Furthermore, when the sample size increases, the bias and the mean squared error are almost constant for the naïve estimator while they decrease clearly for the truncation-based estimator.

Asymptotic properties of the truncation-based estimator were studied. Under sufficient conditions, this parametric truncation-based estimator is consistant and asymptotically normally distributed. The covariance matrix was detailed. When the time-to-onset is exponentially distributed, these sufficient conditions are checked as soon as a condition for the maximum likelihood estimation existence is satisfied. When the time-to-onset is Weibull or log-logistic distributed, a condition for the maximum likelihood estimation existence was conjectured.

The asymptotic distribution of the maximum likelihood estimator makes it possible to derive Wald-type and profile likelihood confidence intervals for the distribution pa-

rameters. A second simulation study showed that the estimated coverage probability of the Wald-type confidence intervals could be far from the expected level because of a bias of the parametric maximum likelihood estimator, a gap from the gaussian distribution and a bias of the asymptotic variance estimator. In these cases, the profile likelihood confidence intervals perform better.

Some goodness-of-fit procedures adapted to right-truncated data are presented. Graphical procedures and goodness-of-fit tests may be distinguished. These procedures make it possible to check the fit of different parametric families to the data.

Illustrating the developed methods, a real dataset of 64 cases of lymphoma, that occurred after anti TNF-α treatment and that were reported to the French pharmacovigilance, was finally analyzed.

Whilst an application to pharmacovigilance was led, the theoretical developments and the results of the simulation study may be used for any retrospective analysis from case registries where data are right-truncated.

Valorisation scientifique

Articles

Leroy F., Dauxois J.Y., Théophile H., Haramburu F., Tubert-Bitter P., Estimating time-to-onset of adverse drug reactions from spontaneous reporting databases. *BMC Medical Research Methodology* 2014 **14** : 17. DOI : 10.1186/1471-2288-14-17.

Leroy F., Dauxois J.Y., Tubert-Bitter P., On the parametric maximum likelihood estimator for independent but non-identically distributed observations with application to truncated data. Soumis pour publication.

Leroy F., Dauxois J.Y., Tubert-Bitter P., Asymptotic confidence intervals based on the parametric maximum likelihood estimation from right-truncated time-to-adverse event data in postmarketing pharmacovigilance. Soumis pour publication.

Communications orales

Leroy F., Dauxois J.Y., Théophile H., Haramburu F., Tubert-Bitter P., Time to onset of adverse drug reactions : parametric modelling for spontaneously reported cases, 32^{th} *Annual Conference of the International Society for Clinical Biostatistics*, Août 2011, Ottawa.

Leroy F., Dauxois J.Y., Théophile H., Haramburu F., Tubert-Bitter P., Propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance pour des données tronquées à droite : Application à la pharmacovigilance, $44^{\text{èmes}}$ *Journées De Statistique de la Société Française de Statistiques*, Mai 2012, Bruxelles.

Table des matières

Liste des tableaux

Liste des figures

Notations

X	Variable aléatoire caractérisant le délai de survenue de l'effet indésirable médicamenteux
f, F	Distribution de X : densité, fonction de répartition
F^*	Partie identifiable de F de manière non-paramétrique
$\widehat{F}^*$	Estimateur non-paramétrique de F^*
τ	Borne supérieure du support de T
θ	Paramètre du modèle
θ^0	Vraie valeur du paramètre θ
$\widehat{\theta}_n$	Estimation de θ
$\widehat{\theta}_{\text{EN}}$	Estimateur naïf
$\widehat{\theta}_{\text{EBT}}$	Estimateur basé sur la troncature
T	Variable aléatoire caractérisant le délai de troncature
g, G	Distribution de T : densité, fonction de répartition
D	Médicament suspecté
E_D	Effet indésirable d'intérêt, imputable au médicament D
t_a	Date d'analyse
N	Taille de la population
n	Taille de l'échantillon observé
w	Probabilité pour une observation de ne pas être tronquée
$p = 1 - w$	Proportion théorique de données tronquées dans la population

t^*	Maximum des délais de troncature observés
κ	Probabilité que X soit inférieur strictement à τ
M	Nombre de délais de troncature distincts observables dans la population
$I\{A\}$	Fonction indicatrice de l'événement A
$\nabla_\theta f(\theta)$	Opérateur gradient : Vecteur des dérivées partielles de la fonction f par rapport aux différentes composantes de θ
$\Gamma(\cdot)$	Fonction gamma
$\xrightarrow[n\to+\infty]{P}$	Convergence en probabilité d'une suite de variables aléatoires indexée par n
$\xrightarrow[n\to+\infty]{d}$	Convergence en distribution d'une suite de variables aléatoires indexée par n

Introduction

Le cadre statistique dans lequel s'inscrit cette thèse est celui de l'estimation d'une distribution en présence de données tronquées à droite.

L'analyse des données de survie est le domaine de la statistique qui s'intéresse à la modélisation du délai jusqu'à la survenue d'un événement d'intérêt. Ce délai de survenue est caractérisé par une variable aléatoire X. A partir de l'observation d'un échantillon de réalisations de cette variable aléatoire X, un des enjeux de l'analyse des données de survie est l'estimation de la distribution de la variable aléatoire X.

Parfois, il n'est pas possible d'observer complètement cet échantillon de réalisations de la variable aléatoire X ; seules des réalisations partielles sont observées. On dit alors qu'on observe des données incomplètes. La censure et la troncature sont les deux causes de données incomplètes les plus répandues en analyse de données de survie. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt. On sait juste que ce délai appartient à un certain intervalle de temps. La troncature survient quand on ne peut observer que les individus de l'échantillon dont le délai de survenue appartient à un certain intervalle de temps. On observe donc un sous-échantillon. Les mécanismes de censure et de troncature peuvent survenir simultanément.

Soit A, A_L et A_U trois réels. La variable aléatoire X est dite tronquée à droite (respectivement tronquée à gauche) quand on ne peut observer une réalisation de X que si cette réalisation est inférieure (respectivement supérieure) au réel A. La variable aléatoire X est dite tronquée par intervalle quand on ne peut observer une réalisation de X que si cette réalisation appartient à l'intervalle $[A_L, A_U]$. Autrement dit, la variable

aléatoire X est tronquée par intervalle si elle est à la fois tronquée à droite et tronquée à gauche. Les réels A, A_L et A_U peuvent varier d'une réalisation à une autre. Ils peuvent être considérés comme déterministes ou aléatoires. Nous considérons dans cette thèse le cadre déterministe. L'observation des délais d'incubation du SIDA chez les patients infectés par une transfusion de sang contaminé dans les années 80 constitue le jeu de données tronquées à droite le plus étudié dans la littérature scientifique. Seuls les individus ayant été transfusés par du sang contaminé et ayant développé le SIDA avant la date de fin d'étude sont observés. Les individus entrent dans l'échantillon lorsque l'événement d'intérêt (le SIDA) survient ; leur délai de survenue est alors observé rétrospectivement. Autrement dit, seuls les individus dont le délai d'incubation du SIDA est inférieur au délai écoulé entre la date de la transfusion de sang contaminé et la date de fin d'étude sont observés.

La pharmacovigilance est la surveillance des médicaments ayant reçu une autorisation de mise sur le marché avec pour objectifs l'identification d'effets indésirables et la prévention du risque d'effet indésirable résultant de leur utilisation. Les essais cliniques réalisés avant la mise sur le marché du médicament échouent à identifier ces effets indésirables, lorsqu'ils sont rares ou de délai de survenue long, en raison des contraintes de réalisation de ces essais, ce qui a conduit à l'essor des méthodes de surveillance post-marketing au cours des dernières décennies. Les données de notifications spontanées constituent une source précieuse pour mener à bien les objectifs de la pharmacovigilance. Il s'agit de bases de données où sont enregistrés les effets indésirables suspectés et notifiés qui surviennent après la mise sur le marché des médicaments. Des méthodes fondées sur ces bases de données ont été développées pour la détection de signal, c'est-à-dire l'identification de potentiels effets indésirables. On distingue des méthodes de détection ciblée (Fourrier *et al.*, 2001; Tubert *et al.*, 1991; Moore *et al.*, 1997; Van der Heijden *et al.*, 2002) et des méthodes de détection automatisée (Bate *et al.*, 1998; DuMouchel, 1999; Szarfman *et al.*, 2002; Evans *et al.*, 2001; Ahmed *et al.*, 2009, 2010).

Il a récemment été suggéré que la modélisation du délai jusqu'à la survenue d'un effet indésirable médicamenteux pourrait être un complément utile aux méthodes de

détection de signal, que celles-ci soit fondées sur les notifications spontanées (Maignen *et al.*, 2010; Van Holle *et al.*, 2012) ou sur des données longitudinales observationnelles (Cornelius *et al.*, 2012). Il s'agit du délai écoulé entre une exposition à un médicament et la survenue d'un effet indésirable imputable à ce médicament. La distribution du délai de survenue d'un effet indésirable médicamenteux est lié au mécanisme d'action sous-jacent. L'estimation de la distribution du délai de survenue permettrait de mieux comprendre le mécanisme de survenue de l'effet indésirable médicamenteux, d'identifier des groupes d'individus à risque ou des fenêtres de temps à risque dans le déroulement d'un traitement et ainsi de prévenir ou diagnostiquer précocément la survenue des effets indésirables. Par ailleurs, l'estimation du délai de survenue permettrait d'améliorer la planification des études pharmacoépidémiologiques mises en place pour confirmer ou non un signal ; par exemple, des fenêtres de temps à risque doivent être choisies *a priori* pour les études épidémiologiques basées sur les cas seuls (Farrington, 1995; Maclure, 1991). L'estimation de la distribution du délai de survenue d'un effet indésirable médicamenteux, basée sur la source de données précieuse que constituent les notifications spontanées, contribue à mener à bien les principaux objectifs de la pharmacovigilance.

On considère dans cet exposé un médicament (ou classe médicamenteuse) et un effet indésirable, imputable à ce médicament, particuliers. Soit E_D l'effet indésirable d'intérêt, survenu consécutivement à la prise du médicament D. On souhaite estimer la distribution du délai jusqu'à la survenue de l'effet indésirable E_D, caractérisé par la variable aléatoire X, à partir d'un échantillon de réalisations de cette variable aléatoire X issu des notifications spontanées. Soit t_a la date à laquelle on effectue l'analyse des données de survie. Cette date coïncide avec la date d'extraction des données à partir des notifications spontanées. Etant donné que ces bases de données incluent les individus exposés à un médicament qui ont présenté un effet indésirable imputable à ce médicament, seuls les délais de survenue des patients exposés au médicament D qui ont développé l'effet indésirable E_D avant la date d'analyse t_a peuvent être observés. Autrement dit, les individus dont le délai jusqu'à la survenue de l'effet indésirable E_D est inférieur ou égal au délai écoulé entre la date d'exposition et la date d'analyse t_a

sont observés. Par conséquent, les données sur le délai jusqu'à la survenue de l'effet indésirable E_D issues des notifications spontanées sont tronquées à droite. La problématique de cette thèse est donc l'estimation d'une distribution en présence de données tronquées à droite.

L'estimateur non-paramétrique du maximum de vraisemblance en présence de données tronquées à droite a été développé et utilisé pour estimer le délai d'incubation du SIDA chez les patients infectés par du sang contaminé (Lagakos *et al.*, 1988; Kalbfleisch et Lawless, 1991). Cependant, cet estimateur ne permet d'estimer qu'une distribution conditionnelle de la variable aléatoire X. Pour mener à bien les objectifs de la pharmacovigilance, nous avons besoin de l'estimation de la distribution du délai jusqu'à la survenue de l'effet indésirable E_D, mais pas à une constante multiplicative inconnue près. Nous nous sommes donc intéressés dans cette thèse à l'estimation paramétrique de la distribution de la variable aléatoire X en présence de données tronquées à droite. Nous supposons un modèle fonction d'un paramètre θ pour la variable aléatoire X et l'objectif est d'estimer ce paramètre θ à partir des données observées. Nous travaillons avec l'estimateur du maximum de vraisemblance.

Les méthodes adaptées aux données tronquées à droite sont méconnues et peu répandues dans la recherche biomédicale. Par conséquent, les méthodes classiques en analyse de données de survie, c'est-à-dire qui ne prennent pas en compte la troncature, sont parfois utilisées à la place des méthodes appropriées. Or, une telle procédure peut conduire à des estimations biaisées. Le premier objectif de cette thèse a été de mettre en évidence les conséquences de la non prise en compte du caractère tronqué à droite des données sur l'estimation du maximum de vraisemblance du paramètre de la distribution supposée.

L'estimateur non-paramétrique du maximum de vraisemblance, bien qu'il ne permette pas d'estimer la distribution non conditionnelle de la variable aléatoire X, a de bonnes propriétés asymptotiques. On souhaiterait que l'estimateur paramétrique ait également de bonnes propriétés asymptotiques. En raison du caractère tronqué à droite des données et du cadre déterministe que nous avons fixé, la théorie asymptotique clas-

sique du maximum de vraisemblance paramétrique ne s'applique pas dans notre cas. Le second objectif de la thèse a donc été d'examiner les propriétés asymptotiques de cet estimateur.

L'établissement des propriétés asymptotiques précédentes permet d'associer un intervalle de confiance à l'estimation obtenue par maximum de vraisemblance. Le troisième objectif de cette thèse a été d'étudier la qualité de l'estimation par intervalle de l'estimateur paramétrique du maximum de vraisemblance.

Enfin, le choix d'un modèle paramétrique approprié aux données observées est une question complexe. La dernière partie de ce travail de thèse a été d'explorer quelques outils adaptés aux données tronquées à droite permettant de justifier le choix d'un modèle par rapport à un autre.

Trois modèles paramétriques classiques en analyse des données de survie ont été considérés : exponentiel, Weibull et log-logistique. Nous avons choisi ces trois familles de lois car elles permettent de traiter la multiplicité des fonctions de risque instantané que l'on peut rencontrer en pharmacovigilance.

Un jeu de données réelles constitué de 64 cas de lymphomes consécutifs à un traitement anti TNF-α issus de la base de pharmacovigilance française et qui nous a été fourni par le centre régional de pharmacovigilance de Bordeaux a été analysé afin d'illustrer les méthodes mises en œuvre ou développées tout au long de cette thèse.

Le chapitre 1 est consacré à la présentation des cadres non-paramétrique et paramétrique. Les conséquences de la non prise en compte de la troncature à droite sont étudiées au chapitre 2 ; une première étude de simulations est présentée. Les chapitres 3 et 4 s'intéressent au développement des propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance. Les intervalles de confiance sont développés dans le chapitre 5 ; une seconde étude de simulations est détaillée. Le chapitre 6 présente quelques outils d'adéquation adaptés aux données tronquées à droite. Ce manuscrit s'achève par une discussion sur le travail réalisé et l'énoncé de quelques perspectives. Par souci de fluidité, les démonstrations de certains théorèmes ainsi que des résultats de simulations sont renvoyés en annexe.

Chapitre 1

Cadres non-paramétrique et paramétrique

1.1 Notations

On considère, à la date d'analyse t_a fixée, la population de taille finie mais inconnue N des individus exposés au médicament D qui ont présenté ou présenteront l'effet indésirable E_D avant de décéder. On souhaite estimer la distribution du délai jusqu'à la survenue de l'effet indésirable médicamenteux E_D dans cette population. On suppose la variable aléatoire X continue, de support $\mathbb{R}^{+*}$, de fonction de répartition F et de densité de probabilité f. Pour estimer la distribution de la variable aléatoire X, nous disposons des notifications spontanées. A l'instant t_a, on observe uniquement l'échantillon de taille n inférieure à N des individus exposés au médicament D ayant déjà présenté l'effet indésirable E_D. La figure 1.1 donne deux exemples hypothétiques d'individus appartenant à la population considérée. Pour chacun de ces deux individus, soit $(x_i)_{i=1,2}$ le délai jusqu'à la survenue de l'effet indésirable médicamenteux E_D et $(t_i)_{i=1,2}$ le délai écoulé entre la date d'exposition au médicament D et la date d'analyse t_a. Le patient n^o 1 est observé car il a développé l'effet indésirable E_D avant la date d'analyse t_a ; on a donc $x_1 \leqslant t_1$. Le patient n^o 2 n'est pas observé parce que l'effet indésirable E_D n'est

pas encore survenu au moment de l'analyse ; on a donc $x_2 > t_2$. Le délai $(t_i)_{i=1,2}$ est défini comme étant le délai de troncature associé à l'individu i. La date d'analyse t_a est commune à tous les individus. En revanche, chaque patient a sa propre date d'exposition au médicament D. Ainsi, chaque patient a son propre délai de troncature.

On désigne par T la variable aléatoire continue caractérisant ce délai de troncature, de fonction de répartition G, de densité de probabilité g et indépendante de X. Le support de la variable aléatoire T est l'intervalle $[0, \tau]$, où τ coïncide avec le délai écoulé entre l'autorisation de mise sur le marché (AMM) du médicament D et la date d'analyse t_a. On définit la quantité w par la probabilité pour une observation de ne pas être tronquée :

$$\begin{aligned} w &= P\left(X < T\right) \\ &= \int_0^{+\infty} \int_x^{+\infty} f(x) g(t) dt dx. \end{aligned}$$

La quantité w est inconnue car entièrement définie par les distributions des variables aléatoires X et T. La probabilité $p = 1 - w$ s'interprète comme étant la proportion théorique de données tronquées dans la population. On ne peut pas savoir à l'avance combien d'individus seront observés. Par conséquent, la taille d'échantillon est aléatoire. Soit Z la variable aléatoire caractérisant le nombre d'individus, parmi N, qui sont observés. L'entier n est une réalisation de cette variable aléatoire Z. Les observations s'écrivent donc $\left(n, (x_i, t_i)_{1 \leqslant i \leqslant n}\right)$, où chaque couple (x_i, t_i) est tel que $x_i \leqslant t_i$. Il s'agit d'une réalisation du vecteur aléatoire $\left(Z, (X_i, T_i | \, X_i \leqslant T_i)_{1 \leqslant i \leqslant Z}\right)$.

L'aléa de la taille d'échantillon n'est pas considéré dans cet exposé ; nous travaillons conditionnellement à la taille d'échantillon n.

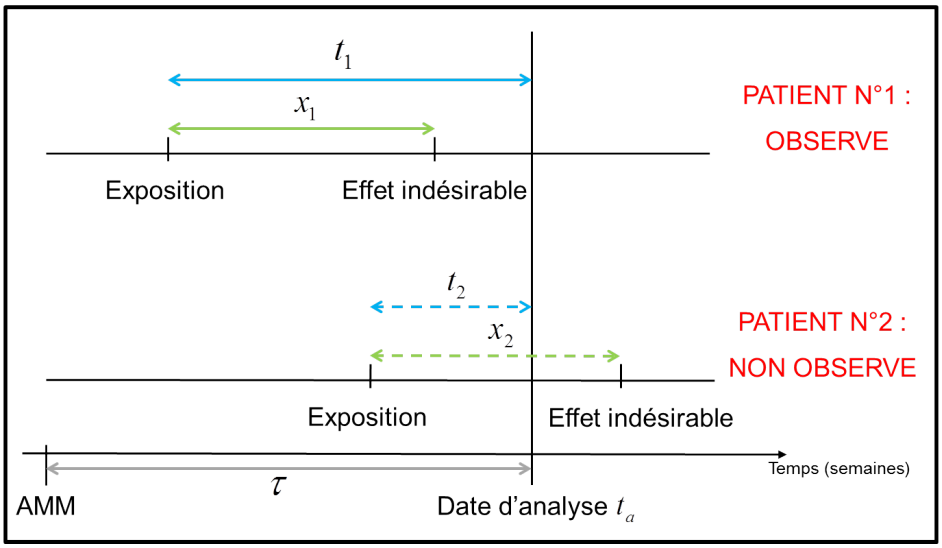

FIGURE 1.1 – Troncature à droite et délais jusqu'à la survenue d'un effet indésirable médicamenteux issus des notifications spontanées.

1.2 Cadre non-paramétrique

La vraisemblance conditionnelle des délais de survenue et de troncature observés sachant que l'échantillon est de taille n s'écrit :

$$\begin{aligned} L\left(x_1,\ldots,x_n,t_1,\ldots,t_n;F,G|\,n\right) &= \prod_{i=1}^{n} P\left(X_i=x_i,T_i=t_i|\,X_i\leqslant T_i\right) \\ &= \frac{1}{w^n}\prod_{i=1}^{n} f(x_i)g(t_i). \end{aligned} \tag{1.1}$$

Cette vraisemblance permet d'estimer la fonction de répartition G. En revanche, la seule partie identifiable de F à partir de la vraisemblance conditionnelle (1.1) est la distribution conditionnelle de X sachant que X est inférieur ou égal à τ, de fonction de répartition

$$F^*(x) = \frac{F(x)}{F(\tau)} I\left\{0 < x \leqslant \tau\right\}, \tag{1.2}$$

où $I\{A\}$ désigne la fonction indicatrice de l'événement A. Le paramètre w n'est pas identifiable. Par principe d'invariance, s'il était possible d'estimer la distribution non conditionnelle de X, l'estimation du maximum de vraisemblance du paramètre w pour-

rait être obtenue. La vraisemblance conditionnelle des délais de survenue observés sachant que l'échantillon est de taille n et sachant les délais de troncature peut également être considérée pour estimer la distribution de X. Elle a pour expression,

$$L\,(x_1,\ldots,x_n;F|\,t_1,\ldots,t_n,n)=\prod_{i=1}^{n}\frac{f(x_i)}{F(t_i)}. \tag{1.3}$$

Comme pour la vraisemblance (1.1), la partie identifiable de F à partir de la vraisemblance (1.3) est la distribution conditionnelle (1.2). Par symétrie, la vraisemblance conditionnelle des délais de troncature observés sachant que l'échantillon est de taille n et sachant les délais de survenue est définie par,

$$L\,(t_1,\ldots,t_n;G|\,x_1,\ldots,x_n,n)=\prod_{i=1}^{n}\frac{g(t_i)}{1-G(x_i)}, \tag{1.4}$$

et on peut en déduire une estimation de la fonction de répartition G.

Woodroofe (1985), Keiding et Gill (1990) et Gross et Huber-Carol (1992) ont obtenu les estimations non-paramétriques du maximum de vraisemblance de F^* et G à partir de la vraisemblance conditionnelle (1.1). Keiding et Gill (1990) ont utilisé une approche par processus Markoviens.

Vardi (1985), Lagakos *et al.* (1988) et Kalbfleisch et Lawless (1991) ont dérivé l'estimateur non-paramétrique du maximum de vraisemblance de F^* à partir de la vraisemblance conditionnelle (1.3). Vardi (1985) se place dans un cadre plus général qui englobe les données tronquées à droite. Lagakos *et al.* (1988) ont utilisé une approche par temps inversé ; l'inversion du temps transforme les données tronquées à droite en données tronquées à gauche, pour lesquelles l'estimateur non-paramétrique du maximum de vraisemblance a déjà été développé (Cox et Oakes, 1984). En considérant les variables aléatoires $S=\tau-X$ et $Y=\tau-T$, les observations transformées $(\tau-x_i,\tau-t_i)_{1\leqslant i\leqslant n}$ sont des réalisations indépendantes et identiquement distribuées du vecteur aléatoire $(S,Y|\,S\geqslant Y)$.

Wang (1987) a prouvé que, dès que les estimateurs non-paramétrique $\widehat{F}^*$ et $\widehat{G}$ issus respectivement des vraisemblances conditionnelles (1.3) et (1.4) existent, alors $\widehat{F}^*$ et $\widehat{G}$ coïncident avec les estimateurs non-paramétriques issus de la vraisemblance conditionnelle (1.1).

L'estimateur non-paramétrique de F^* du maximum des vraisemblances (1.1) ou (1.3) s'écrit :

$$\widehat{F}^*(x) = \prod_{v_j > x} \left(1 - \frac{n_j}{N_j}\right) I\left\{0 < x \leqslant \tau\right\}, \tag{1.5}$$

où l'ensemble $(v_j)_{1 \leqslant j \leqslant s}$ désigne les s valeurs distinctes et ordonnées des réalisations $(x_i)_{1 \leqslant i \leqslant n}$; les quantités $n_j = \sum_{i=1}^{n} I\left\{X_i = v_j\right\}$, pour $1 \leqslant j \leqslant s$, représentent le nombre d'individus de l'échantillon ayant un délai de survenue égal à v_j et les quantités $N_j = \sum_{i=1}^{n} I\left\{X_i \leqslant v_j \leqslant t_i\right\}$, pour $1 \leqslant j \leqslant s$, désignent le nombre d'individus à risque. Comme la fonction de répartition empirique ou l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958), l'estimateur (1.5) est un estimateur produit-limite. Mais la définition des ensembles d'individus à risque est moins intuitive pour ce dernier que pour les deux premiers.

Vardi (1985) a énoncé une condition nécessaire et suffisante pour l'existence et l'unicité de l'estimateur non-paramétrique $\widehat{F}^*$. Woodroofe (1985) propose une alternative à cet estimateur quand la condition de Vardi (1985) n'est pas vérifiée.

Woodroofe (1985) et Keiding et Gill (1990) ont étudié les propriétés asymptotiques de cet estimateur. L'estimateur non-paramétrique (1.5) est un estimateur consistant de F^*. De plus, pour (a, b) un couple de réels tel que $0 < a < b < \tau$, la fonction aléatoire $\sqrt{n}\left(\widehat{F}^*(\cdot) - F^*(\cdot)\right)$ définie sur $[a, b]$ converge faiblement vers le processus gaussien $W^*(\cdot)$ centré et de fonction covariance

$$\mathrm{Cov}(W^*(x), W^*(y)) = F^*(x)F^*(y) \int_{\max(x,y)}^{\tau} \frac{dC(z)}{R^2(z)},$$

où

$$\begin{cases} C(z) = \dfrac{1}{w} \displaystyle\int_{-\infty}^{\min(z,\tau)} \int_{-\infty}^{\infty} I\left\{u \leqslant v\right\} dG(v)dF(u), \\ R(z) = \dfrac{1}{w} \displaystyle\int_{-\infty}^{\min(z,\tau)} \int_{z}^{\infty} I\left\{u \leqslant v\right\} dG(v)dF(u). \end{cases}$$

La covariance peut être estimée par

$$\widehat{\mathrm{Cov}}(W^*(x), W^*(y)) = \widehat{F}^*(x)\widehat{F}^*(y) \sum_{\max(x,y)\leqslant v_j \leqslant \tau} \frac{n_j}{N_j(N_j - n_j)}. \tag{1.6}$$

La loi asymptotique de cet estimateur sera utilisée dans le chapitre 6, pour la présentation des outils d'adéquation adaptés aux données tronquées à droite.

Remarque 1.2.1. Dans la littérature scientifique traitant de l'estimateur non-paramétrique du maximum de vraisemblance, on trouve que la partie identifiable de F à partir des observations est soit la distribution conditionnelle de X sachant que X est inférieur ou égal à τ soit la distribution conditionnelle de X sachant que X est inférieur ou égal au maximum des délais de troncature observés $t^* = \max_{1\leqslant i\leqslant n}(t_i)$. Etant donné que ce maximum t^* converge en probabilité vers τ quand la taille d'échantillon tend vers l'infini, on peut considérer indifféremment l'une ou l'autre de ces distributions conditionnelles.

Remarque 1.2.2. Wang *et al.* (1986) et Tsai *et al.* (1987) suggèrent d'estimer la distribution conditionnelle de X sachant que X est inférieur ou égal à T^*, où T^* est tel qu'il y ait suffisamment d'individus à risque pour tout intervalle d'estimation.

1.3 Cadre paramétrique

Dans un cadre paramétrique, l'estimation d'une distribution se réduit à l'estimation d'un paramètre qui peut être multidimensionnel. Soient $F(\cdot;\theta)$ et $f(\cdot;\theta)$ respectivement les fonction de répartition et densité de probabilité de X, où $\theta = (\theta_1, \ldots, \theta_r)$ est le paramètre vectoriel que l'on cherche désormais à estimer. Soit $G(\cdot;\beta)$ et $g(\cdot;\beta)$ respectivement les fonction de répartition et densité de probabilité de T, dépendantes d'un

paramètre β. La vraisemblance conditionnelle des délais de survenue et de troncature observés sachant que l'échantillon est de taille n s'écrit désormais :

$$L\left(x_1,\ldots,x_n,t_1,\ldots,t_n;\theta,\beta|\, n\right)=\frac{1}{w^n}\prod_{i=1}^{n}f(x_i;\theta)g(t_i;\beta), \tag{1.7}$$

et la vraisemblance conditionnelle des délais de survenue (resp. délais de troncature) sachant que l'échantillon est de taille n et sachant les délais de troncature (resp. délais de survenue) s'écrit :

$$L\left(x_1,\ldots,x_n;\theta|\, t_1,\ldots,t_n,n\right)=\prod_{i=1}^{n}\frac{f(x_i;\theta)}{F(t_i;\theta)}, \tag{1.8}$$

$$\left(\text{resp. }\ L\left(t_1,\ldots,t_n;\beta|\, x_1,\ldots,x_n,n\right)=\prod_{i=1}^{n}\frac{g(t_i;\beta)}{1-G(x_i;\beta)}\quad\right). \tag{1.9}$$

Les références bibliographiques traitant de l'estimateur paramétrique du maximum de vraisemblance en présence de données tronquées à droite sont peu nombreuses. Elles considèrent toutes le problème de l'estimation de la distribution du délai d'incubation du SIDA. Medley *et al.* (1987), Medley *et al.* (1988), Kalbfleisch et Lawless (1988) et Kalbfleisch et Lawless (1989) ont dérivé l'estimateur paramétrique du maximum de vraisemblance des paramètres θ et β à partir de la vraisemblance conditionnelle (1.7). Lui *et al.* (1986), Lagakos *et al.* (1988), Brookmeyer et Gail (1988) et Lawless (2003) ont, quant à eux, obtenu l'estimation du paramètre θ en considérant la vraisemblance conditionnelle (1.8). Certains d'entre-eux ont précisé qu'on ne pouvait pas se fier à cette estimation. Cependant, peu ou pas d'explications ont été données et aucune étude de simulations n'a été mise en œuvre.

Choisir une distribution paramétrique pour la variable aléatoire T n'est pas simple. De plus, les échantillons issus des notifications spontanées sont généralement de taille restreinte, ce qui réduit les capacités de l'échantillon à estimer conjointement les paramètres θ et β de manière convenable. La problématique de ce travail de thèse étant l'estimation de la distribution non conditionnelle de la variable aléatoire X, on choisit

de travailler tout au long de cet exposé à partir de la vraisemblance conditionnelle (1.8), ce qui revient à considérer les délais de troncature comme déterministes. On souhaite désormais estimer le paramètre vectoriel θ appartenant à un ensemble Θ.

L'estimation du maximum de vraisemblance, obtenue à partir de la vraisemblance conditionnelle (1.8) est noté $\widehat{\theta}_n$ et est défini par :

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(x_1, x_2, \ldots, x_n; \theta |\, t_1, t_2, \ldots, t_n, n). \tag{1.10}$$

On a $\widehat{\theta}_n = \left(\widehat{\theta}_{1n}, \ldots, \widehat{\theta}_{rn}\right)$. Un algorithme itératif se révèle souvent nécessaire pour résoudre ce problème d'optimisation. Les algorithmes de Newton-Raphson et de Nelder-Mead ont été utilisés pour les études de simulations présentées aux chapitres suivants.

Trois modèles paramétriques classiques en analyse des données de survie ont été considérés : exponentiel, Weibull et log-logistique. Le tableau 1.1 donne la paramétrisation choisie pour ces trois familles de lois. Nous avons choisi ces trois familles de lois car elles permettent de traiter la multiplicité des risques instantanés que l'on peut rencontrer en pharmacovigilance. En effet, certains effets indésirables ont un délai de survenue très court, de quelques minutes à quelques heures après l'exposition au traitement. D'autres surviennent seulement après quelques jours, semaines, mois voire même quelques années d'exposition. Cette variation dépend de nombreux facteurs tels que la pharmacocinétique du médicament et de ces métabolites, ou bien le mécanisme pathophysiologique de l'effet indésirable (Arme-p, 1992). Le modèle le plus simple suppose un risque instantané constant au cours du temps ; la distribution correspondante est la loi exponentielle de paramètre θ_1. Mais les effets indésirables peuvent aussi survenir précocement ou tardivement comme c'est le cas par exemple quand le risque de survenue de l'effet indésirable dépend de la durée d'exposition au médicament. Deux familles de lois, parmi d'autres, permettent de modéliser ces fonctions de risque : les distributions de Weibull et log-logistique (Tableau 1.1). Ces deux distributions dépendent de deux paramètres (θ_1, θ_2) ; le paramètre θ_1 est le paramètre d'échelle et θ_2 est le paramètre de forme. La fonction de risque du modèle de Weibull est décroissante si θ_2 est strictement

inférieur à 1, croissante si θ_2 est strictement supérieur à 1 et constante si θ_2 est égal à 1. Dans ce dernier cas, la loi de Weibull coïncide avec la distribution exponentielle. La fonction de risque du modèle log-logistique est décroissante si θ_2 est inférieur strictement à 1 et admet un unique maximum si θ_2 est supérieur strictement à 1. La figure 1.2 montre les différentes formes de fonction de risque instantané que l'on peut rencontrer en considérant ces trois familles de distributions.

Les chapitres suivants sont consacrés, tout d'abord, à l'importance de la prise en compte de la troncature à droite dans l'estimation de θ, ensuite à l'établissement des propriétés asymptotiques de cet estimateur du maximum de vraisemblance conditionnelle et enfin à l'obtention d'intervalles de confiance pour le paramètre θ.

TABLEAU 1.1 – Paramètres, supports, densités, fonctions de répartition et espérances des distributions exponentielle, de Weibull et log-logistique.

Distribution	Exponentielle (θ_1)	Weibull (θ_1, θ_2)	Log-logistique (θ_1, θ_2)
Paramètre(s)	$\theta_1 > 0$	$\theta_1 > 0$ $\theta_2 > 0$	$\theta_1 > 0$ $\theta_2 > 0$
Support	$x > 0$	$x > 0$	$x > 0$
Densité	$f(x) = \theta_1 \exp(-\theta_1 x)$	$f(x) = \theta_1\theta_2(\theta_1 x)^{\theta_2 - 1}\exp(-(\theta_1 x)^{\theta_2})$	$f(x) = \dfrac{\theta_1\theta_2(\theta_1 x)^{\theta_2 - 1}}{(1 + (\theta_1 x)^{\theta_2})^2}$
Fonction de répartition	$1 - \exp(-\theta_1 x)$	$1 - \exp(-(\theta_1 x)^{\theta_2})$	$\dfrac{(\theta_1 x)^{\theta_2}}{1 + (\theta_1 x)^{\theta_2}}$
Espérance	$\dfrac{1}{\theta_1}$	$\dfrac{1}{\theta_1}\Gamma\left(1 + \dfrac{1}{\theta_2}\right)$	$\dfrac{\pi/(\theta_1\theta_2)}{\sin(\pi/\theta_2)}$

$\Gamma(\cdot)$ est la fonction gamma.

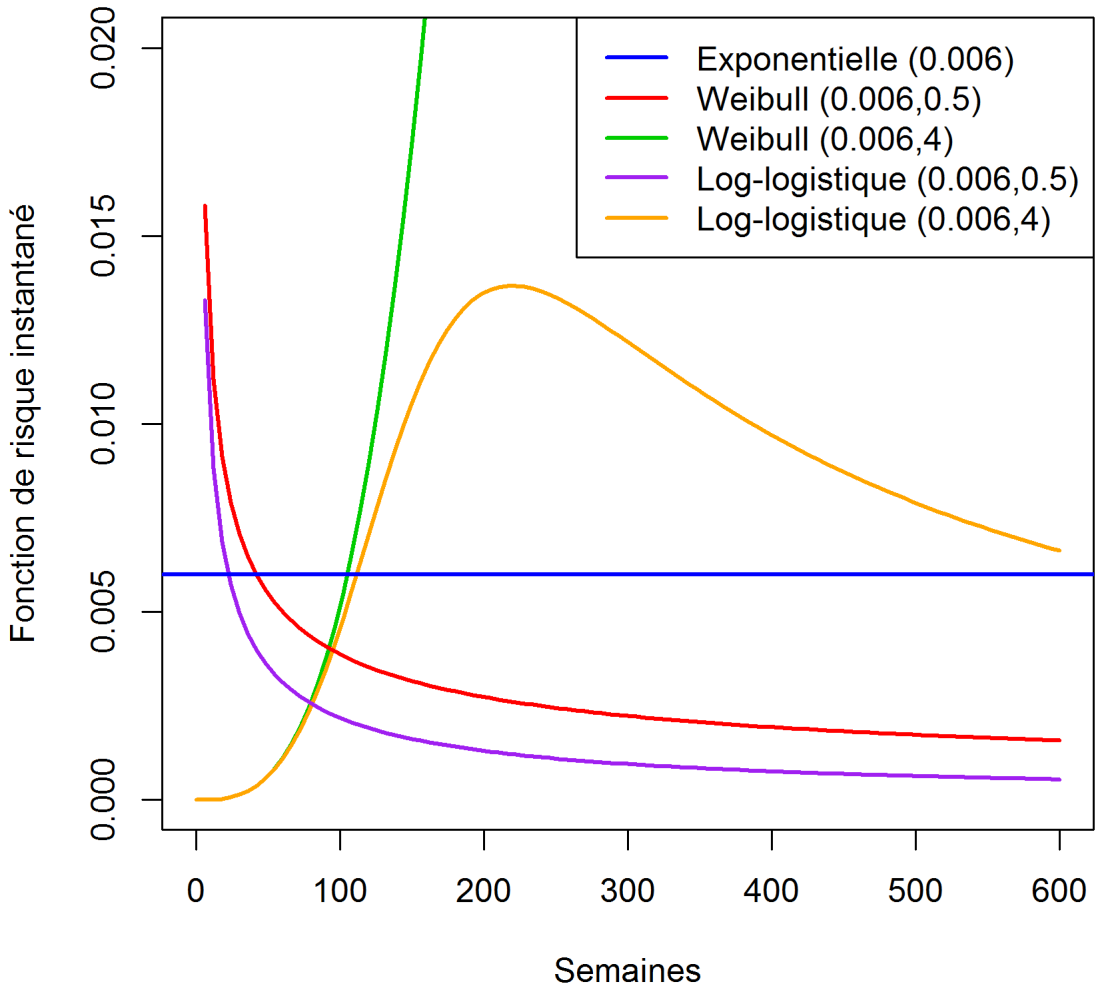

FIGURE 1.2 – Forme des fonctions de risque instantané des distributions exponentielle, de Weibull et log-logistique. Les fonctions de risque instantané de la loi exponentielle de paramètre 0.006, de la loi de Weibull de paramètres (0.006,0.5), de la loi de Weibull de paramètres (0.006,4), de la loi log-logistique de paramètres (0.006,0.5) et de la loi log-logistique de paramètres (0.006,4) sont représentées.

Chapitre 2

De l'importance de prendre en compte la troncature à droite

Dans le chapitre précédent, nous avons rappelé les expressions de la vraisemblance et de l'estimateur du maximum de vraisemblance quand les données sont tronquées à droite. Cependant, le caractère tronqué à droite des données n'est pas toujours pris en compte. Les méthodes classiques en analyse de données de survie, c'est-à-dire qui ne prennent pas en compte la troncature, sont parfois utilisées à la place des méthodes appropriées. Une telle procédure peut conduire à des estimations biaisées. L'objectif de ce chapitre est de mettre en évidence les conséquences, sur l'estimation du maximum de vraisemblance du paramètre de la distribution supposée, de ne pas tenir compte du caractère tronqué à droite des données. Les deux estimations, naïve et basée sur la troncature à droite, sont présentées. Une comparaison de ces estimations à l'aide d'une étude de simulations est mise en œuvre. Enfin, notre jeu de données réelles constitué de 64 cas de lymphomes consécutifs à un traitement anti TNF-α issus de la base de pharmacovigilance française est présenté et analysé.

2.1 Deux estimations paramétriques du maximum de vraisemblance

2.1.1 Estimation naïve

Quand la troncature à droite est ignorée, les données observées sont considérées comme complètes. On observe n délais de survenue, $(x_i)_{1\leqslant i\leqslant n}$, correspondant à n réalisations indépendantes et identiquement distribuées de la variable aléatoire X. La vraisemblance de l'échantillon s'écrit alors :

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta). \tag{2.1}$$

L'estimation du maximum de vraisemblance, $\widehat{\theta}_{\mathrm{EN}} = \left(\widehat{\theta}_{\mathrm{EN},1}, \ldots, \widehat{\theta}_{\mathrm{EN},r}\right)$, obtenue à partir de cette vraisemblance est une estimation naïve de θ et est définie par :

$$\widehat{\theta}_{\mathrm{EN}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(x_1, x_2, \ldots, x_n; \theta).$$

Cette estimation est souvent obtenue en résolvant explicitement les équations normales, même si parfois un algorithme itératif est nécessaire pour résoudre ce problème de maximisation.

2.1.2 Estimation basée sur la troncature

Il s'agit de l'estimation (1.10), obtenue à partir de la vraisemblance (1.8) et présentée au chapitre précédent. La vraisemblance (1.8) est différente de la vraisemblance (2.1). Cette estimation basée sur la troncature à droite (EBT), que nous noterons $\widehat{\theta}_{\mathrm{EBT}} = \left(\widehat{\theta}_{\mathrm{EBT},1}, \ldots, \widehat{\theta}_{\mathrm{EBT},r}\right)$ dans la suite de ce chapitre, est l'estimation pertinente du paramètre θ. Cet estimateur est différent de $\widehat{\theta}_{\mathrm{EN}}$.

Remarque 2.1.1. La vraisemblance (1.8) est bien égale à la vraisemblance (2.1) quand la troncature à droite disparaît.

2.2 Comparaison des estimateurs par étude de simulations

2.2.1 Design

Une première étude de simulations a été mise en œuvre à l'aide du logiciel R (R Core Team, 2012) afin d'évaluer les conséquences, en termes de biais et d'erreur quadratique moyenne, de la non-considération du caractère tronqué à droite des données lors de l'obtention de l'estimation du maximum de vraisemblance, mais aussi afin de visualiser les performances de l'estimateur basé sur la troncature à droite dans des situations fréquemment rencontrées en pharmacovigilance.

Les familles des distributions exponentielle, de Weibull et log-logistique ont été considérées pour la génération du délai jusqu'à la survenue de l'effet indésirable médicamenteux. Les valeurs 0.05 et 1 ont été utilisées pour le paramètre θ_1 de la loi exponentielle et pour le paramètre d'échelle θ_1 des lois de Weibull et log-logistique. Pour le paramètre de forme θ_2 des lois de Weibull et log-logistique, les valeurs 0.5 et 2 ont été choisies. Les fonctions de base *rexp*, *rweibull* et la fonction *rllogis* du package **eha** ont été utilisées pour la génération des délais de survenue. Les délais de troncature ont été générés selon une loi uniforme continue sur l'intervalle $[0, \tau]$, à l'aide de la fonction *runif* du logiciel R. Les délais de survenue et de troncature ont été générés indépendamment. Soit $\kappa = P(X < \tau)$, la probabilité qu'une réalisation de X appartienne à l'intervalle $[0, \tau]$ des valeurs observables de X. La borne supérieure τ de l'intervalle a été déterminée de manière à ce que l'on ait $\kappa = 0.25$, 0.50 ou 0.80. La probabilité $1 - \kappa$ est une borne inférieure de la proportion théorique $1 - w$ de données tronquées dans la population. En effet, la variable aléatoire T est presque-sûrement inférieure ou égale à τ donc on a

$$P(X \leqslant T) \leqslant P(X \leqslant \tau) = \kappa.$$

Et on a ainsi $1 - P(X > T) \geqslant 1 - \kappa$. Cependant, cette probabilité κ est directement

liée au support de la distribution du délai de troncature et sera considérée comme paramètre d'intérêt dans l'interprétation des résultats. Pour chaque réalisation générée du vecteur aléatoire (X, T), si le délai de survenue généré était inférieur ou égal au délai de troncature généré, alors cette réalisation était incluse dans la base. Sinon, une autre réalisation du vecteur aléatoire était simulée. Ce processus a été poursuivi jusqu'à ce que l'on obtienne un échantillon de taille $n = 100$ ou $n = 500$. Le tableau 2.1 regroupe les différents scénarios obtenus en croisant ces paramètres de distribution.

Pour chaque échantillon généré, les estimations naïve et basée sur la troncature à droite du paramètre vectoriel θ de la distribution supposée ont été obtenues. Un algorithme itératif de maximisation s'est révélé nécessaire pour résoudre ce problème, excepté pour obtenir l'estimation naïve de la loi exponentielle. La fonction *maxLik* du package **maxLik** du logiciel R a été utilisée pour la maximisation de la vraisemblance. Nous avons fait appel à l'algorithme de Newton-Raphson pour la distribution exponentielle et à l'algorithme de Nelder-Mead pour les lois de Weibull et log-logistique. Pour chaque scénario étudié, 1000 réplications ont été générées.

Pour chaque scénario, pour chaque composante du paramètre vectoriel θ et pour chacune des deux approches, naïve et basée sur la troncature, le biais et l'erreur quadratique moyenne de l'estimateur ainsi que la probabilité que l'estimateur soit supérieur à la vraie valeur du paramètre ont été estimés. Le biais est estimé par la moyenne empirique des écarts entre l'estimation et la vraie valeur du paramètre :

$$\begin{aligned}
\text{Biais}\left(\widehat{\theta}\right) &= \mathrm{E}\left(\widehat{\theta}\right) - \theta, \\
\widehat{\text{Biais}}\left(\widehat{\theta}\right) &= \widehat{\mathrm{E}}\left(\widehat{\theta}\right) - \theta \\
&= \frac{1}{1000}\sum_{k=1}^{1000}\left(\widehat{\theta}_k - \theta\right).
\end{aligned}$$

L'erreur quadratique moyenne (EQM) est la somme de la variance et du biais au carré de l'estimateur. Il est estimé par la somme de la variance empirique et de l'estimation

TABLEAU 2.1 – Scénarios de l'étude de simulations n° 1 : trente scénarios différents pour une taille d'échantillon, obtenus en considérant trois familles de distributions pour le délai de survenue X, différentes valeurs du paramètre θ de la distribution et trois valeurs de la probabilité κ qu'une réalisation de X appartienne à l'intervalle $[0, \tau]$ des valeurs observables de X. Les délais de troncature ont été simulés selon une loi uniforme continue sur l'intervalle $[0, \tau]$. La borne supérieure τ de l'intervalle a été déterminée de manière à ce que l'on ait $P(X < \tau) = \kappa$.

Distribution	θ	κ	τ
Exponentielle	0.05	0.25	5.75
Exponentielle	0.05	0.50	13.86
Exponentielle	0.05	0.80	32.19
Exponentielle	1	0.25	0.29
Exponentielle	1	0.50	0.69
Exponentielle	1	0.80	1.61
Weibull	(0.05, 0.5)	0.25	1.65
Weibull	(0.05, 0.5)	0.50	9.61
Weibull	(0.05, 0.5)	0.80	51.81
Weibull	(1, 0.5)	0.25	0.083
Weibull	(1, 0.5)	0.50	0.48
Weibull	(1, 0.5)	0.80	2.59
Weibull	(0.05, 2)	0.25	10.73
Weibull	(0.05, 2)	0.50	16.65
Weibull	(0.05, 2)	0.80	25.37
Weibull	(1, 2)	0.25	0.54
Weibull	(1, 2)	0.50	0.83
Weibull	(1, 2)	0.80	1.27
Log-logistique	(0.05, 0.5)	0.25	2.22
Log-logistique	(0.05, 0.5)	0.50	20
Log-logistique	(0.05, 0.5)	0.80	320
Log-logistique	(1, 0.5)	0.25	0.11
Log-logistique	(1, 0.5)	0.50	1
Log-logistique	(1, 0.5)	0.80	16
Log-logistique	(0.05, 2)	0.25	11.55
Log-logistique	(0.05, 2)	0.50	20
Log-logistique	(0.05, 2)	0.80	40
Log-logistique	(1, 2)	0.25	0.58
Log-logistique	(1, 2)	0.50	1
Log-logistique	(1, 2)	0.80	2

du biais au carré :

$$\begin{aligned}
\mathrm{EQM}\left(\widehat{\theta}\right) &= \mathrm{E}\left(\widehat{\theta}-\theta\right)^2 \\
&= \mathrm{Var}\left(\widehat{\theta}\right) + \left(\mathrm{Biais}\left(\widehat{\theta}\right)\right)^2, \\
\widehat{\mathrm{EQM}}\left(\widehat{\theta}\right) &= \widehat{\mathrm{Var}}\left(\widehat{\theta}\right) + \left(\widehat{\mathrm{Biais}}\left(\widehat{\theta}\right)\right)^2 \\
&= \frac{1}{999}\sum_{k=1}^{1000}\left(\widehat{\theta}_k - \frac{1}{1000}\sum_{k=1}^{1000}\widehat{\theta}_k\right)^2 + \left(\widehat{\mathrm{Biais}}\left(\widehat{\theta}\right)\right)^2.
\end{aligned}$$

L'erreur quadratique moyenne est une mesure de la dispersion de l'estimateur autour de la vraie valeur du paramètre. Elle est souvent utilisée pour comparer des estimateurs puisque elle prend en compte à la fois le biais et la variance de l'estimateur. La probabilité que l'estimateur soit supérieur à la vraie valeur du paramètre est estimée par la proportion de réplications où l'estimation est supérieure à la vraie valeur du paramètre. L'estimation de cette probabilité nous permet de voir si l'estimateur a tendance à sous-estimer ou surestimer de manière systématique le paramètre. Ces estimations du biais, de l'erreur quadratique moyenne et de la probabilité que l'estimateur soit supérieur à la vraie valeur du paramètre sont basées sur les 1000 réplications réalisées pour chaque scénario. Cependant, étant donné que l'algorithme itératif peut échouer à trouver l'estimation du maximum de vraisemblance, ces quantités ont en réalité été estimées sur l'ensemble des réplications où il n'y a pas eu de problème de maximisation.

2.2.2 Résultats

2.2.2.1 Biais et erreur quadratique moyenne

Pour les deux approches, pour toutes les distributions considérées et pour chaque composante du paramètre vectoriel θ, plus κ augmente, plus le biais et l'erreur quadratique moyenne sont réduits (Tableaux 2.2, 2.3 et 2.4). Cette diminution avec κ est plus faible pour le paramètre θ_2 que pour le paramètre θ_1. Les deux estimateurs ont tendance à être positivement biaisés. Cependant, ce biais est parfois presque nul pour

l'estimateur basé sur la troncature à droite. Le biais et l'erreur quadratique moyenne de l'estimateur naïf sont toujours plus élevés que ceux de l'estimateur pertinent, bien que l'écart entre les deux estimateurs soit moins net pour le paramètre θ_2 que pour le paramètre θ_1. Quand la taille d'échantillon n augmente, le biais et l'erreur quadratique moyenne sont presque constants pour l'estimateur naïf alors que pour l'estimateur basé sur la troncature, ils diminuent clairement (Tableaux 2.2, 2.3 et 2.4). Enfin, le biais de l'estimateur naïf peut être inacceptablement élevé quelle que soit la valeur de κ tandis que l'estimateur basé sur la troncature à droite montre de bonnes performances quand κ est égal à 0.8, et souvent même pour des valeurs inférieures selon la distribution considérée.

Par ailleurs, dans la dernière colonne des tableaux 2.2, 2.3 et 2.4 apparaît le nombre de problèmes de maximisation que l'on rencontre dans l'approche basée sur la troncature à droite. Il n'y a aucun problème de maximisation quand la troncature à droite est ignorée. En revanche, pour l'estimateur basée sur la troncature, le nombre de problèmes de maximisation augmente quand κ ou n diminuent.

2.2.2.2 Proportion des réplications où l'estimateur est supérieur à la vraie valeur du paramètre

Les tableaux rassemblant les résultats de ce critère apparaissent dans l'annexe A de ce manuscrit. Pour les deux approches, pour toutes les distributions considérées et pour chaque composante du paramètre vectoriel θ, les tableaux A.1, A.2 et A.3 montrent que l'estimateur naïf de θ_1 semble être presque toujours supérieur à la vraie valeur du paramètre et ce n'est pas loin d'être vrai aussi pour le paramètre θ_2. Cela suggère que l'estimateur naïf de θ_1 puisse être presque-sûrement plus élevé que la vraie valeur du paramètre, ce qui serait une caractéristique statistique peu enviable de cet estimateur.

TABLEAU 2.2 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution exponentielle : estimations du biais et de l'erreur quadratique moyenne des estimateurs naïf et basé sur la troncature à droite, pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

			Estimateur naïf		EBT		
θ_1	κ	n	Biais($\widehat{\theta}_{\mathrm{EN},1}$)	EQM($\widehat{\theta}_{\mathrm{EN},1}$)	Biais($\widehat{\theta}_{\mathrm{EBT},1}$)	EQM($\widehat{\theta}_{\mathrm{EBT},1}$)	NPM
0.05	0.25	100	0.498	0.250	0.030	0.005	224
		500	0.498	0.248	0.007	0.001	79
	0.50	100	0.195	0.038	0.008	0.001	85
		500	0.193	0.037	<0.001	<0.001	1
	0.80	100	0.073	0.005	<0.001	<0.001	2
		500	0.072	0.005	<0.001	<0.001	0
1	0.25	100	10.06	102	0.462	2.17	72
		500	9.95	99	0.046	0.48	10
	0.50	100	3.91	15.4	0.126	0.49	29
		500	3.86	14.9	-0.022	0.12	0
	0.80	100	1.45	2.16	0.004	0.11	0
		500	1.45	2.11	0.004	0.02	0

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature à droite ; EQM, erreur quadratique moyenne ; NPM, nombre de problèmes de maximisation.

TABLEAU 2.3 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution de Weibull : estimations du biais et de l'erreur quadratique moyenne des estimateurs naïf et basé sur la troncature à droite, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

				Estimateur naïf				EBT				
				$\widehat{\theta}_{\mathrm{EN},1}$		$\widehat{\theta}_{\mathrm{EN},2}$		$\widehat{\theta}_{\mathrm{EBT},1}$		$\widehat{\theta}_{\mathrm{EBT},2}$		
θ_1	θ_2	κ	n	Biais	EQM	Biais	EQM	Biais	EQM	Biais	EQM	NPM
0.05	0.5	0.25	100	4.04	16.7	0.200	0.044	0.465	0.51	0.046	0.007	312
			500	3.95	15.6	0.195	0.039	0.106	0.04	0.013	0.001	201
		0.50	100	0.762	0.60	0.167	0.031	0.068	0.018	0.024	0.005	172
			500	0.747	0.56	0.164	0.028	0.015	0.003	0.003	0.001	22
		0.80	100	0.160	0.027	0.119	0.017	0.008	0.002	0.009	0.004	9
			500	0.156	0.025	0.113	0.013	0.001	<0.001	0.001	<0.001	0
1	0.5	0.25	100	80.4	6612	0.201	0.044	8.68	183	0.046	0.007	300
			500	78.9	6249	0.194	0.038	2.07	17	0.012	0.001	186
		0.50	100	15.0	233	0.174	0.034	1.53	7.99	0.031	0.006	163
			500	15.0	225	0.164	0.028	0.32	1.17	0.003	0.001	24
		0.80	100	3.20	10.8	0.117	0.017	0.16	0.67	0.007	0.004	13
			500	3.15	10.0	0.112	0.013	0.041	0.15	<0.001	<0.001	0
0.05	2	0.25	100	0.121	0.015	0.354	0.16	<0.001	0.002	0.097	0.075	8
			500	0.120	0.014	0.333	0.12	-0.004	0.001	0.020	0.016	2
		0.50	100	0.065	0.004	0.278	0.11	-0.004	<0.001	0.047	0.074	6
			500	0.064	0.004	0.264	0.08	-0.002	<0.001	0.004	0.016	0
		0.80	100	0.032	0.001	0.182	0.063	<0.001	<0.001	0.046	0.063	1
			500	0.032	0.001	0.157	0.031	<0.001	<0.001	0.008	0.014	0
1	2	0.25	100	2.41	5.84	0.364	0.17	0.090	0.79	0.10	0.075	1
			500	2.41	5.79	0.336	0.12	-0.082	0.38	0.02	0.015	0
		0.50	100	1.29	1.68	0.283	0.12	-0.073	0.33	0.052	0.069	3
			500	1.29	1.65	0.261	0.07	-0.065	0.12	-0.002	0.017	0
		0.80	100	0.638	0.41	0.186	0.065	-0.024	0.086	0.045	0.064	0
			500	0.636	0.40	0.154	0.030	-0.007	0.014	0.004	0.013	0

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature à droite ; EQM, erreur quadratique moyenne ; NPM, nombre de problèmes de maximisation.

TABLEAU 2.4 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution log-logistique : estimations du biais et de l'erreur quadratique moyenne des estimateurs naïf et basé sur la troncature à droite, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

				Estimateur naïf				EBT				
				$\widehat{\theta}_{\mathrm{EN},1}$		$\widehat{\theta}_{\mathrm{EN},2}$		$\widehat{\theta}_{\mathrm{EBT},1}$		$\widehat{\theta}_{\mathrm{EBT},2}$		
θ_1	θ_2	κ	n	Biais	EQM	Biais	EQM	Biais	EQM	Biais	EQM	NPM
0.05	0.5	0.25	100	6.45	44	0.384	0.16	0.258	0.25	0.041	0.008	217
			500	6.33	40	0.372	0.14	0.043	0.01	0.005	0.001	52
		0.50	100	1.05	1.2	0.319	0.108	0.045	0.012	0.020	0.006	22
			500	1.02	1.1	0.308	0.096	0.009	0.001	0.003	0.001	0
		0.80	100	0.165	0.031	0.195	0.041	0.008	0.001	0.008	0.004	0
			500	0.158	0.026	0.189	0.036	0.001	<0.001	0.001	<0.001	0
1	0.5	0.25	100	129	17533	0.383	0.15	5.06	87	0.042	0.008	207
			500	127	16217	0.374	0.14	1.01	6	0.008	0.001	41
		0.50	100	21.0	467	0.317	0.106	0.93	5.0	0.019	0.006	43
			500	20.5	426	0.308	0.096	0.20	0.6	0.004	0.001	0
		0.80	100	3.31	12	0.201	0.044	0.209	0.55	0.016	0.005	0
			500	3.17	10	0.190	0.037	0.037	0.09	0.002	<0.001	0
0.05	2	0.25	100	0.150	0.022	1.06	1.2	<0.001	0.001	0.08	0.085	4
			500	0.149	0.022	1.04	1.1	-0.001	<0.001	0.01	0.018	0
		0.50	100	0.079	0.006	0.932	0.94	<0.001	<0.001	0.06	0.094	5
			500	0.078	0.006	0.903	0.83	<0.001	<0.001	0.01	0.017	0
		0.80	100	0.035	0.001	0.665	0.50	<0.001	<0.001	0.03	0.078	0
			500	0.035	0.001	0.649	0.43	<0.001	<0.001	0.01	0.013	0
1	2	0.25	100	2.99	9.0	1.07	1.2	0.024	0.57	0.08	0.089	0
			500	2.98	8.9	1.04	1.1	-0.028	0.20	0.01	0.020	0
		0.50	100	1.57	2.49	0.943	0.96	0.007	0.19	0.063	0.095	1
			500	1.56	2.45	0.896	0.82	-0.013	0.04	0.004	0.018	0
		0.80	100	0.702	0.50	0.668	0.50	0.004	0.042	0.045	0.072	0
			500	0.693	0.48	0.648	0.43	0.004	0.007	0.015	0.013	0

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature à droite ; EQM, erreur quadratique moyenne ; NPM, nombre de problèmes de maximisation.

2.3 Application : Délai de survenue d'un lymphome

En France, l'Agence Nationale de Sécurité du Médicament (ANSM) gère la base de données des notifications spontanées. Le système de pharmacovigilance national repose sur les 31 centres régionaux de pharmacovigilance répartis sur tout le territoire français. Tous les effets indésirables qui surviennent après la mise sur le marché des médicaments et qui sont déclarés aux centres régionaux par les professionels de santé ou par les patients eux-mêmes sont enregistrés dans cette base.

Le centre régional de pharmacovigilance de Bordeaux nous a fourni un jeu de données réelles constitué de 64 cas de lymphomes survenus consécutivement à la prise d'un traitement anti TNF-α. Ces données ont été extraites de la base de pharmacovigilance française le 1$^{\text{er}}$ février 2010 (Théophile *et al.*, 2011). La population est constitué des patients souffrant d'arthrite rhumatoïde, de la maladie de Crohn, de spondylite ankylosante, d'arthrite psoriatique, de psoriasis, du syndrome de Sjögren, de dermatomyositis, de polymiositis ou de polyarthropathie et exposés à un ou (successivement) plusieurs des trois traitements anti TNF-α disponibles sur le marché au moment de l'extraction de la base : *etanercept*, *adalimumab* et *infliximab*. La survenue d'un lymphome malin a été confirmée par une analyse histopathologique. L'autorisation de mise sur le marché a été obtenue en août 1999 pour le traitement *infliximab*, en septembre 2002 pour le traitement *etanercept* et en septembre 2003 pour le traitement *adalimumab*. Ces 64 cas sont survenus entre juillet 2001 et octobre 2009. Les dates d'exposition au traitement et les dates de survenue de l'effet indésirable étaient connus pour tous les cas donc aucun des délais de survenue ni des délais de troncature n'était manquant. Les 64 cas de lymphomes ont donc été utilisés pour les analyses. Le délai de troncature maximum observé est de $t^* = 529$ semaines.

Tous traitements confondus, les estimations naïve et basée sur la troncature à droite du maximum de vraisemblance du paramètre θ ont été obtenues pour les modèles exponentiel, de Weibull et log-logistique. A partir de ces estimations, nous avons déterminé les délais de survenue moyens correspondants (cf. Tableau 1.1). En complément, nous

avons dérivé l'estimation non-paramétrique (1.5) du maximum de vraisemblance.

Le tableau 2.5 présente les estimations des paramètres pour les trois modèles et les deux approches. Il n'y a eu aucun problème de maximisation. Les estimations naïves sont toujours plus élevées que les estimations basées sur la troncature à droite. D'après les résultats de l'étude de simulations, on peut penser que l'estimateur naïf surestime les vraies valeurs des paramètres θ_1 et θ_2 et que la taille du biais est liée à la probabilité inconnue κ. L'estimation du paramètre θ par l'approche basée sur la troncature à droite permet d'estimer κ en calculant $F(t^* = 529, \widehat{\theta}_{\text{EBT}})$. Cependant, les estimations de κ varient en fonction du modèle étudié. En particulier, pour le modèle de Weibull, l'estimation est proche de 1 ($\widehat{\kappa} = 0.98$). Plus l'estimation se rapproche de 1, plus les estimations naïve et pertinente sont proches. Ce constat n'est pas surprenant étant donné que, comme précisé dans la remarque 2.1.1, la vraisemblance de l'estimateur pertinent est égale à la vraisemblance de l'estimateur naïf quand la troncature disparaît.

La figure 2.1 montre, pour les données, l'estimation non-paramétrique de la fonction de survie conditionnelle, $\widehat{\frac{F(x)}{F(529)}}$, et l'estimation paramétrique du maximum de vraisemblance des fonctions de survie conditionnelle, $\frac{F(x,\widehat{\theta}_{\text{TBE}})}{F(529,\widehat{\theta}_{\text{TBE}})}$, et non conditionnelle, $F(x, \widehat{\theta}_{\text{TBE}})$, obtenue par l'approche basée sur la troncature à droite. Comme attendu, les estimations des fonctions de survie conditionnelles sont toujours plus proches de l'estimation non-paramétrique du maximum de vraisemblance que les estimations des fonctions de survie non conditionnelles. Les fonctions de survie conditionnelle et non conditionnelle de la loi de Weibull sont presque confondues parce que l'estimation de la probabilité κ est presque égale à 1. En effet, κ caractérise la part observable du support de la distribution de X. Avec $\kappa = 0.98$, on peut observer presque toute la distribution. La figure 2.1 montre que l'estimation de la fonction de survie conditionnelle de la loi de Weibull est plus proche de l'estimation non-paramétrique de la fonction de survie que les estimations paramétriques des fonctions de survie conditionnelles des lois exponentielle et log-logistique. Ainsi, le modèle de Weibull semble être un candidat raisonnable pour décrire les données. Ce constat met en évidence la nécessité de procédures d'adéquation adaptées aux données tronquées à droite pour déterminer le modèle décrivant le mieux

les données. Le chapitre 6 présentera quelques outils permettant d'étudier de manière plus approfondie l'adéquation des données aux modèles considérés.

La figure 2.2 montre l'estimation paramétrique du maximum de vraisemblance de la fonction de survie non conditionnelle pour les deux approches. La distance entre les deux approches diminue avec l'estimation $\widehat{\kappa}$ de la probabilité κ (dans cet ordre : exponentiel, log-logistique et Weibull). De plus, les fonctions de survie basées sur les estimations pertinentes sont toujours au-dessus de celles basées sur les estimations naïves, ce qui est cohérent avec le fait que l'estimateur naïf surestime la vraie valeur des paramètres θ_1 et θ_2. Même pour le modèle de Weibull, c'est-à-dire le modèle avec l'estimation de κ la plus proche de 1, le délai de survenue moyen serait de 135 semaines par l'approche naïve et de 193 semaines avec l'estimation pertinente, ce qui constitue un écart important (Tableau 2.5).

Nous avons aussi calculé des intervalles de confiance à 95% du délai de survenue moyen pour l'approche basée sur la troncature à droite. Etant donné que la normalité asymptotique de l'estimateur pertinent n'a pas encore été étudiée et suite à la remarque d'un reviewer, ces intervalles de confiance ont été dérivés à l'aide du bootstrap. Le bootstrap est une méthode de rééchantillonnage ; de nouveaux échantillons sont créés à partir de l'échantillon initialement observé $(x_i, t_i)_{1 \leqslant i \leqslant n}$ (Efron et Tibshirani, 1993). Classiquement, il consiste à tirer aléatoirement avec remise parmi les observations $(x_i, t_i)_{1 \leqslant i \leqslant n}$, où chaque observation a une probabilité $1/n$ d'être tirée. Dans le cas de données tronquées, Gross et Lai (1996) ont étudié une alternative : on considère les estimations non-paramétriques $\widehat{F}$ et $\widehat{G}$ des fonctions de répartition F et G obtenues à partir de la vraisemblance (1.1) des observations. On obtient la première observation x_1^b en tirant aléatoirement avec remise parmi les observations $(x_i)_{1 \leqslant i \leqslant n}$, où chaque observation a une probabilité d'être tirée qui est fonction de $\widehat{F}$. Indépendamment, on obtient la première réalisation t_1^b en tirant aléatoirement avec remise parmi les observations $(t_i)_{1 \leqslant i \leqslant n}$, où chaque observation a une probabilité d'être tirée qui est fonction de $\widehat{G}$. Si $x_1^b \leqslant t_1^b$, alors cette observation est incluse dans l'échantillon. Sinon, un autre couple est tiré aléatoirement. On génère des couples (x_i^b, t_i^b) jusqu'à ce que l'échantillon soit de taille

n. Gross et Lai (1996) ont comparé ces deux procédures de bootstrap pour des données tronquées et censurées et ont conclu qu'il était préférable d'utiliser le bootstrap classique. La méthode du bootstrap classique a donc été utilisée pour générer 5000 nouveaux échantillons de taille $n = 64$ et obtenir un échantillon de 5000 estimations du délai de survenue moyen à partir duquel nous construisons un intervalle de confiance. La méthode corrigée des percentiles pour le biais et accélération (BCa) a été appliquée pour le calcul des intervalles de confiance. Elle consiste à prendre les percentiles de la distribution du bootstrap des 5000 estimations du délai de survenue moyen correspondants à des fractions (par exemple 2.5% et 97.5% pour un intervalle de confiance à 95%) corrigées (Efron et Tibshirani, 1993). Les intervalles de confiance obtenus sont rassemblés dans le tableau 2.5. Quel que soit le modèle étudié, aucun de ces intervalles de confiance n'inclut l'estimation naïve du délai de survenue moyen, et ce bien que les intervalles de confiance soient extrêmement larges.

2.4 Conclusion

Ce chapitre a permis de mettre en évidence l'importance de la prise en compte du caractère tronqué à droite des données. L'estimateur $\widehat{\theta}_{\mathrm{EBT}}$ est l'estimateur pertinent du paramètre θ, fondé sur la vraisemblance (1.8), et doit être utilisé quand les données sont tronquées à droite. Le chapitre suivant étudie les propriétés asymptotiques de cet estimateur, que nous noterons à nouveau $\widehat{\theta}_n$.

TABLEAU 2.5 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : estimations du paramètre et du délai de survenue moyen pour les distributions exponentielle (θ_1), de Weibull (θ_1, θ_2) et log-logistique (θ_1, θ_2).

	Estimateur naïf			EBT				
Distribution	$\widehat{\theta}_{\mathrm{EN},1}$	$\widehat{\theta}_{\mathrm{EN},2}$	Espérance (Semaines)	$\widehat{\theta}_{\mathrm{EBT},1}$	$\widehat{\theta}_{\mathrm{EBT},2}$	$\widehat{\kappa}$	Espérance (Semaines) Estimation	IC
Exponentielle	0.00739	-	135	0.00172	-	0.60	581	[264,7528]*
Weibull	0.00666	1.55	135	0.00468	1.49	0.98	193	[150,432]*
Log-logistique	0.00890	2.06	171	0.00408	1.53	0.76	567	[207,1.8×10^{12}]*

* Intervalles de confiance à 95% calculés en utilisant la méthode BCa du bootstrap classique basé sur 5000 réplications.
$\widehat{\kappa} = F(t^* = 529, \widehat{\theta}_{\mathrm{EBT}})$.
Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature ; IC, intervalle de confiance.

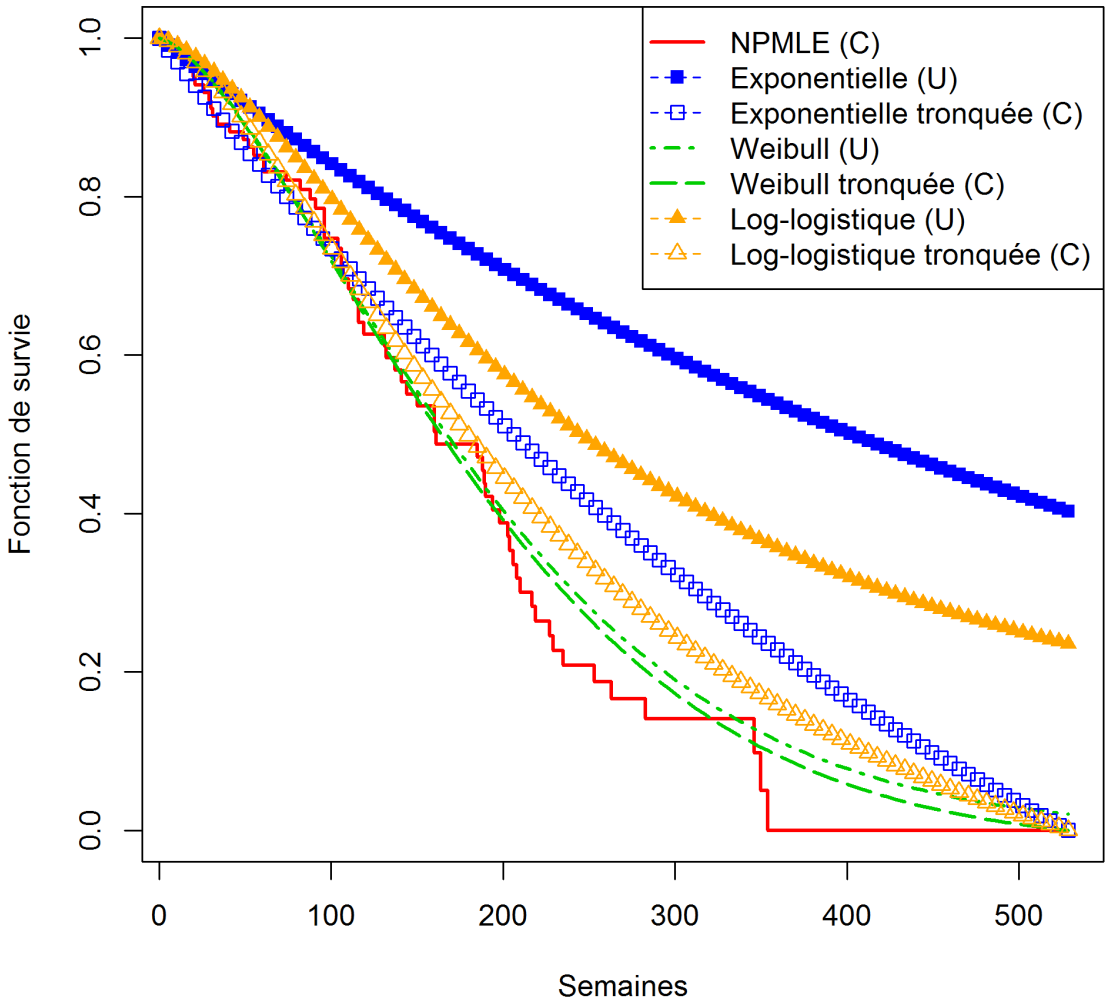

FIGURE 2.1 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : estimations basées sur la troncature à droite de la distribution du délai jusqu'à la survenue d'un lymphome consécutif à la prise d'un traitement anti TNF-α. Les données sont constituées de 64 cas de lymphomes. Les modèles exponentiel, de Weibull et log-logistique sont considérés. Les estimations paramétriques basées sur la troncature à droite de la fonction de survie non conditionnelle (U), de la fonction de survie tronquée à 529 semaines (C) ainsi que l'estimation non-paramétrique du maximum de vraisemblance de la fonction de survie tronquée à 529 semaines (NPMLE) sont représentées.

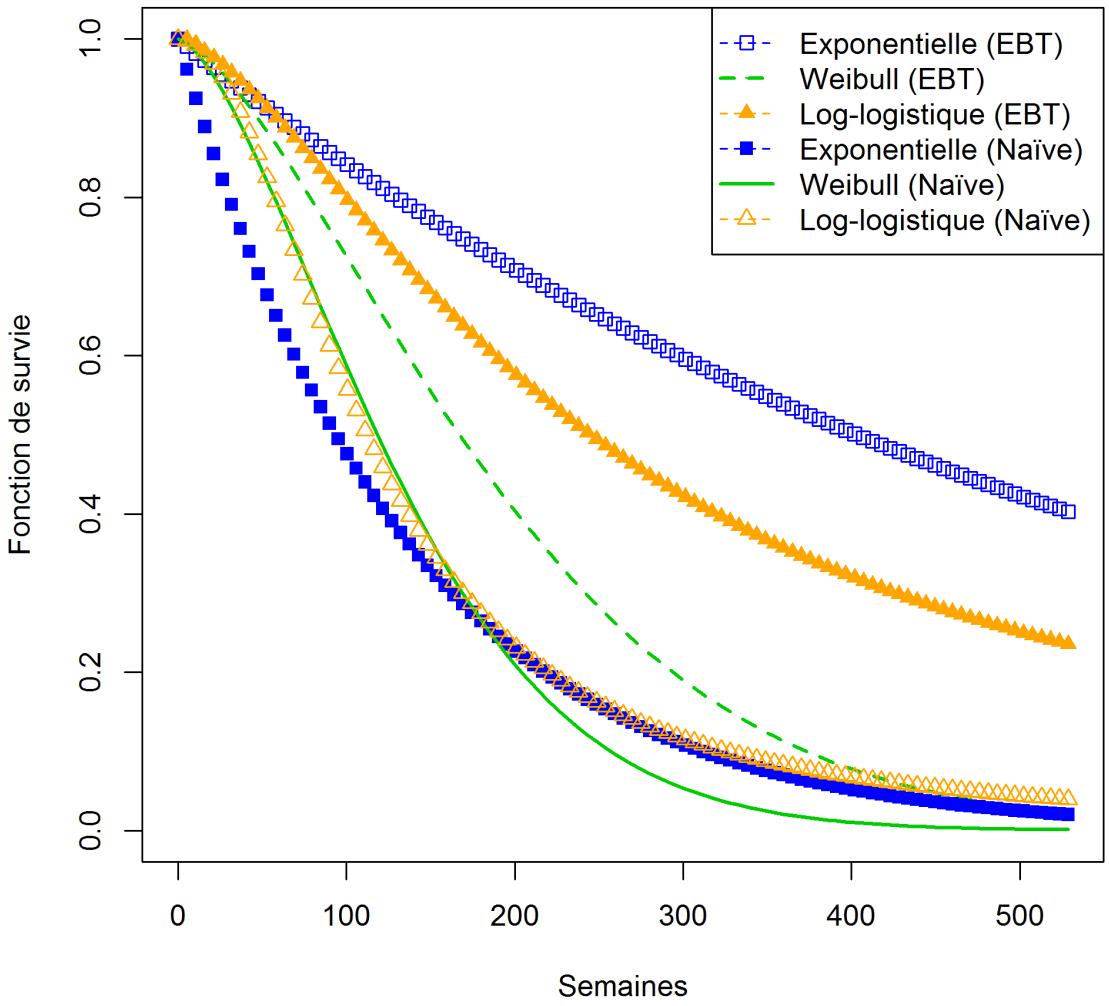

FIGURE 2.2 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : estimations naïves et basées sur la troncature à droite de la distribution du délai jusqu'à la survenue d'un lymphome consécutif à la prise d'un traitement anti TNF-α. Les données sont constituées de 64 cas de lymphomes. Les modèles exponentiel, de Weibull et log-logistique sont considérés. Les estimations paramétriques naïves (Naïve) et basées sur la troncature à droite (TBE) de la fonction de survie non conditionnelle sont représentées.

Chapitre 3

Propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance

3.1 Introduction

Dans le cas où la vraisemblance conditionnelle (1.7) est considérée pour obtenir l'estimation conjointe de X et de T, les données observées sont constituées de n réalisations indépendantes et identiquement distribuées du vecteur aléatoire $(X, T | X \leqslant T)$. Dans ce cas, la théorie asymptotique classique du maximum de vraisemblance s'applique. Cependant, modéliser la variable aléatoire T augmente le nombre de paramètres à estimer, ce qui n'est pas raisonnable avec les petites tailles d'échantillon que l'on rencontre en pharmacovigilance. De plus, choisir une distribution paramétrique pour T n'est pas simple. Nous avons donc considéré la vraisemblance conditionnelle (1.8) et travaillons ainsi dans un cadre déterministe pour les délais de troncature. Dans ce contexte, les délais de survenue observés $(x_i)_{1 \leqslant i \leqslant n}$ constituent un échantillon de réalisations indépendantes des variables aléatoires de distribution respective la distribution conditionnelle de X_i sachant $\{X_i \leqslant t_i\}$, de fonction de répartition $F(\cdot;\theta)/F(t_i;\theta)$ et de densité de

probabilité $f(\cdot;\theta)/F(t_i;\theta)$. On a précisé dans la section 1.1 que chaque patient avait son propre délai de troncature. Par conséquent, nos observations sont des réalisations indépendantes mais non-identiquement distribuées.

La question fondamentale des propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance dans le cas d'un paramètre multidimensionnel et en présence d'observations indépendantes et identiquement distribuées est étudiée depuis plusieurs décennies déjà. Une revue de la littérature est donnée par Hoadley (1971) ou Lehmann (1983). Pour prouver la consistance de l'estimateur, Chanda (1954) résolva les équations normales tandis que Lehmann (1983) étudia le signe d'une fonction de la log-vraisemblance sur une sphère centrée en la vraie valeur du paramètre vectoriel θ.

Les propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance en présence d'observations indépendantes mais non-identiquement distribuées ont déjà été explorées. Hoadley (1971) a considéré le cas de densités de probabilité ayant une part discrète et une part continue. Le cas particulier des modèles formant la famille exponentielle a été étudié par Nordberg (1980). Cependant, certaines distributions largement utilisées en analyse des données de survie ne constituent pas une famille exponentielle ; c'est le cas par exemple de la loi de Weibull. Il existe d'autres approches où les hypothèses que doivent vérifier les densités de probabilité sont énoncées différemment : Ibragimov et Khas'Minskii (1972, 1975a,b, 1981) ont exploré les propriétés asymptotiques de cet estimateur en utilisant le rapport de vraisemblance. Van der Vaart et Wellner (2000) ont développé les propriétés asymptotiques des M-estimateurs dans un cadre très général, incluant les données indépendantes mais non-identiquement distribuées, en utilisant les processus empiriques. Mais les hypothèses proposées dans ces papiers peuvent ne pas être simples à vérifier dans certaines situations. Tandis que Bradley et Gart (1962) ont développé l'extension de la preuve de Chanda (1954) au cas de réalisations indépendantes mais non-identiquement distribuées, il n'existe pas d'extension de la preuve de Lehmann (1983). Dans leur papier, Bradley et Gart (1962) considèrent deux cas de figure : soit le nombre de densités de probabilité distinctes qui peuvent être observées dans la population d'où provient l'échantillon est fini, soit

celui-ci est infini. Dans le cadre statistique initial présenté dans la section 1.1, on a fait l'hypothèse que la variable aléatoire caractérisant le délai de troncature était continue (hypothèse classique pour un délai de survenue). Pour le cadre déterministe dans lequel nous nous plaçons désormais, la distribution de T importe peu. Cependant, elle détermine dans quel cas de figure nous nous trouvons vis-à-vis du nombre de densités de probabilité distinctes qui peuvent être observées dans la population étudiée. En effet, à chaque valeur distincte du délai de troncature correspond une densité de probabilité distincte. Le délai de troncature est le délai écoulé entre la date d'initiation du traitement médicamenteux et la date d'analyse t_a. Le nombre de jours écoulés entre la date d'autorisation de mise sur le marché du médicament et la date d'analyse t_a est fini. Par conséquent, il est réaliste de faire l'hypothèse d'un nombre fini de délais de troncature distincts observables dans la population étudiée.

Dans ce chapitre, nous développons l'extension de la démonstration de Lehmann (1983) aux observations indépendantes mais non-identiquement distribuées. Par souci de généralité, nous considérons les deux cas de figure traités par Bradley et Gart (1962). Pour chaque cas, nous définissons les notations associées, nous énonçons des conditions suffisantes pour que l'estimateur paramétrique du maximum de vraisemblance soit consistant et asymptotiquement normal et nous développons les preuves. Dans le cas d'un nombre infini de densités de probabilité distinctes, les hypothèses sont légèrement différentes de celles énoncées par Bradley et Gart (1962). Les démonstrations de la consistance sont renvoyées en annexe (annexes B et C).

Pour une suite de variables aléatoires indexée par n, les convergences en probabilité et en distribution sont notées respectivement $\xrightarrow[n\to+\infty]{P}$ et $\xrightarrow[n\to+\infty]{d}$.

3.2 Cas d'un nombre infini de distributions de probabilité observables

3.2.1 Notations et hypothèses

Soit $(x_1, x_2, \ldots, x_n)$ un échantillon de réalisations de n variables aléatoires indépendantes $(X_1, X_2, \ldots, X_n)$ de densité de probabilité respective $f_i(\cdot;\theta)$, pour $i = 1, \ldots, n$, où $\theta = (\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_r)$ est un paramètre vectoriel inconnu partagé par toutes ces densités de probabilité non nécessairement identiques. Le vecteur θ appartient à l'ensemble Θ, un ouvert de $\mathbb{R}^r$. Soit $\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_j^0, \ldots, \theta_r^0)$ la vraie valeur du paramètre. Pour tout $i = 1, \ldots, n$, soit $S_i \subset \mathbb{R}$ le support de la densité de probabilité $f_i(\cdot;\theta)$. Le support S_i doit être indépendant du paramètre vectoriel inconnu θ. Soit $\|.\|$ une norme sur Θ ou $\mathbb{R}^r$; l'espace Θ est muni de la topologie induite par cette norme. La vraisemblance de l'échantillon s'écrit

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_i(x_i; \theta)$$

et l'estimateur du maximum de vraisemblance est défini par

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(x_1, x_2, \ldots, x_n; \theta).$$

Les équations normales s'écrivent

$$\nabla_\theta \log L(x_1, x_2, \ldots, x_n; \theta) = 0,$$

où ∇_θ est l'opérateur gradient.

Remarque 3.2.1. Nous supposons que le paramètre vectoriel inconnu θ est partagé par toutes les densités de probabilité parce que c'est le cas pour les données tronquées à droite. Cependant, les théorèmes et leurs démonstrations restent valables quand ce n'est pas le cas.

Nous faisons les hypothèses suivantes :

Hypothèse 3.2.1. *L'estimateur du maximum de vraisemblance est solution des équations normales.*

Hypothèse 3.2.2. *Les équations normales ont une unique solution.*

Hypothèse 3.2.3. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, n$ *et* $(j, p, q) \in \{1, \ldots, r\}^3$, *les dérivées partielles*

$$\frac{\partial log f_i(\cdot;\theta)}{\partial \theta_j}, \ \frac{\partial^2 log f_i(\cdot;\theta)}{\partial \theta_j \partial \theta_p} \quad et \quad \frac{\partial^3 log f_i(\cdot;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}$$

existent pour presque tout x.

Hypothèse 3.2.4. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, n$ *et* $j \in \{1, \ldots, r\}$, *la dérivée partielle* $\dfrac{\partial}{\partial \theta_j} f_i(\cdot;\theta)$ *est une fonction intégrable sur* S_i *et*

$$\int_{S_i} \frac{\partial}{\partial \theta_j} f_i(x;\theta) dx \ = \ \frac{\partial}{\partial \theta_j} \int_{S_i} f_i(x;\theta) dx.$$

Remarque 3.2.2. L'intégrale d'une densité de probabilité sur son support vaut 1. Par conséquent, le membre de droite de l'équation ci-dessus est nul.

Hypothèse 3.2.5. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, n$ *et* $(j, p) \in \{1, \ldots, r\}^2$, *la dérivée partielle* $\dfrac{\partial^2}{\partial \theta_j \partial \theta_p} f_i(\cdot;\theta)$ *est une fonction intégrable sur* S_i *et*

$$\int_{S_i} \frac{\partial^2}{\partial \theta_j \partial \theta_p} f_i(x;\theta) dx \ = \ \frac{\partial}{\partial \theta_j} \int_{S_i} \frac{\partial}{\partial \theta_p} f_i(x;\theta) dx.$$

Hypothèse 3.2.6. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, n$ *et* $(j, p) \in \{1, \ldots, r\}^2$,

$$E_\theta \left(\frac{\partial^2 log f_i(X_i;\theta)}{\partial \theta_j \partial \theta_p} \right) \ et \ \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} E_\theta \left(\frac{\partial^2 log f_i(X_i;\theta)}{\partial \theta_j \partial \theta_p} \right)$$

existent et

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 log f_i(X_i;\theta)}{\partial \theta_j \partial \theta_p} \ \xrightarrow[n \to +\infty]{P} \ \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} E_\theta \left(\frac{\partial^2 log f_i(X_i;\theta)}{\partial \theta_j \partial \theta_p} \right).$$

Hypothèse 3.2.7. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, n$ *et* $(j, p, q) \in \{1, \ldots, r\}^3$,

$$E_\theta \left(\frac{\partial^3 log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right) \text{ et } \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} E_\theta \left(\frac{\partial^3 log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right)$$

existent et

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \xrightarrow[n\to+\infty]{P} \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} E_\theta \left(\frac{\partial^3 log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right).$$

Remarque 3.2.3. La loi faible des grands nombres (Feller, 1968) donne des conditions suffisantes pour la convergence en probabilité des hypothèses 3.2.6 et 3.2.7.

Remarque 3.2.4. Les hypothèses 3.2.4-3.2.6 nous assurent que, pour tout $\theta \in \Theta$ et $j \in \{1, \ldots, r\}$, on a

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f_i(X_i;\theta)}{\partial\theta_j} \xrightarrow[n\to+\infty]{P} \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_\theta \left(\frac{\partial \log f_i(X_i;\theta)}{\partial\theta_j} \right) = 0. \tag{3.1}$$

L'hypothèse 3.2.4 nous assure que l'espérance $\mathrm{E}_\theta \left(\partial \log f_i(X_i;\theta)/\partial\theta_j\right)$ existe et est nulle pour tout $i = 1, \ldots, n$ et $j \in \{1, \ldots, r\}$. Ainsi, la limite dans le membre droit de l'expression (3.1) existe. D'après l'hypothèse 3.2.5, la variance $\mathrm{Var}\left(\partial \log f_i(X_i;\theta)/\partial\theta_j\right)$ coïncide avec l'espérance $\mathrm{E}_\theta \left(-\partial^2 \log f_i(X_i;\theta)/\partial\theta_j \partial\theta_p\right)$ et la limite $\lim_{n\to+\infty} (1/n) \sum_{i=1}^{n} \mathrm{Var}$ $\left(\partial \log f_i(X_i;\theta)/\partial\theta_j\right)$ coïncide avec la limite $\lim_{n\to+\infty} (1/n) \sum_{i=1}^{n} \mathrm{E}_\theta \left(-\partial^2 \log f_i(X_i;\theta)/\partial\theta_j \partial\theta_p\right)$. L'hypothèse 3.2.6 nous assure que ces quantités existent. Ainsi, les conditions suffisantes de la loi faible des grands nombres sont vérifiées et on a bien l'expression (3.1).

Hypothèse 3.2.8. *Il existe* K *tel que pour tout* $\theta \in \Theta$ *et* $(j, p, q) \in \{1, \ldots, r\}^3$,

$$\left| \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} E_\theta \left(\frac{\partial^3 log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right) \right| < K.$$

Hypothèse 3.2.9. *La matrice* $I\left(\theta^0\right) = \left(I_{jp}\left(\theta^0\right)\right)_{1\leqslant j,p\leqslant r}$, *où*

$$I_{jp}\left(\theta^0\right) = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} E_{\theta^0}\left(-\frac{\partial^2 log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right),$$

est définie positive.

Remarque 3.2.5. D'après les hypothèses 3.2.4 et 3.2.5, nous sommes déjà sûrs que $I\left(\theta^0\right)$ existe et est une matrice semi-définie positive.

Hypothèse 3.2.10. *Pour tout* $\epsilon > 0$,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} E_{\theta^0}\left[\sum_{j=1}^{r}\left(\frac{\partial log f_i(X_i;\theta)}{\partial\theta_j}\bigg|_{\theta^0}\right)^2 I\left\{\left(\sum_{j=1}^{r}\left(\frac{\partial log f_i(X_i;\theta)}{\partial\theta_j}\bigg|_{\theta^0}\right)^2\right)^{\frac{1}{2}} > \epsilon\sqrt{n}\right\}\right] = 0.$$

Remarque 3.2.6. L'hypothèse 3.2.10 est l'hypothèse du théorème de la limite centrale multivariée pour des observations indépendantes mais non-identiquement distribuées (Feller, 1971).

3.2.2 Consistance

Le théorème suivant établit la consistance de l'estimateur paramétrique du maximum de vraisemblance.

Théorème 3.2.1. *Si les hypothèses 3.2.1-3.2.9 sont satisfaites, l'estimateur du maximum de vraisemblance* $\widehat{\theta}_n = \left(\widehat{\theta}_{1n},\ldots,\widehat{\theta}_{rn}\right)$ *est un estimateur consistant de* $\theta^0 = (\theta_1^0,\ldots,\theta_r^0)$, *c'est-à-dire pour tout* $\zeta > 0$, $P\left(\left\|\widehat{\theta}_n - \theta^0\right\| < \zeta\right) \underset{n\to+\infty}{\longrightarrow} 1$.

A partir de la formule de Taylor-Lagrange et en considérant séparément chaque terme de ce développement, la démonstration de ce théorème consiste à montrer que la vraisemblance en chaque point de la sphère de centre θ^0 et de rayon ζ - constante arbitraire strictement positive - est inférieure strictement à la vraisemblance au point θ^0, centre de la sphère. Ceci assure alors l'existence d'un maximum local de la fonction de vraisemblance dans l'intérieur de la boule de centre θ^0 et de rayon ζ. Les hypothèses 3.2.1

et 3.2.2 permettent de faire coïncider ce maximum local avec l'estimateur du maximum de vraisemblance, ce qui achève la démonstration de la consistance. Les détails de cette preuve sont renvoyés à l'annexe B.

3.2.3 Normalité asymptotique

Le théorème suivant établit la normalité asymptotique de l'estimateur paramétrique du maximum de vraisemblance.

Théorème 3.2.2. *Si les hypothèses 3.2.1-3.2.10 sont satisfaites, le vecteur aléatoire* $\sqrt{n}(\widehat{\theta}_n - \theta^0)$ *est asymptotiquement normal d'espérance nulle et de matrice de covariance* $\big[I(\theta^0)\big]^{-1}$.

Démonstration. D'après la formule de Taylor-Lagrange et l'hypothèse 3.2.3, nous savons que : pour tout $\theta \in \Theta$, il existe θ' appartient au segment $]\theta^0, \theta[$ tel que pour tout $j = 1, \ldots, r$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\bigg|_{\theta} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\bigg|_{\theta^0} \\ &\quad + \sum_{p=1}^{r}(\theta_p - \theta_p^0)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p}\bigg|_{\theta^0} \\ &\quad + \frac{1}{2}\sum_{p=1}^{r}\sum_{q=1}^{r}(\theta_p - \theta_p^0)(\theta_q - \theta_q^0)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\bigg|_{\theta'}.\end{aligned}$$

Mais pour $\theta = \widehat{\theta}_n$, nous avons

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f_i(x_i;\theta)|_{\widehat{\theta}_n} = 0.$$

Ainsi, pour tout $j = 1, \ldots, r$, nous avons,

$$-\sum_{p=1}^{r}(\widehat{\theta}_p - \theta_p^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0} - \frac{1}{2}\sum_{p=1}^{r}\sum_{q=1}^{r}(\widehat{\theta}_p - \theta_p^0)(\widehat{\theta}_q - \theta_q^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'} = \frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_j}\right|_{\theta^0}.$$

Factorisant par $\sum_{p=1}^{r}(\widehat{\theta}_p - \theta_p^0)$ et multipliant par $\sqrt{n}$, on obtient

$$-\sqrt{n}\sum_{p=1}^{r}(\widehat{\theta}_p - \theta_p^0)\left[\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0} + \frac{1}{2}\sum_{q=1}^{r}(\widehat{\theta}_q - \theta_q^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_j}\right|_{\theta^0}.$$

En notation matricielle, cela donne

$$\begin{pmatrix} a_{11n} & a_{12n} & \cdots & a_{1rn} \\ a_{21n} & a_{22n} & \cdots & a_{2rn} \\ \vdots & & \ddots & \vdots \\ a_{r1n} & a_{r2n} & \cdots & a_{rrn} \end{pmatrix}\begin{pmatrix} \sqrt{n}(\widehat{\theta}_1 - \theta_1^0) \\ \sqrt{n}(\widehat{\theta}_2 - \theta_2^0) \\ \vdots \\ \sqrt{n}(\widehat{\theta}_r - \theta_r^0) \end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} \left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_1}\right|_{\theta^0} \\ \left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_2}\right|_{\theta^0} \\ \vdots \\ \left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_r}\right|_{\theta^0} \end{pmatrix}$$

où

$$a_{jpn} = -\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0} - \frac{1}{2}\sum_{q=1}^{r}(\widehat{\theta}_q - \theta_q^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}.$$

Les vecteurs $\nabla_\theta \log f_i(X_i;\theta)|_{\theta^0}$, pour tout $i = 1, \ldots, n$, sont indépendants mais non-identiquement distribués d'espérance nulle et de matrice de covariance $V_i(\theta^0) = (V_{ijp}(\theta^0))_{1\leqslant j,p\leqslant r}$, où

$$V_{ijp}(\theta^0) = \mathrm{E}_{\theta^0}\left(\left.\frac{\partial \log f_i(X_i;\theta)}{\partial\theta_j}\right|_{\theta^0}\left.\frac{\partial \log f_i(X_i;\theta)}{\partial\theta_p}\right|_{\theta^0}\right).$$

Nous savons d'après l'hypothèse 3.2.5 que

$$V_{ijp}(\theta^0) = \mathrm{E}_{\theta^0}\left(-\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0}\right)$$

et d'après l'hypothèses 3.2.9 que

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} V_i(\theta^0) = I(\theta^0).$$

Ainsi, d'après l'hypothèse 3.2.10 et le théorème de la limite centrale multivariée pour des variables aléatoires indépendantes mais non-identiquement distribuées (Feller, 1971), nous obtenons

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log f_i(X_i;\theta)|_{\theta^0} \xrightarrow[n\to+\infty]{d} N(0, I(\theta^0)). \tag{3.2}$$

D'après les hypothèses 3.2.6 et 3.2.7, la consistance de l'estimateur du maximum de vraisemblance et le théorème de Slutsky, nous avons pour tout $(j,p) \in \{1,\ldots,r\}^2$,

$$a_{jpn} \xrightarrow[n\to+\infty]{P} I_{jp}(\theta^0).$$

Ces convergences en probabilité et la convergence (3.2) conduisent à

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_1 - \theta_1^0) \\ \vdots \\ \sqrt{n}(\widehat{\theta}_r - \theta_r^0) \end{pmatrix} \xrightarrow[n\to+\infty]{d} N(0, \left[I(\theta^0)\right]^{-1}).$$

□

3.3 Cas d'un nombre fini de distributions de probabilité observables

3.3.1 Notations et hypothèses

On note toujours $(x_1, x_2, \ldots, x_n)$ l'échantillon de réalisations de n variables aléatoires indépendantes. Etant donné qu'il y a un nombre fini de densités de probabilité distinctes observables dans la population d'où est issu l'échantillon, celui-ci peut être indexé en fonction de la densité de probabilité de chaque réalisation. Soit M le nombre de densités de probabilité différentes observables dans la population étudiée. Pour tout $i = 1, \ldots, M$, soit $(x_{i1}, x_{i2}, \ldots, x_{ik}, \ldots, x_{in_i})$ le sous-échantillon de réalisations des n_i variables aléatoires $(X_{i1}, X_{i2}, \ldots, X_{ik}, \ldots, X_{in_i})$ indépendantes et identiquement distribuées de densité de probabilité $f_i(\cdot;\theta)$. Alors $\sum_{i=1}^{M} n_i = n$ et l'échantillon peut maintenant s'écrire $(x_{ik})_{1\leqslant i\leqslant M, 1\leqslant k\leqslant n_i}$. La vraisemblance des observations a pour expression :

$$L(x_{11}, x_{12}, \ldots, x_{\mathrm{M}n_{\mathrm{M}}};\theta) = \prod_{i=1}^{M}\prod_{k=1}^{n_i} f_i(x_{ik};\theta)$$

et l'estimateur du maximum de vraisemblance est défini par :

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(x_{11}, x_{12}, \ldots, x_{\mathrm{M}n_{\mathrm{M}}};\theta).$$

Les équations normales s'écrivent

$$\nabla_\theta \log L(x_{11}, x_{12}, \ldots, x_{\mathrm{M}n_{\mathrm{M}}};\theta) = 0.$$

Pour tout $i = 1, \ldots, M$, soit $\mu_i = n_i/n$ la proportion des observations de densité de probabilité $f_i(\cdot;\theta)$ et $0 < \lambda_i < 1$ la limite de la suite μ_i quand n tend vers l'infini. Ces quantités vérifient la relation $\sum_{i=1}^{M} \lambda_i = 1$.

Remarque 3.3.1. Le cas où il existerait q tel que $\lambda_q = 0$ (resp. $\lambda_q = 1$) correspond au cas où il y aurait en réalité seulement $M - 1$ densités de probabilité distinctes dans

la population étudiée (resp. les observations seraient indépendantes et identiquement distribuées).

Remarque 3.3.2. Dans notre contexte de pharmacovigilance, le nombre de délais de troncature distincts observables dans la population peut être élevé mais la taille d'échantillon limitée. Il est alors fort probable que tous les délais de troncature ne soient pas observés dans l'échantillon disponible. Dans ce cas-là, on définit $m(n)$ le nombre de densités de probabilité observées dans l'échantillon et, sans perte de généralité, on suppose que $(f_i(\cdot;\theta))_{1\leqslant i\leqslant m(n)}$ est l'ensemble des $m(n)$ densités de probabilité distinctes observées. Il suffit ensuite de remplacer M par $m(n)$ dans les écritures de l'échantillon et de la vraisemblance. La suite $m(n)$ vérifie alors

$$m(n) \underset{n\to+\infty}{\longrightarrow} M. \tag{3.3}$$

Nous faisons les hypothèses suivantes :

Hypothèse 3.3.1. *L'estimateur du maximum de vraisemblance est solution des équations normales.*

Hypothèse 3.3.2. *Les équations normales ont une unique solution.*

Hypothèse 3.3.3. *Pour tout $\theta \in \Theta$, $i = 1, \ldots, M$ et $(j,p,q) \in \{1,\ldots,r\}^3$, les dérivées partielles*

$$\frac{\partial log f_i(\cdot;\theta)}{\partial\theta_j},\ \frac{\partial^2 log f_i(\cdot;\theta)}{\partial\theta_j\partial\theta_p}\quad et\quad \frac{\partial^3 log f_i(\cdot;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}$$

existent pour presque tout x.

Hypothèse 3.3.4. *Pour tout $\theta \in \Theta$, $i = 1, \ldots, M$ et $j \in \{1,\ldots,r\}$, la dérivée partielle $\dfrac{\partial}{\partial\theta_j} f_i(\cdot;\theta)$ est une fonction intégrable sur S_i et*

$$\int_{S_i} \frac{\partial}{\partial\theta_j} f_i(x;\theta)dx \;=\; \frac{\partial}{\partial\theta_j}\int_{S_i} f_i(x;\theta)dx.$$

Remarque 3.3.3. L'intégrale d'une densité de probabilité sur son support vaut 1. Par conséquent, le membre de droite de l'équation ci-dessus est nul.

Hypothèse 3.3.5. *Pour tout* $\theta \in \Theta$, $i = 1, \ldots, M$ *et* $(j,p) \in \{1,\ldots,r\}^2$, *la dérivée partielle* $\dfrac{\partial^2}{\partial\theta_j\partial\theta_p} f_i(\cdot;\theta)$ *est une fonction intégrable sur* S_i *et*

$$\int_{S_i} \frac{\partial^2}{\partial\theta_j\partial\theta_p} f_i(x;\theta)dx \;=\; \frac{\partial}{\partial\theta_j}\int_{S_i}\frac{\partial}{\partial\theta_p} f_i(x;\theta)dx.$$

Hypothèse 3.3.6. *Il existe* K *tel que pour tout* $\theta \in \Theta$, $i = 1,\ldots,M$ *et* $(j,p,q) \in \{1,\ldots,r\}^3$,

$$E_\theta\left(\frac{\partial^3 log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right)$$

existe et

$$\left|\sum_{i=1}^{M}\lambda_i E_\theta\left(\frac{\partial^3 log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right)\right| < K.$$

Hypothèse 3.3.7. *Pour tout* $\theta \in \Theta$, $i = 1,\ldots,M$ *et* $(j,p) \in \{1,\ldots,r\}^2$,

$$E_\theta\left(\frac{\partial^2 log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right)$$

existe et la matrice $I\left(\theta^0\right) = \left(I_{jp}\left(\theta^0\right)\right)_{1\leqslant j,p\leqslant r}$, *où*

$$I_{jp}\left(\theta^0\right) = \sum_{i=1}^{M}\lambda_i E_{\theta^0}\left(-\left.\frac{\partial^2 log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right),$$

est définie positive.

Remarque 3.3.4. D'après les hypothèses 3.3.4 et 3.3.5, nous sommes déjà sûrs que $I(\theta^0)$ existe et est une matrice semi-définie positive.

Remarque 3.3.5. Pour chaque densité de probabilité, les observations du sous-échantillon correspondant sont indépendantes et identiquement distribuées. Il n'y a donc pas besoin de supposer les convergences en probabilité qui apparaissent dans les hypothèses 3.2.6 et 3.2.7 de la section précédente. De la même manière, il n'y a pas besoin d'hypothèse analogue à l'hypothèse 3.2.10 pour que le théorème de la limite centrale multivariée soit vérifié.

3.3.2 Consistance

Le théorème suivant établit la consistance de l'estimateur paramétrique du maximum de vraisemblance.

Théorème 3.3.1. *Si les hypothèses 3.3.1-3.3.7 sont satisfaites, l'estimateur du maximum de vraisemblance* $\widehat{\theta}_n = \left(\widehat{\theta}_{1n}, \ldots, \widehat{\theta}_{rn}\right)$ *est un estimateur consistant de* $\theta^0 = (\theta_1^0, \ldots, \theta_r^0)$*, c'est-à-dire pour tout* $\zeta > 0$*,* $P\left(\left\|\widehat{\theta}_n - \theta^0\right\| < \zeta\right) \underset{n\to+\infty}{\longrightarrow} 1$.

A partir de la formule de Taylor-Lagrange et en considérant séparément chaque terme de ce développement, la démonstration de ce théorème consiste à montrer que la vraisemblance en chaque point de la sphère de centre θ^0 et de rayon ζ - constante arbitraire strictement positive - est inférieure strictement à la vraisemblance au point θ^0, centre de la sphère. Ceci assure alors l'existence d'un maximum local de la fonction de vraisemblance dans l'intérieur de la boule de centre θ^0 et de rayon ζ. Les hypothèses 3.3.1 et 3.3.2 permettent de faire coïncider ce maximum local avec l'estimateur du maximum de vraisemblance, ce qui achève la démonstration de la consistance. Les détails de cette preuve sont renvoyés à l'annexe C.

3.3.3 Normalité asymptotique

Le théorème suivant établit la normalité asymptotique de l'estimateur paramétrique du maximum de vraisemblance.

Théorème 3.3.2. *Si les hypothèses 3.3.1-3.3.7 sont satisfaites, le vecteur aléatoire* $\sqrt{n}(\widehat{\theta}_n - \theta^0)$ *est asymptotiquement normal d'espérance nulle et de matrice de covariance* $\left[I(\theta^0)\right]^{-1}$.

Démonstration. D'après la formule de Taylor-Lagrange et l'hypothèse 3.3.3, nous savons que : pour tout $\theta \in \Theta$, il existe θ' appartenant au segment $]\theta^0, \theta[$ tel que pour tout

$j = 1, \ldots, r$,

$$\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\bigg|_{\theta} = \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\bigg|_{\theta^0} + \sum_{p=1}^{r} (\theta_p - \theta_p^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p}\bigg|_{\theta^0} + \frac{1}{2} \sum_{p=1}^{r} \sum_{q=1}^{r} (\theta_p - \theta_p^0)(\theta_q - \theta_q^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\bigg|_{\theta'}.$$

Mais pour $\theta = \widehat{\theta}_n$, nous avons

$$\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \nabla_\theta \log f_i(x_{ik};\theta)|_{\widehat{\theta}_n} = 0.$$

Ainsi, pour tout $j = 1, \ldots, r$, nous avons,

$$-\sum_{p=1}^{r} (\widehat{\theta}_p - \theta_p^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p}\bigg|_{\theta^0} - \frac{1}{2} \sum_{p=1}^{r} \sum_{q=1}^{r} (\widehat{\theta}_p - \theta_p^0)(\widehat{\theta}_q - \theta_q^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\bigg|_{\theta'} = \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\bigg|_{\theta^0}.$$

Factorisant par $\sum_{p=1}^{r} (\widehat{\theta}_p - \theta_p^0)$ et multipliant par $\sqrt{n}$, nous avons pour tout $j = 1, \ldots, r$,

$$-\sqrt{n} \sum_{p=1}^{r} (\widehat{\theta}_p - \theta_p^0) \left[\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p}\bigg|_{\theta^0} + \frac{1}{2} \sum_{q=1}^{r} (\widehat{\theta}_q - \theta_q^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\bigg|_{\theta'} \right] = \sqrt{n} \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\bigg|_{\theta^0}.$$

En écrivant $\mu_i = \sqrt{\mu_i}\sqrt{\frac{n_i}{n}}$, nous obtenons pour tout $j = 1, \dots, r$,

$$
\begin{aligned}
&- \sqrt{n} \sum_{p=1}^{r} (\widehat{\theta}_p - \theta_p^0) \left[\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial\theta_j \partial\theta_p} \bigg|_{\theta^0} \right. \\
&\left. + \frac{1}{2} \sum_{q=1}^{r} (\widehat{\theta}_q - \theta_q^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \bigg|_{\theta'} \right] = \sum_{i=1}^{M} \sqrt{\mu_i}\sqrt{n_i} \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial \log f_i(x_{ik};\theta)}{\partial\theta_j} \bigg|_{\theta^0}.
\end{aligned}
$$

En notation matricielle, cela donne

$$
\begin{pmatrix} a_{11n} & a_{12n} & \cdots & a_{1rn} \\ a_{21n} & a_{22n} & \cdots & a_{2rn} \\ \vdots & & \ddots & \vdots \\ a_{r1n} & a_{r2n} & \cdots & a_{rrn} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\theta}_1 - \theta_1^0) \\ \sqrt{n}(\widehat{\theta}_2 - \theta_2^0) \\ \vdots \\ \sqrt{n}(\widehat{\theta}_r - \theta_r^0) \end{pmatrix} = \sum_{i=1}^{M} \sqrt{\mu_i} \frac{1}{\sqrt{n_i}} \sum_{k=1}^{n_i} \begin{pmatrix} \frac{\partial \log f_i(x_{ik};\theta)}{\partial\theta_1} \Big|_{\theta^0} \\ \frac{\partial \log f_i(x_{ik};\theta)}{\partial\theta_2} \Big|_{\theta^0} \\ \vdots \\ \frac{\partial \log f_i(x_{ik};\theta)}{\partial\theta_r} \Big|_{\theta^0} \end{pmatrix}
$$

où

$$
a_{jpn} = - \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial\theta_j \partial\theta_p} \bigg|_{\theta^0} - \frac{1}{2} \sum_{q=1}^{r} (\widehat{\theta}_q - \theta_q^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \bigg|_{\theta'}.
$$

Les vecteurs $\nabla_\theta \log f_i(X_{ik};\theta)|_{\theta^0}$, pour tout $k = 1, \dots, n_i$, sont indépendants et identiquement distribués d'espérance nulle et de matrice de covariance $V_i(\theta^0) = (V_{ijp}(\theta^0))_{1 \leqslant j,p \leqslant r}$, où

$$
V_{ijp}(\theta^0) = \mathrm{E}_{\theta^0} \left(\frac{\partial \log f_i(X_{i1};\theta)}{\partial\theta_j} \bigg|_{\theta^0} \frac{\partial \log f_i(X_{i1};\theta)}{\partial\theta_p} \bigg|_{\theta^0} \right).
$$

Nous savons d'après l'hypothèse 3.3.5 que

$$
V_{ijp}(\theta^0) = \mathrm{E}_{\theta^0} \left(- \frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p} \bigg|_{\theta^0} \right).
$$

Ainsi, d'après le théorème de la limite centrale multivariée pour des variables aléatoires

indépendantes et identiquement distribuées, nous obtenons

$$\frac{1}{\sqrt{n_i}}\sum_{k=1}^{n_i}\nabla_\theta \log f_i(X_{ik};\theta)|_{\theta^0} \xrightarrow[n\to+\infty]{d} N(0, V(\theta^0)).$$

Et d'après le théorème de Slutsky,

$$\sqrt{\mu_i}\frac{1}{\sqrt{n_i}}\sum_{k=1}^{n_i}\nabla_\theta \log f_i(X_{ik};\theta)|_{\theta^0} \xrightarrow[n\to+\infty]{d} N(0, \lambda_i V(\theta^0)).$$

Etant donné que la somme finie de variables aléatoires indépendantes gaussiennes est une variable aléatoire gaussienne, nous avons finalement

$$\sum_{i=1}^{M}\sqrt{\mu_i}\frac{1}{\sqrt{n_i}}\sum_{k=1}^{n_i}\nabla_\theta \log f_i(X_{ik};\theta)|_{\theta^0} \xrightarrow[n\to+\infty]{d} N(0, I(\theta^0)). \tag{3.4}$$

D'après les hypothèses 3.3.6 et 3.3.7, la consistance de l'estimateur du maximum de vraisemblance et le théorème de Slutsky, nous avons pour tout $(j,p) \in \{1,\dots,r\}^2$,

$$a_{jpn} \xrightarrow[n\to+\infty]{P} I_{jp}(\theta^0).$$

Ces convergences en probabilité et la convergence (3.4) conduisent à

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_1 - \theta_1^0) \\ \vdots \\ \sqrt{n}(\widehat{\theta}_r - \theta_r^0) \end{pmatrix} \xrightarrow[n\to+\infty]{d} N(0, \left[I(\theta^0)\right]^{-1}).$$

Remarque 3.3.6. Dans le cas où toutes les densités de probabilité ne sont pas observées dans l'échantillon (cf. remarque 3.3.2), l'adaptation de la démonstration repose sur le résultat (3.3) et sur la proposition suivante :

Proposition 3.3.1. *Soit* $(U_i)_{1\leqslant i\leqslant M}$ *et* $(L_i)_{1\leqslant i\leqslant M}$ *deux suites de variables aléatoires indépendantes. Pour tout* $i = 1, \ldots, M$, *soit* $v_i(n)$ *une suite quelconque et* l_i *un réel. Sous les hypothèses* $m(n) \underset{n\to+\infty}{\longrightarrow} M$, $v_i(n) \underset{n\to+\infty}{\longrightarrow} l_i$ *et* $U_i \underset{n\to+\infty}{\xrightarrow{d}} L_i$, *on a*

$$\sum_{i=1}^{m(n)} v_i(n)U_i \underset{n\to+\infty}{\xrightarrow{d}} \sum_{i=1}^{M} l_i L_i.$$

□

3.4 Application au délai de survenue d'un effet indésirable médicamenteux

Les délais jusqu'à la survenue d'un effet indésirable médicamenteux que nous observons, $(x_i)_{1\leqslant i\leqslant n}$, constituent un échantillon de réalisations indépendantes des variables aléatoires de distribution respective la distribution conditionnelle de X_i sachant $\{X_i \leqslant t_i\}$, de fonction de répartition $F(\cdot;\theta)/F(t_i;\theta)$ et de densité de probabilité $f(\cdot;\theta)/F(t_i;\theta)$. Dans ce cas-là, le paramètre θ est partagé par toutes les densités de probabilité et celles-ci appartiennent toutes à la même famille de distributions : la famille des distributions tronquées de X.

Soit $\{u_q\}_{1\leqslant q\leqslant M}$ l'ensemble des M délais de troncature distincts qui sont observables dans notre population considérée (la population des individus exposés au médicament D qui ont présenté ou présenteront l'effet indésirable E_D avant de décéder). L'ensemble des délais de troncature observés $\{t_i\}_{1\leqslant i\leqslant n}$ est inclus dans l'ensemble $\{u_q\}_{1\leqslant q\leqslant M}$. Sans perte de généralité, nous supposons que l'ensemble $\{u_q\}_{1\leqslant q\leqslant m}$ avec $m < M$ coïncide avec l'ensemble des délais de troncature observés $\{t_i\}_{1\leqslant i\leqslant n}$ (après suppression des doublons).

Pour tout $q = 1, \ldots, M$, soient $f_q(\cdot;\theta) = f(\cdot;\theta)/F(u_q;\theta)$ la densité de probabilité associée au délai de troncature u_q, de support S_q indépendant de θ. Pour tout $q = 1, \ldots, M$, soit n_q le nombre de réalisations de densité de probabilité $f_q(\cdot;\theta)$. Pour $q = m+1, \ldots, M$, on a $n_q = 0$. On a $n = \sum_{q=1}^{m} n_q$. Pour tout $q = 1, \ldots, M$, soit $\mu_q = n_q/n$ la

proportion d'observations de densité de probabilité $f_q(\cdot;\theta)$. Pour tout $q = m+1, \dots M$, on a $\mu_q = 0$. Pour tout $q = 1, \dots, M$, soit λ_q la limite de la suite μ_q quand n tends vers l'infini. Ces quantités vérifient la relation $\sum_{q=1}^{M} \lambda_q = 1$. Pour tout $q = 1, \dots, m$, soit $(x_{q1}, x_{q2}, \dots, x_{qk}, \dots, x_{qn_q})$ le sous-échantillon de réalisations des n_q variables aléatoires $(X_{q1}, X_{q2}, \dots, X_{qk}, \dots, X_{qn_q})$ indépendantes et identiquement distribuées de densité de probabilité $f_q(\cdot;\theta)$. L'échantillon observé peut maintenant s'écrire $(x_{qk})_{1 \leqslant q \leqslant m, 1 \leqslant k \leqslant n_q}$.

Si l'ensemble des densités de probabilités $\{f_q(\cdot;\theta)\}_{1 \leqslant q \leqslant M}$ vérifient les hypothèses 3.3.1-3.3.7, alors l'estimateur paramétrique du maximum de vraisemblance $\widehat{\theta}_n$ est un estimateur consistant de θ^0 et le vecteur $\sqrt{n}\left(\widehat{\theta}_n - \theta^0\right)$ est asymptotiquement normal d'espérance nulle et de matrice de covariance $\left[I(\theta^0)\right]^{-1} = \left[\left(\left[I\left(\theta^0\right)\right]_{jp}\right)_{1 \leqslant j,p \leqslant r}\right]^{-1}$ où

$$\left[I\left(\theta^0\right)\right]_{jp} = \sum_{q=1}^{M} \lambda_q E_{\theta^0}\left(-\left.\frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\right|_{\theta^0}\right). \tag{3.5}$$

3.5 Conclusion

Des conditions suffisantes pour que l'estimateur paramétrique du maximum de vraisemblance soit consistant et asymptotiquement normal ont été énoncées. Mais quelles sont les familles de lois qui vérifient ces conditions suffisantes ? Plus particulièrement, est-ce que les trois familles de lois que l'on considère depuis le début de cet exposé vérifient ces hypothèses ?

Dans le chapitre suivant, nous vérifierons si les hypothèses 3.3.1-3.3.7 sont satisfaites quand la variable aléatoire X caractérisant le délai jusqu'à la survenue de l'effet indésirable médicamenteux suit respectivement une loi exponentielle, de Weibull ou log-logistique. Dans ces cas-là, les densités de probabilité $(f_q(\cdot;\theta))_{1 \leqslant q \leqslant M}$ appartiennent respectivement aux familles des lois exponentielle tronquées, de Weibull tronquées ou log-logistique tronquées, celles-ci étant tronquées par les réels $\{u_q\}_{1 \leqslant q \leqslant M}$.

Chapitre 4

Cas des lois exponentielle, de Weibull ou log-logistique

4.1 Cas de la loi exponentielle

Dans ce cas, pour tout $i = 1, \ldots, n$, la distribution de l'observation x_i est la loi exponentielle tronquée au temps t_i, de densité de probabilité

$$f_{t_i}(x; \theta_1) = \frac{\theta_1 \exp(-\theta_1 x)}{(1 - \exp(-\theta_1 t_i))} I\{0 < x \leqslant t_i\},$$

où θ_1 est un paramètre inconnu strictement positif.

Pour que l'estimateur paramétrique du maximum de vraisemblance soit consistant et asymptotiquement normal, il faut que l'ensemble des lois exponentielle tronquées $\{f_q(\cdot; \theta_1)\}_{1 \leqslant q \leqslant M}$, où

$$f_q(x; \theta_1) = \frac{\theta_1 \exp(-\theta_1 x)}{(1 - \exp(-\theta_1 u_q))} I\{0 < x \leqslant u_q\},$$

vérifient les hypothèses 3.3.1-3.3.7. Pour tout $q = 1, \ldots, M$, le support S_q de la densité de probabilité $f_q(\cdot; \theta_1)$ est l'intervalle semi-ouvert $]0, u_q]$, indépendant de θ_1.

4.1.1 Hypothèses 3.3.1 et 3.3.2

La vraisemblance de l'échantillon s'écrit

$$L(\, x_{11}, x_{12}, \ldots, x_{mn_m}; \theta_1 |\, u_1, u_2, \ldots, u_m, n) = \prod_{q=1}^{m} \prod_{k=1}^{n_q} \frac{\theta_1 \exp(-\theta_1 x_{qk})}{1 - \exp(-\theta_1 u_q)}.$$

Nous considérons cette vraisemblance comme une fonction de θ_1, définie sur $\mathbb{R}^{+*}$. Pour simplifier les notations, nous la notons $L(\theta_1)$ pour la suite. La log-vraisemblance, fonction de θ_1 définie sur $\mathbb{R}^{+*}$, s'écrit

$$\log L(\theta_1) = \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\log \theta_1 - \theta_1 x_{qk} - \log\left(1 - \exp(-\theta_1 u_q)\right)\right). \tag{4.1}$$

La fonction (4.1) est continue et dérivable sur $\mathbb{R}^{+*}$ de dérivée,

$$\frac{\partial \log L(\theta_1)}{\partial \theta_1} = \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\frac{1}{\theta_1} - x_{qk} - \frac{u_q \exp(-\theta_1 u_q)}{1 - \exp(-\theta_1 u_q)} \right). \tag{4.2}$$

La fonction (4.2) s'annule-t-elle en un unique point de $\mathbb{R}^{+*}$? La fonction (4.2) est continue et dérivable sur $\mathbb{R}^{+*}$ de dérivée,

$$\begin{aligned}
\frac{\partial^2 \log L(\theta_1)}{\partial \theta_1^2} &= \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(-\frac{1}{\theta_1^2} + \frac{u_q^2 \exp(-\theta_1 u_q)(1 - \exp(-\theta_1 u_q))}{\left(1 - \exp(-\theta_1 u_q)\right)^2} \right. \\
&\qquad \left. + \frac{u_q \exp(-\theta_1 u_q)(u_q \exp(-\theta_1 u_q))}{\left(1 - \exp(-\theta_1 u_q)\right)^2} \right) \\
&= \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(-\frac{1}{\theta_1^2} + \frac{u_q^2 \exp(-\theta_1 u_q)}{\left(1 - \exp(-\theta_1 u_q)\right)^2} \right) \\
&= \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\frac{u_q^2 \exp(-\theta_1 u_q)}{\left(1 - \exp(-\theta_1 u_q)\right)^2} - \frac{1}{\theta_1^2} \right) \\
&= \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\frac{\theta_1^2 u_q^2 \exp(-\theta_1 u_q) - 1 - \exp(-2\theta_1 u_q) + 2\exp(-\theta_1 u_q)}{\theta_1^2 \left(1 - \exp(-\theta_1 u_q)\right)^2} \right).
\end{aligned} \tag{4.3}$$

Nous voulons étudier le signe de la fonction (4.3). Le dénominateur de chaque membre

de cette double somme étant strictement positif, étudions le signe du numérateur. Pour tout $q = 1, \ldots, M$, on définit la fonction y_q et sa dérivée y'_q sur $\mathbb{R}^{+*}$ par :

$$\begin{cases} y_q(\theta_1) = \theta_1^2 u_q^2 \exp(-\theta_1 u_q) - 1 - \exp(-2\theta_1 u_q) + 2\exp(-\theta_1 u_q), \\ y'_q(\theta_1) = \exp(-\theta_1 u_q)(2\theta_1 u_q^2 - \theta_1^2 u_q^3 + 2u_q \exp(-\theta_1 u_q) - 2u_q). \end{cases}$$

Pour tout $q = 1, \ldots, M$, étant donné que le terme $\exp(-\theta_1 u_q)$ est strictement positif, le signe de la fonction y'_q est celui de la fonction h_q définie sur $\mathbb{R}^{+*}$ par

$$h_q(\theta_1) = 2\theta_1 u_q^2 - \theta_1^2 u_q^3 + 2u_q \exp(-\theta_1 u_q) - 2u_q.$$

Pour tout $q = 1, \ldots, M$, la fonction h_q est dérivable deux fois sur $\mathbb{R}^{+*}$, de dérivées :

$$\begin{cases} h'_q(\theta_1) = 2u_q^2 - 2\theta_1 u_q^3 - 2u_q^2 \exp(-\theta_1 u_q), \\ h''_q(\theta_1) = 2u_q^3(\exp(-\theta_1 u_q) - 1). \end{cases}$$

Une étude de signe de la fonction h''_q donne :

$$\begin{cases} h''_q(\theta_1) = 0 \Longleftrightarrow \theta_1 = 0, \\ h''_q(\theta_1) > 0 \Longleftrightarrow \theta_1 < 0, \\ h''_q(\theta_1) < 0 \Longleftrightarrow \theta_1 > 0. \end{cases}$$

La fonction h''_q étant définie sur $\mathbb{R}^{+*}$, la fonction h'_q est une fonction strictement décroissante sur $\mathbb{R}^{+*}$ pour tout $q = 1, \ldots, M$. Une étude des limites en zéro et en l'infini montre que

$$\begin{cases} \lim\limits_{\theta_1 \to +\infty} h'_q(\theta_1) = -\infty, \\ \lim\limits_{\theta_1 \to 0} h'_q(\theta_1) = 0. \end{cases}$$

Pour tout $q = 1, \ldots, M$, la fonction h'_q est strictement négative sur $\mathbb{R}^{+*}$. La fonction h_q est donc une fonction strictement décroissante sur $\mathbb{R}^{+*}$. Une étude des limites en zéro et en l'infini montre que

$$\begin{cases} \lim\limits_{\theta_1 \to +\infty} h_q(\theta_1) = -\infty, \\ \lim\limits_{\theta_1 \to 0} h_q(\theta_1) = 0. \end{cases}$$

Par conséquent, pour tout $q = 1, \ldots, M$, la fonction h_q est strictement négative sur $\mathbb{R}^{+*}$. Il en va de même de la fonction y'_q. La fonction y_q est donc une fonction strictement décroissante sur $\mathbb{R}^{+*}$. Une étude des limites en zéro et en l'infini montre que

$$\begin{cases} \lim\limits_{\theta_1 \to +\infty} y_q(\theta_1) = -1, \\ \lim\limits_{\theta_1 \to 0} y_q(\theta_1) = 0. \end{cases}$$

La fonction y_q est donc strictement négative sur $\mathbb{R}^{+*}$. Ainsi, la fonction (4.3) est une fonction strictement négative sur $\mathbb{R}^{+*}$ et la fonction (4.2) est strictement décroissante sur $\mathbb{R}^{+*}$. Une étude de limite en l'infini et l'utilisation d'un développement limité en zéro montrent que

$$\begin{cases} \lim\limits_{\theta_1 \to +\infty} \dfrac{\partial \log L(\theta_1)}{\partial \theta_1} = -\sum\limits_{q=1}^{m} \sum\limits_{k=1}^{n_q} x_{qk} \ < 0, \\ \lim\limits_{\theta_1 \to 0} \dfrac{\partial \log L(\theta_1)}{\partial \theta_1} = \sum\limits_{q=1}^{m} \sum\limits_{k=1}^{n_q} \left(\dfrac{u_q}{2} - x_{qk} \right). \end{cases}$$

Finalement,

- Si $\sum\limits_{q=1}^{m} \sum\limits_{k=1}^{n_q} \left(\dfrac{u_q}{2} - x_{qk} \right) \leqslant 0$, alors les équations normales n'ont pas de solution et l'estimateur du maximum de vraisemblance n'existe pas.
- Si $\sum\limits_{q=1}^{m} \sum\limits_{k=1}^{n_q} \left(\dfrac{u_q}{2} - x_{qk} \right) > 0$, alors les équations normales ont une unique solution, l'estimateur du maximum de vraisemblance existe et est unique.

On définit

$$Q_1 = \sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\frac{u_q}{2} - x_{qk} \right). \tag{4.4}$$

La condition $Q_1 > 0$ est une condition d'existence du maximum de vraisemblance pour la loi exponentielle tronquée. Cette condition avait déjà été mentionnée par Bartlett

(1953) dans l'étude des particules instables en chambre noire.

4.1.2 Hypothèse 3.3.3

Soit f_q la fonction de deux variables définie par :

$$\begin{aligned} f_q : \ [0,u_q] \times \mathbb{R}^{+*} &\longrightarrow \mathbb{R}^{+*} \\ (x,\theta_1) &\longmapsto f_q(x,\theta_1) = \frac{\theta_1 \exp(-\theta_1 x)}{1-\exp(-\theta_1 u_q)}. \end{aligned} \tag{4.5}$$

Pour tout x appartenant à l'intervalle $[0,u_q]$, le logarithme de la fonction (4.5) est dérivable par rapport à θ_1 sur $\mathbb{R}^{+*}$, de dérivée

$$\frac{\partial \log f_q(x,\theta_1)}{\partial \theta_1} = \frac{1}{\theta_1} - x - \frac{u_q \exp(-\theta_1 u_q)}{1-\exp(-\theta_1 u_q)}. \tag{4.6}$$

Pour tout x appartenant à l'intervalle $[0,u_q]$, la fonction (4.6) est dérivable par rapport à θ_1 sur $\mathbb{R}^{+*}$, de dérivée

$$\frac{\partial^2 \log f_q(x,\theta_1)}{\partial \theta_1^2} = -\frac{1}{\theta_1^2} + \frac{u_q^2 \exp(-\theta_1 u_q)}{(1-\exp(-\theta_1 u_q))^2}. \tag{4.7}$$

Pour tout x appartenant à l'intervalle $[0,u_q]$, la fonction (4.7) est dérivable par rapport à θ_1 sur $\mathbb{R}^{+*}$, de dérivée

$$\frac{\partial^3 \log f_q(x,\theta_1)}{\partial \theta_1^3} = \frac{2}{\theta_1^3} + \frac{u_q^3 \exp(-\theta_1 u_q)(\exp(-2\theta_1 u_q)-1)}{(1-\exp(-\theta_1 u_q))^4}.$$

4.1.3 Hypothèse 3.3.4

D'après les théorèmes classiques, la fonction (4.5) est une fonction continue sur $[0,u_q] \times \mathbb{R}^{+*}$. La fonction φ_q de deux variables définie par

$$\begin{aligned} \varphi_q : \ [0,u_q] \times \mathbb{R}^{+*} &\longrightarrow \mathbb{R} \\ (x,\theta_1) &\longmapsto \frac{\partial f_q}{\partial \theta_1}(x,\theta_1), \end{aligned} \tag{4.8}$$

existe et est continue sur $[0, u_q] \times \mathbb{R}^{+*}$. Par conséquent, la fonction de θ_1 définie sur $\mathbb{R}^{+*}$ par

$$\theta_1 \longmapsto \int_{[0,u_q]} f_q(x, \theta_1) dx,$$

est dérivable sur $\mathbb{R}^{+*}$ et vérifie

$$\frac{\partial}{\partial \theta_1} \int_{[0,u_q]} f_q(x, \theta_1) dx = \int_{[0,u_q]} \frac{\partial}{\partial \theta_1} f_q(x, \theta_1) dx.$$

4.1.4 Hypothèse 3.3.5

D'après les théorèmes classiques, la fonction (4.8) est continue sur $[0, u_q] \times \mathbb{R}^{+*}$. La fonction de deux variables définie par

$$\begin{aligned} [0, u_q] \times \mathbb{R}^{+*} &\longrightarrow \mathbb{R} \\ (x, \theta_1) &\longmapsto \frac{\partial \varphi_q}{\partial \theta_1}(x, \theta_1), \end{aligned} \tag{4.9}$$

existe et est continue sur $[0, u_q] \times \mathbb{R}^{+*}$. Ainsi, la fonction définie sur $\mathbb{R}^{+*}$ par

$$\theta_1 \longmapsto \int_{[0,u_q]} \varphi_q(x, \theta_1) dx,$$

est dérivable sur $\mathbb{R}^{+*}$ et vérifie

$$\frac{\partial}{\partial \theta_1} \int_{[0,u_q]} \varphi_q(x, \theta_1) dx = \int_{[0,u_q]} \frac{\partial}{\partial \theta_1} \varphi_q(x, \theta_1) dx.$$

4.1.5 Hypothèse 3.3.6

La dérivée troisième du logarithme de la fonction f_q, d'expression

$$\begin{aligned}\frac{\partial^3 \log f_q(x,\theta_1)}{\partial \theta_1^3} &= \frac{2}{\theta_1^3} + \frac{u_q^3 \exp(-\theta_1 u_q)(\exp(-2\theta_1 u_q)-1)}{(1-\exp(-\theta_1 u_q))^4} \\ &= \frac{2(1-\exp(-\theta_1 u_q))^4 + \theta_1^3 u_q^3 \exp(-\theta_1 u_q)(\exp(-2\theta_1 u_q)-1)}{\theta_1^3(1-\exp(-\theta_1 u_q))^4} \\ &= \frac{2(1-\exp(-\theta_1 u_q))^4 + \theta_1^3 u_q^3 \exp(-\theta_1 u_q)(\exp(-\theta_1 u_q)-1)(\exp(-\theta_1 u_q)+1)}{\theta_1^3(1-\exp(-\theta_1 u_q))^4},\end{aligned}$$

est indépendante de x. L'espérance de cette dérivée troisième existe pour tout $q = 1, \ldots, M$, et on cherche donc une constante K telle que

$$\left| \sum_{q=1}^{M} \lambda_q \left(\frac{\partial^3 \log f_q(x,\theta_1)}{\partial \theta_1^3} \right) \right| < K.$$

La dérivée troisième, en tant que fonction de θ_1, est continue sur $\mathbb{R}^{+*}$. De plus, une étude de limite en l'infini montre que

$$\lim_{\theta_1 \to +\infty} \frac{\partial^3 \log f_q(x,\theta_1)}{\partial \theta_1^3} = 0.$$

Un développement limité à l'ordre 8 montre que

$$2(1-\exp(-\theta_1 u_q))^4 + \theta_1^3 u_q^3 \exp(-\theta_1 u_q)(\exp(-\theta_1 u_q)-1)(\exp(-\theta_1 u_q)+1) \underset{0}{\sim} \frac{\theta_1^8 u_q^8}{120},$$

et un autre développement limité montre que

$$\theta_1^3(1-\exp(-\theta_1 u_q))^4 \underset{0}{\sim} \theta_1^7 u_q^4.$$

Ainsi, on a

$$\frac{\partial^3 f_q(x,\theta_1)}{\partial \theta_1^3} \log \underset{0}{\sim} \frac{\theta_1 u_q^4}{120},$$

et la dérivée troisième admet donc une limite nulle en zéro. Par conséquent, cette

fonction, en tant que fonction de θ_1, est bornée. On peut donc trouver une constante K_i, indépendante de θ_1, telle que

$$\left|\frac{\partial^3}{\partial\theta_1^3}\log f_q(x,\theta_1)\right| < K_i.$$

En prenant $K = \sum_{q=1}^{M} K_i$, on obtient

$$\left|\sum_{q=1}^{M}\lambda_q\frac{\partial^3\log f_q(x,\theta_1)}{\partial\theta_1^3}\right| \leqslant \sum_{q=1}^{M}\lambda_q\left|\frac{\partial^3\log f_q(x,\theta_1)}{\partial\theta_1^3}\right| < K.$$

4.1.6 Hypothèse 3.3.7

Le calcul de l'espérance de la dérivée seconde du logarithme de la fonction f_q donne pour tout $q = 1,\ldots,M$,

$$\mathrm{E}_\theta\left(-\frac{\partial^2\log f_q(X_{q1},\theta_1)}{\partial\theta_1^2}\bigg|_{\theta_1}\right) = \frac{1}{(\theta_1)^2} - \frac{u_q^2\exp(-\theta_1 u_q)}{(1-\exp(-\theta_1 u_q))^2}.$$

Etant donné que le paramètre θ_1 est un scalaire, la matrice $I(\theta_1)$ est une constante. On a

$$\begin{aligned} I(\theta_1) &= \sum_{q=1}^{M}\lambda_q\mathrm{E}_\theta\left(-\frac{\partial^2\log f_q(X_{q1},\theta_1)}{\partial\theta_1^2}\right) \\ &= \sum_{q=1}^{M}\lambda_q\left(\frac{1}{\theta_1^2} - \frac{u_q^2\exp(-\theta_1 u_q)}{(1-\exp(-\theta_1 u_q))^2}\right). \end{aligned}$$

Nous avons déjà étudié chaque membre de cette somme dans la section 4.1.1. Chaque élément de cette somme étant strictement positif, la matrice $I(\theta_1)$ est définie positive pour tout θ_1 appartenant à $\mathbb{R}^{+*}$.

4.1.7 Conclusion

En supposant que la condition $Q_1 > 0$ soit vérifiée, l'estimateur du maximum de vraisemblance du paramètre de la loi exponentielle tronquée est un estimateur consistant

de θ^0 et le vecteur $\sqrt{n}(\widehat{\theta}_n - \theta^0)$ est asymptotiquement normal d'espérance nulle et de matrice de covariance

$$\left[I(\theta_1^0)\right]^{-1} = \left[\sum_{q=1}^{M} \lambda_q \left(\frac{1}{(\theta_1^0)^2} - \frac{u_q^2 \exp(-\theta_1^0 u_q)}{(1-\exp(-\theta_1^0 u_q))^2}\right)\right]^{-1}.$$

4.2 Cas des lois de Weibull et log-logistique

Si la variable aléatoire X suit une loi de Weibull, la distribution de l'observation x_i, pour tout $i = 1, \ldots, n$, est la loi de Weibull tronquée au temps t_i, de densité de probabilité

$$f_{t_i}(x; (\theta_1, \theta_2)) = \frac{\theta_1\theta_2(\theta_1 x)^{\theta_2 - 1}\exp(-(\theta_1 x)^{\theta_2})}{(1-\exp(-(\theta_1 t_i)^{\theta_2}))} I(0 < x \leqslant t_i),$$

où $\theta = (\theta_1, \theta_2)$ est un paramètre de dimension 2 inconnu dont les deux composantes sont strictement positives. Pour que l'estimateur paramétrique du maximum de vraisemblance soit consistant et asymptotiquement normal, il faut que l'ensemble des lois de Weibull tronquées $\{f_q(\cdot;\theta)\}_{1 \leqslant q \leqslant M}$, où

$$f_q(x;\theta) = \frac{\theta_1\theta_2(\theta_1 x)^{\theta_2 - 1} exp(-(\theta_1 x)^{\theta_2})}{(1-\exp(-(\theta_1 u_q)^{\theta_2}))} I(0 < x \leqslant u_q),$$

vérifient les hypothèses 3.3.1-3.3.7. Pour tout $q = 1, \ldots, M$, le support S_q de la densité de probabilité $f_q(\cdot;\theta)$ est l'intervalle semi-ouvert $]0, u_q]$, indépendant de θ.

De la même manière, si la variable aléatoire X suit une loi log-logistique, la distribution de l'observation x_i, pour tout $i = 1, \ldots, n$, est la loi log-logistique tronquée au temps t_i, de densité de probabilité

$$f_q(x;\theta) = \frac{\theta_2 x^{\theta_2 - 1}}{t_i^{\theta_2}} \frac{1 + (\theta_1 t_i)^{\theta_2}}{(1 + (\theta_1 x)^{\theta_2})^2} I(0 < x \leqslant t_i),$$

où $\theta = (\theta_1, \theta_2)$ est un paramètre de dimension 2 inconnu dont les deux composantes sont strictement positives. Pour que l'estimateur paramétrique du maximum de vrai-

semblance soit consistant et asymptotiquement normal, il faut que l'ensemble des lois log-logistique tronquées $\{f_q(\cdot;\theta)\}_{1\leqslant q\leqslant M}$, où

$$f_q(x;\theta) = \frac{\theta_2 x^{\theta_2-1}}{u_q^{\theta_2}} \frac{1+(\theta_1 u_q)^{\theta_2}}{(1+(\theta_1 x)^{\theta_2})^2} I(0 < x \leqslant u_q),$$

vérifient les hypothèses 3.3.1-3.3.7. Pour tout $q = 1, \ldots, M$, le support S_q de la densité de probabilité $f_q(\cdot;\theta)$ est l'intervalle semi-ouvert $]0, u_q]$, indépendant de θ.

Montrer que les hypothèses 3.3.1-3.3.7 sont satisfaites pour ces deux familles de distributions semble impossible. Cependant, comme pour la loi exponentielle tronquée, nous avons établi une conjecture sur l'existence de l'estimation paramétrique du maximum de vraisemblance. Cette condition, énoncée ci-dessous, serait la même pour les lois de Weibull tronquées et pour les lois log-logistique tronquées.

Considérons les fonctions h et Q_2 définies par :

$$\begin{cases} h : \theta_1 \longmapsto \widehat{\theta}_2(\theta_1) \text{ tel que } \dfrac{\partial}{\partial\theta_2} \log L\left(x_1, \ldots, x_n; (\theta_1, \widehat{\theta}_2(\theta_1)) \middle| u_1, \ldots, u_m, n\right) = 0, \\ Q_2 : \theta_2 \longmapsto Q_2(\theta_2) = \displaystyle\sum_{q=1}^{m} \sum_{k=1}^{n_q} \left(\frac{{u_q}^{\theta_2}}{2} - x_{qk}^{\theta_2} \right). \end{cases}$$

Nous conjecturons que si $Q_2(\lim\limits_{\theta_1 \to 0} h(\theta_1)) > 0$, alors l'estimation du maximum de vraisemblance existe et est solution des équations normales. Contrairement à la loi exponentielle tronquée, un travail supplémentaire est nécessaire pour prouver théoriquement cette condition. Mais l'étude de simulations présentée dans le chapitre 5 viendra renforcer cette conjecture.

En supposant que les conditions 3.3.1-3.3.7 soient satisfaites (ce qui implique que la condition ci-dessus soit vérifiée), l'estimateur du maximum de vraisemblance du paramètre de la loi de Weibull tronquée (ou de la loi log-logistique tronquée) est un estimateur consistant de θ^0 et le vecteur $\sqrt{n}(\widehat{\theta}_n - \theta^0)$ est asymptotiquement normal d'espérance nulle et de matrice de covariance (3.5).

L'expression formelle de la matrice de covariance asymptotique pour les distribu-

tions de Weibull et log-logistique n'est pas donnée car contrairement à la distribution exponentielle, certaines des espérances de cette matrice ne se calculent pas explicitement.

Chapitre 5

Estimation par intervalle du paramètre

L'établissement des propriétés asymptotiques précédentes permet d'associer à l'estimation du paramètre de la distribution obtenue un intervalle de confiance asymptotique de type Wald. Il est donc légitime de se poser la question de la qualité de l'estimation par intervalle de l'estimateur paramétrique du maximum de vraisemblance. Une seconde étude de simulations est mise en œuvre. Des problèmes de maximisation de la vraisemblance empêchent le calcul de l'intervalle de confiance. Nous faisons le lien entre ces échecs et la condition d'existence du maximum de vraisemblance que nous avons énoncée au chapitre 4. Par ailleurs, certaines estimations des probabilités de couverture des intervalles de confiance de type Wald sont éloignées du niveau attendu. Nous explorons le biais, la normalité et la variance de l'estimateur, qui sont trois sources potentielles d'estimations non satisfaisantes. A la vue des résultats de cette étude de simulations, un autre intervalle de confiance est exploré. Les propriétés asymptotiques établies précédemment permettent de vérifier que le rapport de vraisemblance suit toujours asymptotiquement une distribution du chi-2 quand les données sont tronquées à droite. Nous calculons donc les intervalles de confiance fondés sur le rapport de vraisemblance profilée. Nous avons choisi de présenter séquentiellement les résultats pour

l'intervalle de confiance de Wald et l'intervalle de confiance de la vraisemblance profilée, mais ces résultats ont été obtenus à partir de la même étude de simulations.

5.1 Intervalle de confiance de type Wald

Des régions de confiance pour le paramètre vectoriel inconnu θ^0 et des intervalles de confiance pour chaque composante du paramètre θ^0 peuvent être calculés à partir de la distribution asymptotique de l'estimateur paramétrique du maximum de vraisemblance. L'intervalle de confiance à $100(1-\alpha)\%$ de type Wald pour la $j^{\text{ième}}$ composante du paramètre vectoriel θ^0 est :

$$IC_{100(1-\alpha)\%}(\theta_j^0) = \left[\widehat{\theta}_{jn} \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\left[\widehat{I(\theta^0)}^{-1} \right]_{jj}} \right],$$

où $z_{1-\alpha/2}$ est le quantile à $100(1-\alpha/2)\%$ de la loi normale centrée réduite et $\left[\widehat{I(\theta^0)}^{-1} \right]_{jj}$ est le $j^{\text{ième}}$ élément de la diagonale de l'estimation $\left[\widehat{I(\theta^0)} \right]^{-1}$ de la matrice de covariance asymptotique $[I(\theta^0)]^{-1}$.

Dans l'expression (3.5), en remplaçant θ^0 par $\widehat{\theta}_n$ et les quantités λ_q par n_q/n pour tout $q = 1, \ldots, m$ et par 0 pour tout $q = m+1, \ldots, M$, on estime $[I(\theta^0)]_{jp}$, pour tout $1 \leqslant j, p \leqslant r$, par

$$\sum_{q=1}^{m} \frac{n_q}{n} E_{\widehat{\theta}_n} \left(- \frac{\partial^2 \log\left(f_q(X_{q1};\theta)\right)}{\partial\theta_j \partial\theta_p} \bigg|_{\widehat{\theta}_n} \right). \tag{5.1}$$

Mais en pratique l'espérance E_θ est estimée et la composante de rang (j,p) de la matrice $I(\theta^0)$ est approximée par :

$$\left[\widehat{I(\theta^0)} \right]_{jp} = \frac{1}{n} \sum_{i=1}^{n} \left(- \frac{\partial^2 \log\left(f_{t_i}(x_i;\theta)\right)}{\partial\theta_j \partial\theta_p} \bigg|_{\widehat{\theta}_n} \right). \tag{5.2}$$

Efron et Hinkley (1978) ont montré que, quand l'espérance E_θ à une forme explicite, il valait mieux utiliser l'estimation (5.2) que l'estimation (5.1).

5.2 Qualité de l'estimation par intervalle de type Wald

5.2.1 Design

Une étude de simulations a été mise en œuvre à l'aide du logiciel R (R Core Team, 2012) pour évaluer la qualité de l'estimation par intervalle, en termes de biais et de probabilités de couverture, à taille d'échantillon finie et asymptotiquement. Le biais a été étudié à cause de son influence sur les probabilités de couverture.

Les délais de survenue ont été générés à partir des trois modèles de distribution considérés précédemment : exponentiel, Weibull et log-logistique. Deux valeurs du paramètre θ_1 ont été considérées pour la distribution exponentielle : 0.006 et 0.05. Les mêmes valeurs ont été utilisées pour le paramètre d'échelle θ_1 des lois de Weibull et log-logistique. Pour le paramètre de forme θ_2, les valeurs 0.5 et 2 ont été choisies. Les fonctions de base *rexp*, *rweibull* et la fonction *rllogis* du package **eha** ont été utilisées pour la génération des délais de survenue. Les délais de troncature ont été générés indépendamment des délais de survenue et selon la loi uniforme discrète de support $\mathscr{S}(\tau) = \{0.1, 0.2, 0.3, \ldots, \tau\}$ où τ est un multiple de 0.1. Le nombre décimal τ a été choisi de manière à obtenir la proportion souhaitée p de données tronquées à droite. Alors p et τ sont liés par la relation suivante : $p = (1/\#\mathscr{S}(\tau)) \sum_{u \in \mathscr{S}(\tau)} (1 - F)(u)$, où $1 - F$ est la fonction de survie du délai de survenue. Etant donné que τ est un nombre décimal, cette équation peut ne pas avoir de solution. Dans ce cas, le plus petit nombre décimal τ tel que l'ensemble $\mathscr{S}(\tau) = \{0.1, 0.2, 0.3, \ldots, \tau\}$ vérifie l'inéquation $(1/\#\mathscr{S}(\tau)) \sum_{u \in \mathscr{S}(\tau)} (1 - F)(u) < p$ a été choisi. Les valeurs $\{0.25, 0.50, 0.75\}$ ont été considérées pour la proportion p de données tronquées à droite. Les délais de troncature ont été générés à l'aide de la fonction *sample* du logiciel R. Pour chaque réalisation du vecteur aléatoire (X, T), si le délai de survenue était inférieur au délai de troncature, alors la réalisation était incluse dans la base. Sinon, une autre réalisation du vecteur aléatoire (X, T) était simulée. Des réalisations du vecteur aléatoire (X, T) ont

été générées jusqu'à ce que la taille d'échantillon atteigne la valeur fixée n. Les valeurs $\{100, 200, 500, 1000\}$ ont été considérées pour la taille d'échantillon n. Pour chaque taille d'échantillon n, le tableau 5.1 présente les différents scénarios obtenus en croisant les différents paramètres des simulations.

Pour chaque scénario, 1000 réplications ont été simulées. A partir des résultats de ces 1000 réplications, les estimations du biais et des probabilités de couverture des intervalles de confiance à 95% de type Wald ont été calculées pour chaque composante du paramètre vectoriel θ.

La fonction *maxLik* du package **maxLik** du logiciel R a été utilisée pour la maximisation de la vraisemblance. Nous avons fait appel à l'algorithme de Newton-Raphson pour la distribution exponentielle et à l'algorithme de Nelder-Mead pour les lois de Weibull et log-logistique.

Remarque 5.2.1. La variable aléatoire T avait été simulée selon une loi uniforme continue dans la première étude de simulations. Mais l'examen des propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance nous a conduit à faire l'hypothèse qu'il y avait un nombre fini de délais de troncature distincts dans la population. Nous avons donc changé le processus de simulation et choisi de générer la variable aléatoire T selon une loi uniforme discrète. Dans la première étude de simulations, la probabilité κ qu'une réalisation de X appartienne à l'intervalle $[0, \tau]$ des valeurs observables de X avait été choisie comme paramètre des simulations. Pour cette deuxième étude de simulations, nous avons préféré considérer la proportion p de données tronquées.

5.2.2 Résultats

5.2.2.1 Biais et probabilités de couverture

Quelle que soit la famille de lois, la maximisation a échoué pour certaines réplications. Par conséquent, les estimations du biais et des probabilités de couverture ont été calculées sur l'ensemble des réplications où il n'y a pas eu de problème de maximisation.

TABLEAU 5.1 – Scénarios de l'étude de simulations n° 2 : trente scénarios différents pour une taille d'échantillon, obtenus en considérant trois familles de distributions pour le délai de survenue, différentes valeurs du paramètre θ de la distribution et trois valeurs de la proportion p de données tronquées à droite. Les délais de troncature ont été simulés selon une loi uniforme discrète de support $\{0.1, 0.2, \ldots, \tau\}$. Le nombre décimal τ a été choisi de manière à obtenir la proportion souhaitée p de données tronquées.

Distribution	θ	p	τ
Exponentielle	0.006	0.75	100.9
Exponentielle	0.006	0.50	265.5
Exponentielle	0.006	0.25	653.3
Exponentielle	0.05	0.75	12.1
Exponentielle	0.05	0.50	31.8
Exponentielle	0.05	0.25	78.3
Weibull	(0.006, 0.5)	0.75	32.1
Weibull	(0.006, 0.5)	0.50	198.4
Weibull	(0.006, 0.5)	0.25	898.9
Weibull	(0.05, 0.5)	0.75	3.8
Weibull	(0.05, 0.5)	0.50	23.7
Weibull	(0.05, 0.5)	0.25	107.7
Weibull	(0.006, 2)	0.75	165.2
Weibull	(0.006, 2)	0.50	291.4
Weibull	(0.006, 2)	0.25	590.7
Weibull	(0.05, 2)	0.75	19.8
Weibull	(0.05, 2)	0.50	34.9
Weibull	(0.05, 2)	0.25	70.7
Log-logistique	(0.006, 0.5)	0.75	44.5
Log-logistique	(0.006, 0.5)	0.50	439.2
Log-logistique	(0.006, 0.5)	0.25	4489.1
Log-logistique	(0.05, 0.5)	0.75	5.2
Log-logistique	(0.05, 0.5)	0.50	52.5
Log-logistique	(0.05, 0.5)	0.25	538.3
Log-logistique	(0.006, 2)	0.75	187.7
Log-logistique	(0.006, 2)	0.50	388.4
Log-logistique	(0.006, 2)	0.25	928.7
Log-logistique	(0.05, 2)	0.75	22.5
Log-logistique	(0.05, 2)	0.50	46.5
Log-logistique	(0.05, 2)	0.25	111.3

Distribution exponentielle Dans le tableau 5.2 sont rassemblés les estimations du biais, les estimations des probabilités de couverture et les nombres de problèmes de maximisation pour la distribution exponentielle.

L'estimateur paramétrique du maximum de vraisemblance a tendance à être légèrement positivement biaisé. Quelle que soit la valeur de θ_1, le biais diminue quand la taille d'échantillon augmente ou quand la proportion de données tronquées à droite diminue. Ces résultats sont cohérents avec les résultats de l'étude de simulations présentée au chapitre 2.

Quand la proportion de données tronquées est égale à 0.75, il y a une amélioration des estimations des probabilités de couverture quand la taille d'échantillon augmente. Les estimations sont plus élevées qu'attendu pour les plus petites tailles d'échantillon. Pour les autres proportions de données tronquées, les estimations sont toujours satisfaisantes.

Le nombre de problèmes de maximisation diminue quand la taille d'échantillon augmente ou quand la proportion de données tronquées à droite diminue. La proportion de problèmes de maximisation atteint 13.3% des réplications.

Distributions de Weibull et log-logistique Les résultats sont analogues pour les deux distributions. Dans le tableau 5.3 (resp. tableau 5.4) sont rassemblés les estimations du biais, les estimations des probabilités de couverture et les nombres de problèmes de maximisation pour la distribution de Weibull (resp. log-logistique).

Quel que soit le paramètre, l'estimateur du maximum de vraisemblance a tendance à être positivement biaisé. Le biais diminue quand la taille d'échantillon augmente ou quand la proportion de données tronquées à droite diminue. Le biais est peu élevé pour la composante θ_2 tandis qu'il peut parfois être inacceptablement élevé pour le paramètre θ_1, même pour de grandes tailles d'échantillon. Là encore, ces résultats sont cohérents avec les résultats de l'étude de simulations présentée au chapitre 2.

Quel que soit le scénario, les estimations des probabilités de couverture sont toujours satisfaisantes pour le paramètre θ_2.

En revanche, pour le paramètre θ_1, il y a des triplets (θ_1, θ_2, p) où les probabilités

TABLEAU 5.2 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : biais ($\times 10^4$), probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1) et nombre de problèmes de maximisation (Pbs), pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	p	n	Biais($\widehat{\theta_1}$) $\times 10^4$	$PC_1(\widehat{\theta_1})$	Pbs
0.006	0.75	100	9.492	0.98	114
		200	3.172	0.97	42
		500	0.311	0.96	4
		1000	0.760	0.94	0
	0.50	100	0.667	0.95	0
		200	0.640	0.95	0
		500	-0.100	0.95	0
		1000	0.257	0.94	0
	0.25	100	0.511	0.96	0
		200	0.221	0.95	0
		500	-0.086	0.94	0
		1000	0.103	0.95	0
0.05	0.75	100	93.73	0.98	133
		200	33.29	0.97	43
		500	9.403	0.95	5
		1000	3.560	0.94	0
	0.50	100	7.097	0.95	0
		200	3.962	0.95	0
		500	0.406	0.95	0
		1000	1.022	0.95	0
	0.25	100	8.661	0.95	0
		200	0.366	0.96	0
		500	3.829	0.95	0
		1000	0.867	0.95	0

de couverture sont toujours satisfaisantes (il s'agit des plus petites valeurs de p) mais il y a aussi des triplets (θ_1, θ_2, p) où les probabilités de couverture des intervalles de confiance de type Wald sont anormalement plus basses qu'attendu (il s'agit des plus grandes valeurs de p). Ces probabilités de couverture anormalement faibles surviennent plus souvent pour les petites tailles d'échantillons (dans ces cas-là, il y a une amélioration quand la taille d'échantillon augmente) mais il y a aussi des triplets (θ_1, θ_2, p) où les probabilités de couverture sont inacceptablement peu élevées quelle que soit la taille d'échantillon (il y a même parfois une dégradation quand la taille d'échantillon augmente).

Le nombre de problèmes de maximisation diminue quand la taille d'échantillon augmente ou quand la proportion de données tronquées à droite diminue. La proportion de problèmes de maximisation atteint 25.4% des réplications pour la distribution de Weibull et 13.4% pour la distribution log-logistique.

5.2.3 Etude approfondie

L'objectif de cette section est d'essayer de comprendre pourquoi il y a parfois un grand nombre de problèmes de maximisation ou pourquoi certaines estimations des probabilités de couverture des intervalles de confiance de type Wald pour le paramètre (de forme) sont éloignées du niveau attendu 95%. Pour le premier point, nous approfondissons la condition d'existence du maximum de vraisemblance que nous avons énoncée au chapitre 4. Pour le second point, nous explorons le biais, la normalité et la variance de l'estimateur, qui sont trois sources potentielles d'estimations non satisfaisantes.

5.2.3.1 Problèmes de maximisation

La statistique Q_1 (4.4) a été calculée pour les 1000 réplications de chaque scénario où le délai de survenue X suit une distribution exponentielle. Pour les distributions de Weibull et log-logistique, une valeur θ_{1s} (de θ_1) très proche de 0 a été choisie et, pour chaque réplication de chaque scénario, les quantités $\theta_{2s} = h(\theta_{1s})$ et $Q_2(\theta_{2s})$ ont été

TABLEAU 5.3 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : biais ($\times 10^4$), probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1) et nombre de problèmes de maximisation (Pbs), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Biais$(\widehat{\theta_1})$ $\times 10^4$	$PC_1(\widehat{\theta_1})$	Biais$(\widehat{\theta_2})$ $\times 10^4$	$PC_1(\widehat{\theta_2})$	Pbs
0.006	0.5	0.75	100	229.3	0.93	430.7	0.95	241
			200	124.0	0.92	199.5	0.96	169
			500	48.66	0.91	66.87	0.97	99
			1000	25.43	0.89	31.19	0.95	44
		0.50	100	27.42	0.88	163.7	0.97	82
			200	16.99	0.89	86.34	0.96	18
			500	6.045	0.91	29.02	0.95	0
			1000	2.254	0.93	0.021	0.95	0
		0.25	100	3.208	0.91	79.80	0.94	1
			200	2.767	0.91	27.50	0.94	0
			500	0.638	0.96	6.204	0.96	0
			1000	0.471	0.96	9.792	0.95	0
0.05	0.5	0.75	100	1928	0.94	392.1	0.97	254
			200	1128	0.91	251.6	0.94	199
			500	437.6	0.89	70.46	0.96	120
			1000	219.1	0.89	25.43	0.97	36
		0.50	100	253.3	0.87	141.1	0.96	75
			200	160.0	0.89	92.68	0.95	18
			500	54.43	0.90	23.06	0.95	0
			1000	34.74	0.94	8.302	0.96	0
		0.25	100	36.73	0.93	113.1	0.96	3
			200	8.171	0.93	30.54	0.95	0
			500	7.961	0.94	20.52	0.95	0
			1000	2.829	0.95	8.651	0.94	0
0.006	2	0.75	100	-2.813	0.85	501.6	0.93	16
			200	-3.034	0.89	230.2	0.92	6
			500	-0.961	0.94	104.4	0.95	0
			1000	-0.812	0.94	30.16	0.94	0
		0.50	100	-0.670	0.96	303.7	0.95	0
			200	-0.405	0.96	96.58	0.95	0
			500	-0.092	0.96	67.81	0.95	0
			1000	-0.101	0.95	20.96	0.94	0
		0.25	100	0.343	0.94	223.9	0.94	0
			200	0.129	0.95	170.6	0.96	0
			500	-0.050	0.94	49.35	0.96	0

TABLEAU 5.3 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : biais ($\times 10^4$), probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1) et nombre de problèmes de maximisation (Pbs), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Biais($\widehat{\theta_1}$) $\times 10^4$	$PC_1(\widehat{\theta_1})$	Biais($\widehat{\theta_2}$) $\times 10^4$	$PC_1(\widehat{\theta_2})$	Pbs
			1000	0.070	0.95	20.92	0.96	0
0.05	2	0.75	100	-34.87	0.84	518.8	0.94	2
			200	-22.65	0.89	212.7	0.93	0
			500	-15.59	0.93	72.90	0.93	0
			1000	-7.311	0.95	20.40	0.96	0
		0.50	100	-3.119	0.95	154.1	0.96	1
			200	-2.074	0.97	144.1	0.95	0
			500	-1.741	0.95	62.75	0.95	0
			1000	-1.049	0.97	41.31	0.94	0
		0.25	100	2.044	0.95	299.1	0.94	0
			200	1.416	0.95	48.24	0.95	0
			500	0.440	0.95	44.95	0.95	0
			1000	-0.087	0.95	14.72	0.96	0

TABLEAU 5.4 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : biais ($\times 10^4$), probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1) et nombre de problèmes de maximisation (Pbs), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Biais($\widehat{\theta_1}$) $\times 10^4$	$PC_1(\widehat{\theta_1})$	Biais($\widehat{\theta_2}$) $\times 10^4$	$PC_1(\widehat{\theta_2})$	Pbs
0.006	0.5	0.75	100	140.1	0.88	292.7	0.97	134
			200	74.26	0.88	0.97	0.96	56
			500	27.10	0.87	47.06	0.96	14
			1000	15.49	0.89	26.51	0.95	2
		0.50	100	25.84	0.86	115.4	0.96	5
			200	16.81	0.90	76.32	0.95	1
			500	5.760	0.94	35.05	0.96	0
			1000	1.164	0.93	1.861	0.93	0
		0.25	100	7.102	0.92	98.44	0.95	0
			200	4.093	0.93	38.26	0.95	0
			500	1.742	0.94	37.20	0.94	0
			1000	0.870	0.94	4.580	0.95	0
0.05	0.5	0.75	100	1354	0.91	329.5	0.96	127
			200	705.1	0.89	172.1	0.96	71
			500	258.6	0.90	55.96	0.95	11
			1000	96.04	0.89	1.360	0.95	0
		0.50	100	246.0	0.87	137.7	0.95	8
			200	102.9	0.89	31.08	0.95	0
			500	50.80	0.91	16.96	0.95	0
			1000	33.69	0.94	23.01	0.95	0
		0.25	100	70.48	0.92	112.7	0.96	0
			200	27.30	0.93	61.43	0.94	0
			500	13.41	0.94	34.34	0.95	0
			1000	7.128	0.95	6.683	0.93	0
0.006	2	0.75	100	0.729	0.91	720.9	0.94	9
			200	-0.532	0.95	195.8	0.96	1
			500	-0.326	0.95	96.78	0.94	0
			1000	-0.139	0.95	32.10	0.94	0
		0.50	100	0.438	0.95	487.1	0.95	0
			200	-0.162	0.95	116.2	0.94	0
			500	0.061	0.95	106.3	0.93	0
			1000	0.150	0.95	57.51	0.95	0
		0.25	100	0.617	0.94	430.3	0.94	0
			200	0.103	0.95	143.5	0.96	0
			500	-0.074	0.95	39.81	0.95	0

TABLEAU 5.4 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : biais ($\times 10^4$), probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1) et nombre de problèmes de maximisation (Pbs), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Biais($\widehat{\theta_1}$) $\times 10^4$	$PC_1(\widehat{\theta_1})$	Biais($\widehat{\theta_2}$) $\times 10^4$	$PC_1(\widehat{\theta_2})$	Pbs
			1000	-0.094	0.95	30.85	0.95	0
0.05	2	0.75	100	-4.643	0.92	461.3	0.94	0
			200	6.484	0.96	361.7	0.95	0
			500	-1.035	0.96	88.71	0.96	0
			1000	1.842	0.95	80.44	0.94	0
		0.50	100	-1.733	0.95	327.4	0.94	0
			200	-1.424	0.95	96.25	0.95	0
			500	-0.948	0.95	97.45	0.95	0
			1000	-0.112	0.96	44.67	0.96	0
		0.25	100	2.507	0.95	373.2	0.94	0
			200	1.377	0.95	189.6	0.96	0
			500	-0.086	0.95	53.41	0.94	0
			1000	0.465	0.95	49.53	0.95	0

calculées.

Les tableaux 5.5-5.7 rassemblent, pour chaque scénario, le nombre de réplications où il y a eu un problème de maximisation, le nombre de réplications où la condition d'existence n'était pas satisfaite (donc en réalité le maximum de vraisemblance n'existait pas pour la distribution exponentielle ou était suspecté de ne pas exister pour les distributions de Weibull et log-logistique) et le nombre de réplications où il y avait cohérence entre la survenue d'un problème de maximisation et la vérification de cette condition d'existence (nous avons additionné le nombre de réplications où il y avait un problème de maximisation et où la condition d'existence n'était pas satisfaite et le nombre de réplications où il n'y avait pas de problème de maximisation et où la condition d'existence était satisfaite).

Les tableaux 5.5-5.7 montrent que, quel que soit le scénario, il y a coïncidence presque parfaite entre la survenue d'un problème de maximisation et la non-vérification de cette condition d'existence.

Pour la distribution exponentielle, tous les problèmes de maximisation excepté un (où la statistique Q_1 était positive mais très proche de 0) sont survenus quand la condition d'existence n'était pas satisfaite (Tableau 5.5).

Pour les autres distributions, il y a quelques discordances (Tableaux 5.6-5.7). Pour les scénarios où le paramètre θ_2 était strictement inférieur à 1, il y avait deux types de discordances :

1. Il y avait un problème de maximisation et la condition d'existence était satisfaite mais la statistique Q_2 était presque nulle. Peut-être n'avons-nous pas choisi θ_{1s} assez proche de 0 pour ces réplications ?

2. L'estimation du maximum de vraisemblance a été obtenue et la condition d'existence n'était pas vérifiée mais l'estimation de θ_1 était très inférieure à la vraie valeur du paramètre et le gradient correspondant était loin d'être nul.

Le premier type de discordance est le plus répandu.

Pour les scénarios où le paramètre θ_2 est strictement supérieur à 1, il y avait un seul

type de discordances. L'estimation du maximum de vraisemblance a été obtenue et la condition d'existence n'était pas satisfaite mais l'estimation obtenue était très inférieure à la vraie valeur du paramètre, le gradient correspondant était presque nul, la variance très élevée et la statistique $Q_2(\theta_{2s})$ très élevée en valeur absolue. Ces discordances pourraient provenir de la limite suivante

$$\forall \theta_2 > 1,\ \lim_{\theta_1 \to 0} \frac{\partial \log L}{\partial \theta_1}(\theta_1) = 0.$$

On trouverait vraiment un paramètre vectoriel (θ_1, θ_2), pour lequel les équations normales seraient vérifiées mais où la fonction de vraisemblance n'atteindrait pas son maximum.

5.2.3.2 Biais

Le biais significatif d'un estimateur peut expliquer les estimations non satisfaisantes des probabilités de couverture. Les tableaux 5.2-5.4 montrent que :

- Pour la distribution exponentielle, les triplets (θ_1, p, n) où les estimations des probabilités de couverture sont supérieures à celles attendues sont les triplets (θ_1, p, n) où l'estimateur du maximum de vraisemblance est biaisé (même si le biais correspondant est relativement faible). Les triplets (θ_1, p, n) où les estimations des probabilités de couverture sont satisfaisantes sont ceux où l'estimateur du maximum de vraisemblance n'est pas biaisé.
- Pour les distributions de Weibull et log-logistique, les triplets (θ_1, θ_2, p) où les estimations des probabilités de couverture pour le paramètre θ_1 ont un comportement anormal quand la taille d'échantillon augmente ou bien sont inacceptablement peu élevées pour de petites tailles d'échantillon sont les triplets (θ_1, θ_2, p) où l'estimateur du maximum de vraisemblance de θ_1 est biaisé (et le biais est important). Les triplets (θ_1, θ_2, p) où les estimations des probabilité de couverture sont satisfaisantes pour le paramètre θ_1 (resp. θ_2) sont ceux où l'estimateur du maximum de vraisemblance de θ_1 (resp. θ_2) n'est pas biaisé.

TABLEAU 5.5 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : nombre de réplications où il y a eu un problème de maximisation (Pbs), nombre de réplications où la condition d'existence n'était pas satisfaite (Cond) et nombre de réplications où il y avait coïncidence entre la survenue d'un problème de maximisation et la vérification de la condition d'existence (Conc), pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	p	n	Pbs	Cond	Conc
0.006	0.75	100	114	114	1000
		200	42	41	999
		500	4	4	1000
		1000	0	0	1000
	0.50	100	0	0	1000
		200	0	0	1000
		500	0	0	1000
		1000	0	0	1000
	0.25	100	0	0	1000
		200	0	0	1000
		500	0	0	1000
		1000	0	0	1000
0.05	0.75	100	133	133	1000
		200	43	43	1000
		500	5	5	1000
		1000	0	0	1000
	0.50	100	0	0	1000
		200	0	0	1000
		500	0	0	1000
		1000	0	0	1000
	0.25	100	0	0	1000
		200	0	0	1000
		500	0	0	1000
		1000	0	0	1000

TABLEAU 5.6 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : nombre de réplications où il y a eu un problème de maximisation (Pbs), nombre de réplications où la condition d'existence n'était pas satisfaite (Cond) et nombre de réplications où il y avait coïncidence entre la survenue d'un problème de maximisation et la vérification de la condition d'existence (Conc), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Pbs	Cond	Conc
0.006	0.5	0.75	100	241	244	995
			200	169	170	997
			500	99	99	998
			1000	44	40	996
		0.50	100	82	79	989
			200	18	20	998
			500	0	0	1000
			1000	0	0	1000
		0.25	100	1	1	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
0.05	0.5	0.75	100	254	251	997
			200	199	200	999
			500	120	120	998
			1000	36	38	998
		0.50	100	75	73	998
			200	18	18	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	3	3	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
0.006	2	0.75	100	16	110	906
			200	6	51	955
			500	0	3	997
			1000	0	0	1000
		0.50	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000

TABLEAU 5.6 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : nombre de réplications où il y a eu un problème de maximisation (Pbs), nombre de réplications où la condition d'existence n'était pas satisfaite (Cond) et nombre de réplications où il y avait coïncidence entre la survenue d'un problème de maximisation et la vérification de la condition d'existence (Conc), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Pbs	Cond	Conc
			500	0	0	1000
			1000	0	0	1000
0.05	2	0.75	100	2	106	896
			200	0	48	952
			500	0	9	991
			1000	0	0	1000
		0.50	100	1	1	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000

TABLEAU 5.7 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : nombre de réplications où il y a eu un problème de maximisation (Pbs), nombre de réplications où la condition d'existence n'était pas satisfaite (Cond) et nombre de réplications où il y avait coïncidence entre la survenue d'un problème de maximisation et la vérification de la condition d'existence (Conc), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Pbs	Cond	Conc
0.006	0.5	0.75	100	134	119	985
			200	56	48	992
			500	14	13	999
			1000	2	1	999
		0.50	100	5	2	997
			200	1	0	999
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
0.05	0.5	0.75	100	127	124	997
			200	71	70	999
			500	11	9	998
			1000	0	0	1000
		0.50	100	8	6	998
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
0.006	2	0.75	100	9	24	985
			200	1	3	998
			500	0	0	1000
			1000	0	0	1000
		0.50	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000

TABLEAU 5.7 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : nombre de réplications où il y a eu un problème de maximisation (Pbs), nombre de réplications où la condition d'existence n'était pas satisfaite (Cond) et nombre de réplications où il y avait coïncidence entre la survenue d'un problème de maximisation et la vérification de la condition d'existence (Conc), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	Pbs	Cond	Conc
			500	0	0	1000
			1000	0	0	1000
0.05	2	0.75	100	0	24	976
			200	0	2	998
			500	0	0	1000
			1000	0	0	1000
		0.50	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000
		0.25	100	0	0	1000
			200	0	0	1000
			500	0	0	1000
			1000	0	0	1000

5.2.3.3 Normalité

Nous avons construit les histogrammes des estimations du maximum de vraisemblance pour chaque paramètre et pour chaque scénario. Les scénarios où les estimations des probabilités de couverture sont satisfaisantes sont les scénarios où les histogrammes des estimations ressemblent à un histogramme de loi gaussienne. C'est le cas pour le paramètre θ_2 (Figures 5.1 et 5.2).

Sur les figures 5.3-5.6 sont rassemblés quelques exemples d'histogrammes pour quelques scénarios où les estimations des probabilités de couverture ne sont pas satisfaisantes. Ces histogrammes sont typiques du comportement général de l'estimateur du maximum de vraisemblance dans ce cas-là. On remarque qu'il y a une amélioration quand la taille d'échantillon augmente ou quand la proportion de données tronquées diminue. Un écart entre la distribution de l'estimateur et la distribution gaussienne peut expliquer les estimations non satisfaisantes des probabilités de couverture. Les histogrammes montrent que pour la distribution exponentielle, les triplets (θ_1, p, n) où les estimations des probabilités de couverture sont supérieures à celles attendues sont les triplets (θ_1, p, n) où la distribution de l'estimateur du maximum de vraisemblance est la plus éloignée de la distribution gaussienne (Figures 5.3 et 5.4), c'est-à-dire quand la taille d'échantillon est petite (n=100) et la proportion de données tronquées importante (p=0.75). Les figures 5.5 et 5.6 présentent les histogrammes de quelques scénarios où le délai de survenue est généré selon la distribution de Weibull. Les histogrammes sont semblables pour les scénarios correspondants quand le délai de survenue est généré selon la distribution log-logistique. Pour les distributions de Weibull et log-logistique, les quadruplets (θ_1, θ_2, p, n) où les estimations des probabilités de couverture ne sont pas satisfaisantes pour le paramètre θ_1 sont les quadruplets (θ_1, θ_2, p, n) où la distribution de l'estimateur du maximum de vraisemblance de θ_1 est la plus éloignée de la distribution gaussienne. La figure 5.6 montre que, quand la proportion de données tronquées est élevée (p=0.75), la normalité asymptotique ne semble pas être atteinte pour les plus grandes tailles d'échantillon (n=1000).

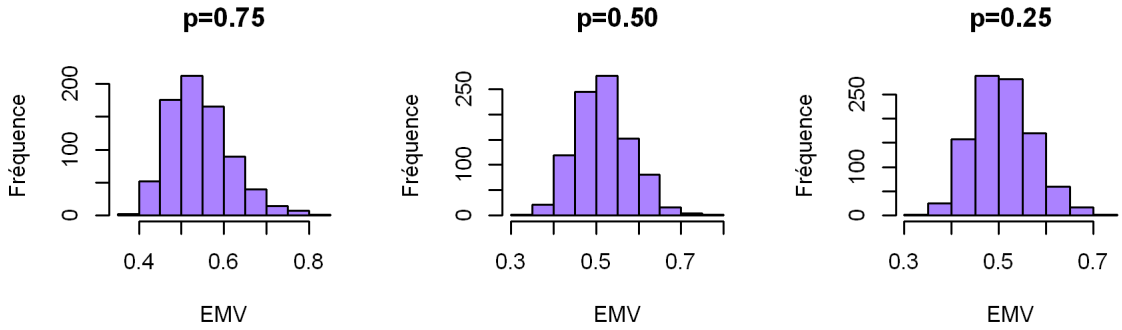

FIGURE 5.1 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : histogrammes des estimations du maximum de vraisemblance du paramètre θ_2 pour n=100, θ_1=0.006, θ_2=0.5 et trois proportions de données tronquées à droite p.
Abréviation : EMV, estimation du maximum de vraisemblance.

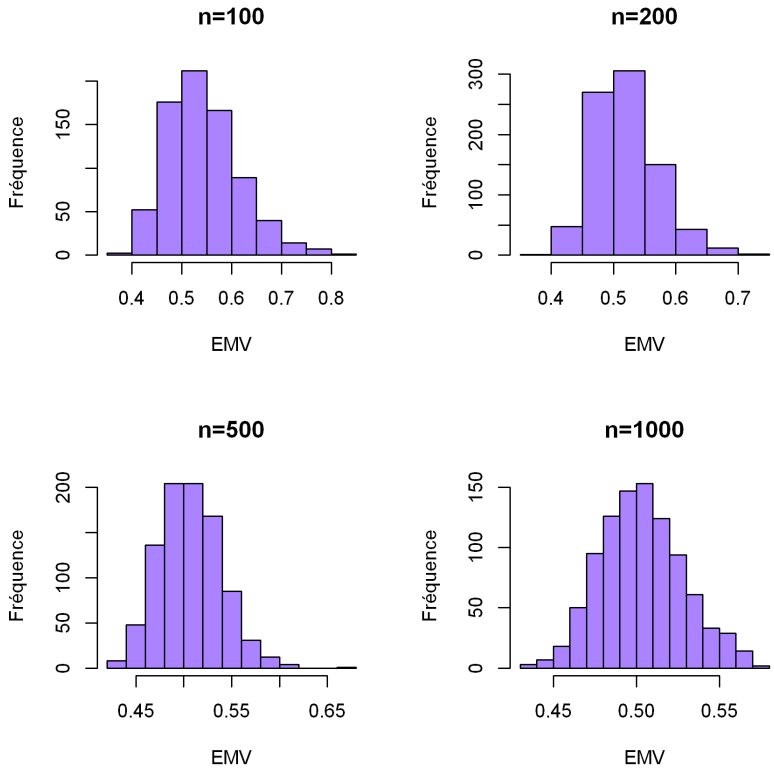

FIGURE 5.2 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : histogrammes des estimations du maximum de vraisemblance du paramètre θ_2 pour p=0.75, θ_1=0.006, θ_2=0.5 et quatre tailles d'échantillon n.
Abréviation : EMV, estimation du maximum de vraisemblance.

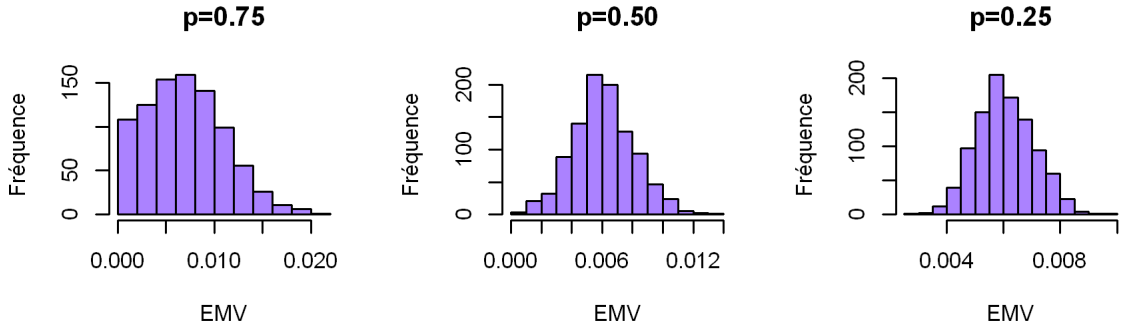

FIGURE 5.3 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : histogrammes des estimations du maximum de vraisemblance pour n=100, θ_1=0.006 et trois proportions de données tronquées à droite p.
Abréviation : EMV, estimation du maximum de vraisemblance.

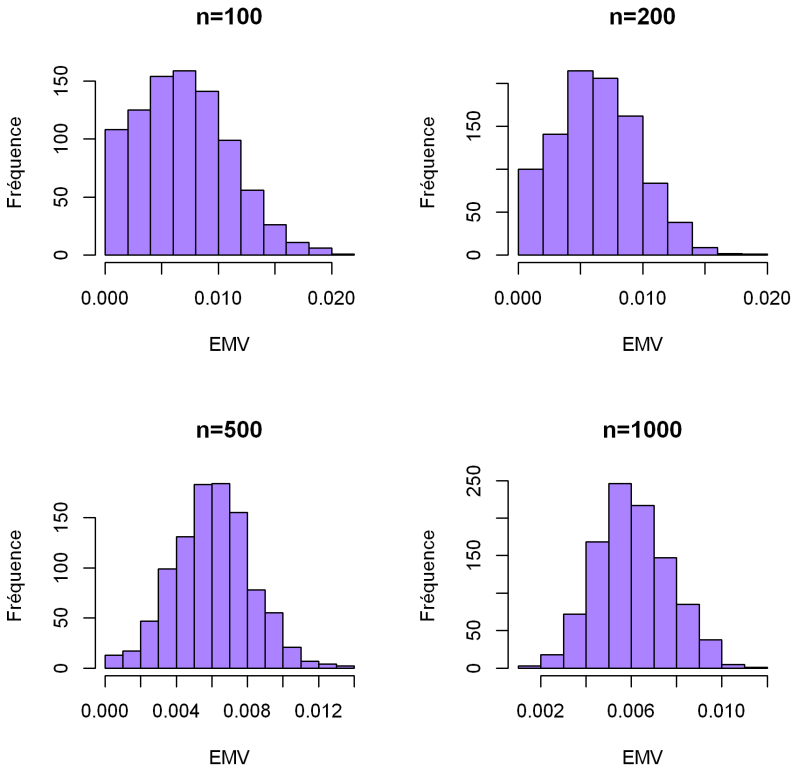

FIGURE 5.4 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : histogrammes des estimations du maximum de vraisemblance pour p=0.75, θ_1=0.006 et quatre tailles d'échantillon n.
Abréviation : EMV, estimation du maximum de vraisemblance.

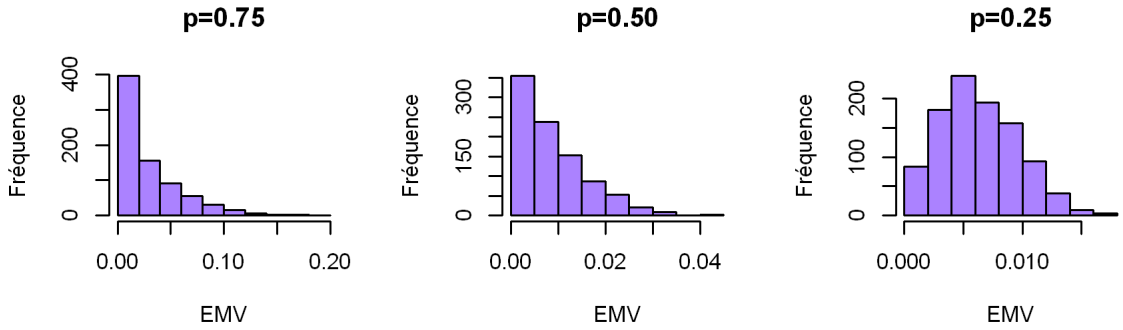

FIGURE 5.5 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : histogrammes des estimations du maximum de vraisemblance du paramètre θ_1 pour n=100, θ_1=0.006, θ_2=0.5 et trois proportions de données tronquées à droite p.
Abréviation : EMV, estimation du maximum de vraisemblance.

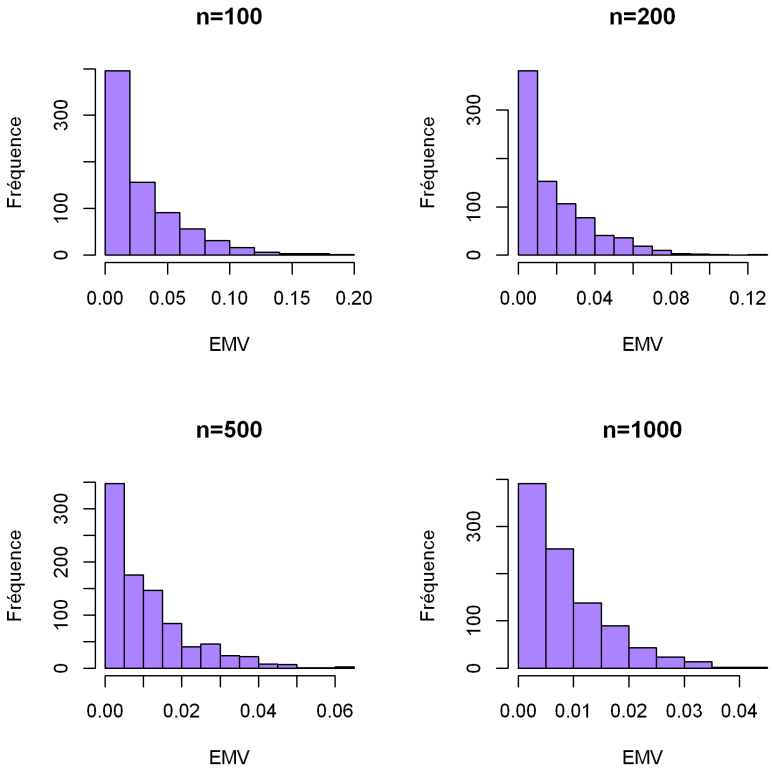

FIGURE 5.6 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : histogrammes des estimations du maximum de vraisemblance du paramètre θ_1 pour p=0.75, θ_1=0.006, θ_2=0.5 et quatre tailles d'échantillon n.
Abréviation : EMV, estimation du maximum de vraisemblance.

5.2.3.4 Matrice de covariance

Remplacer la matrice de covariance asymptotique par son estimation (5.2) pour dériver les intervalles de confiance peut impacter les estimations des probabilités de couverture. Nous souhaitons évaluer cet impact. La vraie valeur des paramètres $(\theta^0, \lambda_q, q = 1, \ldots, M)$ peuvent être utilisés pour obtenir d'autres expressions de la matrice à inverser et des intervalles alternatifs basés sur ces autres expressions de la matrice peuvent être calculés. Nous comparons les estimations des probabilités de couverture des intervalles de confiance avec les estimations des probabilités de couverture de ces intervalles alternatifs qui ne sont pas des intervalles de confiance puisque les vraies valeurs des paramètres sont utilisées. Afin de distinguer l'impact de l'estimation de θ de l'impact de l'estimation des quantités λ_q dans l'estimation de la matrice de covariance, différentes expressions de la matrice à inverser sont considérées. Cependant, la matrice obtenue en substituant simplement θ^0 dans l'expression (5.2) (expression 1) peut ne pas être définie positive et donc inversible. Pour cette raison, une autre estimation de l'espérance E_θ sera utilisée pour minimiser la survenue de ce problème (expressions 2-5). Les expressions suivantes de la matrice sont considérées :

Expression 1 En remplaçant l'estimation $\widehat{\theta}_n$ par sa vraie valeur θ^0, on obtient l'expression $I^{(1)}$:

$$\left[I^{(1)}\right]_{jp} = \sum_{i=1}^{n} \frac{1}{n} \left(- \frac{\partial^2 \log f_{t_i}(x_i;\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0} \right).$$

Cependant, pour les distributions de Weibull et log-logistique, la matrice sous-jacente était rarement définie positive. Par conséquent, les estimations des probabilités de couverture des intervalles alternatifs obtenus à partir de l'expression $I^{(1)}$ ne sont disponibles que pour la distribution exponentielle.

Expression 2 Pour les distributions de Weibull et log-logistique, nous avons utilisé l'estimation (5.1) de $\left[I\left(\theta^0\right)\right]_{jp}$, pour $1 \leqslant j, p \leqslant r$:

$$\sum_{q=1}^{m} \frac{n_q}{n} E_{\widehat{\theta}_n} \left(- \frac{\partial^2 \log\left(f_q(X_{q1};\theta)\right)}{\partial\theta_j \partial\theta_p}\bigg|_{\widehat{\theta}_n} \right).$$

Pour la distribution exponentielle, cela conduit à la même estimation $\left[\widehat{I\left(\theta^0\right)}\right]_{jp}$ et ne produit pas d'expression alternative. Etant donné que l'espérance n'a pas de forme explicite pour les deux autres distributions, nous avons utilisé la fonction *integral* du logiciel R pour approximer l'espérance E_θ. Cette approximation est notée $\widehat{E}_\theta$. Pour s'assurer que les résultats soient comparables, nous avons d'abord déterminé les estimations des probabilités de couverture des intervalles alternatifs basés sur la matrice $I^{(2)}$ où l'approximation $\widehat{E}_{\widehat{\theta}_n}$ de l'espérance, l'estimation du paramètre vectoriel $\widehat{\theta}_n$ et les estimations des quantités λ_q sont utilisées :

$$\left[I^{(2)}\right]_{jp} = \sum_{q=1}^{m} \frac{n_q}{n} \widehat{E}_{\widehat{\theta}_n} \left(- \frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\widehat{\theta}_n} \right).$$

L'expression 2 est une expression intermédiaire permettant d'introduire la troisième expression.

Expression 3 Maintenant, l'estimation du paramètre vectoriel $\widehat{\theta}_n$ est remplacé par sa vraie valeur θ^0 dans l'expression $I^{(2)}$, ce qui nous permet d'obtenir l'expression $I^{(3)}$:

$$\left[I^{(3)}\right]_{jp} = \sum_{q=1}^{m} \frac{n_q}{n} \widehat{E}_{\theta^0} \left(- \frac{\partial^2 \log f_k(X;\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0} \right).$$

Les estimations des probabilités de couverture des intervalles alternatifs calculés à partir des expressions $I^{(2)}$ ou $I^{(3)}$ sont disponibles pour les distributions de Weibull ou log-logistique.

Expression 4 La quatrième expression de la matrice à inverser permet d'étudier l'influence de l'estimation des quantités λ_q. Pour tout $q = 1, \ldots, M$, la vraie valeur de la quantité λ_q est :

$$\lambda_q = \frac{P\left(T = u_q\right) P\left(X < u_q\right)}{P\left(T > X\right)}.$$

Considérant les M délais de troncature distincts qui peuvent être observés dans la population dont est issu l'échantillon et les vraies valeurs des quantités λ_q, on

obtient l'expression $I^{(4)}$ de la matrice :

$$\left[I^{(4)}\right]_{jp} = \sum_{q=1}^{M} \lambda_q E_{\widehat{\theta}_n} \left(- \frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\widehat{\theta}_n} \right)$$

pour la distribution exponentielle et

$$\left[I^{(4)}\right]_{jp} = \sum_{q=1}^{M} \lambda_q \widehat{E}_{\widehat{\theta}_n} \left(- \frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\widehat{\theta}_n} \right)$$

pour les autres distributions. Malheureusement, malgré l'utilisation de l'approximation $E_{\widehat{\theta}_n}$, cette matrice n'était pas systématiquement définie positive pour les distributions de Weibull et log-logistique. Par conséquent, les résultats de l'influence de l'estimation des quantités λ_q sont disponibles seulement pour la distribution exponentielle.

Expression 5 Enfin, l'effet cumulé de l'estimation du paramètre vectoriel θ et des quantités λ_q a été étudié. On obtient ainsi la cinquième matrice $I^{(5)}$, d'expression

$$\left[I^{(5)}\right]_{jp} = \sum_{q=1}^{M} \lambda_q E_{\theta^0} \left(- \frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0} \right)$$

pour la distribution exponentielle et

$$\left[I^{(5)}\right]_{jp} = \sum_{q=1}^{M} \lambda_q \widehat{E}_{\theta^0} \left(- \frac{\partial^2 \log f_q(X_{q1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0} \right)$$

pour les autres distributions. Les estimations des probabilités de couverture des intervalles alternatifs obtenus à partir de cette matrice $I^{(5)}$ sont disponibles pour les trois distributions.

Le tableau 5.8 résume l'obtention de ces expressions. Comme nous l'avons déjà dit, les intervalles calculés à partir de ces expressions de la matrice à inverser ne sont pas des intervalles de confiance puisqu'ils sont basés sur les vraies valeurs des paramètres qui

sont inconnus en pratique ; ce sont des intervalles alternatifs et sont seulement utilisés afin d'étudier l'influence de l'estimation de la matrice de covariance asymptotique dans l'estimation des probabilités de couverture.

En complément, la "vraie" valeur de l'écart-type asymptotique (valeur exacte pour la distribution exponentielle et valeur approximée à l'aide de l'approximation $\widehat{E}_{\theta^0}$ pour les autres distributions) obtenue à partir de l'expression 5, et son estimation basée sur les réplications ont été calculées pour chaque composante du paramètre vectoriel θ.

Les résultats pour la distribution exponentielle sont rassemblés dans le tableau D.1 de l'annexe D. Les estimations des probabilités de couverture supérieures à celles attendues ne viennent pas d'un biais de l'estimateur de la variance asymptotique. Etant donné que ce biais est directement lié au biais de l'estimateur du paramètre θ_1, ces estimations non satisfaisantes des probabilités de couverture des intervalles de confiance de type Wald semblent provenir d'un écart à la normalité. Pour les triplets (θ_1, p, n) où les probabilités de couverture sont satisfaisantes, l'écart-type est bien estimé.

Les résultats sont similaires entre la distribution de Weibull et la distribution log-logistique. Les résultats pour le paramètre θ_1 de la distribution de Weibull (resp. log-logistique) sont présentés dans le tableau D.2 (resp. tableau D.3) de l'annexe D. Pour les triplets (θ_1, θ_2, p) où les estimations des probabilités de couverture des intervalles de confiance de type Wald sont satisfaisantes, l'écart-type est bien estimé. Les variations avec n pour ces triplets semblent provenir des fluctuations d'échantillonnage, d'un biais résiduel ou d'un écart résiduel à la normalité.

Pour les triplets (θ_1, θ_2, p) où les estimations des probabilités de couverture des intervalles de confiance de type Wald sont inférieures au niveau attendu et pour lesquels les variations de ces estimations avec n est anormal, une amélioration des probabilités de couverture quand la taille d'échantillon augmente ou quand la proportion de données tronquées diminue apparaissent quand l'expression 3 est utilisée. Cependant, les estimations des probabilités de couverture sont presque toujours inférieures avec $I^{(3)}$ qu'avec $\widehat{I\left(\theta^0\right)}$, ce qui s'explique par le biais et l'écart à la normalité. Pour les triplets (θ_1, θ_2, p) où les estimations des probabilités de couverture des intervalles de confiance de type

TABLEAU 5.8 – Etude de simulations n° 2 : estimation de l'inverse de la matrice de covariance asymptotique utilisée pour calculer les intervalles de confiance et expressions alternatives utilisées pour calculer des intervalles alternatifs.

	$\widehat{I(\theta^0)}$	$I^{(1)}$	$I^{(2)}$	$I^{(3)}$	$I^{(4)}$	$I^{(5)}$
E_k	$\frac{1}{n}\sum_{i=1}^{n}$	$\frac{1}{n}\sum_{i=1}^{n}$	$\widehat{E}_k$	$\widehat{E}_k$	$\frac{1}{n}\sum_{i=1}^{n}$ or $\widehat{E}_k$	$\frac{1}{n}\sum_{i=1}^{n}$ or $\widehat{E}_k$
θ^0	$\theta = \widehat{\theta}_n$	$\theta = \theta^0$	$\theta = \widehat{\theta}_n$	$\theta = \theta^0$	$\theta = \widehat{\theta}_n$	$\theta = \theta^0$
λ_k	$\frac{n_k}{n}$	$\frac{n_k}{n}$	$\frac{n_k}{n}$	$\frac{n_k}{n}$	λ_k	λ_k

Wald étaient anormalement faibles par rapport au niveau attendu mais pour lesquels la variation avec n est satisfaisante, l'écart entre les estimations observées et le niveau attendu diminue quand l'expression $I^{(3)}$ est utilisée. Par conséquent, les estimations des probabilités de couverture observées viennent d'un biais de l'estimateur de la variance asymptotique (directement lié au biais de l'estimateur du paramètre θ_1), d'un biais de l'estimateur paramétrique du maximum de vraisemblance et d'un écart à la normalité, et ces effets se compensent souvent.

Le tableau D.4 (resp. tableau D.5) de l'annexe D montre qu'il n'y a pas de biais de l'estimateur de la variance asymptotique pour le paramètre θ_2 de la distribution de Weibull (resp. log-logistique).

5.3 Qualité de l'estimation par intervalle de la vraisemblance profilée

5.3.1 Intervalle de confiance de la vraisemblance profilée

La vraisemblance profilée permet de calculer des intervalles de confiance pour chaque composante du paramètre vectoriel θ. La théorie de la vraisemblance profilée a été déve-

loppée dans un cadre multivarié pour des observations indépendantes et identiquement distribuées. Dans ce cas, les propriétés asymptotiques de l'estimateur paramétrique du maximum de vraisemblance suffisent à prouver que le rapport de vraisemblance profilée suit asymptotiquement une distribution du chi-2, ce qui permet le calcul des intervalles de confiance de la vraisemblance profilée (Pawitan, 2001). Pour des observations indépendantes mais non-identiquement distribuées, sachant les propriétés asymptotiques de $\widehat{\theta}_n$ présentées au chapitre 3 et suivant la démonstration de Pawitan (2001), la distribution asymptotique du rapport de vraisemblance profilée est inchangée.

Etant donné que la distribution exponentielle est décrite par un paramètre scalaire et que les distributions de Weibull et log-logistique sont décrites par un paramètre de dimension 2, cette méthode est présentée pour ces deux cas particuliers. Quand le paramètre est un scalaire, le rapport de vraisemblance profilée coïncide avec le rapport de vraisemblance et

$$2\log\frac{L(x_1,\ldots,x_n;\widehat{\theta}_n|t_1,\ldots,t_n,n)}{L(x_1,\ldots,x_n;\theta^0|t_1,\ldots,t_n,n)} \xrightarrow[n\to+\infty]{d} \chi^2_1.$$

L'intervalle de confiance à $100(1-\alpha)\%$ correspondant pour le paramètre θ^0 est donné par :

$$IC_{100(1-\alpha)\%}(\theta^0) = \left\{\theta; 2\log\frac{L(x_1,\ldots,x_n;\widehat{\theta}_n|t_1,\ldots,t_n,n)}{L(x_1,\ldots,x_n;\theta|t_1,\ldots,t_n,n)} < \chi^2_{1,1-\alpha}\right\}, \tag{5.3}$$

où $\chi^2_{1,1-\alpha}$ est le quantile à $100(1-\alpha)\%$ de la distribution du chi-2 à 1 degré de liberté.

Pour un paramètre de dimension 2, $\theta = (\theta_1, \theta_2)$, sa vraie valeur et son estimation du maximum de vraisemblance sont $\theta^0 = (\theta^0_1, \theta^0_2)$ et $\widehat{\theta}_n = (\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$. Pour calculer un intervalle de confiance pour la composante θ^0_1, la méthode de la vraisemblance profilée consiste à considérer θ_1 comme le paramètre d'intérêt et θ_2 comme un paramètre de nuisance. Inversement, pour calculer un intervalle de confiance pour la composante θ^0_2, le paramètre θ_2 est considéré comme paramètre d'intérêt et θ_1 comme un paramètre de nuisance. Plus précisement, pour une valeur fixée de θ_1, l'estimation du maximum de

vraisemblance de θ_2 est décrite par la fonction h déjà introduite au chapitre 4 :

$$h : \theta_1 \longmapsto \widehat{\theta}_2(\theta_1) \text{ tel que } \frac{\partial}{\partial\theta_2}\log L\left(x_1,\ldots,x_n;(\theta_1,\widehat{\theta}_2(\theta_1))\Big|\, t_1,\ldots,t_n,n\right)=0,$$

et la vraisemblance profilée de θ_1 est $L(x_1,\ldots,x_n;(\theta_1,h(\theta_1))|t_1,\ldots,t_n,n)$. Le rapport de la vraisemblance profilée vérifie :

$$2\log\frac{L(x_1,\ldots,x_n;(\widehat{\theta}_{1n},\widehat{\theta}_{2n})|t_1,\ldots,t_n,n)}{L(x_1,\ldots,x_n;(\theta_1^0,h(\theta_1^0))|t_1,\ldots,t_n,n)} \underset{n\to+\infty}{\xrightarrow{d}} \chi^2_1.$$

L'intervalle de confiance asymptotique à $100(1-\alpha)\%$ correspondant pour le paramètre θ_1^0 s'écrit :

$$IC_{100(1-\alpha)\%}(\theta_1^0)=\left\{\theta_1; 2\log\frac{L(x_1,\ldots,x_n;(\widehat{\theta}_{1n},\widehat{\theta}_{2n})\,|\,t_1,\ldots,t_n,n)}{L(x_1,\ldots,x_n;(\theta_1,h(\theta_1))|\,t_1,\ldots,t_n,n)}<\chi^2_{1,1-\alpha}\right\}. \quad (5.4)$$

De manière symétrique, un intervalle de confiance asymptotique à $100(1-\alpha)\%$ pour la composante θ_2 peut être obtenu.

L'intervalle de confiance basé sur la vraisemblance profilée pourrait avoir un meilleur comportement que les intervalles de type Wald car la méthode de la vraisemblance profilée est moins sensible à la paramétrisation choisie (Pawitan, 2001). Nous avons comparé dans l'étude de simulations présentée dans la section 5.2 les estimations des probabilités de couverture des intervalles de confiance de type Wald et les estimations des probabilités de couverture des intervalles de confiance de la vraisemblance profilée.

5.3.2 Etude de simulations

Pour chaque scénario de l'étude de simulations présentée dans la section 5.2 (cf. tableau 5.2), les probabilités de couverture des intervalles de confiance à 95% basés sur le rapport de vraisemblance ont été estimées si le délai de survenue suivait une distribution exponentielle et les estimations des probabilités de couverture des intervalles de confiance à 95% basés sur la vraisemblance profilée ont été calculées si le délai de

survenue suivait une distribution de Weibull ou log-logistique. Pour estimer les probabilités de couverture, il n'a pas été nécessaire de calculer explicitement les intervalles de confiance de la vraisemblance profilée. Pour la distribution exponentielle, si la vraie valeur θ^0 du paramètre vérifie l'inégalité suivante :

$$2\log\frac{L(x_1,\ldots,x_n;\widehat{\theta}_n|t_1,\ldots,t_n,n)}{L(x_1,\ldots,x_n;\theta^0|t_1,\ldots,t_n,n)} < \chi^2_{1,1-\alpha},$$

alors l'intervalle de confiance (5.3) contient la vraie valeur du paramètre. Pour les distributions de Weibull et log-logistique, si la vraie valeur θ^0 du paramètre vérifie l'inégalité suivante :

$$2\log\frac{L(\,x_1,\ldots,x_n;(\widehat{\theta}_{1n},\widehat{\theta}_{2n})\Big|\,t_1,\ldots,t_n,n)}{L(\,x_1,\ldots,x_n;(\theta_1^0,h(\theta_1^0))|\,t_1,\ldots,t_n,n)} < \chi^2_{1,1-\alpha},$$

alors l'intervalle de confiance (5.3) contient la vraie valeur du paramètre. En comptabilisant le nombre de réplications, parmi 1000, où cette inégalité était vérifiée, on obtenait l'estimation de la probabilité de couverture. L'algorithme de Newton-Raphson a été utilisé pour l'étape de maximisation nécessaire pour le calcul de la vraisemblance profilée $L(\,x_1,\ldots,x_n;(\theta_1^0,h(\theta_1^0))|\,t_1,\ldots,t_n,n)$.

Les résultats concernant ces intervalles de confiance sont présentés dans les tableaux 5.9, 5.10 et 5.11.

Pour la distribution exponentielle, les estimations des probabilités de couverture sont similaires pour les deux types d'intervalles de confiance.

Pour le paramètre θ_2 des distributions de Weibull et log-logistique, les estimations des probabilités de couverture sont semblables pour les deux types d'intervalles de confiance. En revanche, pour le paramètre θ_1, les estimations des probabilités de couverture des intervalles de confiance basés sur la vraisemblance profilée sont nettement meilleures que celles des intervalles de confiance de type Wald.

L'étape de maximisation supplémentaire nécessaire pour établir l'intervalle de confiance de la vraisemblance profilée du paramètre θ_2 a échoué pour certaines réplications où l'estimation du maximum de vraisemblance du paramètre vectoriel θ avait pourtant été

obtenue. Les estimations des probabilités de couverture de ces intervalles de confiance ont donc été calculées sur l'ensemble des réplications où il n'y avait pas de problème de maximisation. La proportion des réplications où cette étape de maximisation supplémentaire a échoué, parmi les réplications où la maximisation initiale n'avait pas échoué, a le même comportement en fonction de la taille d'échantillon et de la proportion de données tronquées que le nombre de problèmes de maximisation initiale. Il atteint 11.7% pour la distribution de Weibull et 1.3% pour la distribution log-logistique.

5.4 Application : Délai de survenue d'un lymphome

Revenons à l'analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α issus de la base de pharmacovigilance française. Ces données ont été présentées dans la section 2.3 du chapitre 2.

Il y avait $m = 59$ délais de troncature distincts dans l'échantillon observé. Comme il n'y a eu aucun problème d'approvisionnement durant la période de commercialisation des traitements, le nombre M de délais de troncature distincts qui peuvent être observés dans la population dont est issu l'échantillon coïncide avec le nombre de semaines qui se sont écoulées entre l'autorisation de mise sur le marché du traitement ayant été mis sur le marché en premier, *infliximab*, et la date d'analyse, c'est-à-dire $M = 548$.

Tous traitements confondus, nous avons calculé les quantités Q_1 ou $Q_2(\theta_{2s})$ pour les distributions exponentielle, de Weibull et log-logistique afin de savoir si la condition d'existence du maximum de vraisemblance est satisfaite. Nous avons dérivé, pour ces distributions, les estimations du maximum de vraisemblance (déjà obtenues dans la section 2.3 du chapitre 2) et les intervalles de confiance à 95% associés pour les paramètres des distributions.

Le tableau 5.12 montre que, quelle que soit la distribution, la quantité Q_1 ou $Q_2(\theta_{2s})$ est strictement positive. L'estimation du maximum de vraisemblance du paramètre de la distribution existe pour la distribution exponentielle et semble exister pour les autres distributions. Les estimations des paramètres ainsi que leurs intervalles de confiance à

TABLEAU 5.9 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1, pour rappel), probabilité de couverture des intervalles de confiance à 95% de la vraisemblance profilée (PC_2) et nombre de problèmes de maximisation (Pbs), pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	p	n	$PC_1(\widehat{\theta_1})$	$PC_2(\widehat{\theta_1})$	Pbs
0.006	0.75	100	0.98	0.98	114
		200	0.97	0.97	42
		500	0.96	0.96	4
		1000	0.94	0.94	0
	0.50	100	0.95	0.95	0
		200	0.95	0.94	0
		500	0.95	0.95	0
		1000	0.94	0.94	0
	0.25	100	0.96	0.96	0
		200	0.95	0.95	0
		500	0.94	0.95	0
		1000	0.95	0.95	0
0.05	0.75	100	0.98	0.98	133
		200	0.97	0.97	43
		500	0.95	0.95	5
		1000	0.94	0.94	0
	0.50	100	0.95	0.94	0
		200	0.95	0.95	0
		500	0.95	0.95	0
		1000	0.95	0.95	0
	0.25	100	0.95	0.94	0
		200	0.96	0.96	0
		500	0.95	0.95	0
		1000	0.95	0.95	0

TABLEAU 5.10 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald (CP_1, pour rappel), probabilité de couverture des intervalles de confiance à 95% de la vraisemblance profilée (CP_2), nombre de problèmes de maximisation (Pbs) et nombre de problèmes supplémentaires de maximisation liés à la vraisemblance profilée (Supp), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$PC_1(\widehat{\theta}_1)$	$PC_2(\widehat{\theta}_1)$	$PC_1(\widehat{\theta}_2)$	$PC_2(\widehat{\theta}_2)$	Pbs	Supp
0.006	0.5	0.75	100	0.93	0.95	0.95	0.96	241	89
			200	0.92	0.97	0.96	0.96	169	22
			500	0.91	0.97	0.97	0.98	99	4
			1000	0.89	0.97	0.95	0.95	44	1
		0.50	100	0.88	0.97	0.97	0.97	82	1
			200	0.89	0.96	0.96	0.95	18	0
			500	0.91	0.95	0.95	0.94	0	0
			1000	0.93	0.95	0.95	0.95	0	0
		0.25	100	0.91	0.95	0.94	0.94	1	0
			200	0.91	0.93	0.94	0.94	0	0
			500	0.96	0.97	0.96	0.96	0	0
			1000	0.96	0.96	0.95	0.95	0	0
0.05	0.5	0.75	100	0.94	0.97	0.97	0.98	254	65
			200	0.91	0.96	0.94	0.95	199	45
			500	0.89	0.96	0.96	0.96	120	7
			1000	0.89	0.98	0.97	0.97	36	0
		0.50	100	0.87	0.97	0.96	0.96	75	4
			200	0.89	0.96	0.95	0.95	18	0
			500	0.90	0.94	0.95	0.95	0	0
			1000	0.94	0.95	0.96	0.96	0	0
		0.25	100	0.93	0.97	0.96	0.96	3	0
			200	0.93	0.94	0.95	0.95	0	0
			500	0.94	0.95	0.95	0.95	0	0
			1000	0.95	0.95	0.94	0.95	0	0
0.006	2	0.75	100	0.85	0.96	0.93	0.96	16	23
			200	0.89	0.96	0.92	0.94	6	2
			500	0.94	0.95	0.95	0.95	0	0
			1000	0.94	0.94	0.94	0.94	0	0
		0.50	100	0.96	0.95	0.95	0.95	0	0
			200	0.96	0.95	0.95	0.95	0	0
			500	0.96	0.96	0.95	0.95	0	0
			1000	0.95	0.95	0.94	0.94	0	0
		0.25	100	0.94	0.94	0.94	0.94	0	0
			200	0.95	0.95	0.96	0.95	0	2

TABLEAU 5.10 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1, pour rappel), probabilité de couverture des intervalles de confiance à 95% de la vraisemblance profilée (PC_2), nombre de problèmes de maximisation (Pbs) et nombre de problèmes supplémentaires de maximisation liés à la vraisemblance profilée (Supp), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$PC_1(\widehat{\theta_1})$	$PC_2(\widehat{\theta_1})$	$PC_1(\widehat{\theta_2})$	$PC_2(\widehat{\theta_2})$	Pbs	Supp
			500	0.94	0.95	0.96	0.95	0	0
			1000	0.95	0.95	0.96	0.95	0	0
0.05	2	0.75	100	0.84	0.96	0.94	0.96	2	3
			200	0.89	0.96	0.93	0.95	0	2
			500	0.93	0.93	0.93	0.93	0	0
			1000	0.95	0.95	0.96	0.96	0	0
		0.50	100	0.95	0.96	0.96	0.96	1	0
			200	0.97	0.97	0.95	0.95	0	0
			500	0.95	0.95	0.95	0.95	0	0
			1000	0.97	0.96	0.94	0.94	0	0
		0.25	100	0.95	0.95	0.94	0.94	0	0
			200	0.95	0.94	0.95	0.95	0	0
			500	0.95	0.95	0.95	0.95	0	0
			1000	0.95	0.95	0.96	0.96	0	0

TABLEAU 5.11 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1, pour rappel), probabilité de couverture des intervalles de confiance à 95% de la vraisemblance profilée (PC_2), nombre de problèmes de maximisation (Pbs) et nombre de problèmes supplémentaires de maximisation liés à la vraisemblance profilée (Supp), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$PC_1(\widehat{\theta}_1)$	$PC_2(\widehat{\theta}_1)$	$PC_1(\widehat{\theta}_2)$	$PC_2(\widehat{\theta}_2)$	Pbs	Supp
0.006	0.5	0.75	100	0.88	0.97	0.97	0.97	134	8
			200	0.88	0.97	0.96	0.96	56	0
			500	0.87	0.96	0.96	0.95	14	0
			1000	0.89	0.94	0.95	0.95	2	0
		0.50	100	0.86	0.96	0.96	0.96	5	0
			200	0.90	0.95	0.95	0.95	1	0
			500	0.94	0.96	0.96	0.96	0	0
			1000	0.93	0.95	0.93	0.93	0	0
		0.25	100	0.92	0.95	0.95	0.95	0	0
			200	0.93	0.95	0.95	0.95	0	0
			500	0.94	0.94	0.94	0.94	0	0
			1000	0.94	0.95	0.95	0.95	0	0
0.05	0.5	0.75	100	0.91	0.97	0.96	0.96	127	11
			200	0.89	0.97	0.96	0.96	71	3
			500	0.90	0.96	0.95	0.95	11	0
			1000	0.89	0.95	0.95	0.95	0	0
		0.50	100	0.87	0.95	0.95	0.95	8	0
			200	0.89	0.95	0.95	0.95	0	0
			500	0.91	0.94	0.95	0.95	0	0
			1000	0.94	0.94	0.95	0.95	0	0
		0.25	100	0.92	0.95	0.96	0.96	0	0
			200	0.93	0.94	0.94	0.93	0	0
			500	0.94	0.96	0.95	0.95	0	0
			1000	0.95	0.95	0.93	0.93	0	0
0.006	2	0.75	100	0.91	0.95	0.94	0.95	9	0
			200	0.95	0.94	0.96	0.95	1	0
			500	0.95	0.95	0.94	0.94	0	0
			1000	0.95	0.95	0.94	0.95	0	0
		0.50	100	0.95	0.95	0.95	0.95	0	0
			200	0.95	0.94	0.94	0.95	0	0
			500	0.95	0.95	0.93	0.93	0	0
			1000	0.95	0.96	0.95	0.95	0	0
		0.25	100	0.94	0.94	0.94	0.94	0	0
			200	0.95	0.95	0.96	0.96	0	0

TABLEAU 5.11 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald (PC_1, pour rappel), probabilité de couverture des intervalles de confiance à 95% de la vraisemblance profilée (PC_2), nombre de problèmes de maximisation (Pbs) et nombre de problèmes supplémentaires de maximisation liés à la vraisemblance profilée (Supp), pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$PC_1(\widehat{\theta_1})$	$PC_2(\widehat{\theta_1})$	$PC_1(\widehat{\theta_2})$	$PC_2(\widehat{\theta_2})$	Pbs	Supp
			500	0.95	0.95	0.95	0.94	0	0
			1000	0.95	0.95	0.95	0.95	0	0
0.05	2	0.75	100	0.92	0.95	0.94	0.94	0	0
			200	0.96	0.95	0.95	0.95	0	0
			500	0.96	0.96	0.96	0.96	0	0
			1000	0.95	0.95	0.94	0.95	0	0
		0.50	100	0.95	0.96	0.94	0.94	0	0
			200	0.95	0.95	0.95	0.96	0	0
			500	0.95	0.94	0.95	0.95	0	0
			1000	0.96	0.96	0.96	0.96	0	0
		0.25	100	0.95	0.95	0.94	0.94	0	0
			200	0.95	0.95	0.96	0.96	0	0
			500	0.95	0.95	0.94	0.94	0	0
			1000	0.95	0.95	0.95	0.95	0	0

TABLEAU 5.12 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : quantités Q_1 ou $Q_2(\theta_{2s})$, estimations des paramètres, intervalles de confiance à 95% de type Wald (IC_1) et intervalles de confiance à 95% de la vraisemblance profilée (IC_2).

Distribution	Q_1 ou Q_2	Paramètre	Estimation	IC_1	IC_2
Exponentielle	940	θ_1	0.0017	[0, 0.0044]	[0, 0.0044]
Weibull	1178	θ_1	0.0047	[0.0029, 0.0065]	[0.0019, 0.0062]
		θ_2	1.50	[1.09, 1.90]	[1.09,1.91]
Log-logistique	2358	θ_1	0.0041	[0.0011, 0.0070]	[0.0008, 0.0068]
		θ_2	1.53	[1.03, 2.04]	[1.06, 2.06]

95% sont rassemblés dans le tableau 5.12. Les intervalles de confiance asymptotiques à 95% sont larges et semblables pour les deux méthodes. La figure 5.7 montre les estimations des fonctions de risque instantané pour les trois distributions. La forme du risque instantané est bien différente d'une distribution à une autre. Cependant, déterminer quelle distribution serait un candidat raisonnable pour décrire nos données n'est pas simple.

5.5 Conclusion

Les propriétés asymptotiques énoncées au chapitre 3 permettent de calculer un intervalle de confiance asymptotique de type Wald pour chaque composante du paramètre vectoriel de la distribution étudiée. Une deuxième étude de simulations, présentée dans ce chapitre, a été mise en œuvre afin d'évaluer la qualité de cette estimation par intervalle. Un certain nombre d'échantillons générés ne vérifiaient pas la condition d'existence du maximum de vraisemblance qui a été mise en évidence théoriquement ou conjecturée au cours du chapitre 4. Dans ce cas, le calcul de l'intervalle de confiance était impossible. Le risque de survenue d'un problème d'existence du maximum de vraisemblance est plus élevé quand la proportion de données tronquées est importante ou quand la taille d'échantillon est petite. Lorsque l'intervalle de confiance pouvait être obtenu, cette étude de simulations a montré que les estimations des probabilités de couverture des intervalles de confiance de type Wald pouvaient être éloignées du niveau attendu, en

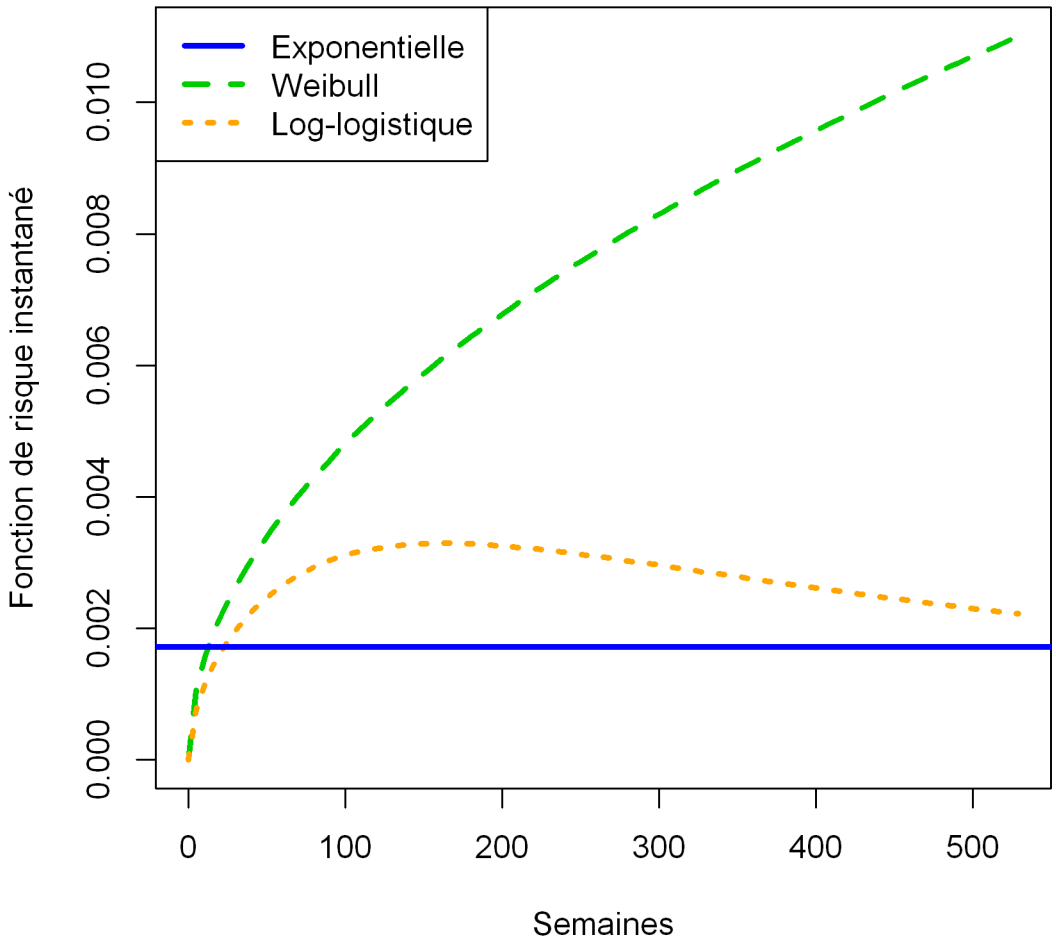

FIGURE 5.7 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : estimations paramétriques des fonctions de risque instantané pour les distributions exponentielle, de Weibull et log-logistique.

raison du biais de l'estimateur paramétrique du maximum de vraisemblance et d'un écart à la normalité. De manière générale, plus la proportion de données tronquées est élevée ou plus la taille d'échantillon est petite, plus l'écart entre l'estimation et le niveau attendu est important. Dans ces cas-là, la qualité de l'estimation par intervalle de confiance issu de la vraisemblance profilée est meilleure. Là encore, les propriétés asymptotiques du chapitre 3 justifient théoriquement le calcul de cet intervalle de confiance. Une étape de maximisation supplémentaire étant nécessaire pour le calcul de cet intervalle de confiance pour les distributions de Weibull et log-logistique (ce problème ne se pose pas pour la distribution exponentielle étant donné que le paramètre est de dimension 1), la survenue d'un échec de cette maximisation rend impossible le calcul de cet intervalle. Un tel échec s'est avéré cependant inexistant pour le calcul de l'intervalle de confiance du paramètre θ_1, extrêmement rare pour le calcul de l'intervalle de confiance du paramètre θ_2 de la loi log-logistique et survenant dans moins de 12% des cas pour le calcul de l'intervalle de confiance du paramètre θ_2 de la distribution de Weibull.

Chapitre 6

Quelques procédures d'adéquation

Pour chaque composante du paramètre vectoriel des distributions exponentielle, de Weibull et log-logistique, nous pouvons, à la lecture des chapitres précédents, énoncer les propriétés asymptotiques de l'estimateur du maximum de vraisemblance de cette composante et associer à l'estimation les intervalles de confiance de type Wald et de la vraisemblance profilée.

Pour notre jeu de données réelles constitué de 64 cas de lymphomes consécutifs à un traitement anti TNF-α, ces travaux ont été réalisés en parallèle pour les trois distributions, sans s'inquiéter jusque-là de savoir quelle était la famille de distributions en adéquation avec les données. Nous nous interrogeons désormais sur les procédures d'aide à la décision de la famille de lois sous-jacente à ces données. Classiquement, les méthodes disponibles se répartissent en deux catégories : les procédures graphiques et les tests d'adéquation.

Les procédures d'adéquation adaptées aux données tronquées à droite sont peu nombreuses. L'objectif de ce chapitre n'est pas pour autant de faire une revue exhaustive de la littérature des procédures d'adéquation qui s'appliqueraient aux données tronquées à droite. Nous présentons seulement quelques outils qui nous permettront de vérifier l'adéquation de nos données à un modèle exponentiel, de Weibull ou log-logistique.

Dans la suite de ce chapitre, on désigne par F^* (resp. S^*) la fonction de répartition

(resp. fonction de survie) de la distribution conditionnelle de X sachant que X est inférieur ou égal à t^* (le maximum des délais de troncature observés), d'expression

$$F^*(x) = \frac{F(x)}{F(t^*)} I\left\{0 < x \leqslant t^*\right\}. \tag{6.1}$$

6.1 Procédures graphiques

Plusieurs procédures graphiques se distinguent. Elles reposent, pour la plupart, sur la comparaison des estimations paramétriques et non-paramétriques du maximum de vraisemblance de la distribution de la variable aléatoire X. En présence de données tronquées à droite, l'estimateur non-paramétrique (1.5) ne permet d'estimer que la fonction de répartition F^* (6.1). L'estimation du maximum de vraisemblance du paramètre de la distribution supposée permet quant à elle d'estimer à la fois la fonction de répartition F et la fonction de répartition F^*. Une solution pour étendre ces procédures graphiques aux données tronquées à droite est donc de comparer les estimations paramétriques et non-paramétriques du maximum de vraisemblance de la fonction de répartition F^*.

6.1.1 Superposition de l'estimation non-paramétrique et de l'estimation paramétrique

Il s'agit de la procédure graphique la plus simple à mettre en œuvre. Elle consiste à superposer sur un même graphique les estimations paramétrique et non-paramétrique du maximum de vraisemblance de F^*. C'est la procédure d'adéquation graphique proposée par Lawless (2003) quand les données sont tronquées à droite. Elle a été mise en œuvre au chapitre 2 (cf. figure 2.1) quand nous voulions juger de l'adéquation des 64 cas de lymphomes aux modèles exponentiel, de Weibull et log-logistique. Nous avions alors noté que le modèle de Weibull semblait être un candidat raisonnable pour décrire les données.

Sur la figure 2.1 du chapitre 2 étaient superposées l'estimation non-paramétrique de

S^* ainsi que les estimations paramétriques de S (fonction de survie de X) et de S^* pour les trois modèles. La figure 6.1 reprend cette procédure graphique mais sans considérer l'estimation de S et en consacrant un graphique à chaque modèle. Cette figure confirme que la distribution de Weibull (resp. exponentielle) semble être, parmi ces trois lois, celle qui décrit le mieux (le moins bien) les données.

L'écart entre les deux estimations, qui repose sur l'évaluation d'un écart de courbure, est plus difficile à apprécier que l'écart entre un nuage de points et une droite. Les deux procédures graphiques suivantes comparent un nuage de points à une droite.

6.1.2 Graphique probabilité-probabilité

Nous proposons d'adapter la procédure graphique probabilité-probabilité de la manière suivante : cela consiste à tracer le nuage des points $\left(\widehat{F}^*\left(v_j\right), F^*\left(v_j;\widehat{\theta}_n\right)\right)$, où $\widehat{F}^*\left(\cdot\right)$ et $F^*\left(\cdot;\widehat{\theta}_n\right)$ sont respectivement les estimations non-paramétrique (1.5) et paramétrique de la fonction de répartition F^* et l'ensemble $(v_j)_{1\leqslant j\leqslant m}$ désigne les m valeurs distinctes et ordonnées des délais de survenue observés dans l'échantillon (comme lors de la présentation de l'estimateur non-paramétrique (1.5)).

Pour une taille d'échantillon assez grande et en cas d'adéquation entre le modèle paramétrique supposé et les données observées, le nuage de points se confond avec la première bissectrice. Cependant, du fait des fluctuations d'échantillonnage, les points du nuage ne sont presque jamais parfaitement alignés avec la première bissectrice.

En présence d'une variable aléatoire continue, une amélioration classique de cette procédure consiste à remplacer $\widehat{F}^*\left(v_j\right)$ par la quantité

$$\widehat{F}_j^{\#} = 0.5\widehat{F}^*\left(v_j\right) + 0.5\widehat{F}^*\left(v_j^-\right),$$

où $\widehat{F}^*\left(v_j^-\right)$ est la limite à gauche au point v_j de la fonction $\widehat{F}^*$.

La figure 6.2 rassemble les graphiques probabilité-probabilité ainsi que la procédure améliorée pour les trois distributions. L'amélioration de la procédure ne modifie que très légèrement les graphiques. Pour les distributions exponentielle et log-logistique, le

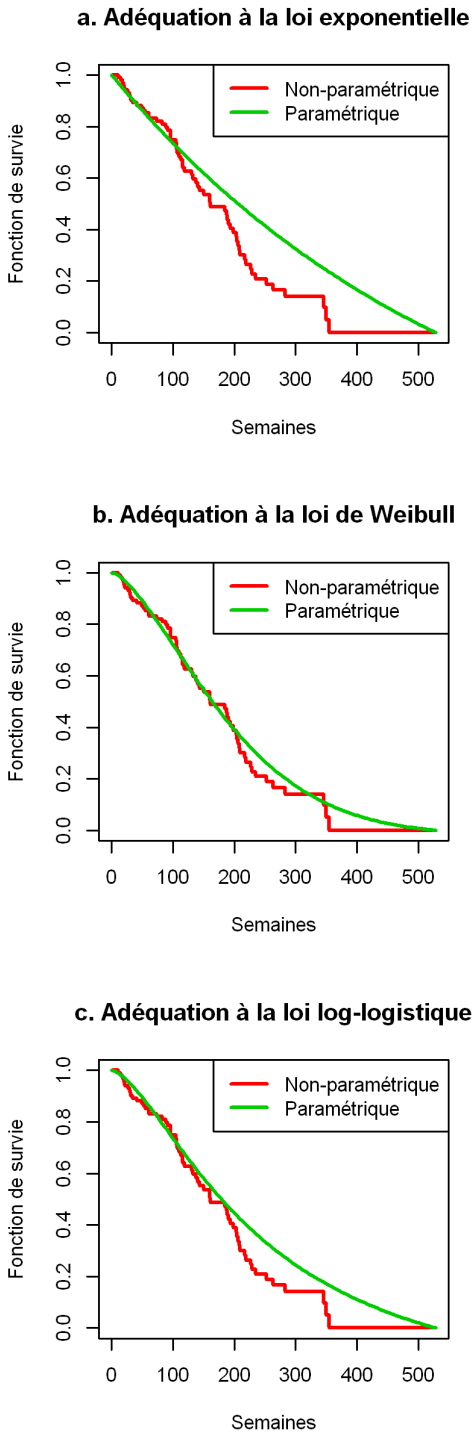

FIGURE 6.1 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : superposition des estimations non-paramétrique et paramétrique de la fonction de survie S^* ($t^* = 529$) pour les distributions exponentielle (a), de Weibull (b) et log-logistique (c).

nuage de points oscille autour de la première bissectrice pour les plus petites valeurs du délai de survenue et s'éloigne nettement de la première bissectrice pour les plus grandes valeurs du délai de survenue. Pour la distribution de Weibull, le nuage de points oscille autour de la première bissectrice pour presque toutes les valeurs du délai de survenue. Cette figure est en accord avec la figure 6.1.

Les deux procédures graphiques que nous venons de voir ont recourt à l'estimation paramétrique du maximum de vraisemblance. Ce sont des procédures qui permettent d'apprécier l'adéquation des données à la distribution particulière $F(\cdot;\widehat{\theta}_n)$. La procédure graphique suivante permet de s'affranchir de l'estimation paramétrique.

6.1.3 Procédure de linéarisation

Supposons que la fonction de répartition $F(\cdot;\theta)$ soit telle qu'il existe des fonctions g_1 et g_2 indépendantes de θ telles que $g_1\left(F\left(x;\theta\right)\right) = \lambda g_2\left(x\right) + \mu$. Cette procédure graphique repose sur la construction du nuage de points $\left(g_1\left(\widehat{F}(v_j)\right); g_2\left(v_j\right)\right)$ où $\widehat{F}$ est l'estimation non-paramétrique de F.

Pour une taille d'échantillon assez grande et en cas d'adéquation entre le modèle paramétrique supposé et les données observées, le nuage de points se confond, aux fluctuations d'échantillonnage près, avec la droite de coefficient directeur λ et d'ordonnée à l'origine μ.

Contrairement aux deux premières procédures graphiques, cette méthode ne fait pas appel à l'estimation du paramètre θ ; elle permet même de l'estimer puisque les coefficients λ et μ dépendent de θ et sont estimables graphiquement. Cette procédure graphique présente l'avantage d'explorer l'adéquation à une famille de lois.

Dans le cadre précis qui nous intéresse et qui est celui des données tronquées à droite, on cherche s'il existe des fonctions g_1 et g_2 indépendantes de θ telles que

$$g_1\left(F^*\left(x;\theta\right)\right) = \lambda g_2\left(x\right) + \mu, \tag{6.2}$$

et on construit le nuage de points $\left(g_1\left(\widehat{F}^*(v_j)\right); g_2\left(v_j\right)\right)$.

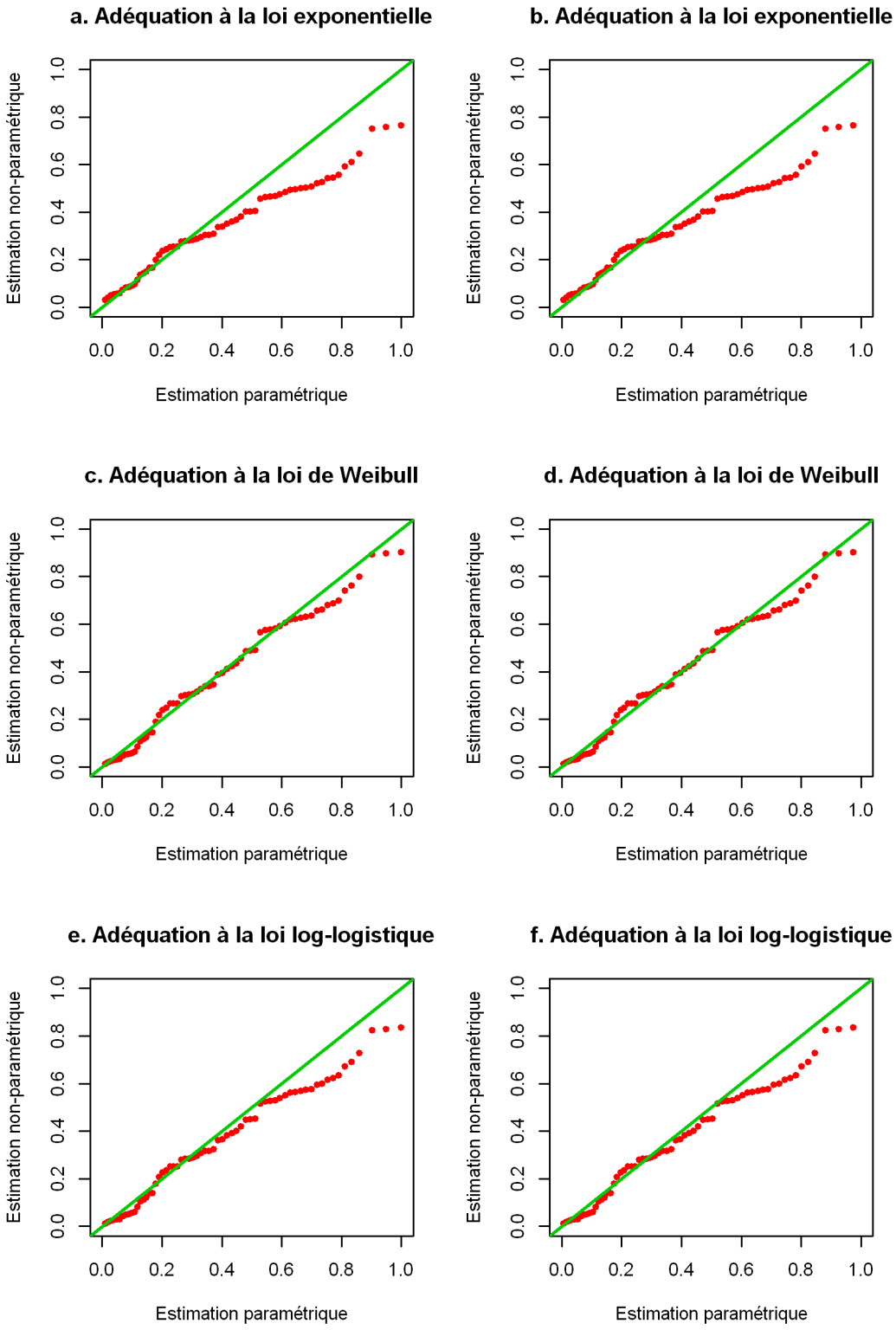

FIGURE 6.2 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : graphique probabilité-probabilité et superposition de la première bissectrice pour les distributions exponentielle (a), de Weibull (c) et log-logistique (e), et procédure améliorée pour les distributions exponentielle (b), de Weibull (d) et log-logistique (e).

Pour les trois familles de lois qui sont considérées dans cet exposé, il n'est pas possible de trouver un couple de fonctions g_1 et g_2 vérifiant l'équation (6.2).

6.1.4 Remarques

Toutes les procédures graphiques présentées ci-dessus sont fondées sur l'estimation de la fonction de répartition F^*. Toute autre fonction caractérisant la variable aléatoire X peut se substituer à la fonction de répartition. Il peut s'agir de la fonction de survie, de la fonction de risque instantané ou de la fonction de risque cumulé.

Ces méthodes sont basées sur les estimations paramétriques (sauf la procédure de linéarisation) et non-paramétriques du maximum de vraisemblance. Cependant, toute autre estimation pourrait se substituer à l'estimation du maximum de vraisemblance, dès lors que cet estimateur a de bonnes propriétés. Le "hazard plotting" par exemple, s'intéresse aux procédures graphiques fondées sur l'estimateur non-paramétrique de Nelson-Aalen de la fonction de risque cumulé et a été considéré pour le cas des données tronquées à gauche (Nelson, 1990). Pour l'estimation paramétrique, on pourrait considérer l'estimateur des moments.

Les deux premières procédures graphiques sont fondées sur l'estimation paramétrique du maximum de vraisemblance. Par conséquent, dans les cas où un problème de maximisation survient, ces méthodes ne peuvent pas être mises en œuvre.

6.1.5 Conclusion

Les procédures graphiques sont simples à mettre en œuvre, facilement interprétables et s'appliquent à toutes les distributions théoriques. Cependant, elles sont fondées sur l'appréciation de l'analyste et donc subjectives. Les tests d'adéquation permettent de discriminer de manière objective entre plusieurs familles de lois.

6.2 Tests d'adéquation

On considère toujours notre variable aléatoire X de fonction de répartition F. On distingue les tests d'une hypothèse nulle simple et les tests d'une hypothèse nulle composite. Les tests d'une hypothèse nulle simple vérifient l'adéquation à une distribution de probabilité particulière, c'est-à-dire

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0, \end{cases} \tag{6.3}$$

tandis que les tests d'une hypothèse nulle composite vérifient l'adéquation à une famille paramétrique de distributions de probabilité, c'est-à-dire

$$\begin{cases} H_0 : F \in \mathscr{F}_0 \\ H_1 : F \notin \mathscr{F}_0, \end{cases} \tag{6.4}$$

où $\mathscr{F}_0$ est un ensemble de distributions définies par un paramètre $\theta \in \Theta$.

6.2.1 Transformation par la fonction de répartition

Cette méthode, détaillée dans le cas général dans le chapitre 6 du livre de D'Agostino et Stephens (1986), permet de tester une hypothèse nulle simple et repose sur le théorème suivant :

Théorème 6.2.1. *Si X est une variable aléatoire continue de fonction de répartition F, alors $U = F(X)$ est une variable aléatoire continue de loi uniforme sur* $[0, 1]$.

Par conséquent, si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires continues indépendantes et identiquement distribuées de fonction de répartition F_0, alors les variables aléatoires $U_1 = F_0(X_1), U_2 = F_0(X_2), \ldots, U_n = F_0(X_n)$ sont des variables aléatoires continues indépendantes et identiquement distribuées de loi uniforme sur $[0, 1]$. On peut alors remplacer le test d'adéquation des données $(X_1, X_2, \ldots, X_n)$ à la fonction de ré-

partition F_0, par le test d'adéquation des données $(U_1, U_2, \ldots, U_n)$ à la loi uniforme sur $[0, 1]$.

Cette méthode s'étend facilement aux données tronquées à droite. Les délais de survenue $(x_1, x_2, \ldots, x_n)$ que nous observons sont des réalisations indépendantes mais non-identiquement distribuées des variables aléatoires continues $(X_1, X_2, \ldots, X_n)$ de distribution respective la fonction de répartition $F^{t_i}(x) = F(x)/F(t_i)$ pour tout $i = 1, \ldots, n$. Alors sous H_0, les variables aléatoires $F_0^{t_1}(X_1), F_0^{t_2}(X_2), \ldots, F_0^{t_n}(X_n)$ sont des variables aléatoires continues indépendantes et identiquement distribuées de loi uniforme sur $[0, 1]$. On a transformé un échantillon de données indépendantes mais non-identiquement distribuées en un échantillon de données indépendantes et identiquement distribuées de loi indépendante des paramètres de la distribution sous-jacente de départ.

On peut désormais utiliser tous les outils existants pour tester l'adéquation de données indépendantes et identiquement distribuées à une loi uniforme continue sur $[0, 1]$.

Mais pour notre application, nous n'avons pas d'idée précise de la distribution à tester. Nous voudrions réaliser les tests suivants :

$$1. \begin{cases} H_0 : X \text{ suit une loi exponentielle} \\ H_1 : X \text{ ne suit pas une loi exponentielle,} \end{cases} \tag{6.5}$$

$$2. \begin{cases} H_0 : X \text{ suit une loi de Weibull} \\ H_1 : X \text{ ne suit pas une loi de Weibull,} \end{cases} \tag{6.6}$$

$$3. \begin{cases} H_0 : X \text{ suit une loi log-logistique} \\ H_1 : X \text{ ne suit pas une loi log-logistique.} \end{cases} \tag{6.7}$$

Il s'agit dans ce cas de tester une hypothèse nulle qui n'est plus simple, mais composite. Le test d'adéquation le plus utilisé dans le cas général est le test du chi-2. La famille des tests d'adéquation du chi-2, lorsque le paramètre doit être estimé, inclut deux types d'approches dans le cas général (hors données tronquées) : soit c'est l'estimation du maximum de vraisemblance de l'échantillon observé qui est utilisée et alors

la distribution asymptotique de la statistique "usuelle" n'est pas une loi du chi-2 et doit être corrigée, soit on utilise comme estimateur la valeur du paramètre qui minimise la statistique et dans ce cas, la statistique est distribuée de manière asymptotique suivant une loi du chi-2. Une autre distinction peut être apportée, suivant que l'on fixe la partition du support, ou bien que l'on fixe les probabilités des intervalles constituant la partition. Le test qui suit, et qui s'applique aux données tronquées à droite, utilise l'estimateur qui minimise la statistique et s'appuie sur les probabilités des intervalles constituant la partition.

6.2.2 Généralisation du test du chi-2 de Pearson-Fisher

Li et Doss (1993) ont généralisé le test du chi-2 de Pearson-Fisher (Fisher, 1922; Fisher *et al.*, 1924) à tous les cas où un estimateur non-paramétrique de la fonction de répartition F de la variable aléatoire X est disponible et admet une certaine loi limite. L'estimateur non-paramétrique du maximum de vraisemblance (1.5) dans le cas de données tronquées à droite rentre dans ce cadre. D'ailleurs, les auteurs mettent en œuvre ce test pour les données sur le délai d'incubation du SIDA. Cependant, dans notre cas, l'estimateur non-paramétrique estime la fonction de répartition F^*. Cette généralisation du test du chi-2 de Pearson-Fisher permet la réalisation des tests suivants :

$$1.\ \begin{cases} H_0 : F^* \text{ appartient à l'ensemble des lois exponentielles tronquées} \\ \qquad \text{au temps } t^* \\ H_1 : F^* \text{ n'appartient pas à l'ensemble des lois exponentielles tronquées} \\ \qquad \text{au temps } t^*, \end{cases} \tag{6.8}$$

$$2.\ \begin{cases} H_0 : F^* \text{ appartient à l'ensemble des lois de Weibull tronquées} \\ \qquad \text{au temps } t^* \\ H_1 : F^* \text{ n'appartient pas à l'ensemble des lois de Weibull tronquées} \\ \qquad \text{au temps } t^*, \end{cases} \tag{6.9}$$

$$3.\begin{cases} H_0 : F^* \text{ appartient à l'ensemble des lois log-logistique tronquées} \\ \qquad \text{au temps } t^* \\ H_1 : F^* \text{ n'appartient pas à l'ensemble des lois log-logistique tronquées} \\ \qquad \text{au temps } t^*. \end{cases} \tag{6.10}$$

Les détails de la mise en œuvre de ce test sont les suivants. Pour tout n, soit $0 = a_0^n < a_1^n < \ldots < a_k^n = t^*$ une partition du support de la distribution supposée $F^*(\cdot;\theta)$ telle que a_i^n soit fonction de $\widehat{F}^*$, estimateur non-paramétrique (1.5) et converge en probabilité vers une constante a_i, avec $0 < a_1$ et $a_{k-1} < t^*$. Pour tout $i = 1, \ldots, k$, on définit

$$\begin{cases} p_i^n(\theta) = F^*(a_i^n;\theta) - F^*\left(a_{i-1}^n;\theta\right), \\ \widehat{p}_i = \widehat{F}^*(a_i^n) - \widehat{F}^*\left(a_{i-1}^n\right). \end{cases}$$

On note $p^n(\theta) = (p_i^n(\theta))_{1\leqslant i\leqslant k}$ et $\widehat{p} = (\widehat{p}_i(\theta))_{1\leqslant i\leqslant k}$ les vecteurs associés à ces quantités. Soit $D_n(\theta)$ une matrice appartenant à l'ensemble des matrices carrées de taille k, fonction de θ et des éléments composant la partition $(a_i^n)_{1\leqslant i\leqslant k}$ du support de $F^*(\cdot;\theta)$. Li et Doss (1993) proposent plusieurs possibilités pour cette matrice $D_n(\theta)$, notamment les deux matrices diagonales suivantes :

$$D_n(\theta) = I, \text{ matrice identité de taille k,} \tag{6.11}$$

$$D_n(\theta) = \operatorname{diag}\left((p_1^n(\theta))^{-1/2}, (p_2^n(\theta))^{-1/2}, \ldots, (p_k^n(\theta))^{-1/2}\right). \tag{6.12}$$

On définit les vecteurs suivants :

$$\begin{cases} \varphi_n(\theta) = \sqrt{n}\left(\widehat{p} - p^n(\theta)\right), \\ \xi_n(\theta) = D_n(\theta)\,\varphi_n(\theta). \end{cases}$$

L'estimation du paramètre θ utilisée dans ce test est la valeur $\widehat{\theta}_d$ qui minimise la quantité $\xi_n'(\theta)\,\xi_n(\theta)$, où $\xi_n'(\theta)$ désigne la transposée du vecteur $\xi_n(\theta)$. Dans le cas où on considère

la matrice (6.12) pour la matrice $D_n(\theta)$, la statistique qui est minimisée pour obtenir l'estimation du paramètre θ est la statistique de Pearson-Fisher.

On désigne par Σ^1 la matrice carrée de taille $k-1$ dont l'élément de rang (j,p), pour tout $1 \leqslant j,p \leqslant k-1$ est l'estimation (1.6) de la variance asymptotique de l'estimateur non-paramétrique (1.5) :

$$\Sigma^1_{jp} = \widehat{\mathrm{Cov}}(W^*(a^n_j), W^*(a^n_p)).$$

Soit la matrice $C_n(\theta)$ de taille $k \times r$ définie par

$$C_n(\theta) = D_n(\theta) \begin{pmatrix} \dfrac{\partial p^n_1(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial p^n_1(\theta)}{\partial \theta_r} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial p^n_k(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial p^n_k(\theta)}{\partial \theta_r} \end{pmatrix}.$$

Soient la matrice carrée $P_n(\theta)$ de taille k et J la matrice de taille $k \times (k-1)$ définies par

$$P_n(\theta) = I - C_n(\theta)[C'_n(\theta)C_n(\theta)]^{-1}C'_n(\theta),$$

$$J = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

On définit la matrice $\Sigma_n(\theta) = P_n(\theta)D_n(\theta)J\Sigma^1 J' D_n(\theta)P_n(\theta)$ et son estimation $\widehat{\Sigma}_n$ obtenue en remplaçant le paramètre θ par son estimation $\widehat{\theta}_d$. Soit $\widehat{\Sigma}^-_n$ l'inverse généralisée de $\widehat{\Sigma}_n$. On définit la statistique de test $Q = {\xi'}_n\left(\widehat{\theta}_d\right)\widehat{\Sigma}^-_n\xi_n\left(\widehat{\theta}_d\right)$. Alors, sous certaines

conditions de régularité, on a

$$Q \xrightarrow[n\to+\infty]{d} \chi^2_{k-r-1}.$$

Ce test d'adéquation a été mis en œuvre pour tester l'adéquation des 64 cas de lymphomes aux modèles exponentiel, de Weibull et log-logistique. Nous avons considéré une partition à 5 classes du support $[0, t^*]$ avec $t^* = 529$ semaines. Les éléments de la partition ont été choisis de manière à ce que chaque intervalle de la partition soit équiprobable vis-à-vis de l'estimation non-paramétrique (1.5), c'est-à-dire que pour tout $i = 1, \ldots, k$, on a $\widehat{p}_i = 0.2$ et on considère la partition $(a_i)_{0\leqslant i\leqslant k}$ telle que :

$$\widehat{F}^*(a_0^n) = 0, \widehat{F}^*(a_1^n) = 0.2, \widehat{F}^*(a_2^n) = 0.4, \widehat{F}^*(a_3^n) = 0.6, \widehat{F}^*(a_4^n) = 0.8, \widehat{F}^*(a_5^n) = 1.$$

On a donc pris les quantiles de la fonction de répartition $\widehat{F}^*$. On obtient la partition

$$a_o^n = 0, a_1^n = 88, a_2^n = 132, a_3^n = 198, a_4^n = 253, a_5^n = 529.$$

On a choisi $D_n(\theta) = I$. La fonction *ginv* du package **MASS** du logiciel R (R Core Team, 2012) a été utilisée pour obtenir l'inverse généralisée de la matrice $\Sigma_n(\theta)$. Le risque de première espèce a été fixé à 5%.

Le tableau 6.1 présente les résultats de ce test d'adéquation. Les estimations $\widehat{\theta}_n$ et $\widehat{\theta}_d$ sont différentes. En particulier, l'estimation $\widehat{\theta}_{2d}$ du paramètre θ_2 de la loi de Weibull est très proche de 1, ce qui n'était pas le cas de l'estimation du maximum de vraisemblance $\widehat{\theta}_{2n}$. En conséquence, les statistiques de test pour les distributions exponentielle et de Weibull sont identiques. Ce test d'adéquation rejette l'hypothèse nulle d'adéquation à une distribution exponentielle ou à une distribution de Weibull. Le test d'adéquation ne rejette pas l'hypothèse nulle d'adéquation à une distribution log-logistique et la p-value associée est bien loin du seuil fixé 0.05.

Le test d'adéquation rejette l'hypothèse nulle d'adéquation avec la distribution exponentielle, ce qui est en accord avec ce qui était suggéré par les procédures graphiques. Mais ce test d'adéquation rejette également l'hypothèse nulle d'adéquation à la dis-

tribution de Weibull et ne rejette pas l'hypothèse nulle d'adéquation à la distribution log-logistique, ce qui n'est pas cohérent avec les conclusions tirées des procédures graphiques.

TABLEAU 6.1 – Analyse des 64 cas de lymphomes consécutifs à un traitement anti TNF-α : estimation du paramètre minimisant une distance du "type" chi-2, statistique de test Q et p-value des tests d'adéquation aux modèles exponentiel, de Weibull et log-logistique.

Distribution	Exponentielle	Weibull	Log-logistique
$\widehat{\theta}_{1d}$	0.0036	0.0033	0.0061
$\widehat{\theta}_{2d}$	-	1.09	2.29
Q	9.31	9.31	1.85
ddl	4	3	3
p-value	0.025	0.009	0.396

Abréviation : ddl, degrés de liberté.

Discussion et perspectives

Nous avons considéré tout au long de cet exposé un médicament D (ou classe médicamenteuse) et un effet indésirable E_D, consécutif à la prise de ce médicament, particuliers. Nous souhaitions estimer la distribution du délai jusqu'à la survenue de l'effet indésirable E_D et évaluer l'intérêt des notifications spontanées pour mener cette estimation. Or, les données sur le délai jusqu'à la survenue de l'effet indésirable E_D issues des notifications spontanées sont tronquées à droite.

Nous avons choisi de travailler avec l'estimateur du maximum de vraisemblance. La présentation du cadre non-paramétrique a mis en évidence le défaut de l'estimateur non-paramétrique du maximum de vraisemblance. Celui-ci ne permet d'estimer la distribution du délai de survenue qu'à une constante multiplicative près inconnue. Nous nous sommes donc concentrés sur l'estimateur paramétrique du maximum de vraisemblance et nous avons exploré plus particulièrement les distributions exponentielle, de Weibull et log-logistique.

Les méthodes adaptées aux données tronquées à droite restent peu répandues dans la recherche biomédicale. Un estimateur naïf, c'est-à-dire qui ne prend pas en compte la troncature, est parfois utilisé à la place de l'estimateur pertinent. Nous avons mis en évidence, à l'aide d'une étude de simulations, les conséquences de la non prise en compte du caractère tronqué à droite des données sur l'estimation du maximum de vraisemblance du paramètre de la distribution supposée. Les deux estimateurs, naïf et basé sur la troncature à droite, peuvent être positivement biaisés mais le biais est bien moindre pour l'estimateur basé sur la troncature et il en va de même pour l'erreur

quadratique moyenne. De plus, le biais et l'erreur quadratique moyenne de l'estimateur pertinent diminuent nettement avec l'augmentation de la taille d'échantillon, ce qui n'est pas le cas de l'estimateur naïf.

Cette première étude de simulations a montré que l'estimateur paramétrique du maximum de vraisemblance semblait avoir de bonnes propriétés asymptotiques. Nous les avons explorées. Les conditions suffisantes pour que cet estimateur soit consistant et asymptotiquement normal ont été énoncées. Nous avons vérifié que ces conditions suffisantes étaient satisfaites quand le délai de survenue de l'effet indésirable médicamenteux suit la distribution la plus répandue en analyse des données de survie : la loi exponentielle. Au passage, nous avons mis en évidence une condition d'existence de l'estimation du maximum de vraisemblance. Nous n'avons pas pu vérifier que les conditions suffisantes étaient satisfaites pour les distributions de Weibull et log-logistique. Cependant, nous avons conjecturé une condition d'existence de l'estimation du maximum de vraisemblance pour ces deux distributions. Il pourrait être intéressant de prouver théoriquement cette conjecture. Nous ne pouvons pas proposer d'interprétation de cette condition d'existence même dans le cas plus simple de la distribution exponentielle. L'étude des discordances entre la survenue d'un problème de maximisation et la vérification de la condition d'existence a montré que calculer les quantités Q_1 ou Q_2 permettait de détecter les estimations qui ne sont pas pertinentes.

L'établissement des propriétés asymptotiques précédentes permet d'associer à l'estimation du paramètre de la distribution obtenue un intervalle de confiance asymptotique de type Wald. Nous avons étudié, à l'aide d'une seconde étude de simulations, la qualité de l'estimation par intervalle de type Wald de l'estimateur paramétrique du maximum de vraisemblance. En raison du biais de l'estimateur du paramètre, d'un écart à la normalité et du biais de l'estimateur de la variance asymptotique, la probabilité de couverture de l'intervalle de confiance de type Wald n'est pas toujours satisfaisante. L'établissement des propriétés asymptotiques précédentes permet de vérifier que le rapport de vraisemblance suit toujours asymptotiquement une distribution du chi-2 quand les données sont tronquées à droite. Nous avons donc étudié les performances des inter-

valles de confiance fondés sur la vraisemblance profilée. Et dans les cas où la couverture des intervalles de confiance de Wald était éloignée du niveau attendu, la probabilité de couverture associée à l'intervalle de confiance de la vraisemblance profilée était nettement plus satisfaisante.

La méthode basée sur la vraisemblance profilée est moins sensible à la paramétrisation que les intervalles de confiance de type Wald. Choisir la meilleure paramétrisation n'est pas simple mais cela pourrait améliorer la normalité de l'estimateur à taille d'échantillon finie et ainsi améliorer la probabilité de couverture des intervalles de confiance de type Wald. Par exemple, pour que les supports des distributions supposées soient en accord avec l'échelle des temps que l'on rencontre en pharmacovigilance, des valeurs très proches de zéro ont été considérées pour le paramètre (d'échelle pour les distributions de Weibull et log-logistique). Par conséquent, les estimations sont proches de zéro, qui est une des bornes de l'intervalle des valeurs possibles pour le paramètre. Changer de paramétrisation afin que le paramètre (d'échelle) soit plus éloigné de cette borne de l'intervalle pourrait réduire l'écart à la normalité de l'estimateur.

Dans la dernière partie de cet exposé, nous avons présenté quelques procédures d'adéquation adaptées aux données tronquées à droite. Nous avons examiné trois procédures graphiques, un test d'adéquation pour une hypothèse nulle simple et un test d'adéquation pour une hypothèse nulle composite. Ce dernier est une généralisation du test du chi-2 de Pearson-Fisher. Ces procédures d'adéquation ont été appliquées à notre base de données constituée de 64 cas de lymphomes consécutifs à un traitement anti TNF-α. Les procédures graphiques désignent le modèle de Weibull comme étant le meilleur modèle pour décrire les données mais la généralisation du test du chi-2 rejette l'hypothèse nulle d'adéquation au modèle exponentiel et l'hypothèse nulle d'adéquation au modèle de Weibull mais ne rejette pas l'hypothèse nulle d'adéquation au modèle log-logistique.

Ce test d'adéquation est complexe à mettre en œuvre. Une nouvelle estimation du paramètre vectoriel θ est nécessaire. Cette estimation est basée sur les données groupées par intervalle de la partition et minimise une distance du "type" chi-2. Les

estimations du maximum de vraisemblance et du “type” chi-2 sont très différentes pour notre application. En particulier, l'estimation du “type” chi-2 du paramètre de forme est très proche de 1 pour la loi de Weibull, ce qui n'était pas le cas de l'estimation du maximum de vraisemblance et ce qui conduit au rejet du modèle de Weibull par ce test. Dans le cas des données indépendantes et identiquement distribuées, il existe un test du chi-2 corrigé qui utilise l'estimation du maximum de vraisemblance du paramètre de la distribution testée. Ce test est plus simple à mettre en œuvre. L'extension de ce chi-2 corrigé aux données tronquées à droite est envisagée. Il pourrait être intéressant de mener une étude de simulations afin de comparer les performances de ces deux tests ou de faire une étude de sensibilité afin d'étudier l'influence du choix de la partition.

De manière générale, les tests du chi-2 sont souvent moins puissants que d'autres types de tests (D'Agostino et Stephens, 1986). Cependant, les tests d'adéquation adaptés aux données tronquées à droite sont peu nombreux. Hwang et Wang (2008) ont proposé un autre test du chi-2 basé sur l'estimation du maximum de vraisemblance de l'échantillon observé. Guilbaud (1988) a étendu le test de Kolmogorov-Smirnov aux données tronquées à gauche et censurées à droite. Nous envisageons d'évaluer la procédure de test consistant à effectuer le test de Guilbaud après avoir préalablement appliqué la méthode d'inversion du temps de Lagakos *et al.* (1988) pour se ramener à des données tronquées à gauche. Une perspective de ce travail serait donc la construction de tests d'adéquation adaptés aux données tronquées à droite.

Tout au long de cet exposé, nous nous sommes concentrés sur les distributions exponentielle, de Weibull et log-logistique mais il serait utile d'examiner d'autres modèles comme les distributions gamma ou log-normale ou encore les modèles de mélange. Dans des situations plus complexes, le traitement peut être composé de plusieurs médicaments, chacun d'entre-eux induisant l'effet indésirable mais dans une fenêtre de temps différente. Dans ce cas, les variations au cours du temps de la fonction de risque instantané peuvent être multiples et une famille de distributions plus complexes pourrait se révéler utile.

Quand la condition d'existence est satisfaite mais qu'aucune estimation du maxi-

mum de vraisemblance est obtenue, la mise en œuvre de l'algorithme EM (Dempster *et al.*, 1977) résoudrait peut-être le problème. Quand la condition d'existence n'est pas satisfaite, il n'est pas possible d'obtenir l'estimation du maximum de vraisemblance. Dans ces cas-là, une approche alternative comme le bootstrap peut être explorée. De manière générale, l'estimateur du maximum de vraisemblance pouvant être positivement biaisé et de biais élevé à taille d'échantillon finie, il serait intéressant de considérer d'autres estimateurs paramétriques, tels que l'estimateur des moments ou l'estimateur des quantiles, et de comparer les performances de ces approches alternatives avec celles de l'estimateur du maximum de vraisemblance.

De manière non-paramétrique, on ne peut estimer la distribution du délai de survenue qu'à une constante multiplicative près inconnue. Sans autre source de données, cette constante multiplicative n'est pas identifiable. On pourrait imaginer utiliser les données de vente ou de prescription du médicament D pour estimer la taille N de la population exposée au médicament D. En considérant tous les conditionnements du médicament D disponibles sur le marché, à partir du nombre d'unités vendues pour chacun des conditionnements depuis la mise sur le marché et après avoir estimé la durée moyenne d'un traitement, une estimation de la taille N de la population exposée au médicament D peut être obtenue. Cette estimation de N permet d'estimer la constante multiplicative inconnue et ainsi estimer de manière non-paramétrique la fonction de répartition F que l'on souhaitait estimer initialement. En examinant les propriétés asymptotiques de l'estimateur de N, on pourra étudier les propriétés asymptotiques de ce nouvel estimateur non-paramétrique. Cependant, la population de taille N est constituée des individus exposés au médicament D qui ont présenté ou présenteront l'effet indésirable avant de décéder. Les données de vente nous permettent d'estimer la taille de la population exposée au médicament D. Mais il est peu probable que tous les individus exposés au médicament D présenteront l'effet indésirable avant de décéder. Pour pouvoir utiliser cette estimation de N, peut-être faudrait-il compléxifier notre modèle et intégrer la proportion inconnue de la population qui ne présentera jamais l'effet indésirable. Autrement dit, un modèle à taux de guérison devrait peut-être être considéré. La question

de l'identifiabilité de ce modèle se poserait alors.

Les délais de survenue et de troncature sont calculés à partir de la date d'exposition au médicament. Quand les effets indésirables surviennent à long terme, comme c'est le cas par exemple pour des lymphomes, la date d'initiation du traitement peut ne pas être connue exactement mais au mois près ou à l'année près. Dans ces cas-là, les délais de survenue et de troncature, en plus d'être tronqués, sont censurés par intervalle. Une extension de ce travail aux données tronquées et censurées par intervalle est envisagée. Mais il faudrait peut-être plutôt s'intéresser à l'estimateur semi-paramétrique du maximum de vraisemblance du vecteur aléatoire (X, T).

Nos données sont constituées uniquement des individus exposés au médicament D qui ont présenté l'effet indésirable E_D avant la date d'analyse. Cependant, la base de pharmacovigilance française contient tous les effets indésirables de tous les médicaments. Il serait donc envisageable de collecter tous les individus exposés au médicament D qui ont présenté un effet indésirable, qu'il s'agisse de l'effet indésirable E_D ou d'un autre. Les cas d'effets indésirables autres que E_D devraient être traités comme des données censurées et tronquées. La vraisemblance se compliquerait mais cela permettrait d'augmenter la taille d'échantillon et de mieux estimer la queue de distribution du délai de survenue.

Les notifications spontanées résultent de trois procédés consécutifs : la survenue de l'effet indésirable, son diagnostic et sa notification. La sous-notification est importante, même pour les effets indésirables graves. Parmi les causes de la sous-notification, on compte la sévérité de l'effet indésirable, l'âge du patient et l'ancienneté de l'effet indésirable mais aussi des facteurs dépendant du temps comme la durée écoulée depuis la mise sur le marché ou la durée d'exposition (Weber, 1986; Tubert-Bitter *et al.*, 1998; Haramburu *et al.*, 1997; Moride *et al.*, 1997; Bégaud *et al.*, 2002). Dans l'approche proposée ici, la sous-notification est supposée uniforme. Une telle hypothèse peut ne pas toujours être acceptable. Cependant, avec des effets indésirables à long terme comme les lymphomes et une surveillance homogène durant toute la période de mise sur le marché du médicament, cette hypothèse d'uniformité est vraisemblable.

Le délai de notification peut parfois être très long. Au moment de l'analyse, certaines observations tronquées peuvent l'être à cause du délai de notification et non pas à cause du délai de survenue. Il pourrait être intéressant de modéliser ce délai de notification et de le prendre en compte dans la modélisation du délai de survenue.

A plus long terme, il serait bien d'aborder deux autres des enjeux de l'analyse des données de survie : la comparaison des distributions du délai de survenue de deux médicaments ou l'établissement des facteurs de risque à l'aide d'un modèle de régression.

Bien que ces travaux aient été menés dans le cadre de la pharmacovigilance, les développements théoriques et les résultats des simulations peuvent être utilisés pour toute analyse rétrospective réalisée à partir d'un registre de cas, où les données sur un délai de survenue sont aussi tronquées à droite.

Références

Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P. et Tubert-Bitter, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*, 66(1):301–309.

Ahmed, I., Haramburu, F., Fourrier-Réglat, A., Thiessard, F., Kreft-Jais, C., Miremont-Salamé, G., Bégaud, B. et Tubert-Bitter, P. (2009). Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Statistics in Medicine*, 28(13):1774–1792.

Arme-p (1992). *Analyse d'incidence en pharmacovigilance. Application à la notification spontanée. 2nde Edition*. ARME-PHARMACOVIGILANCE Editions, Bordeaux, France.

Bartlett, M. S. (1953). On the statistical estimation of mean life-times. *Philosophical Magazine*, 44(350):249–262.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A. et De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321.

Bégaud, B., Martin, K., Haramburu, F. et Moore, N. (2002). Rates of spontaneous reporting of adverse drug reactions in France (Letter). *Journal of the American Medical Association*, 288(13):1588.

Bradley, R. et Gart, J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214.

Brookmeyer, R. et Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, 83(402):301–308.

Chanda, K. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, 41(1/2):56–61.

Cornelius, V. R., Sauzet, O. et Evans, S. J. W. (2012). A signal detection method to detect adverse drug reactions using a parametric time-to-event model in simulated cohort data. *Drug Safety*, 35(7):599–610.

Cox, D. R. et Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall.

D'Agostino, R. B. et Stephens, M. A. (1986). *Goodness-of-fit techniques*. Marcel Dekker.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190.

Efron, B. et Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator : observed versus expected fisher information. *Biometrika*, 65:457–482.

Efron, B. et Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Evans, S. J. W., Waller, P. C. et Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486.

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, 51:228–235.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I, 3rd Edition*. John Wiley & Sons.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume II, 2nd Edition*. John Wiley & Sons.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fisher, R. A. *et al.* (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society*, 87:442–450.

Fourrier, A., Bégaud, B., Alpérovitch, A., Verdier-Taillefer, M.-H., Touzé, E., Decker, N. et Imbs, J.-L. (2001). Hepatitis B vaccine and first episodes of central nervous system demyelinating disorders : a comparison between reported and expected number of cases. *British Journal of Clinical Pharmacology*, 51(5):489–490.

Gross, S. T. et Huber-Carol, C. (1992). Regression models for truncated survival data. *Scandinavian Journal of Statistics*, pages 193–213.

Gross, S. T. et Lai, T. L. (1996). Bootstrap methods for truncated and censored data. *Statistica Sinica*, 6:509–530.

Guilbaud, O. (1988). Exact kolmogorov-type tests for left-truncated and/or right-censored data. *Journal of the American Statistical Association*, 83(401):213–221.

Haramburu, F., Bégaud, B. et Moride, Y. (1997). Temporal trends in spontaneous reporting of unlabelled adverse drug reactions. *British Journal of Clinical Pharmacology*, 44(3):299–301.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics*, 42(6):1977–1991.

Hwang, Y.-T. et Wang, C.-c. (2008). A goodness of fit test for left-truncated and right-censored data. *Statistics & Probability Letters*, 78(15):2420–2425.

Ibragimov, I. A. et Khas'Minskii, R. Z. (1972). Asymptotic behavior of statistical estimators in the smooth case. I. study of the likelihood ratio. *Theory of probability and its applications*, 17(3):445–462.

Ibragimov, I. A. et Khas'Minskii, R. Z. (1975a). Local asymptotic normality for non-identically distributed observations. *Theory of probability and its applications*, 20(2):246–260.

Ibragimov, I. A. et Khas'Minskii, R. Z. (1975b). Properties of maximum likelihood and bayes' estimators for non-identically distributed observations. *Theory of probability and its applications*, 20(4):689–697.

Ibragimov, I. A. et Khas'Minskii, R. Z. (1981). *Statistical Estimation : Asymptotic Theory*. Springer-Verlag.

Kalbfleisch, J. et Lawless, J. (1989). Inference based on retrospective ascertainment : an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, 84(406):360–372.

Kalbfleisch, J. D. et Lawless, J. F. (1988). Estimating the incubation period for AIDS patients. *Nature*, 333:504–505.

Kalbfleisch, J. D. et Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32.

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

Keiding, N. et Gill, R. D. (1990). Random truncation models and Markov processes. *The Annals of Statistics*, 18(2):582–602.

Lagakos, S. W., Barraj, L. M. et De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data, 2nd Edition*. John Wiley & Sons.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley & Sons.

Li, G. et Doss, H. (1993). Generalized Pearson-Fisher chi-square goodness-of-fit tests, with applications to models with life history data. *The Annals of Statistics*, 21(2):772–797.

Lui, K.-J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W. et Bregman, D. J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proceedings of the National Academy of Sciences*, 83(May):3051–3055.

Maclure, M. (1991). The case-crossover design : a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133(2):144–153.

Maignen, F., Hauben, M. et Tsintis, P. (2010). Modelling the time to onset of adverse reactions with parametric survival distributions. *Drug Safety*, 33(5):417–434.

Medley, G. F., Anderson, R. M., Cox, D. R. et Billard, L. (1987). Incubation period of AIDS in patients infected via blood transfusion. *Nature*, 328:719–721.

Medley, G. F., Billard, L., Cox, D. R. et Anderson, R. M. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS). *Proceedings of the Royal Society of London. Series B. Biological sciences*, 233(1272):367–377.

Moore, N., Kreft-Jais, C., Haramburu, F., Noblet, C., Andrejak, M., Ollagnier, M. et Bégaud, B. (1997). Reports of hypoglycaemia associated with the use of ACE inhibitors and other drugs : a case/non-case study in the French pharmacovigilance system database. *British Journal of Clinical Pharmacology*, 44(5):513–518.

Moride, Y., Haramburu, F., Requejo, A. A. et Bégaud, B. (1997). Underreporting of adverse drug reactions in general practice. *British Journal of Clinical Pharmacology*, 43(2):177–181.

Nelson, W. (1990). Hazard plotting of left truncated life data. *Journal of Quality Technology*, 22(3):230–238.

Nordberg, L. (1980). Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scandinavian Journal of Statistics*, 7:27–32.

Pawitan, Y. (2001). *In All Likelihood : Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

R Core Team (2012). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Szarfman, A., Machado, S. G. et O'Neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety*, 25(6):381–392.

Théophile, H., Schaeverbeke, T., Miremont-Salamé, G., Abouelfath, A., Kahn, V., Haramburu, F. et Bégaud, B. (2011). Sources of information on lymphoma associated with anti-tumour necrosis factor agents. *Drug Safety*, 34(7):577–585.

Tsai, W.-Y., Jewell, N. P. et Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886.

Tubert, P., Bégaud, B., Haramburu, F. et Péré, J. C. (1991). Spontaneous reporting : how many cases are required to trigger a warning ? *British Journal of Clinical Pharmacology*, 32(4):407–408.

Tubert-Bitter, P., Haramburu, F., Bégaud, B., Chaslerie, A., Abraham, E. et Hagry, C. (1998). Spontaneous reporting of adverse drug reactions : who reports and what ? *Pharmacoepidemiology and Drug Safety*, 7(5):323–329.

Van der Heijden, P. G., Van Puijenbroek, E. P., Van Buuren, S. et Van der Hofstede, J. W. (2002). On the assessment of adverse drug reactions from spontaneous reporting systems : the influence of under-reporting on odds ratios. *Statistics in Medicine*, 21(14):2027–2044.

Van der Vaart, A. W. et Wellner, J. A. (2000). *Weak convergence and Empirical Processes*. Springer-Verlag.

Van Holle, L., Zeinoun, Z., Bauchau, V. et Verstraeten, T. (2012). Using time-to-onset for detecting safety signals in spontaneous reports of adverse events following immunization : a proof of concept study. *Pharmacoepidemiology and Drug Safety*, 21(6):603–610.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, 13(1):178–203.

Wang, M. C. (1987). Product-limit estimates : A generalized maximum likelihood study. *Communication in Statistics. Theory and Methods*, 16:3117–3132.

Wang, M.-C., Jewell, N. P. et Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, 14(4): 1597–1605.

Weber, J. C. P. (1986). *Mathematical models in adverse drug reaction assessment*. Oxford University Press.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177.

Annexe A

Tableaux A.1, A.2 et A.3

Tableau A.1 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution exponentielle : proportion des réplications où l'estimateur du maximum de vraisemblance est supérieur à la vraie valeur du paramètre, pour les estimateurs naïf et basé sur la troncature à droite, pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

θ_1	κ	n	Estimateur naïf	EBT
0.05	0.25	100	100%	61.6%
		500	100%	55.3%
0.05	0.50	100	100%	55.3%
		500	100%	50.4%
0.05	0.80	100	100%	51.1%
		500	100%	51.7%
1	0.25	100	100%	54.8%
		500	100%	50.7%
1	0.50	100	100%	53.2%
		500	100%	48.0%
1	0.80	100	100%	50.0%
		500	100%	51.0%

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature.

TABLEAU A.2 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution de Weibull : proportion des réplications où l'estimateur du maximum de vraisemblance est supérieur à la vraie valeur du paramètre, pour les estimateurs naïf et basé sur la troncature à droite, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

				Estimateur naïf		EBT	
θ_1	θ_2	κ	n	$\widehat{\theta}_{\text{EN},1} > \theta_1$	$\widehat{\theta}_{\text{EN},2} > \theta_2$	$\widehat{\theta}_{\text{EBT},1} > \theta_1$	$\widehat{\theta}_{\text{EBT},2} > \theta_2$
0.05	0.5	0.25	100	100%	100%	81.4%	71.9%
			500	100%	100%	64.6%	64.5%
0.05	0.5	0.50	100	100%	100%	63.3%	60.1%
			500	100%	100%	53.4%	51.0%
0.05	0.5	0.80	100	100%	99.6%	52.0%	53.3%
			500	100%	100%	48.6%	51.6%
1	0.5	0.25	100	100%	100%	79.3%	76.0%
			500	100%	100%	62.0%	61.2%
1	0.5	0.50	100	100%	100%	65.9%	64.6%
			500	100%	100%	53.8%	51.8%
1	0.5	0.80	100	100%	99.5%	52.7%	52.2%
			500	100%	100%	51.9%	50.6%
0.05	2	0.25	100	100%	98.1%	52.1%	61.6%
			500	100%	100%	52.2%	53.7%
0.05	2	0.50	100	100%	94.2%	51.6%	53.3%
			500	100%	100%	50.6%	51.0%
0.05	2	0.80	100	100%	85.4%	56.1%	55.8%
			500	100%	97.9%	52.2%	49.6%
1	2	0.25	100	100%	98.2%	56.2%	62.5%
			500	100%	99.9%	50.1%	54.8%
1	2	0.50	100	100%	94.3%	53.9%	54.2%
			500	100%	99.9%	47.1%	48.1%
1	2	0.80	100	100%	85.3%	54.1%	54.2%
			500	100%	97.9%	52.7%	52.2%

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature.

TABLEAU A.3 – Résultats de l'étude de simulations n° 1 (1000 réplications) pour la distribution log-logistique : proportion des réplications où l'estimateur du maximum de vraisemblance est supérieur à la vraie valeur du paramètre, pour les estimateurs naïf et basé sur la troncature à droite, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois probabilités κ qu'une réalisation du délai de survenue X appartienne à l'intervalle des valeurs observables de X et deux tailles d'échantillon n.

				Estimateur naïf		EBT	
θ_1	θ_2	κ	n	$\widehat{\theta}_{\text{EN},1} > \theta_1$	$\widehat{\theta}_{\text{EN},2} > \theta_2$	$\widehat{\theta}_{\text{EBT},1} > \theta_1$	$\widehat{\theta}_{\text{EBT},2} > \theta_2$
0.05	0.5	0.25	100	100%	100%	67.2%	67.7%
			500	100%	100%	53.6%	52.0%
0.05	0.5	0.50	100	100%	100%	55.4%	57.5%
			500	100%	100%	51.1%	52.0%
0.05	0.5	0.80	100	100%	100%	51.1%	53.2%
			500	100%	100%	50.8%	51.5%
1	0.5	0.25	100	100%	100%	67.7%	66.1%
			500	100%	100%	55.9%	56.1%
1	0.5	0.50	100	100%	100%	54.9%	57.2%
			500	100%	100%	53.4%	53.4%
1	0.5	0.80	100	100%	100%	55.1%	56.5%
			500	100%	100%	51.9%	52.0%
0.05	2	0.25	100	100%	100%	53.2%	55.9%
			500	100%	100%	51.8%	51.8%
0.05	2	0.50	100	100%	100%	55.0%	54.2%
			500	100%	100%	53.3%	52.2%
0.05	2	0.80	100	100%	100%	50.3%	51.5%
			500	100%	100%	53.9%	54.4%
1	2	0.25	100	100%	100%	52.7%	56.1%
			500	100%	100%	53.3%	51.0%
1	2	0.50	100	100%	100%	54.3%	56.4%
			500	100%	100%	50.1%	49.5%
1	2	0.80	100	100%	100%	52.0%	53.7%
			500	100%	100%	52.9%	55.0%

Abréviations : EN, estimateur naïf ; EBT, estimateur basé sur la troncature.

Annexe B

Démonstration du théorème 3.2.1

Démonstration. D'après la formule de Taylor-Lagrange, d'après l'hypothèse 3.2.3, pour tout $i = 1, \ldots, n$ et pour tout $\theta \in \Theta$, on peut écrire

$$\begin{aligned}\log f_i(x_i;\theta) = \log f_i(x_i;\theta^0) &+ \sum_{j=1}^{r} \left.\frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\right|_{\theta^0} (\theta_j - \theta_j^0) \\ &+ \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} \left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \\ &+ \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} \left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0),\end{aligned}$$

où θ' appartient au segment $]\theta^0, \theta[$.

Remarque B.0.1. L'écriture de cette formule nécessite que le segment $]\theta^0, \theta[$ appartienne à Θ, ce qui est toujours vrai si Θ est un ensemble convexe.

En sommant pour i allant de 1 à n, on a pour tout $\theta \in \Theta$,

$$\begin{aligned}\sum_{i=1}^{n} \log f_i(x_i;\theta) = \sum_{i=1}^{n} \log f_i(x_i;\theta^0) &+ \sum_{i=1}^{n}\sum_{j=1}^{r} \left.\frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\right|_{\theta^0} (\theta_j - \theta_j^0) \\ &+ \sum_{i=1}^{n} \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} \left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \\ &+ \sum_{i=1}^{n} \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} \left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0).\end{aligned}$$

Par inversion des sommes et multiplication des deux membres par $1/n$, on obtient pour tout $\theta \in \Theta$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log f_i(x_i;\theta) - \frac{1}{n}\sum_{i=1}^{n} \log f_i(x_i;\theta^0) = \sum_{j=1}^{r} (\theta_j - \theta_j^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\right|_{\theta^0} \\ + \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} \\ + \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'}.\end{aligned}$$

De plus, on a $(1/n)\sum_{i=1}^{n} \log f_i(x_i;\theta) = (1/n)\mathrm{logL}(x_1,\ldots,x_n;\theta)$, que nous noterons $(1/n)$ $\mathrm{logL}(\theta)$ pour la suite de la démonstration. Ainsi, pour tout $\theta \in \Theta$,

$$\begin{aligned}\frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) = \sum_{j=1}^{r} (\theta_j - \theta_j^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial \log f_i(x_i;\theta)}{\partial \theta_j}\right|_{\theta^0} \\ + \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} \\ + \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0) \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'}.\end{aligned}$$

En ajoutant et en retranchant un même terme dans le membre de droite, on obtient pour tout $\theta \in \Theta$,

$$\frac{1}{n}\mathrm{logL}(\theta)-\frac{1}{n}\mathrm{logL}(\theta^0)=\sum_{j=1}^{r}(\theta_j-\theta_j^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \mathrm{log}f_i(x_i;\theta)}{\partial\theta_j}\right|_{\theta^0}$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2\mathrm{log}f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}-\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]$$
$$+\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^3\mathrm{log}f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}\right]. \quad \text{(B.1)}$$

Considérons séparément chaque terme du membre droit de l'équation (B.1). D'après la remarque 3.2.4, nous avons déjà, pour tout $j=1,\ldots,r$,

$$\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \mathrm{log}f_i(X_i;\theta)}{\partial\theta_j}\right|_{\theta^0}\xrightarrow[n\to+\infty]{P}\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial \mathrm{log}f_i(X_i;\theta)}{\partial\theta_j}\right|_{\theta^0}\right)=0.$$

D'après l'hypothèse 3.2.6, nous avons pour tout $(j,p)\in\{1,\ldots,r\}^2$,

$$\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}-\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\xrightarrow[n\to+\infty]{P}0.$$

Pour tout $(j,p)\in\{1,\ldots,r\}^2$, nous avons pour tout $\theta\in\Theta$,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)$$
$$=-\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(-\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right).$$

D'après l'hypothèse 3.2.9, il s'agit d'une forme quadratique définie négative. Ainsi nous avons, pour tout $(j,p)\in\{1,\ldots,r\}^2$ et pour tout $\theta\in\Theta$ différent de θ^0,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)<0.$$

D'après l'hypothèse 3.2.7, pour tout $(j,p,q) \in \{1,\ldots,r\}^3$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\bigg|_{\theta'}\right| \xrightarrow[n\to+\infty]{P} \left|\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta'}\left(\frac{\partial^3 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\bigg|_{\theta'}\right)\right|,$$

et d'après l'hypothèse 3.2.8, nous avons pour tout $(j,p,q) \in \{1,\ldots,r\}^3$,

$$\left|\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta'}\left(\frac{\partial^3 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\bigg|_{\theta'}\right)\right| < K.$$

Soit (ζ,ϵ) un vecteur de constantes positives arbitraires. Les résultats précédents permettent d'écrire trois groupes d'inégalités. Pour tout ζ, pour tout ϵ, il existe n_0 tel que pour tout n supérieur ou égal à n_0 et pour tout $(j,p,q) \in \{1,\ldots,r\}^3$, les probabilités suivantes

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f_i(X_i;\theta)}{\partial\theta_j}\bigg|_{\theta^0}\right| \geqslant \zeta^2\right),$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0} - \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right| \geqslant \zeta\right),$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\bigg|_{\theta'}\right| \geqslant 2K\right),$$

sont majorées par $\epsilon/\left(r\left(1+r+r^2\right)\right)$. Soit S l'événement satisfaisant les $r(1+r+r^2)$ inégalités suivantes :

$$\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f_i(X_i;\theta)}{\partial\theta_1}\bigg|_{\theta^0}\right| < \zeta^2, \cdots, \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_i(X_i;\theta)}{\partial\theta_r\partial\theta_r\partial\theta_r}\bigg|_{\theta'}\right| < 2K\right\}.$$

Les majorations des probabilités précédentes nous donnent $P(S^*) < \epsilon$, où S^* est le complémentaire de S, c'est-à-dire l'événement où au moins une des inégalités précédentes n'est pas satisfaite. Nous avons ainsi $P(S) > 1-\epsilon$.

Maintenant, étudions le signe de la quantité $(1/n)\mathrm{logL}(\theta) - (1/n)\mathrm{logL}(\theta^0)$ pour les éléments de S et pour θ appartenant à la sphère $S\left(\theta^0,\zeta\right)$ de centre θ^0 et de rayon ζ. Puisque θ appartient à $S\left(\theta^0,\zeta\right)$, il existe $j \in \{1,\ldots,r\}$ tel que $|\theta_j - \theta_j^0| < \zeta$. Ainsi nous

avons,

$$\left|\sum_{j=1}^{r}(\theta_j-\theta_j^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \log f_i(x_i;\theta)}{\partial\theta_j}\right|_{\theta^0}\right| < \sum_{j=1}^{r}\zeta\zeta^2$$

et

$$\left|\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right.\right.$$
$$\left.\left.-\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]\right| < \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\zeta^2\zeta.$$

De plus, étant donné que la matrice d'une forme quadratique est symétrique et donc diagonalisable en base orthonormée, nous avons

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]=\sum_{j=1}^{r}\gamma_j\beta_j^2,$$

où $\sum_{j=1}^{r}\beta_j^2=\sum_{j=1}^{r}(\theta_j-\theta_j^0)^2=\zeta^2$. D'après l'hypothèse 3.2.9,

$$\sum_{j=1}^{r}\gamma_j\beta_j^2 \leqslant \max_j(\gamma_j)\sum_{j=1}^{r}\beta_j^2=\max_j(\gamma_j)\zeta^2<0.$$

Ainsi,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]\leqslant \max_j(\gamma_j)\zeta^2<0.$$

Une étude du signe de la fonction $(1/2)r^2\zeta^3+\max_j(\gamma_j)\zeta^2$ prouve que l'on peut trouver ζ_0 et a positifs tels que pour tout ζ plus petit que ζ_0,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0} - \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \log f_i(X_i;\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right] < -a\zeta^2.$$

Enfin,

$$\left|\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^3 \log f_i(x_i;\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}\right]\right|$$
$$< \frac{2}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}\zeta^3 M = b\zeta^3,$$

où $b = r^3M/3$. Rassemblant toutes les inégalités précédentes, nous avons pour les éléments de S et pour tout θ appartenant à $S(\theta^0,\zeta)$:

$$\frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) < r\zeta^3 - a\zeta^2 + b\zeta^3.$$

Sous l'hypothèse $\zeta < a/(r+b)$, nous avons pour tout élément de S, $(1/n)\mathrm{logL}(\theta) - (1/n)\mathrm{logL}(\theta^0) < 0$ pour tout $\theta \in S(\theta^0,\zeta)$. Ainsi, l'événement C dont chaque élément vérifie, pour tout $\theta \in S(\theta^0,\zeta)$,

$$\frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) < 0,$$

contient S et vérifie $P(C) \geqslant P(S) > 1-\epsilon$. Finalement nous avons, pour tout ζ plus petit que $\min(\zeta_0, a/(r+b))$, la probabilité

$$P\left(\forall\theta \in S(\theta^0,\zeta),\ \frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) < 0\right)$$

tend vers 1 quand n tend vers l'infini. Il existe $\widehat{\theta}_n$ appartenant à l'intérieur de la boule $B(\theta^0,\zeta)$ de centre θ^0 et de rayon ζ, c'est-à-dire tel que $\left\|\widehat{\theta}_n - \theta^0\right\| < \zeta$, tel que $\mathrm{logL}(\theta)$

ait un maximum local en $\widehat{\theta}_n$. Par conséquent,

$$\forall \zeta \leqslant \min\left(\zeta_0, \frac{a}{r+b}\right),\ P\left(\left\|\widehat{\theta}_n - \theta_0\right\| < \zeta\right) \underset{n\to+\infty}{\longrightarrow} 1.$$

D'après les hypothèses 3.2.1 et 3.2.2, $\widehat{\theta}_n$ est l'estimateur du maximum de vraisemblance.

□

Annexe C

Démonstration du théorème 3.3.1

Démonstration. D'après la formule de Taylor-Lagrange, d'après l'hypothèse 3.3.3, on peut écrire pour tout $i = 1, \ldots, M$, pour tout $k = 1, \ldots, n_i$ et pour tout $\theta \in \Theta$,

$$\begin{aligned}\log f_i(x_{ik}, \theta) = \log f_i(x_{ik}, \theta^0) &+ \sum_{j=1}^{r} \left.\frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\right|_{\theta^0} (\theta_j - \theta_j^0) \\ &+ \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} \left.\frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \\ &+ \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} \left.\frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0),\end{aligned}$$

où θ' appartient au segment $]\theta^0, \theta[$.

En sommant pour i allant de 1 à M et pour k allant de 1 à n_i, nous avons pour tout $\theta \in \Theta$,

$$\begin{aligned}\sum_{i=1}^{M}\sum_{k=1}^{n_i} \log f_i(x_{ik}, \theta) = \sum_{i=1}^{M}\sum_{k=1}^{n_i} \log f_i(x_{ik}, \theta^0) &+ \sum_{i=1}^{M}\sum_{k=1}^{n_i}\sum_{j=1}^{r} \left.\frac{\partial \log f_i(x_{ik};\theta)}{\partial \theta_j}\right|_{\theta^0} (\theta_j - \theta_j^0) \\ &+ \sum_{i=1}^{M}\sum_{k=1}^{n_i} \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r} \left.\frac{\partial^2 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p}\right|_{\theta^0} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0) \\ &+ \sum_{i=1}^{M}\sum_{k=1}^{n_i} \frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r} \left.\frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial \theta_j \partial \theta_p \partial \theta_q}\right|_{\theta'} (\theta_j - \theta_j^0)(\theta_p - \theta_p^0)(\theta_q - \theta_q^0).\end{aligned}$$

Par inversion des sommes, multiplication des deux membres par $1/n$ et en remarquant que $1/n = \mu_i/n_i$, nous avons pour tout $\theta \in \Theta$,

$$\begin{aligned}\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\log f_i(x_{ik},\theta) = &\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\log f_i(x_{ik},\theta^0)\\ &+\sum_{j=1}^{r}(\theta_j-\theta_j^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial\log f_i(x_{ik};\theta)}{\partial\theta_j}\right|_{\theta^0}\\ &+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^2\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\\ &+\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^3\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}.\end{aligned}$$

De plus, $\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}(1/n_i)\log f_i(x_{ik},\theta) = (1/n)\log\mathrm{L}(x_1,\ldots,x_n;\theta)$ que nous noterons $(1/n)$ $\log\mathrm{L}(\theta)$ pour la suite de cette démonstration. Ainsi, pour tout $\theta \in \Theta$,

$$\begin{aligned}\frac{1}{n}\log\mathrm{L}(\theta)-\frac{1}{n}\log\mathrm{L}(\theta^0) = &\sum_{j=1}^{r}(\theta_j-\theta_j^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial\log f_i(x_{ik};\theta)}{\partial\theta_j}\right|_{\theta^0}\\ &+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^2\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\\ &+\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^3\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}.\end{aligned}$$

En ajoutant et en retranchant un même terme dans le membre de droite, nous obtenons pour tout $\theta \in \Theta$,

$$\frac{1}{n}\mathrm{logL}(\theta)-\frac{1}{n}\mathrm{logL}(\theta^0)=\sum_{j=1}^{r}(\theta_j-\theta_j^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial \mathrm{log}f_i(x_{ik};\theta)}{\partial\theta_j}\right|_{\theta^0}$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^2\mathrm{log}f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}-\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)\right]$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2\mathrm{log}f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right)$$
$$+\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^3\mathrm{log}f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\right|_{\theta'}. \quad \text{(C.1)}$$

Considérons séparément chaque terme du membre droit de l'équation (C.1). Etant donné que les observations du sous-échantillon associé à la densité de probabilité $f_i(\cdot;\theta)$ sont indépendantes et identiquement distribuées, nous avons d'après la loi faible des grands nombres (Feller, 1968), pour tout $i=1,\ldots,M$ et $j=1,\ldots,r$,

$$\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial \mathrm{log}f_i(X_{ik};\theta)}{\partial\theta_j}\right|_{\theta^0}\xrightarrow[n\to+\infty]{P}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial \mathrm{log}f_i(X_{i1};\theta)}{\partial\theta_j}\right|_{\theta^0}\right)=0.$$

D'après le théorème de Slutsky, pour tout $j=1,\ldots,r$

$$\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial \mathrm{log}f_i(X_{ik};\theta)}{\partial\theta_j}\right|_{\theta^0}\xrightarrow[n\to+\infty]{P}0.$$

D'après la loi faible des grands nombres, nous avons pour tout $(j,p)\in\{1,\ldots,r\}^2$,

$$\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^2 \mathrm{log}f_i(X_{ik};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\xrightarrow[n\to+\infty]{P}\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \mathrm{log}f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right).$$

D'après le théorème de Slutsky, nous avons pour tout $(j,p)\in\{1,\ldots,r\}^2$,

$$\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\left.\frac{\partial^2 \mathrm{log}f_i(X_{ik};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\xrightarrow[n\to+\infty]{P}\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\left.\frac{\partial^2 \mathrm{log}f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\right|_{\theta^0}\right).$$

Ainsi, on a pour tout $(j,p) \in \{1,\dots,r\}^2$,

$$\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^2 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0} - \sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta^0}\left(\frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0}\right) \xrightarrow[n\to+\infty]{P} 0.$$

Pour tout $(j,p) \in \{1,\dots,r\}^2$ et pour tout $\theta \in \Theta$, on a

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta^0}\left(\frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0}\right) \\ = -\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta^0}\left(-\frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0}\right).\end{aligned}$$

D'après l'hypothèse 3.3.7, il s'agit d'une forme quadratique définie négative. Ainsi nous avons, pour tout $(j,p) \in \{1,\dots,r\}^2$ et pour tout $\theta \in \Theta$ différent de θ^0,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta^0}\left(\frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p}\bigg|_{\theta^0}\right) < 0.$$

D'après la loi faible des grands nombres, nous avons pour tout $(j,p,q) \in \{1,\dots,r\}^3$,

$$\sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'} \xrightarrow[n\to+\infty]{P} \mathrm{E}_{\theta'}\left(\frac{\partial^3 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'}\right).$$

D'après le théorème de Slutsky, nous avons pour tout $(j,p,q) \in \{1,\dots,r\}^3$,

$$\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'} \xrightarrow[n\to+\infty]{P} \sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta'}\left(\frac{\partial^3 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'}\right).$$

Etant donné que la fonction valeur absolue est une fonction continue, nous avons pour tout $(j,p,q) \in \{1,\dots,r\}^3$,

$$\left|\sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \frac{\partial^3 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'}\right| \xrightarrow[n\to+\infty]{P} \left|\sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta'}\left(\frac{\partial^3 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q}\bigg|_{\theta'}\right)\right|,$$

et d'après l'hypothèse 3.3.6, nous avons pour tout $(j,p,q) \in \{1,\ldots,r\}^3$,

$$\left| \sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta'} \left(\left. \frac{\partial^3 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right|_{\theta'} \right) \right| < K.$$

Soit (ζ,ϵ) un vecteur de constantes positives arbitraires. Les résultats précédents permettent d'écrire trois groupes d'inégalités. Pour tout ζ, pour tout ϵ, il existe n_0 tel que pour tout n supérieur ou égal à n_0 et pour tout $(j,p,q) \in \{1,\ldots,r\}^3$, les probabilités suivantes

$$P\left(\left| \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial \log f_i(X_{ik};\theta)}{\partial\theta_j} \right|_{\theta^0} \right| \geqslant \zeta^2 \right),$$

$$P\left(\left| \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial^2 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p} \right|_{\theta^0} - \sum_{i=1}^{M} \lambda_i \mathrm{E}_{\theta^0} \left(\left. \frac{\partial^2 \log f_i(X_{i1};\theta)}{\partial\theta_j \partial\theta_p} \right|_{\theta^0} \right) \right| \geqslant \zeta \right),$$

$$P\left(\left| \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial^3 \log f_i(X_{ik};\theta)}{\partial\theta_j \partial\theta_p \partial\theta_q} \right|_{\theta'} \right| \geqslant 2K \right),$$

sont majorées par $\epsilon/\left(r\left(1+r+r^2\right)\right)$. Soit S l'événement satisfaisant les $r(1+r+r^2)$ inégalités suivantes :

$$\left\{ \left| \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial \log f_i(X_{ik};\theta)}{\partial\theta_1} \right|_{\theta^0} \right| < \zeta^2, \cdots, \left| \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial^3 \log f_i(X_{ik};\theta)}{\partial\theta_r \partial\theta_r \partial\theta_r} \right|_{\theta'} \right| < 2M \right\}.$$

Les majorations des probabilités précédentes nous donnent $P(S^*) < \epsilon$, où S^* est le complémentaire de S, c'est-à-dire l'événement où au moins une des inégalités précédentes n'est pas satisfaite. Nous avons ainsi $P(S) > 1-\epsilon$.

Maintenant, étudions le signe de la quantité $(1/n)\mathrm{logL}(\theta) - (1/n)\mathrm{logL}(\theta^0)$ pour les éléments de S et pour θ appartenant à la sphère $S\left(\theta^0,\zeta\right)$ de centre θ^0 et de rayon ζ. Puisque θ appartient à $S\left(\theta^0,\zeta\right)$, il existe $j \in \{1,\ldots,r\}$ tel que $|\theta_j - \theta_j^0| < \zeta$. Ainsi nous avons,

$$\left| \sum_{j=1}^{r} (\theta_j - \theta_j^0) \sum_{i=1}^{M} \mu_i \sum_{k=1}^{n_i} \frac{1}{n_i} \left. \frac{\partial \log f_i(x_{ik};\theta)}{\partial\theta_j} \right|_{\theta^0} \right| < \sum_{j=1}^{r} \zeta\zeta^2$$

et

$$\left|\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\frac{\partial^2\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right.\right.$$
$$\left.\left.-\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\frac{\partial^2\log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right]\right| < \frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\zeta^2\zeta.$$

De plus, étant donné que la matrice d'une forme quadratique est symétrique et donc diagonalisable en base orthonormée, nous avons

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\frac{\partial^2\log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right]=\sum_{j=1}^{r}\gamma_j\beta_j^2,$$

où $\sum_{j=1}^{r}\beta_j^2=\sum_{j=1}^{r}\left(\theta_j-\theta_j^0\right)^2=\zeta^2$. D'après l'hypothèse 3.3.7,

$$\sum_{j=1}^{r}\gamma_j\beta_j^2\leqslant\max_j(\gamma_j)\sum_{j=1}^{r}\beta_j^2=\max_j(\gamma_j)\zeta^2<0.$$

Ainsi,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\frac{\partial^2\log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right]\leqslant\max_j(\gamma_j)\zeta^2<0.$$

Une étude du signe de la fonction $(1/2)r^2\zeta^3+\max_j(\gamma_j)\zeta^2$ prouve que l'on peut trouver ζ_0 et a positifs tels que pour tout ζ plus petit que ζ_0,

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\left[\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\frac{\partial^2\log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}-\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\frac{\partial^2\log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right]$$
$$+\frac{1}{2}\sum_{j=1}^{r}\sum_{p=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)\sum_{i=1}^{M}\lambda_i\mathrm{E}_{\theta^0}\left(\frac{\partial^2\log f_i(X_{i1};\theta)}{\partial\theta_j\partial\theta_p}\bigg|_{\theta^0}\right)\right]<-a\zeta^2.$$

Enfin,

$$\left|\frac{1}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}\left[(\theta_j-\theta_j^0)(\theta_p-\theta_p^0)(\theta_q-\theta_q^0)\sum_{i=1}^{M}\mu_i\sum_{k=1}^{n_i}\frac{1}{n_i}\frac{\partial^3 \log f_i(x_{ik};\theta)}{\partial\theta_j\partial\theta_p\partial\theta_q}\bigg|_{\theta'}\right]\right| < \frac{2}{6}\sum_{j=1}^{r}\sum_{p=1}^{r}\sum_{q=1}^{r}\zeta^3 K = b\zeta^3,$$

où $b = r^3K/3$. Rassemblant toutes les inégalités précédentes, nous avons pour les éléments de S et pour tout θ appartenant à $S(\theta^0,\zeta)$:

$$\frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) < r\zeta^3 - a\zeta^2 + b\zeta^3.$$

Sous l'hypothèse $\zeta < a/(r+b)$, nous avons pour tout élément de S, $(1/n)\mathrm{log}L(\theta) - (1/n)\mathrm{log}L(\theta^0) < 0$ pour tout $\theta \in S(\theta^0,\zeta)$. Ainsi, l'événement C dont chaque élément vérifie, pour tout $\theta \in S(\theta^0,\zeta)$,

$$\frac{1}{n}\mathrm{logL}(\theta) - \frac{1}{n}\mathrm{logL}(\theta^0) < 0,$$

contient S et vérifie $P(C) \geqslant P(S) > 1-\epsilon$. Finalement nous avons, pour tout ζ plus petit que $\min(\zeta_0, a/(r+b))$, la probabilité

$$P\left(\forall\theta \in S(\theta^0,\zeta),\ \frac{1}{n}\mathrm{log}L(\theta) - \frac{1}{n}\mathrm{log}L(\theta^0) < 0\right)$$

tend vers 1 quand n tend vers l'infini. Il existe $\widehat{\theta}_n$ appartenant à l'intérieur de la boule $B(\theta^0,\zeta)$ de centre θ^0 et de rayon ζ, c'est-à-dire tel que $\left\|\widehat{\theta}_n - \theta^0\right\| < \zeta$, tel que $\mathrm{log}L(\theta)$ ait un maximum local en $\widehat{\theta}_n$. Par conséquent,

$$\forall\zeta \leqslant\ \min\ \left(\zeta_0, \frac{a}{r+b}\right),\ P\left(\left\|\widehat{\theta}_n - \theta_0\right\| < \zeta\right) \underset{n\to+\infty}{\longrightarrow} 1.$$

D'après les hypothèses 3.3.1 et 3.3.2, $\widehat{\theta}_n$ est l'estimateur du maximum de vraisemblance.

□

Remarque C.1. Dans le cas où toutes les densités de probabilité ne sont pas observées dans l'échantillon (cf. remarque 3.3.2), l'adaptation de la démonstration repose sur le résultat (3.3) et sur la proposition suivante :

Proposition C.1. *Soit* $(U_i)_{1\leqslant i\leqslant M}$ *une suite de variables aléatoires indépendantes. Pour tout* $i = 1, \ldots, M$, *soit* $v_i(n)$ *une suite quelconque. Soient* $(L_i)_{1\leqslant i\leqslant M}$ *et* $(l_i)_{1\leqslant i\leqslant M}$ *deux suites de réels. Sous les hypothèses* $m(n) \underset{n\to+\infty}{\longrightarrow} M$, $v_i(n) \underset{n\to+\infty}{\longrightarrow} l_i$ *et* $U_i \xrightarrow[n\to+\infty]{P} L_i$, *on a*

$$\sum_{i=1}^{m(n)} v_i(n)U_i \xrightarrow[n\to+\infty]{P} \sum_{i=1}^{M} l_i L_i.$$

Annexe D

Tableaux D.1, D.2, D.3, D.4 et D.5

TABLEAU D.1 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour la distribution exponentielle : probabilité de couverture des intervalles de confiance à 95% de type Waldb($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(1,4,5)}$) et vraie valeur (σ) et estimation ($\widehat{\sigma}$) de l'écart-type asymptotique, pour deux valeurs du paramètre θ_1 de la distribution exponentielle, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	p	n	$\widehat{I(\theta^0)}$	$I^{(1)}$	$I^{(4)}$	$I^{(5)}$	σ	$\widehat{\sigma}$
0.006	0.75	100	0.98	0.98	0.98	0.98	0.0050	0.0050
		200	0.97	0.97	0.97	0.97	0.0035	0.0035
		500	0.96	0.96	0.95	0.95	0.0022	0.0022
		1000	0.94	0.94	0.94	0.94	0.0016	0.0016
	0.50	100	0.95	0.94	0.95	0.94	0.0020	0.0020
		200	0.95	0.94	0.94	0.94	0.0014	0.0014
		500	0.95	0.95	0.95	0.95	0.00090	0.00090
		1000	0.94	0.94	0.94	0.94	0.00064	0.00064
	0.25	100	0.96	0.95	0.97	0.95	0.0010	0.0010
		200	0.95	0.95	0.95	0.95	0.00071	0.00072
		500	0.94	0.95	0.95	0.94	0.00045	0.00045
		1000	0.95	0.96	0.95	0.96	0.00032	0.00032
0.05	0.75	100	0.98	0.97	0.98	0.97	0.041	0.042
		200	0.97	0.97	0.97	0.97	0.029	0.029
		500	0.95	0.95	0.95	0.95	0.018	0.018
		1000	0.94	0.94	0.94	0.94	0.013	0.013
	0.50	100	0.95	0.94	0.95	0.94	0.017	0.017
		200	0.95	0.95	0.95	0.95	0.012	0.012
		500	0.95	0.95	0.95	0.95	0.0075	0.0075
		1000	0.95	0.95	0.95	0.95	0.0053	0.0053
	0.25	100	0.95	0.94	0.95	0.94	0.0084	0.0085
		200	0.96	0.96	0.96	0.96	0.0059	0.0060
		500	0.95	0.96	0.95	0.96	0.0038	0.0038
		1000	0.95	0.95	0.95	0.95	0.0027	0.0027

TABLEAU D.2 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_1 de la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_1) et estimation ($\widehat{\sigma}_1$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_1	$\widehat{\sigma}_1$
0.006	0.5	0.75	100	0.93	0.93	0.80	0.80	0.023	0.033
			200	0.92	0.91	0.84	0.84	0.016	0.021
			500	0.91	0.91	0.89	0.89	0.010	0.011
			1000	0.89	0.89	0.91	0.91	0.0073	0.0074
		0.50	100	0.88	0.88	0.94	0.94	0.0085	0.0077
			200	0.89	0.89	0.94	0.94	0.0060	0.0056
			500	0.91	0.91	0.95	0.95	0.0038	0.0036
			1000	0.93	0.93	0.95	0.96	0.0027	0.0026
		0.25	100	0.91	0.91	0.96	0.96	0.0034	0.0032
			200	0.91	0.91	0.94	0.94	0.0024	0.0023
			500	0.96	0.96	0.97	0.96	0.0015	0.0015
			1000	0.96	0.96	0.96	0.96	0.0011	0.0011
0.05	0.5	0.75	100	0.94	0.94	0.80	0.80	0.19	0.29
			200	0.91	0.91	0.83	0.83	0.14	0.18
			500	0.89	0.89	0.89	0.89	0.086	0.094
			1000	0.89	0.89	0.93	0.93	0.061	0.063
		0.50	100	0.87	0.87	0.94	0.94	0.071	0.066
			200	0.89	0.89	0.93	0.93	0.050	0.047
			500	0.90	0.90	0.95	0.95	0.032	0.030
			1000	0.94	0.94	0.95	0.95	0.022	0.022
		0.25	100	0.93	0.94	0.96	0.96	0.028	0.027
			200	0.93	0.93	0.96	0.96	0.020	0.019
			500	0.94	0.94	0.96	0.96	0.013	0.012
			1000	0.95	0.95	0.94	0.94	0.0089	0.0088
0.006	2	0.75	100	0.85	0.85	0.90	0.90	0.0027	0.0031
			200	0.89	0.91	0.92	0.92	0.0019	0.0023
			500	0.94	0.94	0.94	0.94	0.0012	0.0013
			1000	0.94	0.94	0.94	0.94	0.00085	0.00088
		0.50	100	0.96	0.96	0.94	0.94	0.00077	0.00082
			200	0.96	0.96	0.95	0.95	0.00054	0.00057
			500	0.96	0.96	0.96	0.96	0.00034	0.00035
			1000	0.95	0.95	0.95	0.95	0.00024	0.00025
		0.25	100	0.94	0.94	0.93	0.93	0.00037	0.00037
			200	0.95	0.95	0.95	0.95	0.00026	0.00026

TABLEAU D.2 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_1 de la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_1) et estimation ($\widehat{\sigma}_1$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_1	$\widehat{\sigma}_1$
			500	0.94	0.94	0.94	0.94	0.00016	0.00016
			1000	0.95	0.95	0.94	0.94	0.00012	0.00012
0.05	2	0.75	100	0.84	0.85	0.89	0.89	0.022	0.026
			200	0.89	0.90	0.91	0.91	0.016	0.018
			500	0.93	0.93	0.92	0.92	0.010	0.011
			1000	0.95	0.95	0.93	0.93	0.0071	0.0073
		0.50	100	0.95	0.95	0.94	0.94	0.0064	0.0069
			200	0.97	0.97	0.96	0.95	0.0045	0.0047
			500	0.95	0.96	0.94	0.95	0.0029	0.0029
			1000	0.97	0.97	0.96	0.96	0.0020	0.0020
		0.25	100	0.95	0.95	0.94	0.94	0.0031	0.0031
			200	0.95	0.95	0.94	0.94	0.0022	0.0022
			500	0.95	0.95	0.94	0.94	0.0014	0.0014
			1000	0.95	0.95	0.95	0.95	0.00097	0.00097

TABLEAU D.3 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_1 de la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_1) et estimation ($\widehat{\sigma}_1$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_1	$\widehat{\sigma}_1$
0.006	0.5	0.75	100	0.88	0.88	0.81	0.81	0.015	0.024
			200	0.88	0.88	0.85	0.85	0.011	0.014
			500	0.87	0.87	0.90	0.90	0.0068	0.0075
			1000	0.89	0.89	0.90	0.90	0.0048	0.0051
		0.50	100	0.86	0.86	0.90	0.90	0.0069	0.0072
			200	0.90	0.90	0.92	0.92	0.0049	0.0051
			500	0.94	0.94	0.95	0.95	0.0031	0.0031
			1000	0.93	0.93	0.95	0.95	0.0022	0.0022
		0.25	100	0.92	0.92	0.94	0.94	0.0036	0.0037
			200	0.93	0.93	0.94	0.94	0.0026	0.0026
			500	0.94	0.94	0.93	0.93	0.0016	0.0016
			1000	0.94	0.94	0.95	0.95	0.0011	0.0011
0.05	0.5	0.75	100	0.91	0.91	0.79	0.79	0.13	0.22
			200	0.89	0.89	0.83	0.83	0.091	0.12
			500	0.90	0.90	0.89	0.89	0.057	0.065
			1000	0.89	0.89	0.93	0.93	0.041	0.042
		0.50	100	0.87	0.87	0.90	0.90	0.058	0.062
			200	0.89	0.89	0.92	0.92	0.041	0.041
			500	0.91	0.91	0.94	0.94	0.026	0.026
			1000	0.94	0.94	0.95	0.95	0.018	0.018
		0.25	100	0.92	0.92	0.94	0.94	0.030	0.031
			200	0.93	0.93	0.94	0.94	0.021	0.021
			500	0.94	0.94	0.95	0.95	0.013	0.013
			1000	0.95	0.95	0.94	0.94	0.0096	0.0096
0.006	2	0.75	100	0.91	0.92	0.96	0.96	0.0022	0.0022
			200	0.95	0.95	0.94	0.94	0.0016	0.0016
			500	0.95	0.95	0.95	0.95	0.00099	0.0010
			1000	0.95	0.95	0.94	0.94	0.00070	0.00070
		0.50	100	0.95	0.95	0.95	0.95	0.0010	0.0010
			200	0.95	0.95	0.95	0.95	0.00074	0.00075
			500	0.95	0.95	0.95	0.95	0.00047	0.00047
			1000	0.95	0.95	0.95	0.95	0.00033	0.00033
		0.25	100	0.94	0.94	0.94	0.94	0.00065	0.00065
			200	0.95	0.95	0.95	0.95	0.00046	0.00046

TABLEAU D.3 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_1 de la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_1) et estimation ($\widehat{\sigma}_1$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_1	$\widehat{\sigma}_1$
			500	0.95	0.95	0.95	0.95	0.00029	0.00029
			1000	0.95	0.95	0.95	0.95	0.00021	0.00021
0.05	2	0.75	100	0.92	0.94	0.95	0.95	0.018	0.019
			200	0.96	0.96	0.95	0.95	0.013	0.013
			500	0.96	0.96	0.96	0.96	0.0083	0.0083
			1000	0.95	0.95	0.95	0.95	0.0058	0.0058
		0.50	100	0.95	0.95	0.96	0.95	0.0088	0.0088
			200	0.95	0.95	0.94	0.94	0.0062	0.0062
			500	0.95	0.94	0.94	0.95	0.0039	0.0039
			1000	0.96	0.96	0.96	0.96	0.0028	0.0028
		0.25	100	0.95	0.95	0.95	0.95	0.0054	0.0055
			200	0.95	0.96	0.95	0.95	0.0038	0.0038
			500	0.95	0.95	0.94	0.94	0.0024	0.0024
			1000	0.95	0.95	0.96	0.96	0.0017	0.0017

TABLEAU D.4 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_2 de la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_2) et estimation ($\widehat{\sigma}_2$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_2	$\widehat{\sigma}_2$
0.006	0.5	0.75	100	0.95	0.95	0.93	0.93	0.083	0.082
			200	0.96	0.96	0.95	0.95	0.058	0.058
			500	0.97	0.97	0.97	0.97	0.037	0.037
			1000	0.95	0.95	0.95	0.95	0.026	0.026
		0.50	100	0.97	0.97	0.96	0.96	0.072	0.073
			200	0.96	0.95	0.95	0.96	0.051	0.051
			500	0.95	0.95	0.95	0.95	0.032	0.032
			1000	0.95	0.95	0.95	0.95	0.023	0.023
		0.25	100	0.94	0.94	0.94	0.93	0.060	0.061
			200	0.94	0.94	0.93	0.94	0.042	0.042
			500	0.96	0.96	0.96	0.96	0.027	0.027
			1000	0.95	0.95	0.95	0.95	0.019	0.019
0.05	0.5	0.75	100	0.97	0.96	0.96	0.96	0.083	0.082
			200	0.94	0.94	0.94	0.94	0.058	0.058
			500	0.96	0.96	0.97	0.97	0.037	0.037
			1000	0.97	0.97	0.97	0.97	0.026	0.026
		0.50	100	0.96	0.96	0.96	0.96	0.072	0.072
			200	0.95	0.95	0.95	0.95	0.051	0.051
			500	0.95	0.95	0.95	0.95	0.032	0.032
			1000	0.96	0.96	0.95	0.95	0.023	0.023
		0.25	100	0.96	0.96	0.95	0.95	0.060	0.060
			200	0.95	0.95	0.95	0.95	0.042	0.042
			500	0.95	0.95	0.95	0.95	0.027	0.027
			1000	0.94	0.94	0.95	0.95	0.019	0.019
0.006	2	0.75	100	0.93	0.92	0.95	0.95	0.28	0.27
			200	0.92	0.93	0.95	0.95	0.19	0.19
			500	0.95	0.94	0.94	0.94	0.12	0.12
			1000	0.94	0.94	0.94	0.93	0.087	0.087
		0.50	100	0.95	0.96	0.95	0.94	0.22	0.22
			200	0.95	0.95	0.95	0.95	0.15	0.15
			500	0.95	0.95	0.95	0.95	0.097	0.097
			1000	0.94	0.94	0.94	0.94	0.069	0.069
		0.25	100	0.94	0.94	0.93	0.93	0.17	0.17
			200	0.96	0.95	0.95	0.95	0.12	0.12

TABLEAU D.4 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_2 de la distribution de Weibull : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_2) et estimation ($\widehat{\sigma}_2$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution de Weibull, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_2	$\widehat{\sigma}_2$
			500	0.96	0.96	0.96	0.96	0.075	0.075
			1000	0.96	0.95	0.96	0.95	0.053	0.053
0.05	2	0.75	100	0.94	0.92	0.95	0.96	0.28	0.27
			200	0.93	0.93	0.95	0.95	0.19	0.19
			500	0.93	0.93	0.93	0.93	0.12	0.12
			1000	0.96	0.96	0.95	0.96	0.087	0.087
		0.50	100	0.96	0.96	0.95	0.95	0.22	0.22
			200	0.95	0.95	0.95	0.95	0.15	0.15
			500	0.95	0.95	0.95	0.95	0.097	0.098
			1000	0.94	0.94	0.94	0.94	0.069	0.069
		0.25	100	0.94	0.94	0.94	0.93	0.17	0.17
			200	0.95	0.95	0.95	0.95	0.12	0.12
			500	0.95	0.95	0.96	0.95	0.075	0.075
			1000	0.96	0.96	0.96	0.96	0.053	0.053

TABLEAU D.5 – Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_2 de la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_2) et estimation ($\widehat{\sigma}_2$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_2	$\widehat{\sigma}_2$
0.006	0.5	0.75	100	0.97	0.97	0.95	0.95	0.084	0.086
			200	0.96	0.96	0.95	0.95	0.059	0.060
			500	0.96	0.96	0.96	0.96	0.038	0.038
			1000	0.95	0.95	0.95	0.95	0.027	0.027
		0.50	100	0.96	0.96	0.95	0.95	0.075	0.076
			200	0.95	0.95	0.95	0.95	0.053	0.053
			500	0.96	0.96	0.95	0.95	0.034	0.034
			1000	0.93	0.93	0.94	0.94	0.024	0.024
		0.25	100	0.95	0.95	0.94	0.94	0.064	0.065
			200	0.95	0.95	0.95	0.95	0.045	0.046
			500	0.94	0.94	0.94	0.94	0.029	0.029
			1000	0.95	0.95	0.95	0.95	0.020	0.020
0.05	0.5	0.75	100	0.96	0.96	0.95	0.95	0.084	0.086
			200	0.96	0.96	0.94	0.94	0.059	0.060
			500	0.95	0.95	0.95	0.95	0.038	0.038
			1000	0.95	0.95	0.95	0.95	0.027	0.027
		0.50	100	0.95	0.95	0.95	0.95	0.075	0.076
			200	0.95	0.95	0.95	0.95	0.053	0.053
			500	0.95	0.95	0.95	0.95	0.034	0.034
			1000	0.95	0.95	0.95	0.95	0.024	0.024
		0.25	100	0.96	0.96	0.96	0.96	0.064	0.065
			200	0.94	0.94	0.94	0.94	0.045	0.046
			500	0.95	0.95	0.95	0.95	0.029	0.029
			1000	0.93	0.93	0.94	0.94	0.020	0.020
0.006	2	0.75	100	0.94	0.95	0.93	0.93	0.29	0.29
			200	0.96	0.96	0.95	0.95	0.20	0.21
			500	0.94	0.94	0.94	0.94	0.13	0.13
			1000	0.94	0.95	0.95	0.95	0.091	0.091
		0.50	100	0.95	0.95	0.94	0.94	0.25	0.25
			200	0.94	0.94	0.94	0.94	0.18	0.18
			500	0.93	0.93	0.93	0.93	0.11	0.11
			1000	0.95	0.95	0.94	0.95	0.079	0.079
		0.25	100	0.94	0.94	0.93	0.93	0.21	0.21
			200	0.96	0.96	0.96	0.96	0.15	0.15

TABLEAU D.5 – (Suite et fin) Résultats de l'étude de simulations n° 2 (1000 réplications) pour le paramètre θ_2 de la distribution log-logistique : probabilité de couverture des intervalles de confiance à 95% de type Wald ($\widehat{I(\theta^0)}$), probabilités de couverture des intervalles alternatifs ($I^{(2,3,5)}$) et "vraie" valeur (σ_2) et estimation ($\widehat{\sigma}_2$) de l'écart-type asymptotique, pour quatre valeurs du paramètre (θ_1, θ_2) de la distribution log-logistique, trois proportions p de données tronquées à droite et quatre tailles d'échantillon n.

θ_1	θ_2	p	n	$\widehat{I(\theta^0)}$	$I^{(2)}$	$I^{(3)}$	$I^{(5)}$	σ_2	$\widehat{\sigma}_2$
			500	0.95	0.95	0.94	0.94	0.095	0.095
			1000	0.95	0.95	0.95	0.95	0.067	0.067
0.05	2	0.75	100	0.94	0.95	0.95	0.95	0.29	0.29
			200	0.95	0.95	0.93	0.93	0.20	0.21
			500	0.96	0.96	0.96	0.96	0.13	0.13
			1000	0.94	0.95	0.95	0.95	0.091	0.092
		0.50	100	0.94	0.94	0.95	0.95	0.25	0.25
			200	0.95	0.96	0.95	0.95	0.18	0.18
			500	0.95	0.95	0.95	0.95	0.11	0.11
			1000	0.96	0.96	0.96	0.96	0.079	0.079
		0.25	100	0.94	0.94	0.93	0.93	0.21	0.21
			200	0.96	0.96	0.96	0.96	0.15	0.15
			500	0.94	0.94	0.94	0.94	0.095	0.095
			1000	0.95	0.95	0.95	0.95	0.067	0.067