N° d'ordre : **04/2013-M/MT**

REPUBLIQUE ALGÉRIENNE DEMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE-ALGER

FACULTÉ DE MATHÉMATIQUES

جامعة هواري بومدين للعلوم والتكنولوجيا
U S T H B

MÉMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

EN : MATHÉMATIQUES

Spécialité : Probabilités et Statistiques

par : BELKADI Souad

Sujet :

ANALYSE PAR BOOTSTRAP DES DONNÉES CENSURÉES

Soutenu publiquement, le 20 novembre 2013, devant le jury composé de :

M. K. BOUKHETALLA	Professeur	à l'U.S.T.H.B	Président.
Mme. O. SADKI	Maître de Conférences/A	à l'U.O.E.BOUAGHI	Directrice de mémoire.
Mme. Z. GUESSOUM	Maître de Conférences/A	à l'U.S.T.H.B	Examinatrice.
M. T. KERNANE	Maître de Conférences/B	à l'U.S.T.H.B	Invité.

Résumé

La fonction de survie constitue une caractéristique importante des modèles de durée de vie, en particulier lorsque les données ne sont pas complètement observées. Nous nous intéressons aux données aléatoirement censurées à droite. L'estimateur de Kaplan-Meier de la fonction de survie est un estimateur non paramétrique ; aucune hypothèse n'est faite sur la distribution des durées de survie. Les propriétés de cette estimateur ont suscité l'intérêt d'un grand nombre d'auteurs. Son comportement asymptotique en terme de convergence et de normalité asymptotique est encore d'actualité pour différents types de données.

Le présent travail se propose d'étudier certaines caractéristiques de l'estimateur de Kaplan-Meier de la fonction de survie dans le cas de données indépendantes identiquement distribuées censurées à droite par des techniques de ré-échantillonnage par la méthode du bootstrap.

La méthode du bootstrap basée sur l'idée d'exploiter toute l'information apportée par l'échantillon initial, nous permet d'étudier les propriétés des estimateurs.

Dans ce mémoire, nous nous intéressons, en particulier aux moments de la fonction de survie, aux intervalles de confiance pour les probabilités de survie et aux quantiles des durées de survie. Afin d'illustrer notre étude, nous avons consacré une bonne partie de notre travail à des simulations.

Mots-clés : bootstrap, données censurées, estimateur de Kaplan-Meier, fonction de survie, fonction de répartition, fonction de hasard comulée, intervalle de confiance, moments, normalité asymptotique, quantile.

Dédicaces

MERCI du fond du coeur à mes parents, mes soeurs et mes frères, merci pour tout le temps que vous avez consacré à m'aider, merci pour votre soutien inconditionnel tout long de ces années d'étude, merci parce que vous m'encouragez et vous permettez ce que je suis aujourd'hui. Je vous dédie ce mémoire.

Remerciements

Je dis merci de tout mon coeur et toutes mes forces à DIEU tout puissance qui a permis la réalisation de ce travail.

Je dis un grand merci à mon encadreur Mme. ***SADKI Ourida*** *de m'avoir proposée ce sujet et de m'avoir encadré pendant toutes ces années. Je la remercie vivement pour son aide, ses conseils, sa patience, sa disponibilité, ses orientations et ses indications. Mes reconnaissances pour lui de m'avoir fait confiance et de m'avoir mis à l'aise tout au long de l'avancement de ce mémoire. Je ne pourrai jamais assez remercier.*

J'ai été très honoré que M. ***K. BOUKHETALLA*** *accepte la présidence du jury, et j'aimerais lui adresser de ce fait de vifs remerciments.*

Je remercie sincèrement Mme. ***Z. GUESSOUM****, pour avoir accepté de faire partie du jury, et pour le temps accordé à la lecture attentive du mémoire.*

Je remercie M. ***T. KERNANE****, pour avoir accepté de faire partie de mon jury.*

Je tiens à exprimer ma reconnaissance à toute ma famille, mes amis et tous ceux qui m'ont aidé de près ou de loin.

Table des matières

Abréviations et symboles

Nous regroupons ici, les différentes notations et abréviations utilisées tout au long du document.

- i.i.d. : indépendantes et identiquement distribuées.
- f.d.r. : fonction de répartition.
- $\mathbb{1}_A(.)$: fonction indicatrice sur l'ensemble A.
- B : nombre de Bootstraps.
- N : loi normale.
- $\xrightarrow{L}$: convergence en loi.
- $\xrightarrow{\mathbf{P}}$: convergence en probabilité.
- I.C. : intervalle de confiance.
- $x \wedge y := min(x, y)$.
- v.a. : variable aléatoire.

Introduction générale

On observe des durées de vie dans une vaste gamme de champs d'application. Il est donc très important de disposer de bonnes méthodes d'analyse pour ce type de données. Sur le plan mathématique, une durée de vie n'est rien d'autre qu'une variable aléatoire (v.a) non négative. Ce type de variable est fréquent, notamment en médecine, en biologie, en épidémioligie, en finance, en actuariat et en fiabilité. La tâche est d'autant plus complexe lorsque certaines de ces observations sont incomplètes : la censure et la troncature sont des mécanismes inhérents des données de survie. Nous concentrerons notre étude, sur la variable censurée. Il existe plusieurs mécanismes de censures. Particulièrement la censure aléatoire à droite : une v.a de durée X est censurée à droite par une v.a de censure C si on n'observe X que lorsque $X < C$. Une fonction essentielle en analyse de survie est la fonction de survie définit par $S(t) = \mathbf{P}(X > t)$ qui exprime la probabilité que l'individu "survive" à l'instant t. L'analyse de caractéristiques de la fonction de survie liées à la variable X, pouvant favoriser l'identification du modèle, apparaît donc importante. En particulier l'estimation non paramétrique de telles caractéristiques, libre d'hypothèses sur la distribution de X, peut constituer un instrument utile d'interprétation. Différents estimateurs non paramétriques de la fonction de survie ont été proposés, qui tous conduisent à des propriétés de nature asymptotique. L'estimateur de la fonction de survie le plus utilisé lorsqu'aucune hypothèse ne peut être faite sur la distribution des temps de survie est l'estimateur de Kaplan-Meier (1958), où l'objectif est de caractériser et estimer la survie de patients en fonction du temps et de leurs caractéristiques.

L'estimation précise de la probabilité de survie se conduit souvent dans les études médicales, en particulier avec les patients au cours de la surveillance. La probabilité de survie est une estimation ponctulle de la survie, indépendante de la taille de la population étudiée. En revanche, cette probabilité de survie est toujours associée à un intrvalle de confiance (IC)

dont les limites sont liées au nombre de patients inclus dans l'étude. Plus les patients sont nombreux, plus l'IC est petit et plus la probabilité de survie est une bonne estimation de la survie réelle de chaque patient. La précision de l'estimation de la survie pose donc un problème lorsque les patients ont un faible effectif. Des méthodes statistiques de simulation ont été développées pour améliorer la précision de l'estimation Kaplan-Meier, en offrant un IC plus étroit. L'une d'entre elles est le Bootstrap ; c'est une méthode qui a montré ses performances en analyse statistique en générale et en analyse de survie en particulier. Cette technique statistique permet d'obtenir des résultats plus précis dans une étude limitée par une petite taille de l'échantillon.

Ce mémoire est organisé comme suit :

Dans le chapitre 1, nous rappelons les principales fonctions en analyse de survie : fonction de survie-taux de survie et les différents types de censure (censure à droite, censure à gauche, censure par intervalle). Dans le cas d'observations censurées à droite, nous introduisons l'estimateur de Kaplan-Meier de la fonction de survie. Nous donnons leurs principales propriétés de convergence, lois limites et variances asymptotiques. Nous présentons quelques exemples de données censurées avec les graphes des estimateurs correspondants.

Dans le chapitre 2, nous présentons la méthode du bootstrap qui sera analysé et implémenté dans le logiciel R, d'abord nous développons les résultats sur l'estimation de la fonction de la survie dans le cas de données censurées à droite par la méthode du bootstrap de l'estimatur de Kaplan-Meier. Nous donnons les expressions asymptotiques de la variance de cet estimateur, la plus couramment utilisée pour estimer la variance est d'utiliser la formule de Greenwood. L'IC de cette probabilité de survie a ensuite été calculé selon la méthode classique de Greenwood puis en utilisant le rééchantillonnage par bootstrap.

Dans le chapitre 3, nous présentons des simulations numériques illustrant les comportements des estimateurs des fonctions de survie de Kaplan-Meier et Kaplan-Meier bootstrapé. Les deux méthodes sont évaluées et comparées en utilisant des données réelles (de la littérature dans le thème) et des données simulées (différents modèles sont proposés).

Chapitre 1

L'analyse des données de survie

Introduction

L'analyse de survie est un domaine des statistiques qui trouve sa place dans tous les champs d'application où l'on étudie la survenue d'un évènement. L'objectif de cette analyse réside dans l'analyse du délai de survenue d'un évènement dans un ou plusieurs groupes d'individus. Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, dans les enquêtes épidémiologiques, les données sont souvent recueillies de façons incomplètes. La censure et la troncature font partie de processus générant ce type de données. Dans ce chapitre nous allons rappeler quelques notions de base dans l'analyse de survie pour rendre la lecture plus facile.

1.1 Définitions et Notations

1.1.1 Les données de survie

Dans de nombreuses affections, la survie est le principal critère que l'on observe. En statistique, le terme de survie se généralise à l'apparition d'un évènement. Dans le domaine médical, on peut bien sûr observer le décès d'un patient, mais aussi l'apparition d'une pathologie ou la réponse à un traitement, alors qu'en industrie, cette variable peut être la durée séparant deux pannes successives d'une machine. En économie, on s'intéresse à la durée de chômage ou le temps passé dans un emploi, la durée de vie d'une entreprise. Dans le domaine bancaire, on peut s'intéresser à l'attrition d'un client et en assurances on s'intéresse à la durée de cotisation avant le premier remboursement. Ceci implique donc de tenir compte de la durée de surveillance de chaque sujet ainsi que du moment auquel chacun des évènements s'est produit. Comme dans toute analyse statistique, on veut décrire les observations, donc fournir les estimations de paramètres pertinents, comparer la survie de plusieurs groupes de sujets, expliquer et prédire la durée de survie en fonction de certains facteurs.

1.1.2 La durée de survie et la date d'origine

La durée de survie, notée par X, définie comme le délai écoulé entre deux états (états 0 et 1). Pour définir ce délai il est nécessaire de définir une date d'origine qui est la date de début du phénomène étudié. Par exemple, dans l'étude de l'évolution d'une maladie, la date d'origine X_0 est la date de début de la maladie et si on s'intéresse à l'âge du sujet à la survenue de l'évènement, la date d'origine sera la date de naissance du sujet $X_0 = 0$.

La principale source de difficulté dans l'analyse des durées de vie et pour diverses raisons, est la présence de données incomplètes ; pour lesquelles la variable d'intérêt n'est pas complètement observée pour toute les données de l'échantillon. Nous présentons brièvement deux cas de données incomplètes ; les données censurées et les données tronquées.

1.1.3 Les données censurées

En analyse de survie, les données recueillies sont la plupart du temps censurées ; ils se présentent quand le chercheur ne dispose pas d'assez de temps pour attendre que toutes les observations atteignent l'événement d'intérêt.

Dans ce cas la durée de survie X est définie comme le délai écoulé entre la date d'origine et la date de survenue de l'évènement, elle n'est pas observée sur toute la durée de l'étude pour tous les sujets, on définit alors l'évènement par le couple (T, δ), δ étant l'indicatrice d'observation de l'évènement et T le temps correspondant à l'information connue (le temps d'évènement ou le temps de censure). Il existe trois catégories de censure qu'on nomme censure à droite, censure à gauche et censure par intervalle.

Censure à droite

La durée de survie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de survie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue. Ce modèle est adapté au cas où l'événement considéré est le décès du patient et où la date de fin d'étude est préalablement fixée. Les différents modes de censure, décrits dans le cadre d'une censure à droite, sont les suivants,

1. **La censure de type I : censure fixée**

   Soit C une valeur fixée, au lieu d'observer les variables X_1,...,X_n qui nous intéressent, on observe X_i uniquement lorsque $X_i \leq C$.

   Les observations sont : $T_i = X_i \wedge C = \min(X_i, C)$. Ce mécanisme de censure est fréquemment rencontré dans les études épidémiologiques, correspond à l'observation de la durée de survie de n patients au cours d'une expérience de durée prédéterminée C et les applications industrielles. Par exemple, on peut tester la durée de vie de n objets identiques (ampoules) sur un intervalle d'observation fixé $[0, u]$.

2. **La censure de type II : censure aléatoire**

   Soient C_1,...,C_n des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) telle que les observations consistent en $\{(T_i, \delta_i), i \geq 1\}$,

   où $T_i = X_i \wedge C_i = min(X_i, C_i)$ et $\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$: l'indicateur de censure, sert en fait à connaître la nature de l'observation, il indique si $\delta_i = 1$ c'est une survie, l'événement est complètement observée c'est à dire $T_i = X_i$. Si $\delta_i = 0$ c'est une censure, l'individu n'est pas observée c'est à dire $T_i = C_i$.

   La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par

   - **la perte de vue :** le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients " perdus de vue ".
   - **l'arrêt ou le changement du traitement :** les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.
   - **la fin de l'étude :** l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement).

3. **La censure de type III : attente**

   On décide d'observer les durées de survie des n patients jusqu'à ce que r d'entre eux soient décédés et d'arrêter l'étude à ce moment là.

   Soit $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ les statistiques d'ordre associées à $X_1, X_2, ..., X_n$. Les observations consistent en $\{(T_i, \delta_i), i \geq 1\}$ où $T_i = X_i \wedge X_{(r)}$ et $\delta_i = \mathbb{1}_{\{X_i \leq X_{(r)}\}}$. Ce modèle souvent utilisé dans les études de fiabilité, correspond à l'observation de la durée de fonctionnement de n machines tant que r d'entre elles ne sont pas tombées en panne.

Exemple 1.1. On s'intéresse au temps de survie de personnes atteintes d'une maladie. On fixe le temps d'étude et à la fin de ce temps certaines personnes sont encore vivantes. Pour ces personnes, tout ce que l'on sait est que leur temps de survie dépasse le temps observé, ce sont des données censurées à droite de type I.

Censure à gauche

Une durée de survie est dite censurée à gauche si l'individu a déjà connu l'évènement d'intérêt avant l'entrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple (T, δ) :

$$T = (X \vee C) = max(X, C),$$
$$\delta = \mathbb{1}_{\{X \geq C\}}$$

Avec la durée de vie et le temps de censure C supposés indépendants. Si $\delta = 1$, le sujet subit l'évènement et est observé. Si $\delta = 0$, le sujet est dit censuré à gauche : au lieu d'observer X, on observe une valeur C avec pour seule information le fait que X soit inférieur à C.

Exemple 1.2. Sur le même (exemple 1.1) que précédemment, on ne peut pas toujours savoir le moment exact du déclenchement de la maladie, pour certaines personnes, on sait seulement que leur âge est inférieur à leur âge au moment de l'étude. Ces données sont censurées à gauche.

Censure par intervalle

Dans ce cas l'information apportée par l'expérience se traduit par une appartenance de la durée de survie à un intervalle de temps $(C_1 < X < C_2)$. Ceci est le cas lorsque les patients dans les essais cliniques ont des suivis périodiques, par exemple chaque six mois, si une maladie surgit, on sait seulement qu'elle est produite dans l'intervalle de temps séparant deux visites. Ce type de censure peut aussi apparaitre dans les expériences industrielles où il y a des inspections périodiques des machines.

1.1.4 Les données tronquées

Un autre cas où les données incomplètes apparaissent est celui des données tronquées. Lors d'une étude pratique sur les durées de vie, il n'est pas rare que la variable d'intérêt

X ne soit pas observable quand elle est inférieure à un seuil U ou une v.a. U , nous dirons qu'on est en présence d'une troncature gauche. Dans le cas contraire, on parle de troncature aléatoire à droite. La troncature élimine donc de l'étude une partie de l'échantillon, ce qui aura pour conséquence que l'analyse pourra porter seulement sur la loi conditionnelle de X sachant que $X \geq U$. La variable U, appelée variable de troncature droite (ou gauche), est supposée indépendante de la variable X. Il est également possible d'avoir la combinaison des deux mécanismes, on parle alors de troncature par intervalle. De nombreux travaux ont été effectués sur l'analyse de données tronquées (Klein et Moeschberger,1997).

Exemple 1.3. La troncature gauche apparaît, par exemple dans la recherche d'objets cachés qui devront être assez grands pour être détectés comme les réserves de petrole, les astres lointains,... Dans ce cas, nous disposons de N objets dans l'échantillon, mais nous ne sommes capables d'observer que les n objets suffisamment grands (supérieurs à U).

Exemple 1.4. Lagakos en (1998) présentent des données sur les temps d'infection et l'induction pour 258 adultes et 37 enfants qui ont été infectés par le virus de SIDA. Ici, le nombre de personnes infectés est inconnu et l'information est disponible seulement pour ceux qui ont été infectés et qui ont développés le SIDA dans un certain laps de temps. Ainsi, les personnes qui n'ont pas encore développé le SIDA ne sont pas connues de l'enquêteur et ne sont pas inclus dans l'échantillon. C'est le cas de troncature à droite.

1.2 Fonctions de base en analyse de survie

L'analyse de données de survie se résume finalement à l'étude d'une variable aléatoire positive continue X. Cette variable modélise la durée de survie, soit le délai entre la date d'origine et le temps de survenue de l'événement. La loi de X peut être décrite par plusieurs fonctions :

1.2.1 Fonction de survie

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

$$S(t) = \mathbf{P}(X > t), t \geq 0. \tag{1.1}$$

C'est une fonction continue monotone non croissante telle que $S(0) = 1$ et $\lim_{n \longrightarrow \infty} S(t) = 0$.

1.2.2 Fonction de répartition

La fonction de répartition $F(t)$ représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à-dire

$$F(t) = \mathbf{P}(X \leq t) = 1 - S(t). \tag{1.2}$$

1.2.3 Densité de probabilité

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$

$$F(t) = \int_0^t f(u)du.$$

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

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t < X \leq t + \Delta t)}{\Delta t} = F^{'}(t) = -S^{'}(t). \tag{1.3}$$

Pour t fixé, $f(t)\Delta t$ représente la probabilité de mourir entre t et $t + \Delta t$ pour un sujet.

1.2.4 Taux de hasard

Cette fonction est aussi appelée fonction de risque instantané de décés. Le taux de hasard, notée h est définie par

$$h(t) = \begin{cases} 0 \quad si\ S(t) = 0 \\ \frac{f(t)}{S(t)} \quad si\ S(t) \neq 0 \end{cases} \tag{1.4}$$

La terminologie de taux de hasard se justifie par le fait que si f est continue, alors pour $t > 0$ tel que $P(X > t) > 0$, on a

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbf{P}(t < X \leq t + \Delta t | X > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathbf{P}(t < X \leq t + \Delta t)}{\mathbf{P}(X > t)\Delta t} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{\mathbf{P}(t < X \leq t + \Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{-S^{'}(t)}{S(t)} \end{aligned}$$

Pour t fixé, $h(t)\Delta t$ représente la probabilité pour un sujet décéde entre t et $t + \Delta t$ sachant que ce sujet est encore vivant juste avant l'instant t.

1.2.5 La fonction de hasard cumulée

Supposons que h est intégrable sur tout compact (ce qui est le cas si f est continue), on définit la fonction de hasard cumulée, notée H par :

$$H(t) = \int_0^t h(u)du = -\ln(S(t)). \tag{1.5}$$

Pour t fixé, $H(t)$ représente la somme des risques de subir l'évènement dans l'intervalle $]0, t]$. On peut déduire de l'équation (1.5) une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

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

On en déduit que

$$f(t) = h(t)\exp(-\int_0^t h(u)du).$$

1.3 Quantités associées à la distribution de survie

Moyenne et variance de la durée de survie

Le temps moyen de survie $E(X)$ et la variance de la durée de survie $Var(X)$ sont définis par les quantités suivantes :

$$E(X) = \int_0^\infty S(t)dt, \quad Var(X) = 2\int_0^\infty tS(t)dt - (E(X))^2$$

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

Quantiles de la durée de survie

– La médiane de la durée de survie est le temps t pour lequel la probabilité de survie $S(t)$ est égale à 0.5, c'est-à-dire, la valeur t_m qui satisfait $S(t_m) = 0.5$. Dans le cas où l'estimateur est une fonction en escalier (ex : Kaplan-Meier), il se peut qu'il y est un intervalle de temps vérifiant $S(t_m) = 0.5$. Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[B_i, B_s]$ un intervalle de confiance de niveau α de $S(t_m)$, alors un intervalle de confiance de niveau α du temps médian t_m est

$$[S^{-1}(B_s), S^{-1}(B_i)]$$

1.4 Estimation de la fonction de survie

Si l'on ne peut pas supposer a priori que la loi de la durée de survie obéit à un modèle paramétrique, on peut estimer la fonction de survie S grâce à plusieurs méthodes non-paramétriques dont la plus intéressante est celle de Kaplan-Meier.

1.4.1 Estimateur de Kaplan-Meier

Kaplan et *Meier* (1958), ([15]) ont obtenu un estimateur de la fonction de survie pour des données aléatoirement censurées à droite. Cet estimateur permet d'intégrer l'information provenant de toutes les observations disponibles, tant censurées que non censurées, parce que la survie jusqu'à tout récemment est considérée comme une série d'étapes définies par les durées de survie et les durées censurées observées.

Soit $X_1, X_2, ..., X_n$ une suite de v.a. positives indépendantes et identiquement distribuées (i.i.d.) de fonction de répartition (f.d.r.) F représentant les durées de survie et $C_1, C_2..., C_n$ est la suite de v.a i.i.d. représentant les censures de f.d.r. G.
Pour des raisons d'identifiabilité de modèle on suppose que la suite (X_i) est indépendante de la suite (C_i) et on observe $\{(T_i, \delta_i), i \geq 1\}$ où $T_i = X_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$. On ordonne les T_i par ordre croissant $T_{(1)} \leq T_{(2)} \leq ... \leq T_{(n)}$ et on prend $\delta_{(1)}, \delta_{(2)}, ..., \delta_{(n)}$ correspondants.

L'idée de construction de l'estimateur de Kaplan-Meier (Kaplan et Meier [1958]) s'appuie sur la remarque suivante : si $t' < t$, la probabilité de survivre au-delà de l'instant t est égale au produit suivant :

$$\begin{aligned} S(t) &= \mathbf{P}(X > t) \\ &= \mathbf{P}(X > t/X > t')\mathbf{P}(X > t') \\ &= \mathbf{P}(X > t/X > t')S(t') \end{aligned}$$

Si l'on renouvelle l'opération en choisissant une date $t'' < t$ antérieure à $t' < t$ on aura de même

$$\begin{aligned} S(t') &= \mathbf{P}(X > t') \\ &= \mathbf{P}(X > t'/X > t'')\mathbf{P}(X > t'') \\ &= \mathbf{P}(X > t'/X > t'')S(t'') \end{aligned}$$

En réitérant, nous distinguons des produits d'éléments en $\mathbf{P}(X > t/X > t^{'})$. Le but est alors d'observer les instants où se produit un évènement, qu'il s'agisse d'une mort ou d'une censure, et de conditionner par rapport à ces moments. Nous définissons alors les probabilités conditionnelles suivantes :

$$p_i = \mathbf{P}(X > T_{(i)}/X > T_{(i-1)})$$

où p_i est la probabilité de survivre sur l'intervalle $I_i =]T_{(i-1)}, T_{(i)}]$ sachant que l'individu était vivant à l'instant $T_{(i-1)}$. On peut alors estimer $q_i = 1 - p_i$ qui est la probabilité de mourir pendant l'intervalle I_i sachant que l'on était vivant au début de cet intervalle. Alors l'estimateur naturel de q_i est :

$$\widehat{q_i} = \frac{M_i}{R_i} \tag{1.6}$$

M_i est le nombre des sujets décédés sur l'intervalle $]T_{(i-1)}, T_{(i)}]$ et R_i représente le nombre des sujets qui sont vivants (donc " à risque" de mourir) juste avant l'instant $T_{(i)}$. Si $\delta_{(i)} = 1$, c'est qu'il y a eu un mort en $T_{(i)}$ et donc $M_i = 1$. Si $\delta_{(i)} = 0$, c'est qu'il y a eu une censure en $T_{(i)}$ et donc $M_i = 0$.

Par suite, $\widehat{p_i} = \begin{cases} 1 - \frac{1}{R_i} & \text{en cas de mort en } T_{(i)} \\ 1 & \text{en cas de censure en } T_{(i)} \end{cases}$

On a $R_i = n - (i-1)$(car il y'a eu $(i-1)$ "sujets" ou censures avant $T_{(i)}$ et il y'a n individus dans l'étude)

L'estimateur de Kaplan Meier ([21]) peut donc être défini de la façon suivante :

$$\widehat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} \tag{1.7}$$

Remarque 1.1. Cas où il y a des ex-æquo :

1. *Si ces ex-æquo sont tous des morts, la seule différence tient à ce que M_i n'est plus égal à 1 mais au nombre des morts et l'estimateur de Kaplan-Meier devient :*

$$\widehat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{M_i}{R_i}\right)^{\delta_{(i)}} \tag{1.8}$$

2. *Si ces ex-æquo sont des deux sortes, on considère que les observations non censurées ont lieu juste avant les censurées.*

Remarque 1.2. *1. L'estimateur de Kaplan-Meier est une fonction en escaliers qui fait des sauts à chaque instant t_i représentant un décès. La valeur du saut dépend du nombre d'évènements au temps t_i et aussi du nombre de censures à ce temps là.*

2. Pour un échantillon i.i.d. de durées non censurées $(X_i)_{i=1,\dots,n}$, un estimateur naturel de la survie de la variable X est la survie empirique

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{k} \mathbb{1}_{\{X_i > x\}} \tag{1.9}$$

1.4.2 Exemples

Exemple 1. Données de Freireich.

En 1963, le Docteur Freireich a fait un essai thérapeutique pour comparer les durées de rémission (des rechutes), en semaines de 21 patients atteints de leucémie aiguë selon qu'ils ont reçu ou non un médicament appelé 6-mercaptopurine (6-MP), le groupe témoin a reçu un placebo. Les résultats obtenus sont les suivantes :

6 M-P	6	6	6	6^+	7	9^+	10	10^+	11^+	13	16
	17^+	19^+	20^+	22	23	25^+	32^+	32^+	34^+	35^+	
placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	

Les nombres avec + correspondent à des patients perdus de vue à la date considérée. Ils sont exclus "vivants" de l'étude et leur "durée de survie" est supérieure à celle indiquée. Ils sont censurés et correspond au type de censure II (aléatoire). Par exemple le 4ème patient est perdu de vue au bout de 6 semaines de traitement avec le 6-MP : il a donc une durée de rémission supérieure à 6 semaines.

Dans l'analyse de survie on tient compte de toutes les observations censurées ou non. En effet dans les problèmes d'estimations statistiques si on élimine les observations censurés du groupe traité par le 6 M-P (12 patients) on perd de l'information puisque on ne tient pas compte des patients ayant des durées de rémission plus longues.

– L'estimateur de Kaplan-Meier (1.8) de la fonction de survie S du groupe de 21 malades traité par le traitement 6-MP (existence d'ex-aequo) donne le tableau suivant :

Temps t_i	R_i	M_i	$\widehat{S}_{(6-MP)}(t_i)$	Intervalle
0	21	0	1	[0, 6[
6	21	3	(1-3/21)$*$1 = 0.857	[6, 7[
7	17	1	(1-1/17)$*$0.857 = 0.807	[7, 10[
10	15	1	(1-1/15)$*$0.807 = 0.753	[10, 13[
13	12	1	(1-1/12)$*$0.753 = 0.690	[13, 16[
16	11	1	(1-1/11)$*$0.690 = 0.627	[16, 22[
22	7	1	(1-1/7)$*$0.627 = 0.538	[22, 23[
23	6	1	(1-1/6)$*$0.538 = 0.448	[23, 1[

– L'estimateur (1.9) pour le groupe traité par un placebo (pas de censure) donne le tableau suivant :

Semaine i	Nombre de rémissions à la semaine i	$\widehat{S}_{(placebo)}(t_i)$
0	21	1
1	19	(19/21)=0.90
2	17	(17/21)=0.81
3	16	0.76
4	14	0.66
5	12	0.57
8	8	0.38
11	6	0.28
12	4	0.19
15	3	0.14
17	2	0.09
22	1	0.05
23	0	0

Représentations graphiques : Le graphe suivant présente les estimateurs de la fonction de survie par Kaplan-Meier (1.8) et par (1.9) pour les deux traitements :

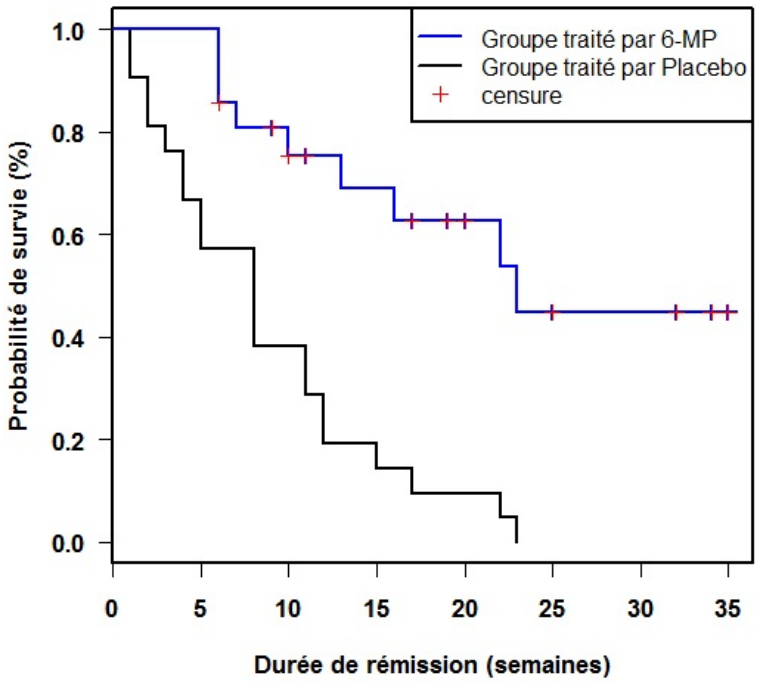


FIGURE 1.1 – Estimations de la fonction de survie pour les deux traitements

Remarque :

- Il existe des durées supérieures à 23 semaines pour le groupe traité par le traitement 6-MP mais elles sont toutes censurées. En conséquence dans le groupe traité par un placebo, et contrairement au précédent, la valeur estimée de la fonction de survie correspondant au temps d'évènement maximal observé (soit 23 semaines) ne s'annule pas. Précédemment nous avions $S(23) = 0$ du fait que la durée maximale était non censurée. En d'autres termes, le fait que l'estimateur de la fonction de survie s'annule ne signifie pas que tous les individus ont connu l'évènement étudié mais seulement que la durée maximale ne correspond pas à une censure.
- En ce qui concerne la représentation graphique de la fonction de survie beaucoup vont la tracer jusqu'au temps $t = 23$, ce qui est raisonnable puisque l'estimateur K.M n'est pas défini au-delà du temps d'évènement maximal. Toutefois, vous trouverez aussi des présentations qui vont la prolonger jusqu'au temps $t = 35$, maximum des temps censurés qui est ici supérieur au plus grand temps d'évènement connu, avec une horizontale d'ordonnée 0.448.

Exemple 2. Données des ventilateurs.

Le tableau suivant donne la durées de fonctionnement, en heures de 70 ventilateurs parmi lesquels certains ont présnté une défaillance et les autres qui tombent en panne pendant une durée de fonctionnement, correspondent à des données censurées.

4.5	4.6^+	11.5	11.5	15.6^+	16.0	16.6^+	18.5^+	18.5^+	18.5^+
18.5^+	18.5^+	20.3^+	20.3^+	20.3^+	20.7	20.7	20.8	22.0^+	30.0^+
30.0^+	30.0^+	30.0^+	31.0	32.0^+	34.5	37.5^+	37.5^+	41.5^+	41.5^+
41.5^+	41.5^+	43.0^+	43.0^+	43.0^+	43.0^+	46.0	48.5^+	48.5^+	48.5^+
48.5^+	50.0^+	50.0^+	50.0^+	61.0^+	61.0	61.0^+	61.0^+	63.0^+	64.5^+
64.5^+	67.0^+	74.5^+	78.0^+	78.0^+	81.0^+	81.0^+	82.0^+	85.0^+	85.0^+
85.0^+	87.5^+	87.5	87.5^+	94.0^+	99.0^+	101.0^+	101.0^+	101.0^+	115.0^+

Représentations graphiques : Le graphe suivant donne l'estimateurs de Kaplan-Meier de la fonction de survie S des données de ventilateurs.

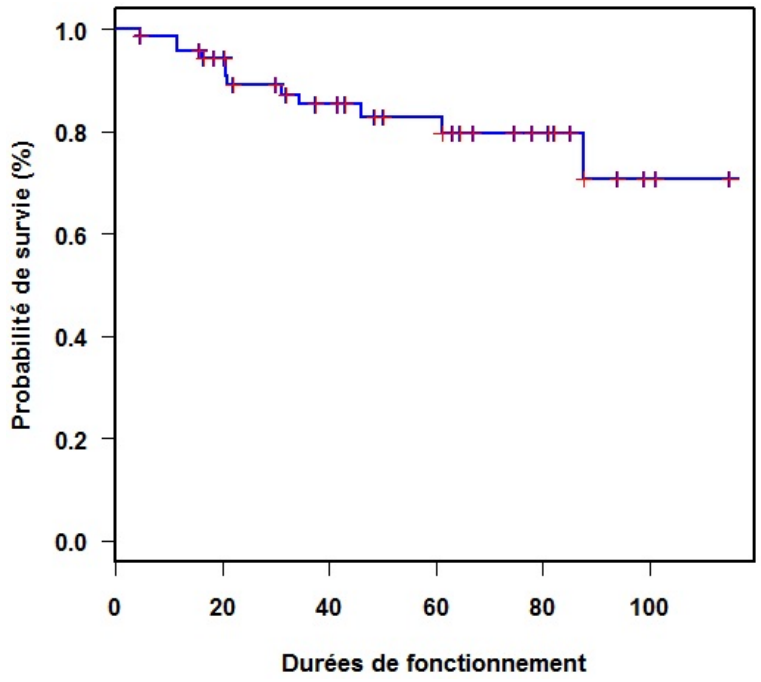


FIGURE 1.2 – L'estimateur de Kaplan-Meier des données de ventilateurs

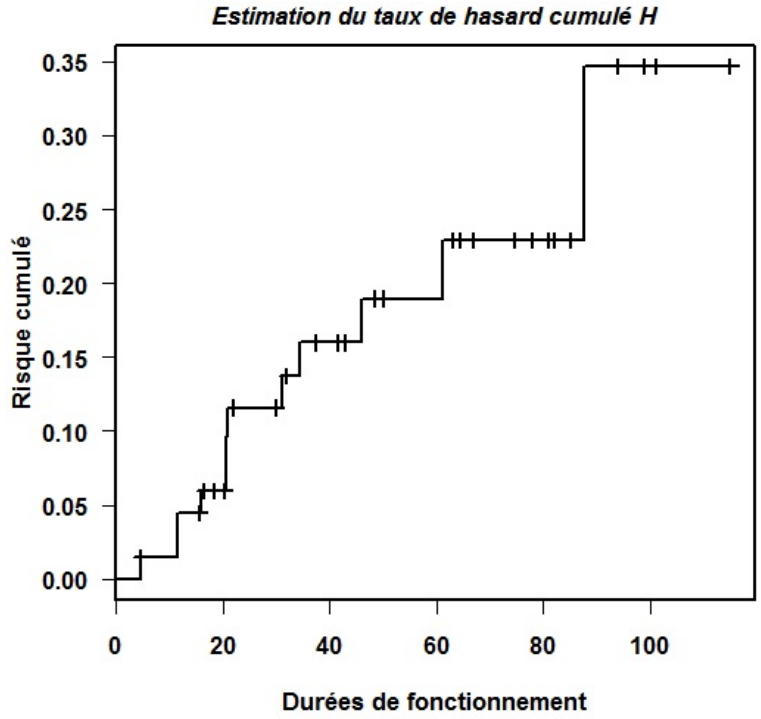
Par exemple $\widehat{S}(60 \times 10^3) = 0.8$ ce qui donne 80% des ventilateurs vont fonctionner au delà de 60×10^3 heures.

L'estimateur de Kaplan-Meier de S dans cet exemple se calcule par la formule (1.8) comme suit :

Temps t_i	R_i	M_i	$\widehat{S}(t_i)$	Intervalle
0	70	0	1	[0, 4.5[
4.5	70	1	(1-1/70)∗1 = 0.986	[4.5, 11.5[
11.5	68	2	(1-2/68)∗0.986 = 0.957	[11.5, 16.0[
16.0	65	1	(1-1/65)∗0.957 = 0.942	[16.0, 20.7[
20.7	55	2	(1-2/55)∗0.942 = 0.908	[20.7, 20.8[
20.8	53	1	(1-1/53)∗0.908 = 0.891	[20.8, 31.0[
31.0	47	1	(1-1/47)∗0.891 = 0.872	[31.0, 34.5[
34.5	45	1	(1-1/45)∗0.872 = 0.852	[34.5, 46.0[
46.0	34	1	(1-1/34)∗0.852 = 0.827	[46.0, 61.0[
61.0	26	1	(1-1/26)∗0.827 = 0.795	[61.0, 87.5[
87.5	9	1	(1-1/9)∗0.795 = 0.707	[87.5, ∞[

Estimation du taux de hasard cumulé H

Le graphe suivant donne l'estimateur $\widehat{H} = -ln(\widehat{S}(t))$ du taux de hasard cumulé des données des ventilateurs.



Par exemple $H(60) = 0.18 = 18\%$, on a 18% de risque que le lot des ventilateurs tombe en panne pendant une durée de fonctionnement de 60.10^3 heures.

1.5 Processus ponctuels associés

Au milieu des années 1970, Aalen présente une théorie des martingales pour le processus de comptage qui offre un cadre unifié pour les méthodes statistiques de l'analyse de survie. Les processus de comptage sont utilisés dans les représentations intégrales des statistiques des données censurées. La théorie fournit des formes simples et unifiées des estimateurs, des statistiques de test et des méthodes de régression. Ces méthodes permettent aussi d'obtenir des expressions simples de statistiques compliquées et des distributions asymptotiques des test et des estimateurs.

Dans l'analyse des durées de survie, chaque "individu" a une durée de vie X, de densité f, de fonction de répartition F et de fonction de survie $S = 1 - F$ et les v.a. X et la censure C sont supposées indépendantes. Pour chaque "individu" on lui associe un couple de processus ponctuel $((Y(t), N(t)), t \geq 0)$ donné par (où $T = X \wedge C$)

$$\begin{aligned} Y(t) &= \mathbb{1}_{\{T \geq t\}} = \mathbb{1}_{\{X \wedge C \geq t\}} \\ &= \begin{cases} 1 & \text{si l'individu est en observation et à risque à l'instant t} \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

Le processus $(Y(t), t \geq 0)$ est dit processus risque : il indique la présence du risque juste avant l'instant t. Le processus $N(t)$ est défini par :

$N(t)$ = l'indicateur de l'"évènements d'intérêt" dans l'intervalle $]0, t]$

$N(t) = \mathbb{1}_{\{T \leq t, \delta = 1\}} = \mathbb{1}_{\{X \leq t, \delta = 1\}} = \mathbb{1}_{\{X \leq t\}} \mathbb{1}_{\{\delta = 1\}} = \mathbb{1}_{\{X \leq t\}} \mathbb{1}_{\{X \leq C\}}$

Dans ce mémoire on se place dans le cadre d'un mécanisme de censure aléatoire à droite seulement.

Notations : Si dans l'étude nous avons n "individus" (machine, patients ...ect) indexés par $\{i = 1, ..., n\}$, on note $T_1, T_2, ..., T_n$ les instants de survenue de "l'évènement d'intérêt". On observe $T_i = X_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$. Donc on dispose des observations $(T_i, \delta_i), i = 1, ..., n$. A chaque "individu i" correspond le processus risque $Y_i(t)$ défini par

$$Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$$

et le processus ponctuel $N_i(t)$ indicateur de "survenue de l'évènement" dans $]0, t]$

$$N_i(t) = \mathbb{1}_{\{T_i \leq t, \delta_i = 1\}}$$

de même le processus ponctuel $M_i(t)$ indicateur de "censure de l'évènement" dans $]0,t]$

$$M_i(t) = \mathbb{1}_{\{T_i \leq t, \delta_i = 0\}}$$

On définit les processus ponctuels suivants définis sur $[0,\infty[$:

Le processus risque $\overline{Y}_n(t)$ par

$$\overline{Y}_n(t) = \sum_{i=1}^{n} Y_i(t)$$

qui est le nombre d'"individus" à risque à l'instant t et en observation.

Le processus $\overline{N}_n(t)$ par

$$\overline{N}_n(t) = \sum_{i=1}^{n} N_i(t)$$

qui est le nombre d'"évènements" observés dans l'intervalle $]0,t]$.

Le processus $M_n(t)$ défini par $M_n(t) = \sum_{i=1}^{n} M_i(t)$ qui est le nombre d'observations censurées inférieures ou égales à t.

On définit aussi les processus suivants :

$$\Lambda(t) = \int_0^t \overline{Y}_n(s) dH(s)$$

le processus intensité cumulée

$$\overline{M}_n(t) = \overline{N}_n(t) - \Lambda(t) = \overline{N}_n(t) \int_0^t \overline{Y}_n(s) h(s) ds$$

L'estimateur de Nelson-Aalen $\widehat{H}_{NA}(t)$ du taux de risque cumulé $H(t)$ a été proposé par Nelson en 1972 et ensuite par Aalen en 1978, ([3]).

$$\widehat{H}_{NA}(t) = \int_0^t \frac{d\overline{N}_n(u)}{\overline{Y}_n(u)}$$

Remarque 1.3. *Dans le cas général où F est une f.d.r. quelconque, la fonction de survie $S(t) = \exp(-H(t))$ est l'unique solution de l'équation intégrale suivante :*

$$S(t) = 1 - F(t) = 1 - \int_0^t S(s^-)dH(s)$$

et l'estimateur de Kaplan-Meier Vérifie

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(s^-)d\widehat{H}_{NA}(s) := 1 - \widehat{F}_n(t)$$

1.6 Propriétés de l'estimateur de Kaplan-Meier

L'E.K.M. possède beaucoup de propriétés analogues à celles de la fonction de répartition empirique. Nous commençons par ses propriétés de convergence. Selon les références [9] et [20], on a :

Théorème 1.1. *(Shorack et Wellner,(1986))*
Si pour $t > 0$, $F(t) < 1$ et $\overline{Y}_n(t) \xrightarrow[n\longrightarrow+\infty]{P} \infty$ alors

$$\sup_{0\leq s\leq t} |\widehat{S}(s) - S(s)| \xrightarrow[n\longrightarrow+\infty]{P} 0$$

Preuve :

L'estimateur de Kaplan-Meier vérifie $\widehat{S}(s) = 1 - \widehat{F}_n(s)$ où $\widehat{F}_n$ est définie par :

$$\widehat{F}_n(s) = \int_0^t \widehat{S}(s^-)d\widehat{H}_{NA}(s)$$

Il suffit de montrer que

$$\sup_{0\leq s\leq t} |\widehat{F}_n(s) - F(s)| \xrightarrow[n\longrightarrow\infty]{P} 0.$$

Soit $t > 0$, $F(t) < 1$ et $\overline{Y}_n(t) \xrightarrow[n\longrightarrow\infty]{P} \infty$, le th.3.2.3 p. 97 de [10], donne

$$\mathbf{P}\left(\frac{\widehat{F}_n(s) - F(s)}{1 - F(s)} = Z_n(s), s \in [0, t]\right) \xrightarrow[n\longrightarrow\infty]{} 1$$

où

$$Z_n(s) = \int_0^t \frac{1 - \widehat{F}_n(u^-)}{1 - F(u)} \frac{\mathbb{1}_{\{\overline{Y}_n(u)>0\}}}{\overline{Y}_n(u)} d\overline{M}_n(u)$$

Il suffit maintenant de montrer que

$$\sup_{0\leq s\leq t}|Z_n(s)| \xrightarrow[n\longrightarrow\infty]{P} 0$$

pour avoir

$$\sup_{0\leq s\leq t}|\widehat{F}_n(s)-F(s)| \xrightarrow[n\longrightarrow\infty]{P} 0$$

Par l'inégalité de Lenglart, pour η et ε positifs nous avons

$$\begin{aligned}
\mathbf{P}\left(\sup_{0\leq s\leq t}|Z_n(s)|^2\geq\varepsilon\right) &\leq \mathbf{P}\left(\sup_{0\leq s\leq t}|\int_0^s\frac{1-\widehat{F}_n(u^-)}{1-F(u)}\frac{\mathbb{1}_{\{\overline{Y}_n(u)>0\}}}{\overline{Y}_n(u)}d\overline{M}_n(u)|^2\geq\varepsilon\right)\\
&\leq \frac{\eta}{\varepsilon}+\mathbf{P}\left(\int_0^t\left(\frac{1-\widehat{F}_n(u^-)}{1-F(u)}\frac{\mathbb{1}_{\{\overline{Y}_n(u)>0\}}}{\overline{Y}_n(u)}\right)^2 d<\overline{M}_n>_u\geq\eta\right)\\
&\leq \frac{\eta}{\varepsilon}+\mathbf{P}\left(\int_0^t\left(\frac{1-\widehat{F}_n(u^-)}{1-F(u)}\right)^2\frac{\mathbb{1}_{\{\overline{Y}_n(u)>0\}}}{\overline{Y}_n(u)}dH(u)\geq\eta\right)\\
&\leq \frac{\eta}{\varepsilon}+\mathbf{P}\left(\frac{1}{(1-F(t))^2}\frac{1}{\overline{Y}_n(t)}H(t)\geq\eta\right)\\
&= \frac{\eta}{\varepsilon}+\mathbf{P}\left(\overline{Y}_n(t)\leq\frac{H(t)}{\eta(1-F(t))^2}\right)
\end{aligned}$$

Or

$$\overline{Y}_n(t) \xrightarrow[n\longrightarrow\infty]{P} \infty$$

donc

$$\mathbf{P}\left(\overline{Y}_n(t)\leq\frac{H(t)}{\eta(1-F(t))^2}\right)\xrightarrow[n\longrightarrow\infty]{} 0$$

Comme ε et η sont arbitraires on en déduit

$$\sup_{0\leq s\leq t}|Z_n(s)| \xrightarrow[n\longrightarrow\infty]{P} 0$$

Par conséquent

$$\sup_{0\leq s\leq t}|\widehat{F}_n(s)-F(s)| \xrightarrow[n\longrightarrow\infty]{P} 0$$

D'où le résultat.

Rappelons qu'on se place toujours dans le cadre de la censure aléatoire droite, on suppose que C est indépendante de X, la variable observée n'est plus X mais $T = X \wedge C$, de f.d.survie Ψ, vérifie $\Psi = S.G$ (G la f.d.survie de la censure droite C), on a le théorème suivant

Théorème 1.2. *(Gill(1980))* [9] *Dans l'espace $D([0,\tau], \| \ . \ \|_\infty, \mathcal{P})$ des fonctions continues à droite possédant des limites à gauche en tout point de $[0,\tau]$, muni de la norme infinie et de sa tribu de projection, si $\Psi(\tau^-) > 0$ et si S est continue sur $[0,\tau]$, alors*

$$\sqrt{n}(\widehat{S}(t) - S(t)) \overset{L}{\Longrightarrow} Z,$$

où Z est un processus gaussien centré, de fonction de covariance

$$Cov(Z(s), Z(t)) = S(s)S(t) \int_0^{s\wedge t} \frac{dF(u)}{S^2(u)(1-G(u))}.$$

Cela signifie que pour toute fonction Φ de $D([0,\tau], \| \ . \ \|_\infty, \mathcal{P})$ vers $\mathbb{R}$, mesurable continue et bornée, on a lorsque $n \to +\infty$,

$$E[\Phi\left(\sqrt{n}(\widehat{S}(t) - S(t))\right)] \longrightarrow E[\Phi(Z)].$$

En particulier, on obtient la normalité asymptotique de $\widehat{S}(t_0)$ en tout point $t_0 \in [0,\tau]$.

Théorème 1.3. [9] *En tout point t_0 de continuité de S, $t_0 \in [0,\tau]$ et $S(\tau^-) > 0$,*

$$\sqrt{n}(\widehat{S}(t_0) - S(t_0)) \underset{n\longrightarrow+\infty}{\overset{L.}{\longrightarrow}} \mathcal{N}(0, V^2(t_0)),$$

avec

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{S(du)}{S^2(u)G(u)},$$

où $G(t)$ la fonction de survie de la variable C.

Preuve :

Considérons les quantités $\Psi(t) = \mathbf{P}(T > t)$ et $\Psi^{(1)}(u) = \mathbf{P}(T > u, \delta = 1)$. D'après l'hypothèse d'indépendance, on obtient les égalités suivantes

$$\Psi(t) = \mathbf{P}(T > t) = \mathbf{P}(X > t, C > t) = S(t)G(t)$$

$$\begin{aligned}\Psi^{(1)}(t) &= \mathbf{P}(T > t, \delta = 1) = E[\mathbb{1}_{\{X \leq C, X > t\}}] = E(G(X^-)\mathbb{1}_{\{X>t\}}) \\ &= \int_t^{+\infty} G(u^-)f(u)du = -\int_t^{+\infty} G(u^-)S(du).\end{aligned}$$

Par conséquent, $\Psi^{(1)}(dt) = G(t^-)S(dt)$ et on peut ainsi écrire

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0} \frac{\Psi^{(1)}(du)}{\Psi(u)G(u^-)S(u)}.$$

En remplaçant les fonctions $\Psi(u)$ et $\Psi^{(1)}(u)$ par leurs équivalents empiriques (calculables car les variables T et δ sont observées), c-à-dire $\Psi(u)$ par $\Psi_n(u) = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{T_i>u\}}$ et $\Psi^{(1)}(u)$ par $\Psi_n^{(1)}(u) = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{T_i>u,\delta_i=1\}} = 1 - n^{-1}\sum_{i=1}^n \mathbb{1}_{\{T_i\leq u,\delta_i=1\}}$, et S par $\widehat{S}$, on obtient l'estimateur suivant :

$$\widehat{V}^2(t_0) = -\widehat{S}^2(t_0)\int_0^{t_0} \frac{\Psi_n^{(1)}(du)}{\Psi_n(u)\Psi_n(u^-)}$$

Un estimateur de la variance de l'estimateur de Kaplan-Meier (qui converge presque sûrement, lorsque $n \to +\infty$, vers la variance asymptotique de $\widehat{S}$) est

$$\widehat{Var}(\widehat{S}(t)) = \frac{1}{n}V^2(t).$$

Autre écriture de l'estimateur de Greenwood : Avec les notations, $M_{(i)}$ le nombre de décès en $T_{(i)}$ et R_i le nombre d'individus à risque de subir l'évènement juste avant le temps $T_{(i)}$, soit encore $R_i = \sum_{j=1}^n \mathbb{1}_{\{T_j \geq T_{(i)}\}}$.

On remarque que :

$$\Psi_n(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i>u\}} = \frac{M_i - R_i}{n},$$

$$\Psi_n(u^-) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq u\}} = \frac{M_i}{n},$$

$$\Psi_n^{(1)}(du) = -\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \in [u;u+du], \delta_i = 1\}} = -\frac{R_i}{n}.$$

Alors, on estime la variance de $\widehat{S}(t)$ par :

$$\begin{aligned}
\widehat{Var}(\widehat{S}(t)) &= n^{-1}\widehat{V}^2(t) \\
&= n^{-1}\widehat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{M_i/n}{(R_i - M_i)/n(R_i/n)} \\
&= \widehat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{M_i}{(R_i - M_i)R_i}
\end{aligned}$$

Remarque 1.4. *L'intérêt de résultats de convergence au niveau du processus lui même plutôt que pour un instant fixé est que l'on peut en déduire des bandes de confiance asymptotique pour l'estimateur de Kaplan-Meier. On peut trouver dans Gill [1980] une démonstration de la normalité asymptotique de* $\widehat{S}(t)$, *fondée sur la théorie des processus ponctuels.*

1.7 Estimateur de la variance de $\widehat{S}(t)$

Si les observations censurées sont les (T_i, δ_i), $i = 1, 2, ..., n$, avec éventuellement des ex-æquo, et $T_{(1)} < T_{(2)} < ... < T_{(k)}$ la suite des T_i réordonnés par ordre croissant, M_i le nombre des morts à l'instant $T_{(i)}$, R_i le nombre des sujets à risque à $T_{(i)}$, l'estimateur de la variance de l'estimateur de Kaplan-Meier à un temps t fixé est donné par la formule de Greenwood (Greenwood, 1926)[3] :

$$\widehat{Var}\widehat{S}(t) = \widehat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{M_i}{R_i(R_i - M_i)}$$

En particulier, s'il n'y a pas d'ex-æquo

$$\widehat{Var}\widehat{S}(t) = \widehat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{\delta_i}{(n-i)(n-i+1)}$$

Remarque 1.5.

L'estimateur de Greenwood a été obtenu grâce aux considérations suivantes[3], [25] :

On part de la formule de calcul de l'estimateur Kaplan Meier :

$$\widehat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{M_i}{R_i}\right)$$

et donc :

$$\log \widehat{S}(t) = \sum_{T_{(i)} \leq t} \log \widehat{p}_i$$

En utilisant **la méthode delta**, [25]
soit à calculer $Var[g(Z)]$ où $g()$ est une fonction continue dérivable. Un développement de Taylor à l'ordre 1 au voisinage de z_0 donne :

$$g(Z) = g(z_0) + (Z - z_0)\frac{\partial g(Z)}{\partial z_0} = g(z_0) + (Z - z_0)g^{'}(z_0)$$

et donc :

$$Var[g(Z)] = [g^{'}(z_0)]^2 Var(Z) \tag{1.10}$$

Ici $g = Fonction\ \ logarithme$ et $Z = \widehat{p}_i$

$$Var(\log \widehat{S}(t)) = \sum_{T_{(i)} \leq t} Var(\log \widehat{p}_i)$$

$$Var(\log \widehat{p}_i) = Var(\widehat{p}_i)\left(\frac{\partial}{\partial p_i} \log p_i\right)^2 = Var(\widehat{p}_i)\frac{1}{p_i^2}$$

On a $p_i = \mathbf{P}(X > T_{(i)}/X > T_{(i-1)})$ et $\widehat{p}_i = [1 - \frac{M_i}{R_i}]$ est l'estimateur de p_i
Lorsque $R_i = r$ (r un entier $\leq n$) alors

$$M_i \sim \mathcal{B}(r, 1 - p_i) \Rightarrow Var(M_i) = rp_i(1 - p_i)$$

Enfin, $Var(\widehat{p}_i) = Var(1 - \frac{M_i}{R_i}) = \frac{1}{R_i^2}Var(M_i) = \frac{R_i p_i(1-p_i)}{R_i^2} = \frac{p_i(1-p_i)}{R_i}$ que l'on peut estimer par :

$$Var(\widehat{p}_i) = \frac{\widehat{p}_i(1 - \widehat{p}_i)}{R_i}$$

donc

$$Var(\log \widehat{p}_i) = \frac{\widehat{p}_i(1 - \widehat{p}_i)}{R_i}\frac{1}{p_i^2} = \frac{(1 - \widehat{p}_i)}{R_i \widehat{p}_i}$$

$$Var(\log \widehat{S}(t)) = \sum_{T_{(i)} \leq t} \frac{(1-\widehat{p}_i)}{R_i \widehat{p}_i} = \sum_{T_{(i)} \leq t} \frac{1 - \frac{R_i - M_i}{R_i}}{R_i \frac{R_i - M_i}{R_i}}$$

Appliquant encore une fois la formule (1.10) avec $g = Fonction \ \ logarithme$ et $Z = \widehat{S}(t)$,

$$Var(\log \widehat{S}(t)) \approx \frac{1}{\widehat{S}(t)^2} Var(\widehat{S}(t))$$

on a finalement l'approximation suivante :

$$Var(\widehat{S}(t)) \approx \widehat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{M_i}{R_i(R_i - M_i)} \tag{1.11}$$

Remarque 1.6. **Estimation de la fonction quantile**

La fonction quantile de la durée de survie est définie par

$F(t_p) = p$ où $S(t_p) = 1 - p, 0 < p < 1$. Lorsque la fonction de répartition F est strictement croissante et continue alors $t_p = F^{-1}(p)$ où $t_p = S^{-1}(1-p)$.

L'estimateur de fonction quantile ([21]) de la durée de survie est définie par

$$\widehat{t_p} = \inf(t : F(t) \geq p), 0 < p < 1$$

$$\widehat{t_p} = \inf(t : S(t) \leq 1 - p)$$

La variance du quantile de survie est donnée par :

$$\widehat{Var}(\widehat{t_p}) = \frac{\widehat{Var}(\widehat{S}(\widehat{t_p}))}{(\widehat{f}(\widehat{t_p}))^2} \quad ([25])$$

où $\widehat{Var}(\widehat{S}(\widehat{t_p}))$ est l'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier, et $\widehat{f}(\widehat{t_p})$ c'est l'estimation de la densité de probabilité au point $\widehat{t_p}$. Elle est défini par : $\widehat{f}(\widehat{t_p}) = \frac{\widehat{S}(\widehat{u}_p) - \widehat{S}(\widehat{l}_p)}{\widehat{l}_p - \widehat{u}_p}$ où $\widehat{u}_p = max\{t_i | \widehat{S}(t_i) \geq 1 - p + \epsilon\}$, et $\widehat{l}_p = max\{t_i | \widehat{S}(t_i) \geq 1 - p - \epsilon\}$, pour $i = 1, ..., r \leq n$ avec r le nombre des instants complètement observés et en géneral on prend $\epsilon = 0.05$. (Voir exemple 1.8.2)

1.8 Intervalles de confiance

Pour évaluer la confiance que l'on peut avoir en une estimation, il est nécessaire de lui associer un intervalle qui contient, avec une certaine probabilité, la vraie valeur du paramètre.

Définition d'un intervalle de confiance

L'estimation par intervalle de confiance d'un paramètre θ consiste à associer à un échantillon, un intervalle aléatoire I, choisi de telle façon que la probabilité pour qu'il contienne la valeur inconnue du paramètre soit égale à un nombre fixé à l'avance, aussi grand que l'on veut. On écrit :

$$\mathbf{P}(\theta \in I) = 1 - \alpha$$

$1 - \alpha$ est la probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre, c'est le seuil de confiance. Cette probabilite est fixée par le risque d'erreur α choisi a priori (en general on choisit $\alpha = 5\%$).

1.8.1 Construction d'intervalles de confiance pour la survie

L'estimation ponctuelle de la survie à un moment donné doit impérativement être accompagné d'un intervalle de confiance (IC), gage de la précision de l'estimation, habituellement au seuil de $\alpha = 0.05$ (I.C à 95%).

Il s'agit de trouver deux bornes a et b telles que $\forall\ t > 0$ on a : $\mathbf{P}[a \leq S(t) \leq b] = 1 - \alpha$, ou α est un seuil de risque fixe a priori. La connaissance des bornes de confiance pour la fonction de survie $S(t)$ constitue en effet une information utile dans les études de survie.

Il a été démontré que $\sqrt{n}(\widehat{S}(t) - S(t))$ converge vers une martingale gaussienne centrée (Théorème 1.3). Une des conséquences est que la distribution asymptotique de $\widehat{S}(t)$ est gaussienne et centrée sur $S(t)$, pour de grands échantillons. Compte-tenu des resultats precedents, son ecart-type estime, noté $\widehat{\sigma}$, est donne par :

$$\widehat{\sigma}(\widehat{S}(t)) = \sqrt{Var(\widehat{S}(t))} \tag{1.12}$$

et donc un intervalle de confiance au seuil $100(1 - \alpha)\%$ ([2], [4]) peut être construit selon :

$$[\widehat{S}(t) - \widehat{\sigma}(\widehat{S}(t))z_{1-\alpha/2}; \widehat{S}(t) + \widehat{\sigma}(\widehat{S}(t))z_{1-\alpha/2}] \tag{1.13}$$

où $z_{1-\alpha/2}$ est le fractile de rang $100 \times (1 - \alpha/2)$ de la distribution normale standarisée.

1.8.2 Exemple

Pour illustrer les points précédents, on reprend les données de Freireich utilisée en 1.3.2 ci-dessus du groupe de 21 malades traité par le traitement 6-MP (existence d'ex-aequo), et on s'intéresse à l'estimations de l'erreur standard et les intervalles de confiance de la survie $S(t)$. Le seuil de risque par defaut est utilisé ($\alpha = 5\%$) et les résultats sont presentés dans la figure (1.3).

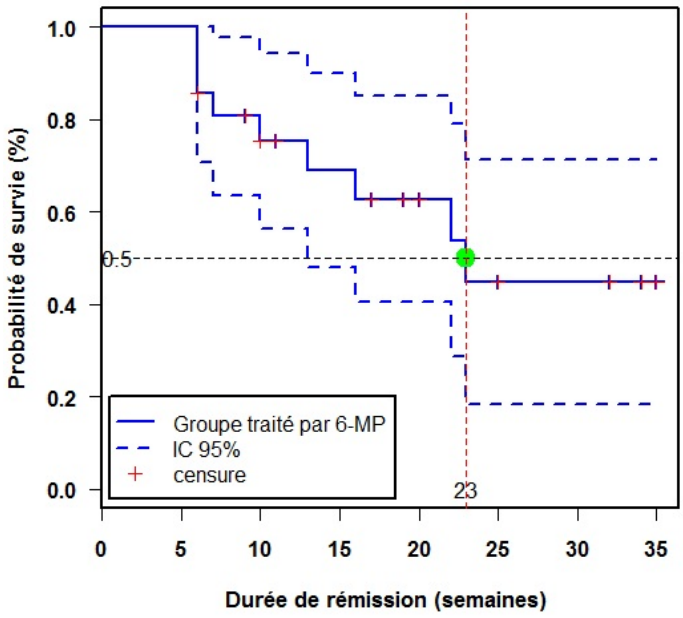


FIGURE 1.3 – Estimation d'intervalles de confiance pour la survie à 95%

La courbe principale est l'estimation de la courbe de survie, les courbes en pointillés correspondent aux limites inférieures et supérieures de l'intervalle de confiance à 95%.

Le tableau ci-dessous affiche les probabilités de survie par palier, et traduit les données de la courbe.

Temps t_i	R_i	M_i	$\widehat{S}(t_i)$	$\widehat{\sigma}(\widehat{S}(t))$	I.C 95%
6	21	3	0.857	0.0764	[0.707, 1.000]
7	17	1	0.807	0.0869	[0.636, 0.977]
10	15	1	0.753	0.0963	[0.564, 0.942]
13	12	1	0.690	0.1068	[0.481, 0.900]
16	11	1	0.627	0.1141	[0.404, 0.851]
22	7	1	0.538	0.1282	[0.286, 0.789]
23	6	1	0.448	0.1346	[0.184, 0.712]

Estimation de quantile de survie

D'après le graphique, la ligne verticale pointillé indique la valeur de la médiane de survie, est de 23 semaines. Par défaut le logiciel **R** affiche l'estimation de la médiane.

La médiane $\widehat{t}_{0.5} = 23$ semaines.

$$\widehat{u}_{0.5} = max\{t_i|\widehat{S}(t_i) \geq 1 - 0.5 + 0.05\} = max\{t_i|\widehat{S}(t_i) \geq 0.55\} = 16,$$

$$\widehat{l}_p = max\{t_i|\widehat{S}(t_i) \geq 1 - 0.5 - 0.05\} = max\{t_i|\widehat{S}(t_i) \geq 0.45\} = 23$$

et

$$\widehat{f}(23) = \frac{\widehat{S}(16) - \widehat{S}(23)}{23 - 16} = \frac{0.627 - 0.448}{7} = 0.0255$$

La variance de $\widehat{t}_{0.5}$ est :

$$Var(23) = \frac{Var(\widehat{S}(23))}{(\widehat{f}(23))^2} = (\frac{0.1346}{0.0255})^2 = 27.86$$

On a $\widehat{\sigma}(23) = 5.27$

L'intervalle de confiance approximative de la médiane à 95% est obtenu comme suit :

$$23 \pm 1.96 \times 5.27 \Longrightarrow [12.67, 33.33]$$

Chapitre 2

Bootstrap des données censurées

Introduction

Pour l'analyse statistique de nombreux problèmes d'utilisation fréquente, nous ne disposons que de la théorie asymptotique. Le statisticien appliqué est alors confronté au problème de savoir jusqu'à quel point il peut se fier aux résultats fondés sur la théorie asymptotique lorsqu'il met réellement en oeuvre une analyse sur des données réelles, le plus souvent en nombre restreint. Il est alors demandeur de méthodes qui soient plus performantes que celles issues de la simple application des résultats asymptotiques du premier ordre, conséquence immédiate du théorème central limite. Il souhaite évidemment que ces méthodes s'appliquent à des modèles statistiques aussi généraux que possible. C'est alors qu'interviennent les techniques de rééchantillonnage, dont l'efficacité est intimement liée aux développements de l'outil informatique, qui fournissent des méthodes pour résoudre de façon pratique ce problème.

À l'origine les méthodes de rééchantillonnage ont été développées pour estimer le biais et la variabilité d'un estimateur sans faire d'hypothèses sur la distribution de probabilité de la population dont on estime un paramètre.

Le but des méthodes de rééchantillonnage est de tirer des réalisations d'une distribution inconnue. Ce problème paradoxal se résout en tirant dans une distribution proche de la distribution inconnue, cette distribution proche est en fait la distribution empirique. Le principe des méthodes de rééchantillonage est de tirer au hasard des observations dans l'échantillon dont on dispose. Dans le bootstrap, on effectue des tirages avec remise, alors que le jackknife procéde par tirages sans remise.

2.1 Le bootstrap

La méthode du bootstrap a été proposée par Bradley Efron comme un outil de calcul utilisé pour résoudre des problèmes d'inférence statistique. Ce dernier a écrit beaucoup au sujet de la méthode et de ses généralisations (voir Efron,1982, 1985, etc.). Des milliers de papiers ont été rédigés sur le bootstrap dans les dernières décennies.
Son principe, à la fois simple et révolutionnaire, salué parfois comme "la plus importante idée nouvelle en statistique des trois ou quatre dernières décennies", reprend en fait des idées plus anciennes (jackknife, validation croisée,...). Le bootstrap s'appuie sur le fait de pouvoir en rééchantillonnant dans les données proprement dites, estimer les caractéristiques du phénomène aléatoire qui a engendré ces données.

On utilise le bootstrap quand nous avons peu d'informations à propos des statistiques des données ou que nous avons une faible quantité de données qui nous empêchent d'utiliser des résultats asymptotiques.

Soit X une variable de loi inconnue F, on dipose d'un échantillon $(x_1, x_2, ..., x_n)$ et on veut étudier la distribution d'un estimateur $\widehat{\theta}$ d'un certain paramètre θ (moyenne, médiane,...), calculer sa variance et donner un intervalle de confiance (I.C). Efron (1979) a introduit l'idée de bootstrap ([11], [18], [8]), on se basant sur le principe élémentaire suivant :

Si n est grand, la distribution empirique F_n^* est proche de F, on aura donc une bonne approximation de la loi de T en utilisant F_n^* à la place de F. On est donc amener à tirer des échantillons de n valeurs dans la loi de F_n^*, ce qui revient à rééchantillonner dans l'échantillon $(x_1, x_2, ..., x_n)$. On effectue des tirages avec remise de n valeurs parmi les n valeurs observées : les valeurs observées $x_1, x_2, ..., x_n$ sont donc répétées selon les réalisations d'un vecteur multinomial $K_1, K_2, ..., K_n$ d'effectif n et de probabilité $p_i = \frac{1}{n}$. Lorsque n n'est pas très élevé, on peut énumérer tous les échantillons possibles équiprobables (il y en a n^n), sinon on se contente d'en tirer un nombre B suffisament grand. Si le nombre de réplications B tend vers l'infini, la moyenne de toutes les estimations bootstrap converge vers l'estimateur du maximum de vraisemblance empirique (i.e. utilisant la loi de F_n^*) et permet d'estimer sa variance. En pratique, on se contentera de quelques centaines de tirages au plus.

Estimation bootstrap

Soit X un caractère dans une population Ω, à valeurs dans $(E, \mathcal{T}_E)$. On veut estimer une caractéristique de la loi de X, $\theta = \theta(P_X)$.
Supposons avoir observé un n-échantillon i.i.d. de la loi de X dans Ω $X_1 = x_1, ..., X_n = x_n$. On peut estimer la loi inconnue P_X par la loi empirique de l'échantillon :

$$\hat{P}_{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i}$$

$\hat{P}_{\underline{x}}$ est la probabilité discrète dans E qui charge les points $x_1, ..., x_n$ avec équiprobabilité. Le bootstrap intervient alors au moment où ceci n'est pas faisable à la main.

Procédure bootstrap

Considérons $\hat{\theta} = \varphi(\underline{X}) = \varphi(X_1, ..., X_n)$ est un estimateur calculable du paramètre θ.

- On choisit un entier B "assez grand".
- On simule B n-échantillons i.i.d. suivant la loi $\hat{P}_{\underline{X}}$.
- Pour chaque n-échantillons $x_{j_1}, ..., x_{j_n},\ j = 1, ..., B$, on calcule

$$\hat{\theta}_j = \varphi(x_{j_1}, ..., x_{j_n}) = \varphi(\underline{x}_j)$$

  qui est une réalisation de $\hat{\theta}$.
- On estime θ par $\overline{\theta}_B = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j$ et la variance de $\hat{\theta}$ par $var(\hat{\theta}) = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}_j - \overline{\theta}_B)^2$.

D'après le théorème central limite :

$$\sqrt{B}(\hat{\theta}_n - \overline{\theta}_B) \sim \mathcal{N}(0, var(\hat{\theta})) \quad [1]$$

2.2 Type de bootstrap

Il y a deux formes de bootstrap : le bootstrap paramétrique et le bootstrap non paramétrique.

1. ***bootstrap paramétrique***

   On remplace d'abord F par un modèle paramétrique $\widehat{F}$ qui semble bien ajuster les données. On simule ensuite B échantillons $x_1^*, ..., x_B^*$ de taille n, indépendamment les

uns des autres qui proviennent de la distribution $\widehat{F}$. Enfin, on calcule les estimateurs $\widehat{\theta}(x_1^*), ..., \widehat{\theta}(x_B^*)$ de θ obtenues sur les échantillons bootstrap simulés de $X = (X_1, ..., X_1)$.

2. ***bootstrap non-paramétrique***

   Si nous ne pouvons pas attribuer un modèle paramétrique aux données, nous utilisons $\widehat{F} = F_n$ comme approximation de F, où F_n est la distribution empirique ; On génère alors B échantillons de taille n provenant de F_n. De nouveau, on calcule les B valeurs $\widehat{\theta}(x_1^*), ..., \widehat{\theta}(x_B^*)$ de θ pour ces B échantillons simulés. Dans les deux cas, la distribution empirique des valeurs simulées $\widehat{\theta}(x_1^*), ..., \widehat{\theta}(x_B^*)$ une approximation de la distribution d'échantilonnage de θ. On appelle cette approximation la *distribution bootstrap* de θ.

Remarque 2.1. Choix du B, le nombre de Bootstraps :

Selon Efron, il est rarement nécessaire d'utiliser plus de $B = 200$ échantillons bootstrap pour estimer une variance ; dans bien des cas, $B = 50$ ou 100 sont suffisants. L'importance des valeurs extrêmes de la statistique $\widehat{\theta}$ étudiée est un facteur important dans la détermination du choix de B : plus ces valeurs sont fréquentes, plus B devrait être grand. On notera cependant que certaines autres applications du bootstrap exigent un B beaucoup plus grand ; ce sera en particulier le cas pour l'application à la construction d'intervalles de confiance.

2.2.1 Utilisations fondamentales de Bootstrap

Les deux applications fondamentales du bootstrap sont la réduction du biais et la détermination d'intervalles de confiance [12].

1. Réduction du biais :

On peut utiliser le bootstrap pour réduire le biais. Supposons que l'on veuille estimer $\theta(F) = [\int x dF(x)]^r$ à partir d'un échantillon auquel est associé F_0. On choisit l'estimateur

$$\widehat{\theta}(F) = \theta(F_0) = [\int x dF_0(x)]^r$$

$$Biais = E\{\theta(F) - \theta(F_0)|F\}$$

Comme on ne connaît pas F, on utilise le principe du bootstrap en remplaçant dans cette équation F par F_0 et F_0 par F_1 où F_1 est la loi associée à un n-échantillon d'une variable de loi F_0 :

$$\widehat{Biais} = E\{\theta(F_0) - \theta(F_1)|F_0\}$$

Donc l'estimateur sans biais de θ s'obtient en retranchant à $\theta(F_0)$ cet estimateur de son biais, soit :
Estimateur sans biais de $\theta = \theta(F_0) - \widehat{Biais}$

Pour obtenir un estimateur sans biais, on doit donc ajouter à $\theta(F_0)$ où t est défini par

$$E(\theta(F_0) - \theta(F) + t) = 0$$

On a donc remplacé l'équation initiale qui donne la correction t que l'on devrait faire pour supprimer le biais de l'estimateur $\widehat{\theta}(F_0)$ par une équation bootstrap qui donne une correction t^*, en principe calculable, et dont on espère qu'elle est une bonne estimation de t. On remarque que t est un paramètre qui dépend de F alors que t^* est une statistique dépendant de F_0. De cette équation se déduit la correction $t^* = \theta(F_0) - E(\theta(F_1)|F_0)$. On doit donc calculer la quantité : $E(\theta(F_1)|F_0)$, et l'estimateur sans biais est alors égal à :

$$\theta(F_0) + t^* = 2\theta(F_0) - E(\theta(F_1)|F_0).$$

2. Intervalle de confiance :

Soit F la loi inconnue, dont on veut estimer le paramètre $\theta(F)$ par un intervalle de confiance à $0,95$ et F_0 la loi associée à l'échantillon observé. $\theta(F_0)$ est l'estimateur de $\theta(F)$. Soit F la loi inconnue, dont on veut estimer le paramètre $\theta(F)$ par un intervalle de confiance à 0.95 et F_0 la loi associée à l'échantillon observé. $\theta(F_0)$ est un estimateur de θ. Pour obtenir, à partir de $\theta(F_0)$, un intervalle de confiance (en général asymétrique) pour $\theta(F)$, on a besoin de connaître la loi de $\theta(F) - \theta(F_0)$, sous F (alors que F est inconnue) ou une approximation pour cette loi. Si c'est le cas, on prend pour bornes de l'intervalle, en notant $t_1 = t_0.025$ et $t_2 = t_0.975$ les quantiles 0.025 et 0.975 de cette loi : $[\theta(F_0) + t_1; \theta(F_0) + t_2]$. En effet :

$$P(\theta(F) - \theta(F_0) < t_1) = 0.025$$

$$P(\theta(F) - \theta(F_0) > t_2) = 0.025$$

$$P(\theta(F_0) + t_1 \leq \theta(F) \leq \theta(F_0) + t_2) = 0.975.$$

Si on ne connaît pas cette loi, et si on n'a pas d'approximation pour celle-ci, ou tout simplement si on en dispose mais que les calculs sont très compliqués, le bootstrap permet de lui substituer la loi de $\theta(F_0) - \theta(F_1)$ sous F_0.

2.3 Application de bootstrap sur des données censurées

Le bootstrap est bien validée par de nombreuses études statistiques et une des premières applications du bootstrap a été faite dans le contexte d'analyse de la survie (Efron, 1981) pour répondre à certaines questions notamment pour construire les bandes de confiance ([7],[23], [24]).

2.3.1 Bootstrap de l'estimateur de Kaplan Meier

Soient $X_1, ..., X_n$ n variables représentant les durées de vie de n sujets, sont des variables aléatoires positives, indépendantes et de fonction de répartition F, et indépendamment d'elles, et $C_1, ..., C_n$ les instants de censures associés, positives, de f.d.r. G.

On note $\{(T_1, \delta_1), ..., (T_n, \delta_n)\}$ l'échantillon réellement observé et $(t_{(i)}, \delta_{(i)})$ l'échantillon ordonné suivant les valeurs de t_i.
Efron (1981)([5],[6], [11]) suggère le plan du rééchantillonnage suivant :

On tire un échantillon bootstrapé $(Z_j^* = (T_j^*, \delta_j^*))_{j=1,...,n}$, en tirant chaque couple aléatoirement et avec remise dans l'échantillon observé $(Z_j = (t_{(j)}, \delta_{(j)}))_{j=1,...,n}$.

On définit, pour $j = 1, ..., n$, les variables aléatoires :
$m_j^* = \sum_{k=1}^n \mathbb{1}_{(T_k^*, \delta_k^*) = (t_{(j)}, \delta_{(j)})}$ égale au nombre de couples bootstrapés égaux à $(\mathrm{t}_{(j)}, \delta_{(j)}))$.
$\mathrm{M}_j^* = m_j^* + ... + m_n^*$ égale au nombre de couples dont les durées sont supèrieurs où égales à $t_{(j)}$.

L'estimateur de Kaplan-Meier construit avec les données $(T_j^*, \delta_j^*)_{j=1,...,n}$ s'écrit :

$$\widehat{S}^*(t_{(k)}) = 1 - \widehat{F}_n^*(t_{(k)}) = \prod_{i=1}^{k} \left(1 - \frac{m_i^*}{M_i^*}\right)^{\delta_{(i)}} \tag{2.1}$$

2.3.2 Estimation de la variance de l'estimateur de Kaplan-Meier par le bootstrap d'Efron

De (2.1) on a :

$$\log(\widehat{S}^*(t_{(k)})) = \log\left(\prod_{i=1}^{k} \left(1 - \frac{m_i^*}{M_i^*}\right)^{\delta_{(i)}}\right)$$

$$\log(\widehat{S}^*(t_{(k)})) = \sum_{i=1}^{k} \delta_{(i)} \log\left(1 - \frac{m_i^*}{M_i^*}\right)$$

$$Var^*(\log(\widehat{S}^*(t_{(k)}))) = \sum_{i=1}^{k} \delta_{(i)} Var^*\left(\log\left(1 - \frac{m_i^*}{M_i^*}\right)\right)$$

En appliquant la formule (1.10) avec $g = Fonction\ \ logarithme$ et $X = \widehat{S}^*(t_{(k)})$.
On a $M_i^* \sim \mathcal{B}(n, n_i/n)$, avec $n_j = n - j + 1$.
La loi conditionnelle de m_j^* sachant $M_i^* = N$ est $\mathcal{B}(N, p)$, avec $p = 1/n - j + 1$ (c'est-à-dire : P(avoir T_i^*=$t_{(j)}$)).

$$Var^*(m_j^*/M_j^* = N) = Np(1-p) = N(1/n - j + 1)(1 - 1/n - j + 1)$$

$$Var^*(m_j^*/M_j^*) = M_j^* p(1-p) = M_j^*(1/n - j + 1)(1 - 1/n - j + 1)$$

$$Var^*(\frac{m_j^*}{M_j^*}/M_j^*) = (1/M_j^*)^2 Var^*(m_j^*|M_j^*) = (1/M_j^*)(1/n - j + 1)(1 - 1/n - j + 1)$$

Donc

$$Var^*(\log(\widehat{S}^*(t_{(k)}))) = \sum_{i=1}^{k} \frac{\delta_{(i)}}{n_i^2}$$

quand $n \to \infty$ on a :

$$Var^*(\widehat{S}^*(t_{(k)})) \sim S^2(t_{(k)}) \sum_{i=1}^{k} \frac{\delta_{(i)}}{(n-i+1)^2} \quad [11] \tag{2.2}$$

Remarque 2.2. *On sait que la loi d'un estimateur bootstrapé est proche de celle de l'estimateur initial, quand n est grand. On a donc en particulier :*

$$Var^*(\widehat{S}^*(t_{(k)})) \sim Var(\widehat{S}(t)).$$

2.3.3 Intervalle de confiance

"Le percentile bootstrap" est la méthode proposée par Efron [5], pour le calcul de l'intervalle de confiance $100(1-\alpha)\%$ pour l'analyse de survie. Elle débute par la construction de B répliques obtenus par retirage des (T_i^*, δ_i^*) de l'échantillon d'origine observé $\{(t_{(j)}, \delta_{(j)}), j = 1, ..., n\}$, ensuite on ordonne les B survies Kaplan-Meier $\widehat{S}^*(t_{(k)})$.

On en déduit les limites de l'intervalle de confiance à $(1-\alpha)100\%$ pour la survie $\widehat{S}^*(t_{(k)})$ de la forme :

$$[\widehat{S}^*(t_{(k)})(\alpha/2); \widehat{S}^*(t_{(k)})(1-\alpha/2)] \tag{2.3}$$

$(\alpha/2)$ et $(1-\alpha/2)$ sont les quantiles empiriques des B courbes de survie.

2.3.4 Exemple

Dans l'exemple des données de Freireich, nous présentons les graphes de l'estimateur de Kaplan-Meier (1.8), (km) est en bleu avec celle obtenue par rééchantillonnage bootstrap (2.1), (bootkm) en noir de la fonction de survie S avec des intervalles de confiance à 95%.

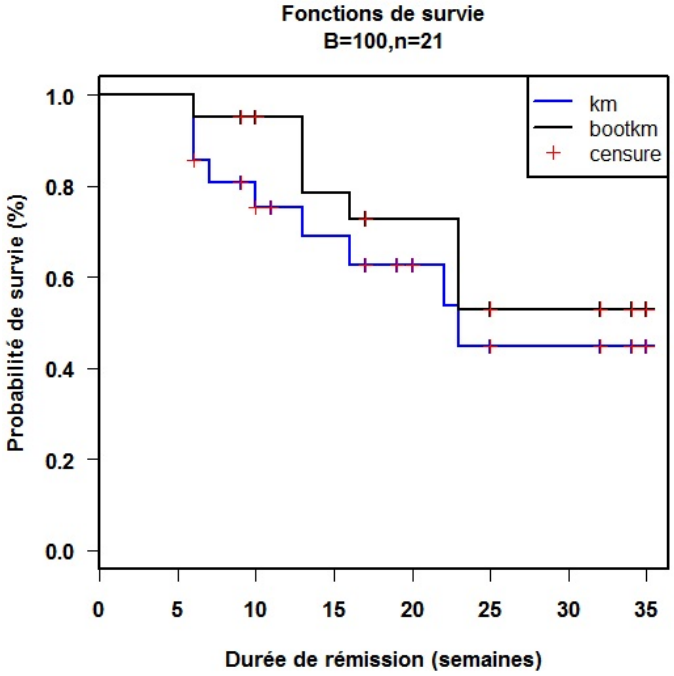


FIGURE 2.1 – Estimation de survie selon la méthode de Bootstrap

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=21, B=100 et comparaison avec la variance asymptotique.

Temps	6	7	10	13	16	22	23
$\widehat{\sigma}(\widehat{S}(t))$	0.0764	0.0869	0.0963	0.1068	0.1141	0.1282	0.1346
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0641	0.0778	0.0857	0.0960	0.1042	0.1213	0.1237

Illustrations et comparaisons pour la fonction de survie

Malgré que la taille de l'échantillon soit petite (n= 21), les deux courbes ; la courbe de survie de Kaplan Meier et la courbe "bootstrap" ; sont très proches. Les estimations de la variance sont presque égales.

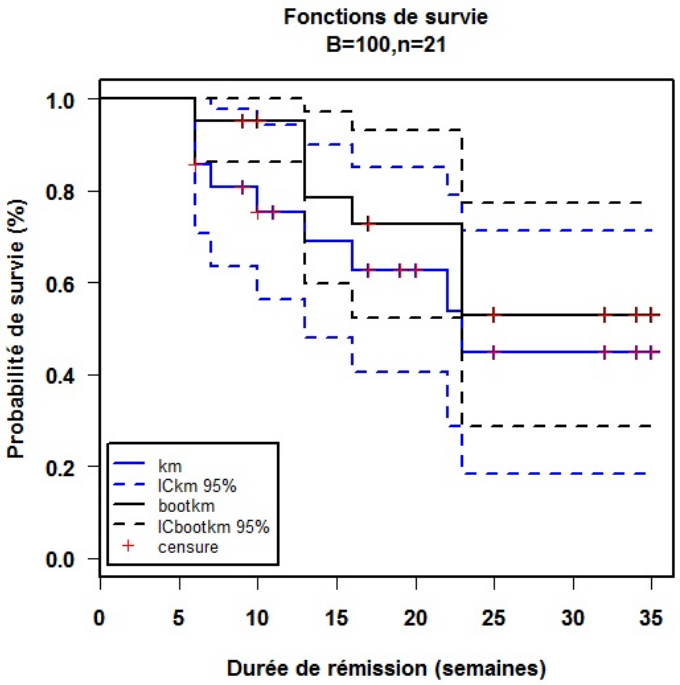


FIGURE 2.2 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
6	0.857	[0.707, 1.000]	[0.707, 1.000]
7	0.807	[0.636, 0.977]	[0.782, 1.000]
10	0.753	[0.564, 0.942]	[0.642, 0.977]
13	0.690	[0.481, 0.900]	[0.596, 0.972]
16	0.627	[0.404, 0.851]	[0.524, 0.933]
22	0.538	[0.286, 0.789]	[0.308, 0.802]
23	0.448	[0.184, 0.712]	[0.287, 0.772]

On peut dire que les intervalles de confiance obtenue par la méthode du bootstrap pour le groupe 6-MP sont meilleurs que ceux à l'aide de l'approximation de Greewood.

Conclusion

Dans notre série de 21 patients, la faible puissance liée à la petite taille de l'échantillon rend l'extrapolation des résultats à la population générale très délicate. L'utilisation du Bootstrap a permis d'obtenir des IC plus étroits que sur l'échantillon d'origine et finalement assez proche des données de la littérature, donc vraisemblablement plus précis. Il s'agit donc d'un outil statistique intéressant dans ce cas précis dont on ne doit pas ignorer l'existence.

2.4 Estimateurs de moment des quantité de survie

Nous développons les expressions des quantités importants, telles, $E(T^k)$, $E(\widehat{S}_0(t)^k)$ et $E([\widehat{F}^{-1}(u)]^k)$, correspondant au moments de la durée de vie, moments de la survie et les moments des quantiles de survie respectivement. Pour un échantillon (iid) continue de taille n à partir d'une distribution ayant un support positif, on définit les théorème suivant :

Théorème 2.1. [13] *L'estimateur bootstrap du moment d'ordre k ($k \geq 1$) de la variable aléatoire T est donnée par*

$$E_{\widehat{S}(t)}(T^k) = \begin{cases} \sum_{i=1}^{m} X_{(i)}^k (\widehat{S}(X_{(i-1)}) - \widehat{S}(X_{(i)})) & si\ \delta_{(n)} = 1 \\ \sum_{i=1}^{m} X_{(i)}^k (\widehat{S}(X_{(i-1)}) - \widehat{S}(X_{(i)})) + T_{(n)}^k \widehat{S}(X_{(i)}) & si\ \delta_{(n)} = 0 \end{cases} \tag{2.4}$$

où $\widehat{S}(t)$ est l'estimateur de kaplan meier défini à l'équation (1.7) et $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(m)}$ sont les m valeurs distinctes ordonnées non censurées parmi $(X_1, X_2, ..., X_n)$ avec $m \leq n$, (m : le nombre d'observations non censurées), par définition $X_{(0)} = 0$ et $\widehat{S}(X_{(0)}) = 1$.

Preuve

Pour le cas $\delta_{(n)} = 1$, l'espérance de T^k est donné par

$$\begin{aligned} E(T^k) &= \int_0^\infty T^k dF & (2.5) \\ &= \int_0^1 [F^{-1}(u)]^k du & (2.6) \\ &= \sum_{i=1}^{m} \int_{1-\widehat{S}(X_{(i-1)})}^{1-\widehat{S}(X_{(i)})} [F^{-1}(u)]^k du. & (2.7) \end{aligned}$$

Ensuite, en substituant $\widehat{F} = 1 - \widehat{S}$ de l'équation (2.4) pour $F^{-1}(u)$ dans l'équation (2.7), où $F^{-1}(u)$ est constante dans l'intervalle fermé, $[1 - \widehat{S}(X_{(i-1)}), 1 - \widehat{S}(X_{(i)})]$ est égal à $X_{(i)}$.

Remarque 2.3. *La moyenne et la variance bootstrap de T sont donnés par*

$$\mu_{\widehat{S}}(T) = E_{\widehat{S}(T)}(T)$$

et

$$\sigma^2_{\widehat{S}(T)} = E_{\widehat{S}(T)}(T^2) - [E_{\widehat{S}(T)}(T)]^2.$$

On peut construire un intervalle de confiance "naif" de la durée de vie, est donné par la relation suivant :

$$\mu_{\widehat{S}}(T) \pm t_{n-1,\alpha/2}\sigma^2_{\widehat{S}(T)}, \tag{2.8}$$

où $t_{n-1,\alpha/2}$ désigne le quantile d'ordre $\alpha/2$ d'une distribution de Student.

Théorème 2.2. [13] *L'estimateur bootstrap du moment d'ordre k ($k \geq 1$) pour $F^{-1}(u)$ dans le cas $\delta_n = 1$ est donnée par :*

$$E_{\widehat{S}(T)}[(F^{-1}(u))^k] = \sum_{j=1}^{m} X(j)^k w(j)_{p,q}, \tag{2.9}$$

où $w(j)_{p,q} = \beta_{p,q}(1 - \widehat{S}(X_{(j)})) - \beta_{p,q}(1 - \widehat{S}(X_{(j-1)}))$,

$\beta_{p,q}(.)$ désigne la distribution cumulative de la loi beta avec les paramètres $p = [nu] + 1$ et $q = n - [nu] + 2$.

dans le cas $\delta_n = 0$, l'estimateur devient

$$E_{\widehat{S}(T)}[(F^{-1}(u))^k] = \sum_{j=1}^{m} X(j)^k w(j)_{p,q} + T^k_{(n)}\beta_{p,q}(1 - \widehat{S}(X_{(j)}))$$

On utilise ce théorème pour construire I.C pour la median, $u = 1/2$.

Preuve

La preuve suit la même façon que le théorème précedent avec une simple substitution de $\widehat{S}$ par S dans $E[F^{-1}(u)]^k$ où $F^{-1}(u) = Y_{([nu]+1)}$ désigne la $([nu] + 1)$, statistique d'ordre, à partir d'un échantillon de taille n.

Remarque 2.4. *Un intervalle de confiance pour le quantile de survie, est donnée par*

$$\mu_{\widehat{F}^{-1}(u)} \pm t_{n-1,\alpha/2}\sigma_{\widehat{F}^{-1}(u)}, \tag{2.10}$$

où $\mu_{\widehat{F}^{-1}(u)} = E_{\widehat{S}(T)}(\widehat{F}^{-1}(u))$ et $\sigma_{\widehat{F}^{-1}(u)} = E_{\widehat{S}(T)}[(\widehat{F}^{-1}(u))^2] - [E_{\widehat{S}(T)}(\widehat{F}^{-1}(u))^2)]^2$,

2.4.1 Exemple

On reprend les données de Freireich utilisée en 1.3.2 du groupe de 21 malades traité par le traitement 6-MP (existence d'ex-aequo), et on s'intéresse à l'estimations d'intervalles de confiance de la médiane de survie.

Estimation de quantile de survie par bootstrap

Dans cette partie on utilise le theorème 2.2. pour construire I.C à 95% pour la mediane, $u = 1/2$. La valeur de la médiane de survie, est de 23 semaines,

$\mu_{\widehat{F}^{-1}(0.5)} = 23.29$ et $\sigma_{\widehat{F}^{-1}(0.5)} = 2.83$

Un intervalle de confiance pour la médiane de survie, est donnée par :

$23.29 \pm t_{20,\alpha/2} \times 2.83 = 23.29 \pm 2.086 \times 2.83 = [17.38, 29.19]$

(de la table de student on a $t_{20,\alpha/2} = 2.086$, $\alpha = 0.05$.)

Les resultats sont presentés dans la figure(2.3).

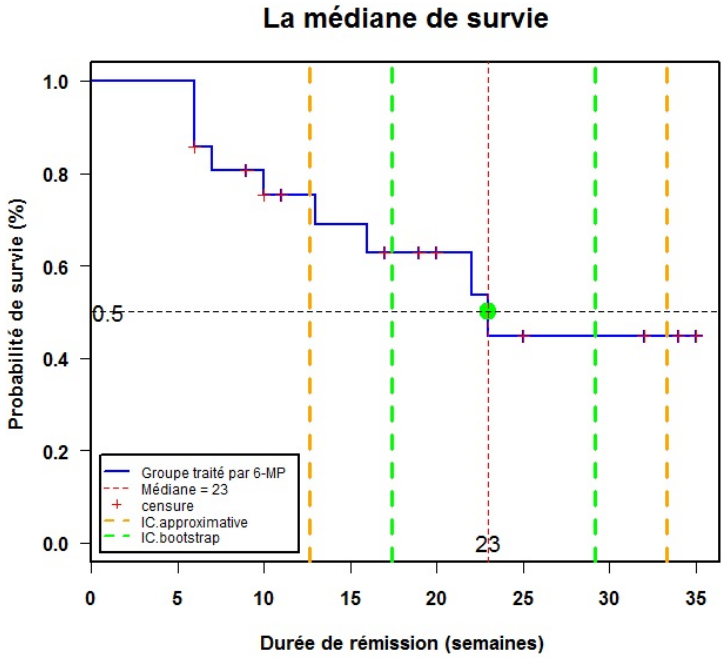


FIGURE 2.3 – La médiane de survie

Intèprétation :

La courbe principale est l'estimation de la courbe de survie, les ligne en pointillés correspondent aux limites inférieures et supérieures de l'intervalle de confiance approximative de la médiane à 95% est égale : $[12.67, 33.33]$(en orange) avec celle obtenue par bootstrap est égale : [17.38, 29.19](en vert).
L'IC (95%) de la médiane obtenu par la méthode de Greenwood (IC approximative) dans l'exemple 1.8.2 était beaucoup plus large et imprécis et celle obtenue par bootstrap (IC bootstrap) plus étroit et plus précis.

Chapitre 3

Simulations

3.1 Introduction

Afin de nous assurer de la performance des estimateurs étudiés, nous générons artificiellement des données à l'aide du logiciel R. Nous présentons les graphes des estimateurs de Kaplan-Meier et de Kaplan Meier bootstrappé pour des données simulées. Nous donnons aussi le taux de survie de l'estimateur de la fonction de survie. Cette section propose une comparaison des résultats obtenus par bootstrap, variance de $S(t)$ et intervalles de confiance pour $S(t)$, avec les résultats de nature asymptotique. Nous effectuerons des comparaisons de résultats pour différentes valeurs de B (le nombre de répliques du bootstrap) et n (la taille de l'échantillon).

3.2 Données simulées

Nous avons simulé différents échantillons de taille ($n = 50$, $n = 100$, $n = 500$, et $n = 1000$), extraits pour trois modèles, d'abord un modèle exponentielle puis un modèle Weibull et un modèle normal qui sont censurées à droite : (X_i, C_i) avec $T_i = X_i \wedge C_i$.

3.2.1 Estimation de la fonction de survie S

1. Modèle 1

$X_i \hookrightarrow \mathcal{E}(0.2)$ et $C_i \hookrightarrow \mathcal{E}(0.5)$, $n = 50$, 100, 500, 1000.

La figure suivante donne l'estimateur de Kaplan-Meier (km) est en bleu avec celle obtenue par bootstrap (bootkm) en noir de la fonction de survie S avec des intervalles de confiance à 95%.

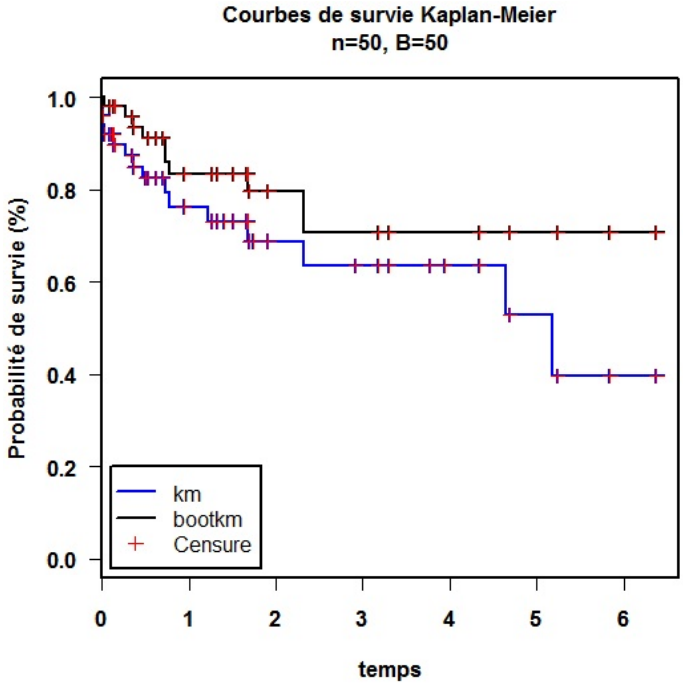


FIGURE 3.1 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 80%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=50 et comparaison avec la variance asymptotique.

Temps	0.0244	0.2693	0.3642	0.4700	0.7378	0.7733	1.6846	2.3212
$\widehat{\sigma}(\widehat{S}(t))$	0.0388	0.0486	0.0530	0.0570	0.0625	0.0672	0.0792	0.0890
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0198	0.0290	0.0365	0.0427	0.0538	0.0582	0.0667	0.0836

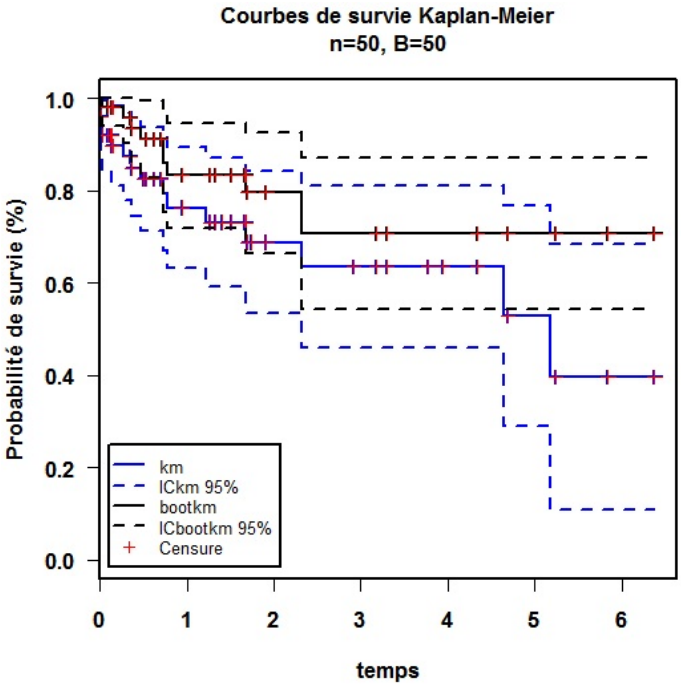


FIGURE 3.2 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0244	0.980	[0.843, 0.995]	[0.941, 1.000]
0.2693	0.958	[0.778, 0.968]	[0.901, 1.000]
0.3642	0.935	[0.745, 0.953]	[0.863, 1.000]
0.4700	0.911	[0.712, 0.936]	[0.827, 0.995]
0.7378	0.859	[0.671, 0.916]	[0.753, 0.964]
0.7733	0.833	[0.631, 0.894]	[0.719, 0.947]
1.6846	0.795	[0.533, 0.843]	[0.664, 0.926]
2.3212	0.707	[0.416, 0.810]	[0.543, 0.870]

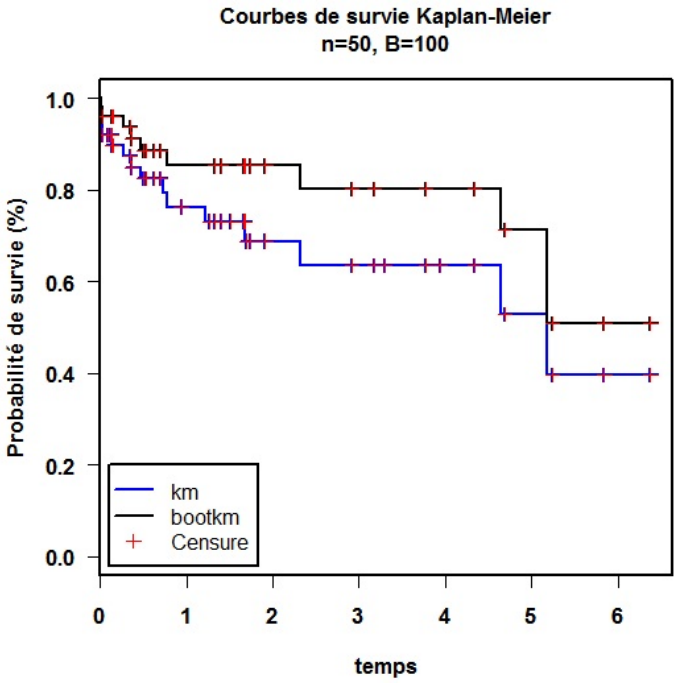


FIGURE 3.3 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 80%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=100 et comparaison avec la variance asymptotique.

Temps	0.0244	0.2693	0.3642	0.4700	0.7733	2.3212
$\widehat{\sigma}(\widehat{S}(t))$	0.0388	0.0486	0.0530	0.0570	0.0672	0.0890
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0277	0.0356	0.0423	0.0488	0.0570	0.0724

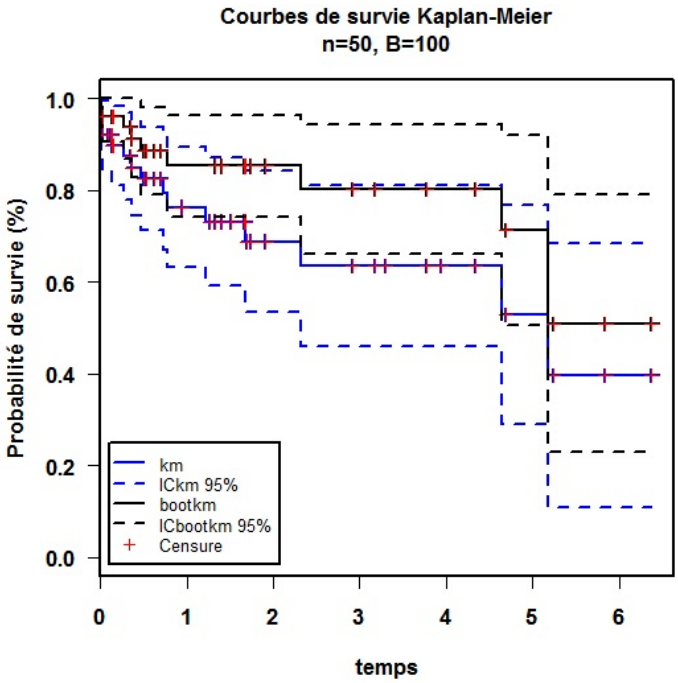


FIGURE 3.4 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0244	0.980	[0.843, 0.995]	[0.906, 1.000]
0.2693	0.958	[0.778, 0.968]	[0.867, 1.000]
0.3642	0.935	[0.745, 0.953]	[0.829, 0.995]
0.4700	0.911	[0.712, 0.936]	[0.789, 0.981]
0.7733	0.833	[0.631, 0.894]	[0.741, 0.964]
2.3212	0.707	[0.416, 0.810]	[0.660, 0.944]

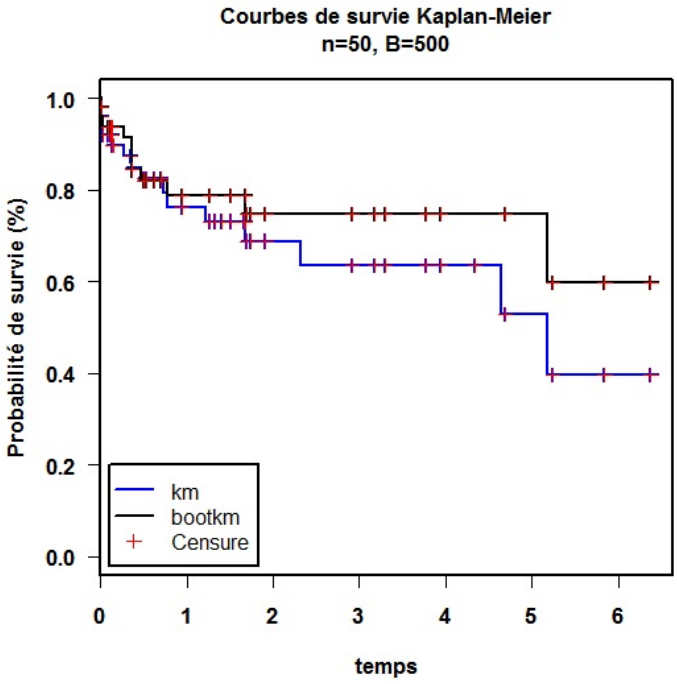


FIGURE 3.5 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 80%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=500 et comparaison avec la variance asymptotique.

Temps	0.0244	0.2693	0.3642	0.4700	0.7733	2.3212
$\widehat{\sigma}(\widehat{S}(t))$	0.0388	0.0486	0.0530	0.0570	0.0672	0.0890
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0350	0.0413	0.0530	0.0570	0.0570	0.0724

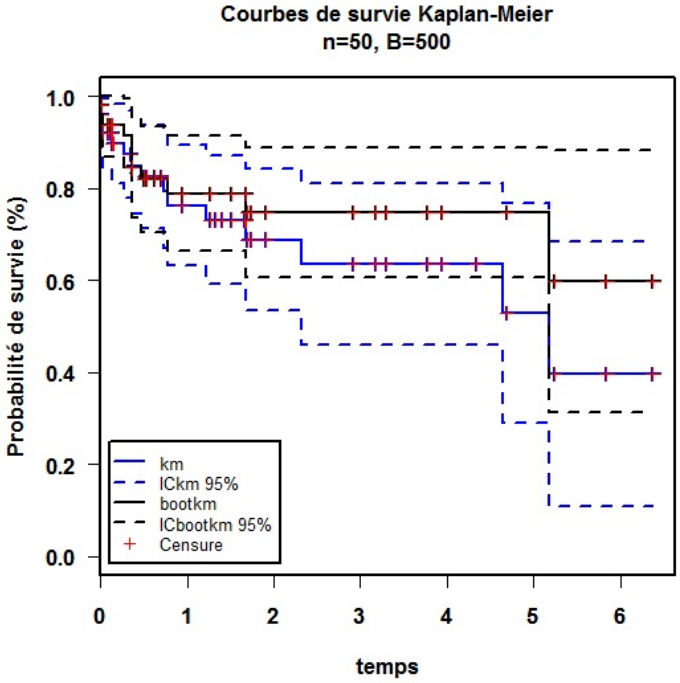


FIGURE 3.6 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0244	0.980	[0.843, 0.995]	[0.869, 1.000]
0.2693	0.958	[0.778, 0.968]	[0.833, 0.995]
0.3642	0.935	[0.745, 0.953]	[0.737, 0.950]
0.4700	0.911	[0.712, 0.936]	[0.705, 0.933]
0.7733	0.833	[0.631, 0.894]	[0.664, 0.913]
2.3212	0.707	[0.416, 0.810]	[0.660, 0.944]

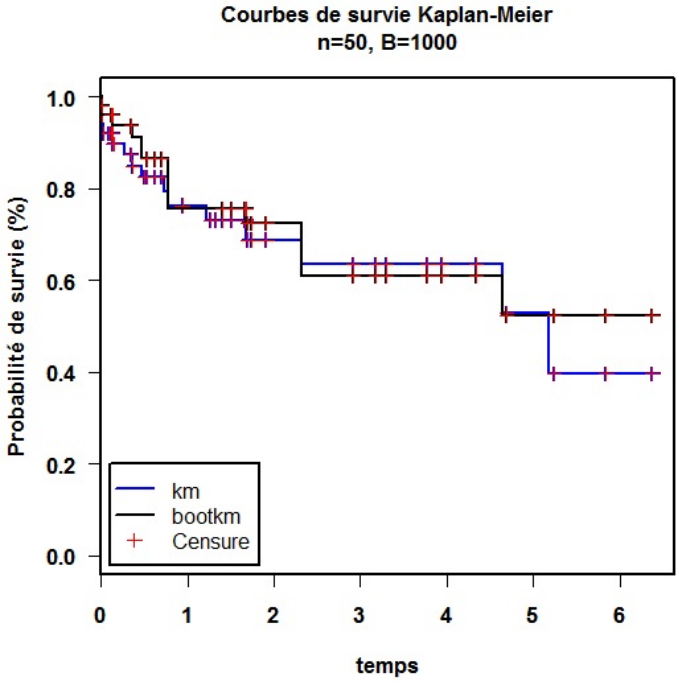


FIGURE 3.7 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 84%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=1000 et comparaison avec la variance asymptotique.

Temps	0.0244	0.2693	0.3642	0.4700	0.7733	2.3212
$\widehat{\sigma}(\widehat{S}(t))$	0.0388	0.0486	0.0530	0.0570	0.0672	0.0890
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0350	0.0413	0.0423	0.0520	0.0672	0.0857

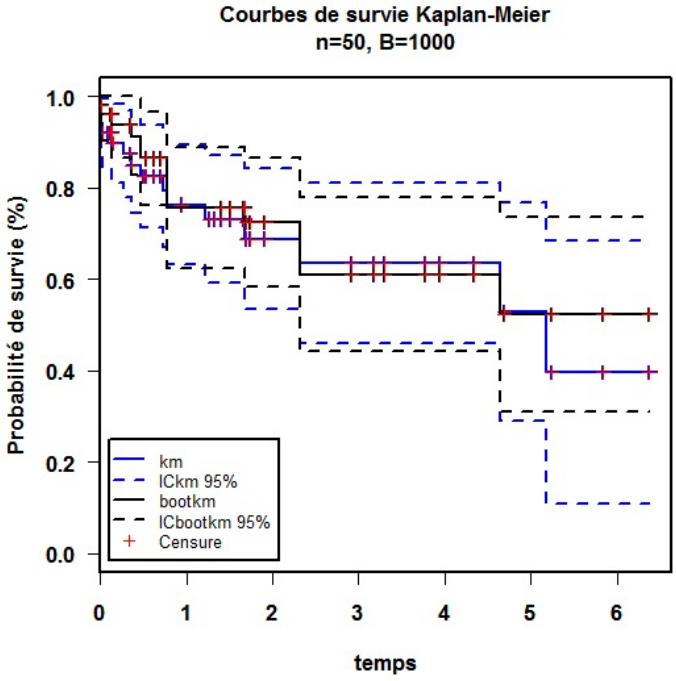


FIGURE 3.8 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0244	0.980	[0.843, 0.995]	[0.869, 1.000]
0.2693	0.958	[0.778, 0.968]	[0.833, 0.995]
0.3642	0.935	[0.745, 0.953]	[0.829, 0.995]
0.4700	0.911	[0.712, 0.936]	[0.762, 0.966]
0.7733	0.833	[0.631, 0.894]	[0.623, 0.889]
2.3212	0.707	[0.416, 0.810]	[0.442, 0.778]

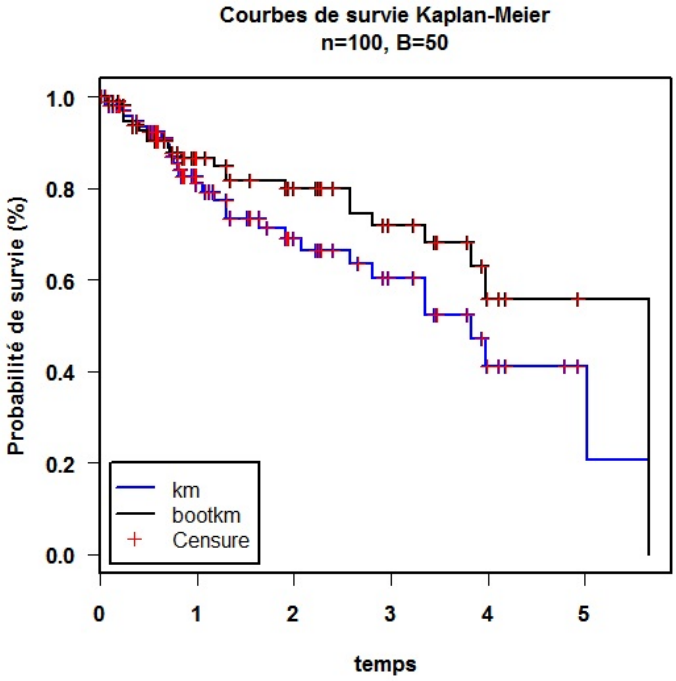


FIGURE 3.9 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 79%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=50 et comparaison avec la variance asymptotique.

Temps	0.0821	0.2322	0.7042	1.3001	1.3063	1.9117	2.5728	2.8090	3.3481
$\widehat{\sigma}(\widehat{S}(t))$	0.0144	0.0180	0.0333	0.0522	0.0543	0.0593	0.0665	0.0703	0.0762
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0100	0.0144	0.0331	0.0418	0.0441	0.464	0.0566	0.0605	0.0681

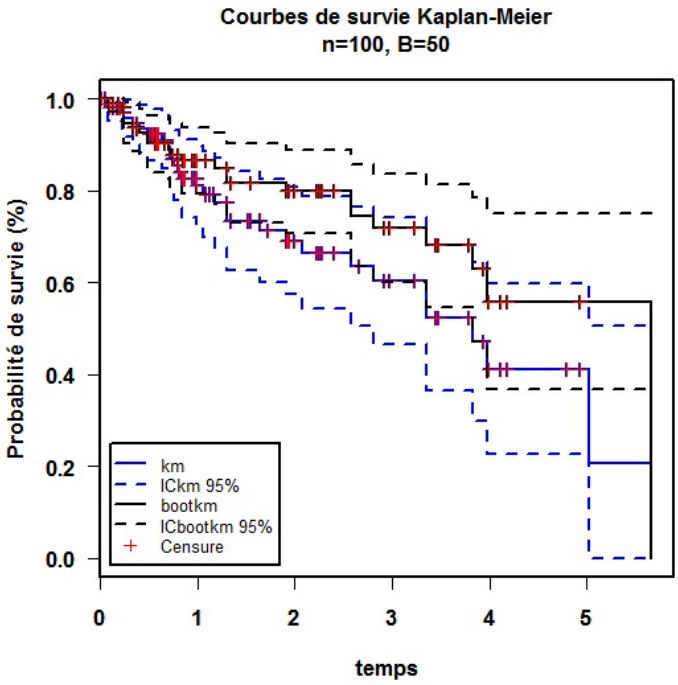


FIGURE 3.10 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0821	0.979	[0.951, 1.000]	[0.970, 1.000]
0.2322	0.968	[0.933, 1.000]	[0.951, 1.000]
0.7042	0.895	[0.830, 0.960]	[0.824, 0.954]
1.3001	0.753	[0.651, 0.856]	[0.750, 0.914]
1.3063	0.734	[0.628, 0.841]	[0.729, 0.902]
1.9117	0.690	[0.574, 0.807]	[0.613, 0.889]
2.5728	0.635	[0.504, 0.765]	[0.624, 0.856]
2.8090	0.603	[0.465, 0.741]	[0.600, 0.837]
3.3481	0.563	[0.413, 0.712]	[0.547, 0.814]

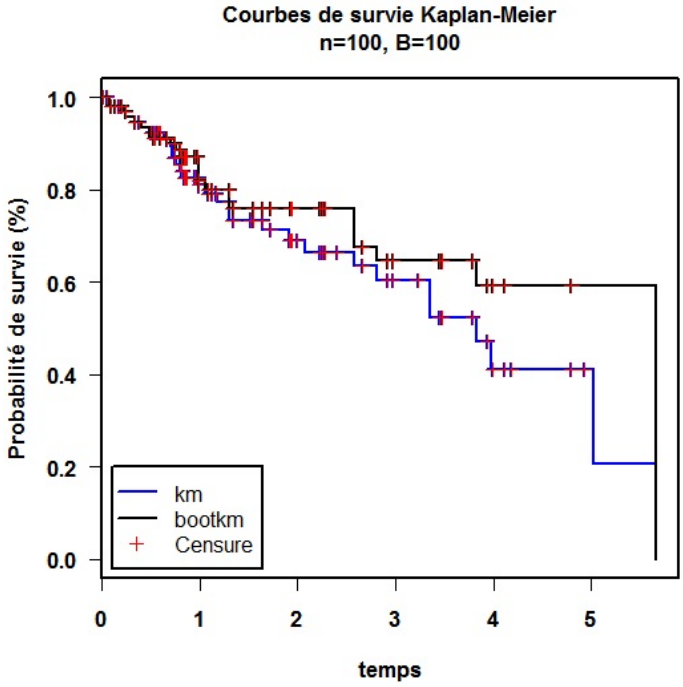


FIGURE 3.11 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 80%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=100 et comparaison avec la variance asymptotique.

Temps	0.0821	0.2322	0.7042	1.3001	1.3063	1.9117	2.5728	2.8090	3.3481
$\widehat{\sigma}(\widehat{S}(t))$	0.0144	0.0180	0.0333	0.0522	0.0543	0.0593	0.0665	0.0703	0.0762
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0141	0.0174	0.0331	0.0418	0.0441	0.464	0.0566	0.0605	0.0681

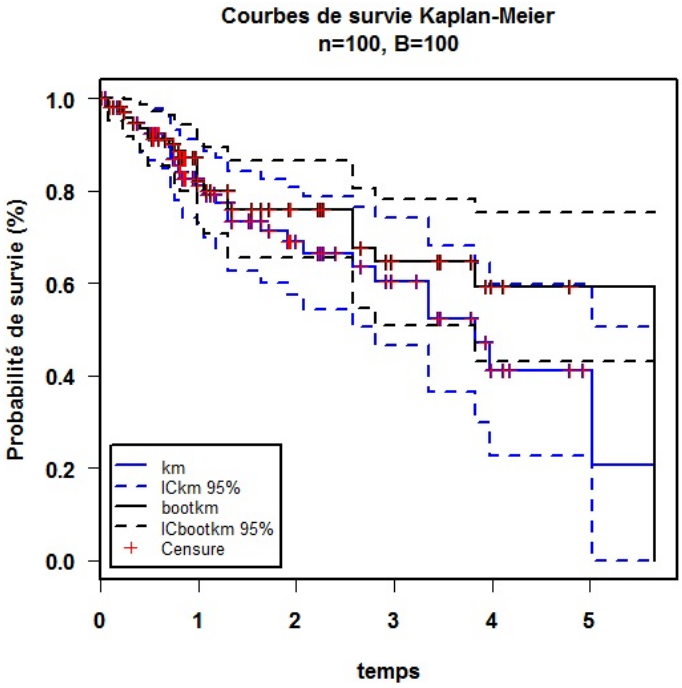


FIGURE 3.12 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0821	0.979	[0.951, 1.000]	[0.952, 1.000]
0.2322	0.968	[0.933, 1.000]	[0.935, 1.000]
0.7042	0.895	[0.830, 0.960]	[0.836, 0.962]
1.3001	0.753	[0.651, 0.856]	[0.679, 0.879]
1.3063	0.734	[0.628, 0.841]	[0.654, 0.864]
1.9117	0.690	[0.574, 0.807]	[0.613, 0.889]
2.5728	0.635	[0.504, 0.765]	[0.545, 0.804]
2.8090	0.603	[0.465, 0.741]	[0.509, 0.781]
3.3481	0.563	[0.413, 0.712]	[0.431, 0.752]

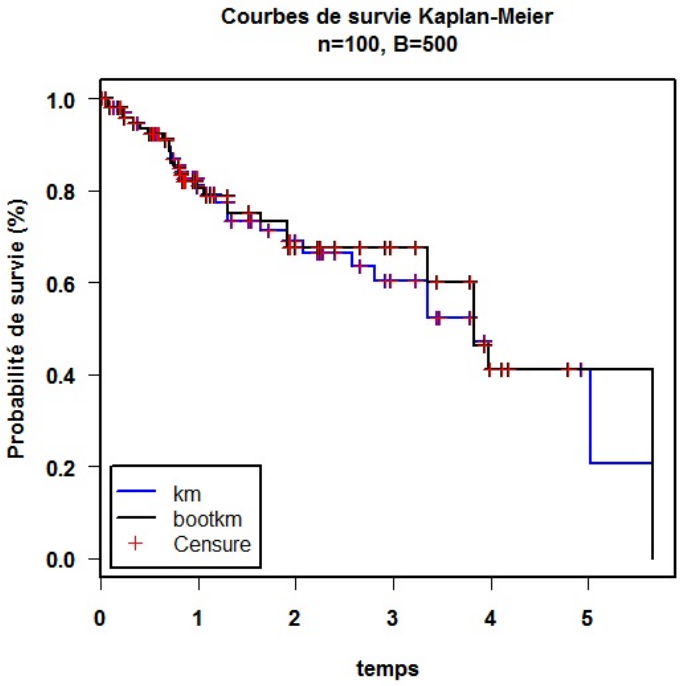


FIGURE 3.13 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 82%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=500 et comparaison avec la variance asymptotique.

Temps	0.0821	0.2322	0.7042	1.3001	1.3063	1.9117	2.5728	2.8090	3.3481
$\widehat{\sigma}(\widehat{S}(t))$	0.0144	0.0180	0.0333	0.0522	0.0543	0.0593	0.0665	0.0703	0.0762
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0144	0.0180	0.0331	0.0489	0.0510	0.0591	0.0661	0.0702	0.0759

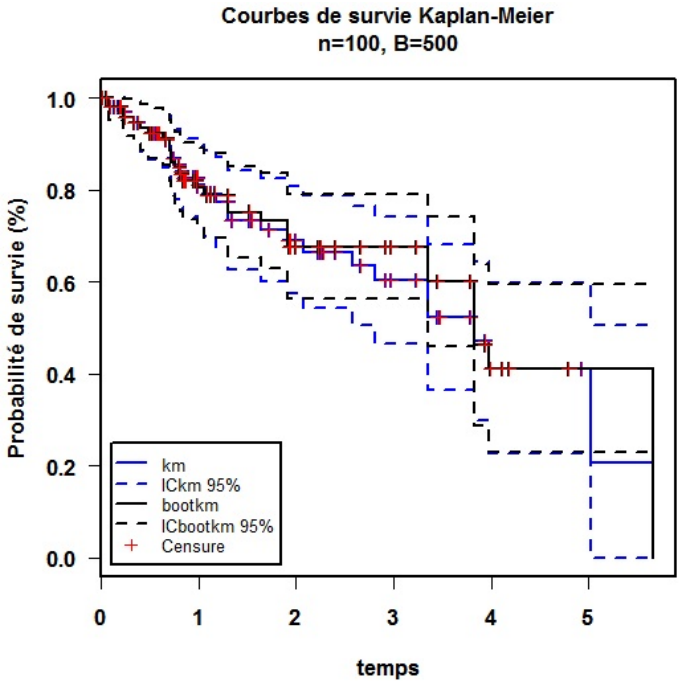


FIGURE 3.14 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0821	0.979	[0.951, 1.000]	[0.952, 1.000]
0.2322	0.968	[0.933, 1.000]	[0.935, 1.000]
0.7042	0.895	[0.830, 0.960]	[0.836, 0.962]
1.3001	0.753	[0.651, 0.856]	[0.674, 0.865]
1.3063	0.734	[0.628, 0.841]	[0.651, 0.851]
1.9117	0.690	[0.574, 0.807]	[0.562, 0.790]
2.5728	0.635	[0.504, 0.765]	[0.545, 0.804]
2.8090	0.603	[0.465, 0.741]	[0.509, 0.781]
3.3481	0.563	[0.413, 0.712]	[0.509, 0.768]

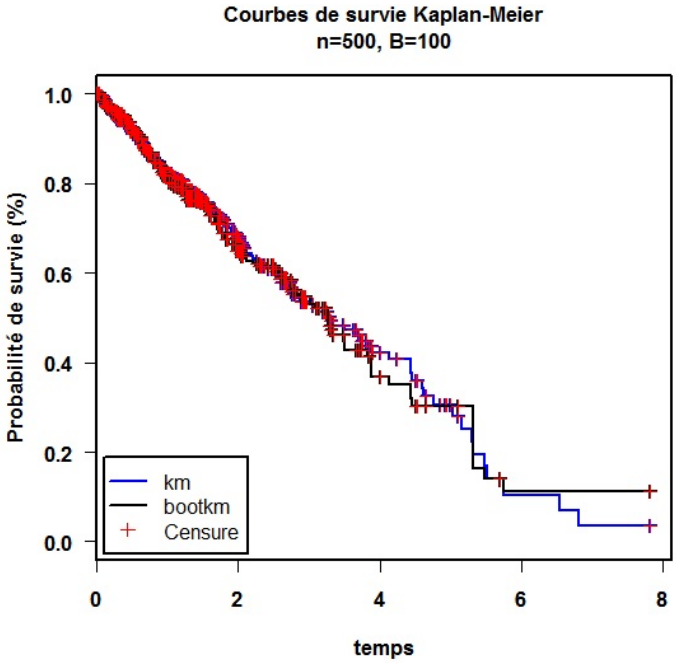


FIGURE 3.15 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 65%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=500, B=100 et comparaison avec la variance asymptotique.

Temps	0.1310	0.1523	0.2149	0.2785	0.5513	0.9034	1.6710	2.6007	3.8836
$\widehat{\sigma}(\widehat{S}(t))$	0.0073	0.0081	0.0089	0.0103	0.0139	0.0186	0.0254	0.0320	0.0443
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0073	0.0078	0.0084	0.0093	0.0136	0.0176	0.0245	0.0315	0.0341

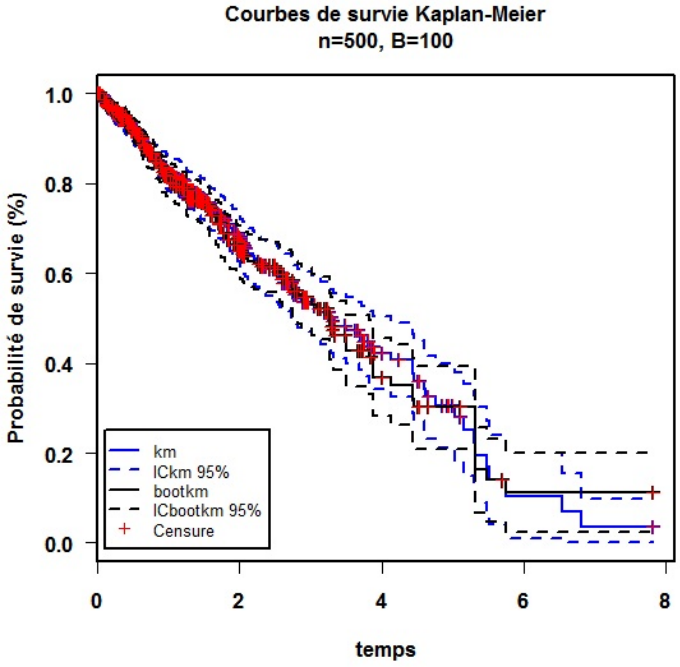


FIGURE 3.16 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.1310	0.9730	[0.958, 0.987]	[0.958, 0.988]
0.1523	0.9667	[0.950, 0.982]	[0.953, 0.984]
0.2149	0.9602	[0.926, 0.967]	[0.948, 0.981]
0.2785	0.9470	[0.922, 0.965]	[0.937, 0.974]
0.5513	0.9057	[0.878, 0.933]	[0.883, 0.937]
0.9034	0.8359	[0.787, 0.863]	[0.799, 0.872]
1.6710	0.7314	[0.661, 0.762]	[0.683, 0.779]
2.6007	0.5789	[0.516, 0.641]	[0.527, 0.651]
3.8836	0.4220	[0.281, 0.456]	[0.341, 0.502]

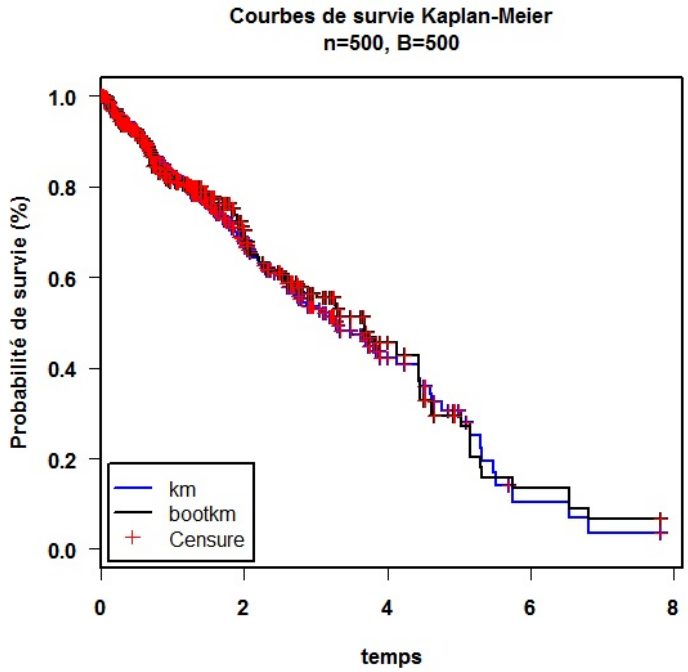


FIGURE 3.17 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 67%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=500, B=500 et comparaison avec la variance asymptotique.

Temps	0.1310	0.1523	0.2149	0.2785	0.5513	0.9034	1.6710	2.6007	3.8836
$\widehat{\sigma}(\widehat{S}(t))$	0.0073	0.0081	0.0089	0.0103	0.0139	0.0186	0.0254	0.0320	0.0443
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0073	0.0078	0.0089	0.0099	0.0137	0.0186	0.0252	0.0314	0.0341

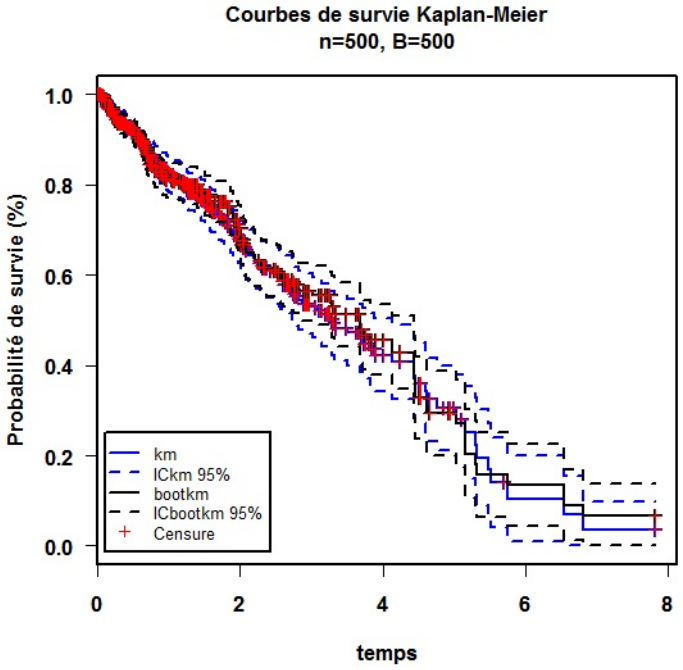


FIGURE 3.18 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.1310	0.9730	[0.958, 0.987]	[0.958, 0.987]
0.1523	0.9667	[0.950, 0.982]	[0.953, 0.984]
0.2149	0.9602	[0.926, 0.967]	[0.934, 0.973]
0.2785	0.9470	[0.922, 0.965]	[0.939, 0.962]
0.5513	0.9057	[0.878, 0.933]	[0.881, 0.935]
0.9034	0.8359	[0.787, 0.863]	[0.799, 0.872]
1.6710	0.7314	[0.661, 0.762]	[0.683, 0.779]
2.6007	0.5789	[0.516, 0.641]	[0.528, 0.651]
3.8836	0.4220	[0.281, 0.456]	[0.341, 0.502]

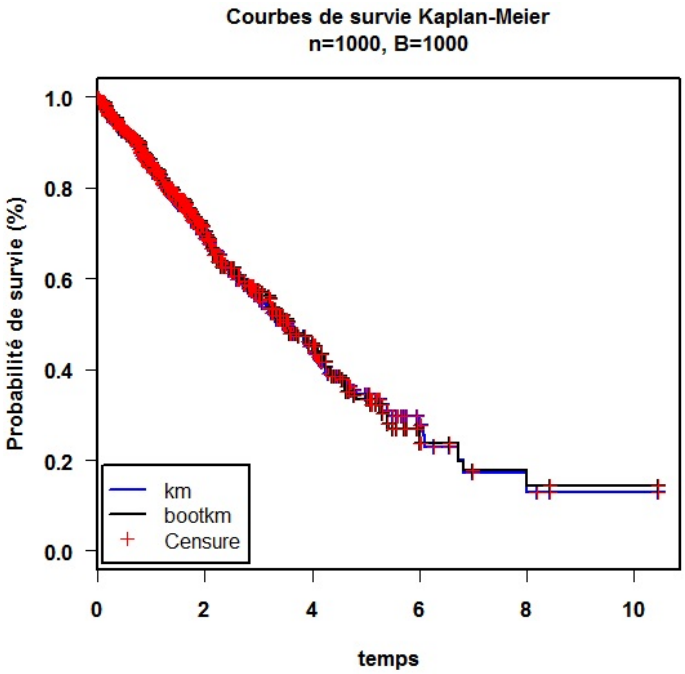


FIGURE 3.19 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 73%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=1000, B=1000 et comparaison avec la variance asymptotique.

Temps	0.0284	0.0469	0.0526	0.0841	0.1028	0.1216	1.5018	1.5510	2.3656
$\widehat{\sigma}(\widehat{S}(t))$	0.0022	0.0030	0.0033	0.0039	0.0046	0.0051	0.0165	0.0171	0.0225
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0014	0.0017	0.0022	0.0027	0.0035	0.0038	0.0162	0.0168	0.0223

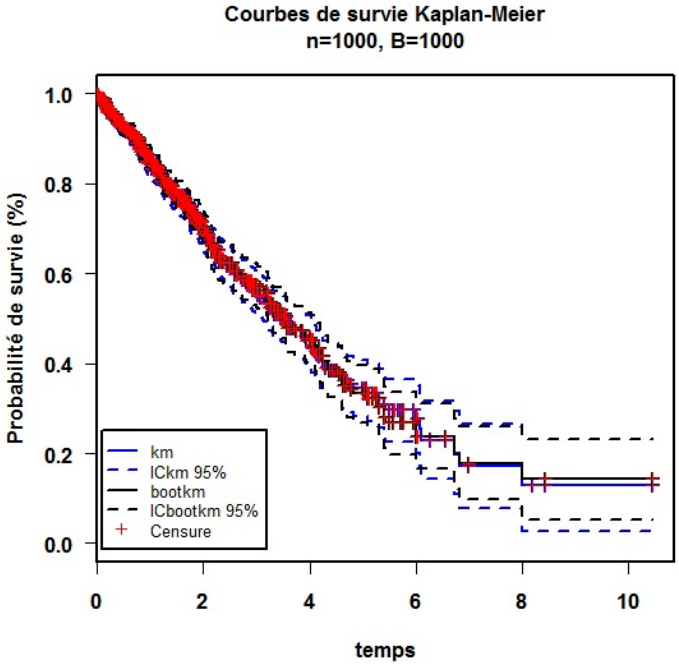


FIGURE 3.20 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0284	0.995	[0.991, 0.999]	[0.994, 1.000]
0.0469	0.991	[0.985, 0.997]	[0.990, 1.000]
0.0526	0.989	[0.982, 0.995]	[0.987, 0.995]
0.0841	0.985	[0.977, 0.992]	[0.984, 0.998]
0.1028	0.978	[0.969, 0.988]	[0.977, 0.995]
0.1216	0.974	[0.964, 0.984]	[0.972, 0.993]
1.5018	0.758	[0.725, 0.790]	[0.749, 0.779]
1.5510	0.743	[0.709, 0.776]	[0.735, 0.801]
2.3656	0.636	[0.571, 0.659]	[0.592, 0.680]

2. Modèle 2

$X_i \hookrightarrow W(0.5, 1)$ et $C_i \hookrightarrow \mathcal{E}(1/2)$, $n = 50,\ 100,\ 500$.

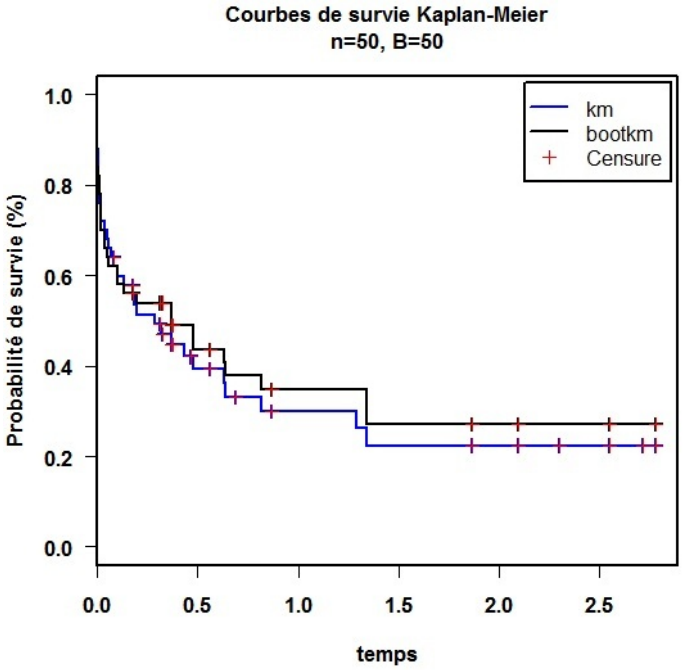


FIGURE 3.21 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 42%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=50 et comparaison avec la variance asymptotique.

Temps	0.0003	0.0010	0.0099	0.0154	0.0366	0.0978	0.4770	0.8170
$\widehat{\sigma}(\widehat{S}(t))$	0.0277	0.0424	0.0586	0.0620	0.0670	0.0698	0.0737	0.0742
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0277	0.0384	0.0566	0.0604	0.0648	0.0695	0.0727	0.0729

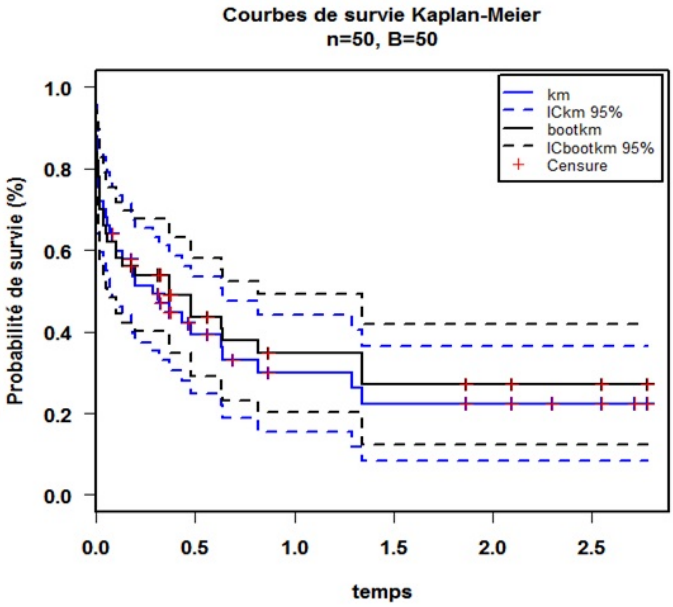


FIGURE 3.22 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0003	0.960	[0.905, 1.000]	[0.906, 1.000]
0.0010	0.900	[0.816, 0.983]	[0.845, 0.995]
0.0099	0.780	[0.665, 0.895]	[0.689, 0.911]
0.0154	0.740	[0.618, 0.862]	[0.642, 0.878]
0.0366	0.700	[0.529, 0.791]	[0.573, 0.827]
0.0978	0.599	[0.462, 0.735]	[0.446, 0.717]
0.4770	0.392	[0.291, 0.580]	[0.249, 0.535]
0.8170	0.299	[0.203, 0.494]	[0.155, 0.442]

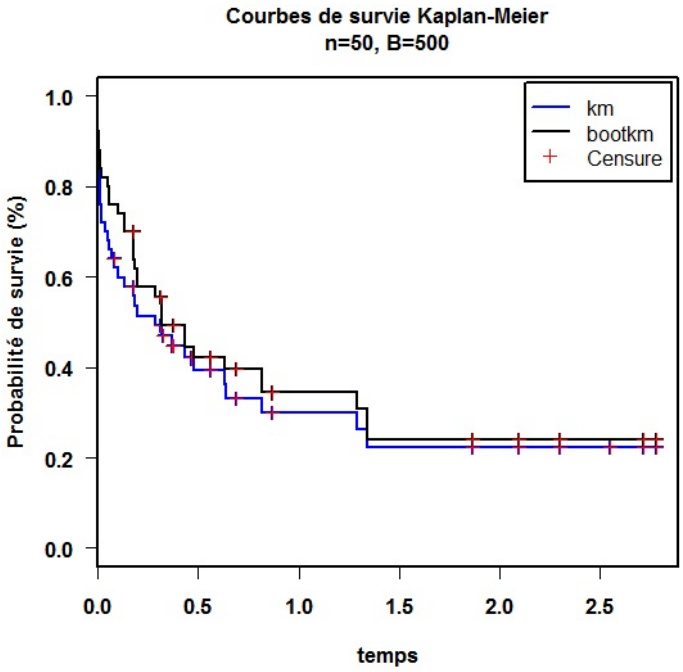


FIGURE 3.23 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 52%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=500 et comparaison avec la variance asymptotique.

Temps	0.0003	0.0010	0.0099	0.0154	0.0366	0.0978	0.4770	0.8170
$\widehat{\sigma}(\widehat{S}(t))$	0.0277	0.0424	0.0586	0.0620	0.0670	0.0698	0.0737	0.0742
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0198	0.0420	0.0566	0.0604	0.0620	0.0686	0.0720	0.0714

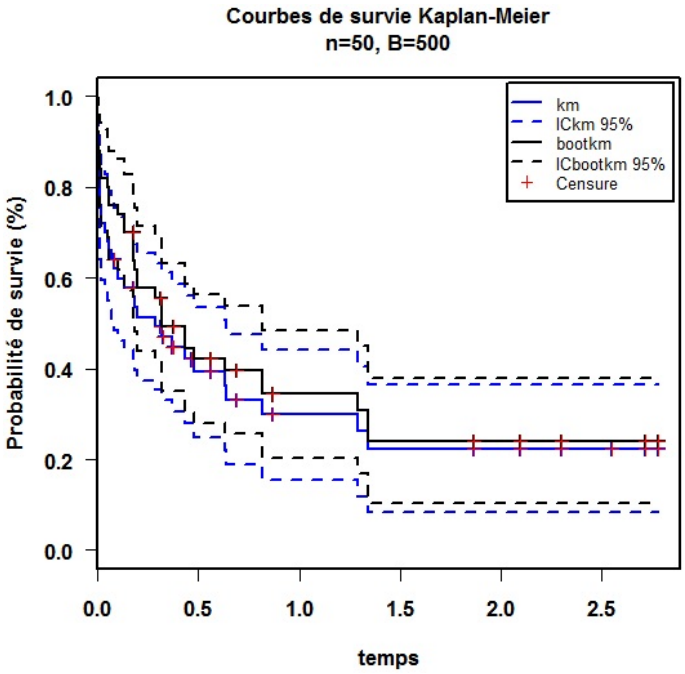


FIGURE 3.24 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0003	0.960	[0.905, 1.000]	[0.941, 1.000]
0.0010	0.900	[0.816, 0.983]	[0.845, 0.995]
0.0099	0.780	[0.665, 0.895]	[0.689, 0.911]
0.0154	0.740	[0.618, 0.862]	[0.714, 0.926]
0.0366	0.700	[0.529, 0.791]	[0.618, 0.862]
0.0978	0.599	[0.462, 0.735]	[0.485, 0.755]
0.4770	0.392	[0.291, 0.580]	[0.280, 0.563]
0.8170	0.299	[0.203, 0.494]	[0.204, 0.484]

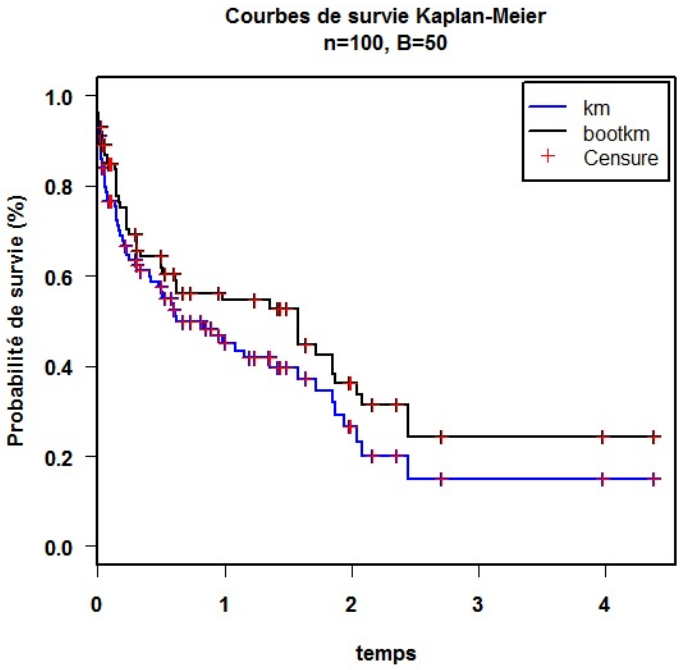


FIGURE 3.25 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 38%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=50 et comparaison avec la variance asymptotique.

Temps	0.0038	0.0247	0.0404	0.0557	0.0757	0.1500	0.1734	0.2531	0.9836
$\widehat{\sigma}(\widehat{S}(t))$	0.0217	0.0313	0.0369	0.0389	0.0428	0.0450	0.04683	0.0495	0.0553
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0217	0.0255	0.0315	0.0328	0.0363	0.0413	0.0444	0.0489	0.0547

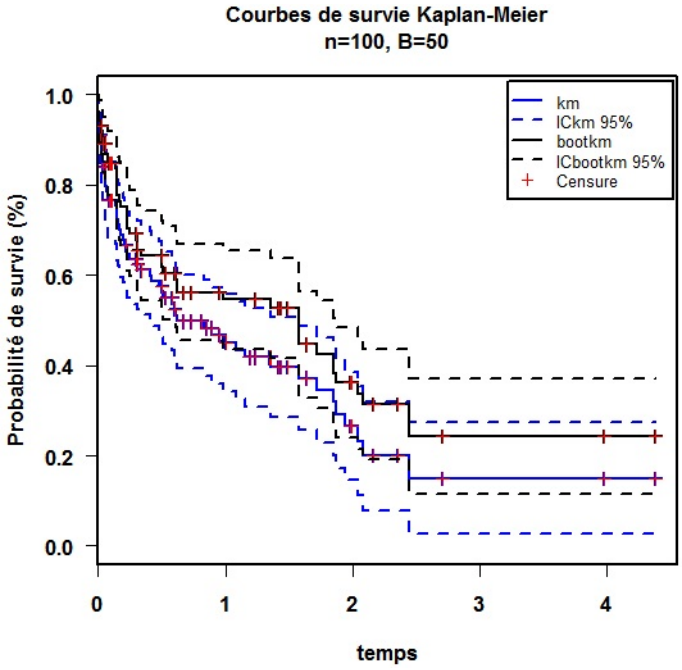


FIGURE 3.26 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0038	0.950	[0.907, 0.993]	[0.907, 0.993]
0.0247	0.890	[0.828, 0.951]	[0.880, 0.980]
0.0404	0.839	[0.766, 0.911]	[0.827, 0.951]
0.0557	0.818	[0.741, 0.894]	[0.814, 0.943]
0.0757	0.765	[0.681, 0.849]	[0.776, 0.918]
0.1500	0.732	[0.644, 0.821]	[0.718, 0.881]
0.1734	0.700	[0.607, 0.791]	[0.677, 0.851]
0.2531	0.634	[0.536, 0.731]	[0.596, 0.788]
0.9836	0.451	[0.437, 0.654]	[0.343, 0.558]

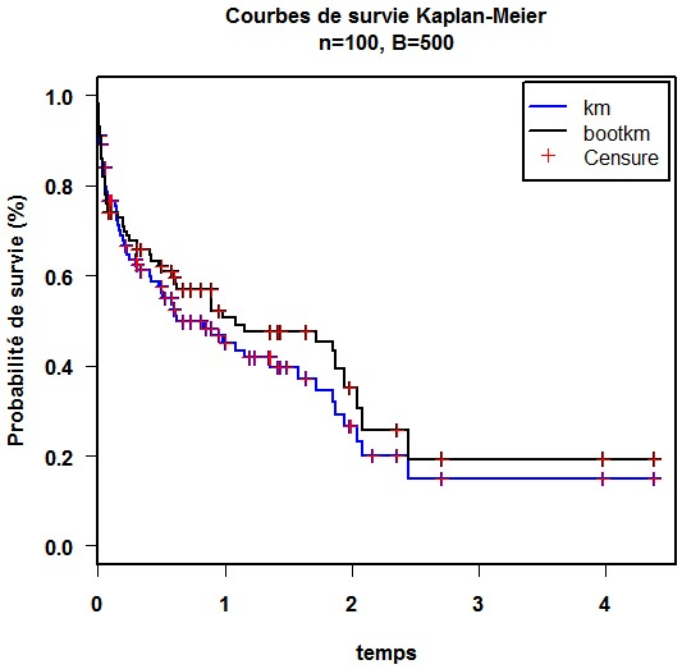


FIGURE 3.27 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 40%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=500 et comparaison avec la variance asymptotique.

Temps	0.0038	0.0247	0.0404	0.0557	0.0757	0.1500	0.1734	0.2531	0.9836
$\widehat{\sigma}(\widehat{S}(t))$	0.0217	0.0313	0.0369	0.0389	0.0428	0.0450	0.04683	0.0495	0.0553
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0217	0.0310	0.0315	0.0328	0.0363	0.0413	0.0444	0.0470	0.0547

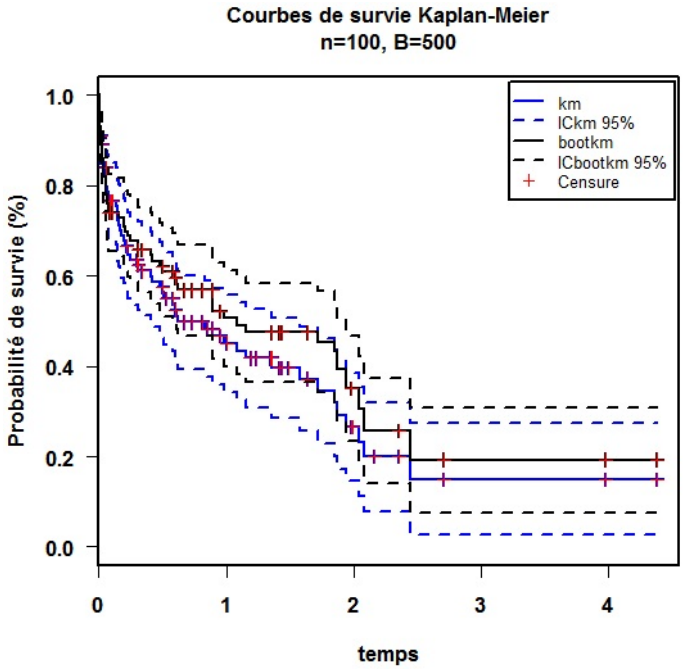


FIGURE 3.28 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0038	0.950	[0.907, 0.993]	[0.907, 0.993]
0.0247	0.890	[0.828, 0.951]	[0.880, 0.980]
0.0404	0.839	[0.766, 0.911]	[0.827, 0.951]
0.0557	0.818	[0.741, 0.894]	[0.814, 0.943]
0.0757	0.765	[0.681, 0.849]	[0.776, 0.918]
0.1500	0.732	[0.644, 0.821]	[0.718, 0.881]
0.1734	0.700	[0.607, 0.791]	[0.677, 0.851]
0.2531	0.634	[0.536, 0.731]	[0.585, 0.770]
0.9836	0.451	[0.437, 0.654]	[0.398, 0.613]

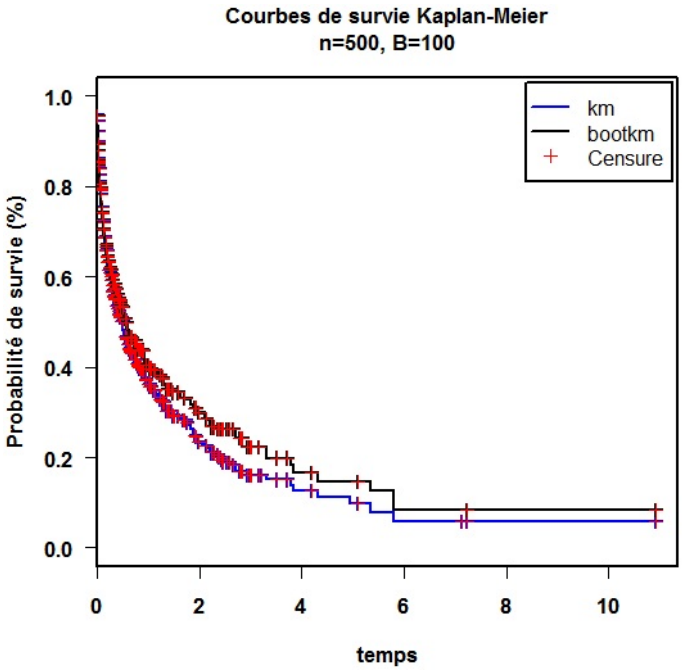


FIGURE 3.29 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 32%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=500, B=100 et comparaison avec la variance asymptotique.

Temps	0.0162	0.0309	0.0462	0.0810	0.139	0.260	0.599	0.633	0.839
$\widehat{\sigma}(\widehat{S}(t))$	0.0020	0.0056	0.0071	0.0076	0.0087	0.0091	0.0115	0.0118	0.0124
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0020	0.0048	0.0068	0.0074	0.0085	0.0089	0.0106	0.0111	0.0118

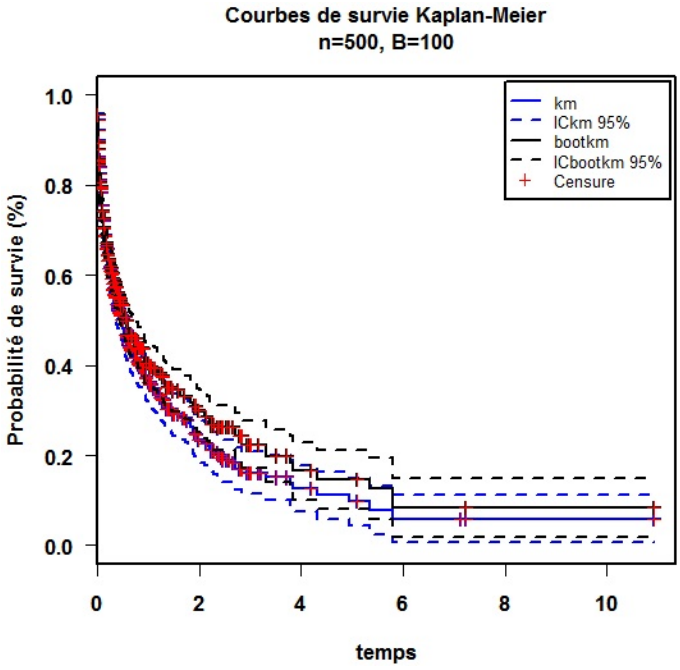


FIGURE 3.30 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0162	0.998	[0.994, 1.000]	[0.994, 1.000]
0.0309	0.984	[0.973, 0.995]	[0.978, 0.998]
0.0462	0.976	[0.962, 0.989]	[0.962, 0.988]
0.0810	0.974	[0.955, 0.985]	[0.960, 0.987]
0.139	0.962	[0.942, 0.977]	[0.945, 0.978]
0.260	0.956	[0.937, 0.974]	[0.940, 0.976]
0.599	0.927	[0.905, 0.950]	[0.919, 0.961]
0.633	0.923	[0.900, 0.947]	[0.912, 0.956]
0.839	0.915	[0.891, 0.940]	[0.900, 0.947]

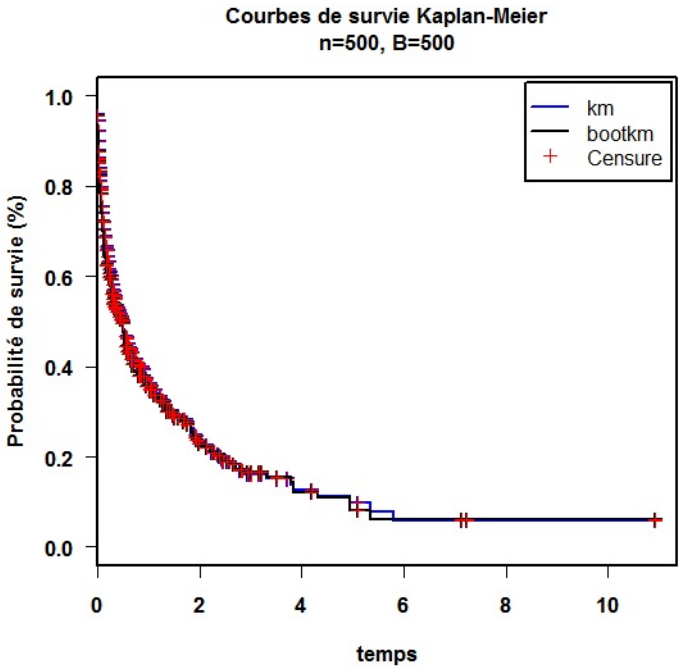


FIGURE 3.31 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 33%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=500, B=500 et comparaison avec la variance asymptotique.

Temps	0.0162	0.0309	0.0462	0.0810	0.139	0.260	0.599	0.633	0.839
$\widehat{\sigma}(\widehat{S}(t))$	0.0020	0.0056	0.0071	0.0076	0.0087	0.0091	0.0115	0.0118	0.0124
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0020	0.0055	0.0065	0.0071	0.0089	0.0091	0.0104	0.0111	0.0120

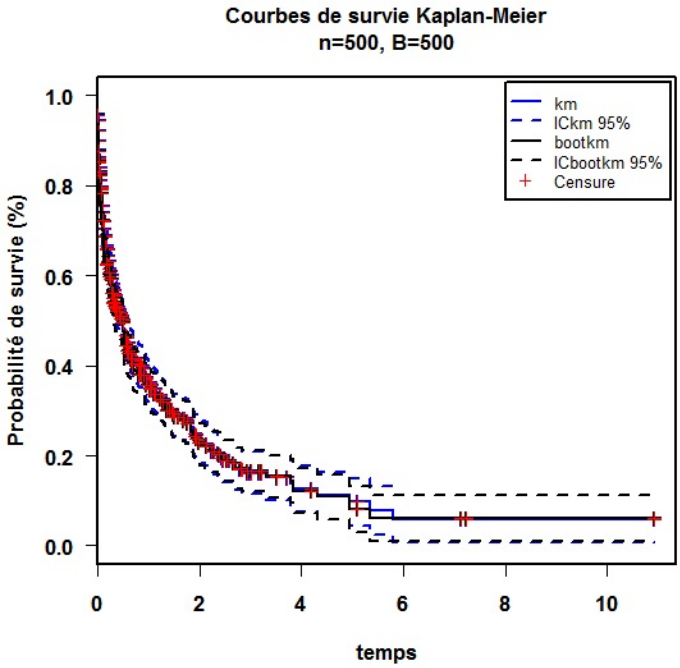


FIGURE 3.32 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
0.0162	0.998	[0.994, 1.000]	[0.994, 1.000]
0.0309	0.984	[0.973, 0.995]	[0.978, 0.994]
0.0462	0.976	[0.962, 0.989]	[0.967, 0.988]
0.0810	0.974	[0.955, 0.985]	[0.960, 0.988]
0.139	0.962	[0.942, 0.977]	[0.945, 0.976]
0.260	0.956	[0.937, 0.974]	[0.938, 0.974]
0.599	0.927	[0.905, 0.950]	[0.921, 0.962]
0.633	0.923	[0.900, 0.947]	[0.912, 0.956]
0.839	0.915	[0.891, 0.940]	[0.898, 0.945]

3. Modèle 3

$X_i \hookrightarrow \mathcal{N}(2, 0.3)$ et $C_i \hookrightarrow \mathcal{N}(3, 0.5)$, $n = 50,\ 100,\ 500$.

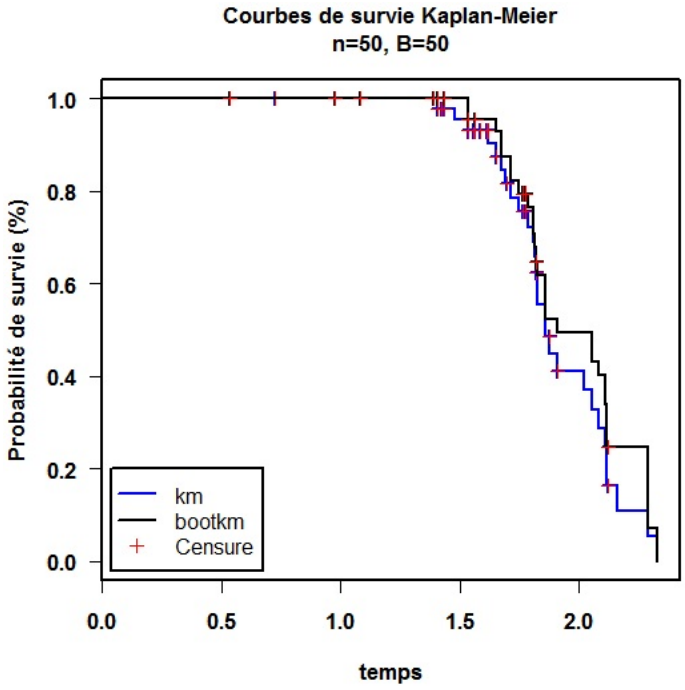


FIGURE 3.33 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 34%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=50 et comparaison avec la variance asymptotique.

Temps	1.54	1.65	1.75	1.81	1.82	1.86	1.91	2.05
$\widehat{\sigma}(\widehat{S}(t))$	0.0390	0.0532	0.0718	0.0792	0.0819	0.0890	0.0900	0.0918
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0321	0.0407	0.0650	0.0751	0.0796	0.0831	0.854	0.851

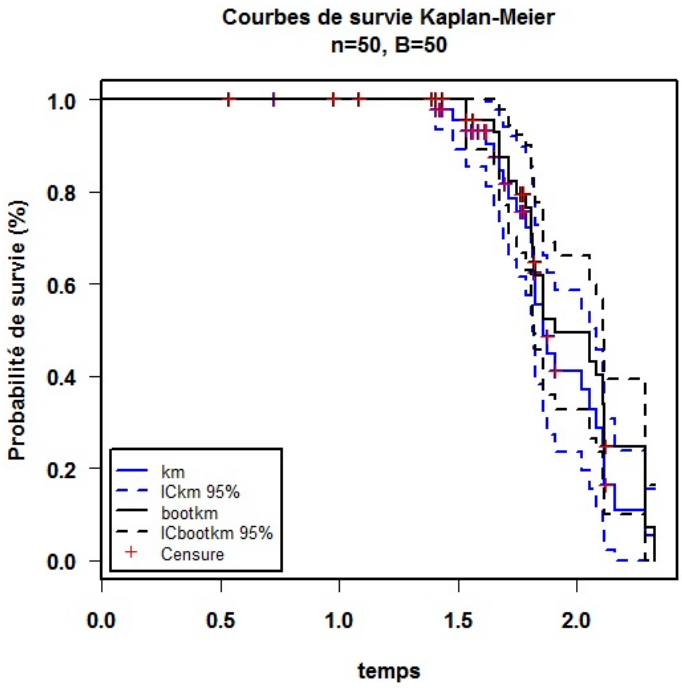


FIGURE 3.34 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.54	0.930	[0.853, 1.000]	[0.890, 1.000]
1.65	0.873	[0.769, 0.978]	[0.847, 1.000]
1.75	0.755	[0.614, 0.896]	[0.667, 0.922]
1.81	0.689	[0.505, 0.809]	[0.625, 0.854]
1.82	0.656	[0.496, 0.817]	[0.524, 0.829]
1.86	0.519	[0.345, 0.694]	[0.357, 0.691]
1.91	0.410	[0.234, 0.587]	[0.325, 0.661]
2.05	0.328	[0.154, 0.502]	[0.264, 0.599]

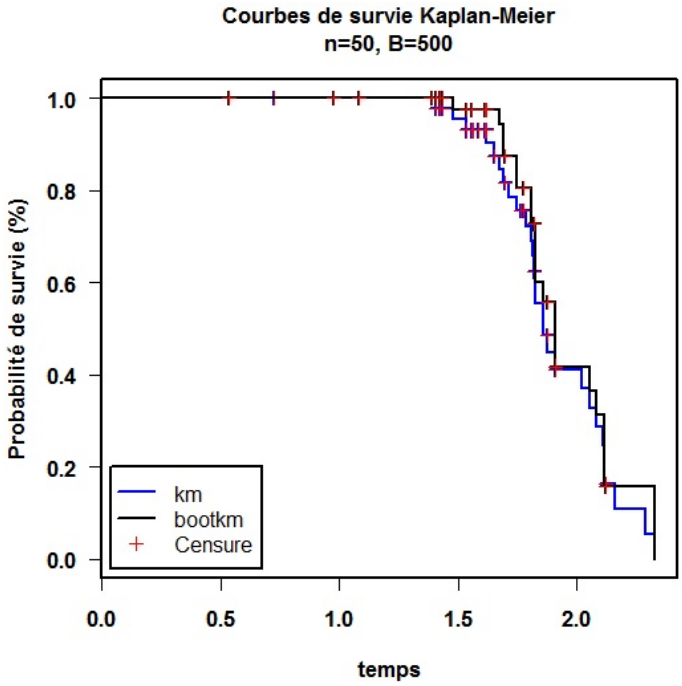


FIGURE 3.35 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 34%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=500 et comparaison avec la variance asymptotique.

Temps	1.54	1.65	1.75	1.81	1.82	1.86	1.91	2.05
$\widehat{\sigma}(\widehat{S}(t))$	0.0390	0.0532	0.0718	0.0792	0.0819	0.0890	0.0900	0.0918
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0225	0.0330	0.0682	0.0774	0.0810	0.0825	0.0831	0.0841

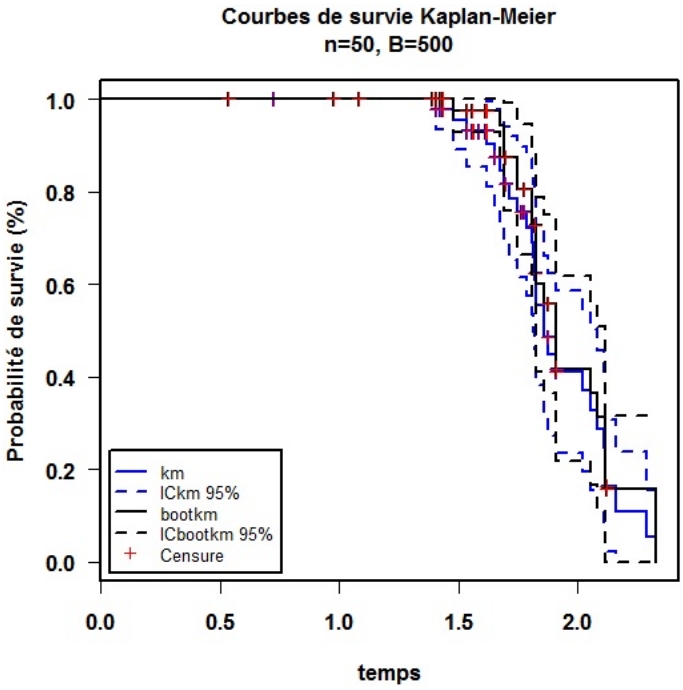


FIGURE 3.36 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.54	0.930	[0.853, 1.000]	[0.923, 1.000]
1.65	0.873	[0.769, 0.978]	[0.867, 1.000]
1.75	0.755	[0.614, 0.896]	[0.633, 0.901]
1.81	0.689	[0.505, 0.809]	[0.618, 0.862]
1.82	0.656	[0.496, 0.817]	[0.416, 0.734]
1.86	0.519	[0.345, 0.694]	[0.356, 0.679]
1.91	0.410	[0.234, 0.587]	[0.324, 0.650]
2.05	0.328	[0.154, 0.502]	[0.287, 0.617]

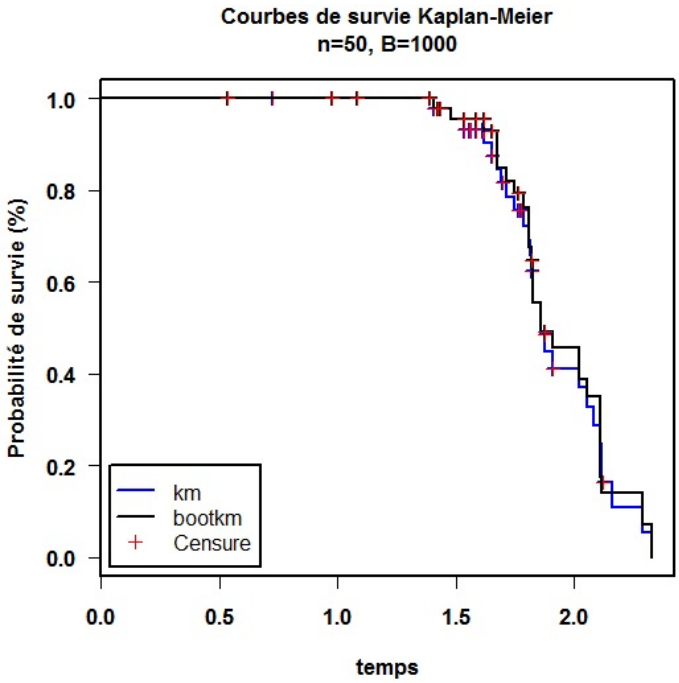


FIGURE 3.37 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 38%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=50, B=1000 et comparaison avec la variance asymptotique.

Temps	1.54	1.65	1.75	1.81	1.82	1.86	1.91	2.05
$\widehat{\sigma}(\widehat{S}(t))$	0.0390	0.0532	0.0718	0.0792	0.0819	0.0890	0.0900	0.0918
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0315	0.0398	0.0658	0.0780	0.0799	0.0851	0.0861	0.0863

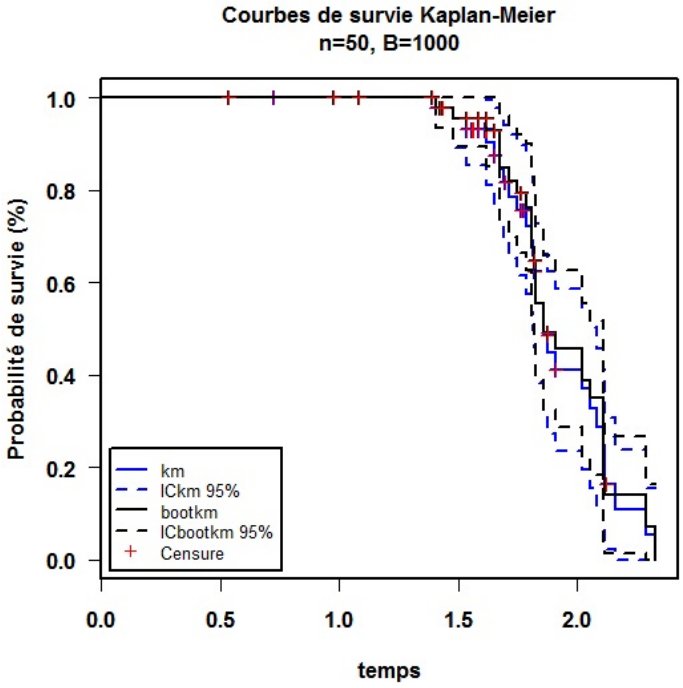


FIGURE 3.38 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.54	0.930	[0.853, 1.000]	[0.892, 1.000]
1.65	0.873	[0.769, 0.978]	[0.850, 1.000]
1.75	0.755	[0.614, 0.896]	[0.663, 0.921]
1.81	0.689	[0.505, 0.809]	[0.522, 0.828]
1.82	0.656	[0.496, 0.817]	[0.488, 0.802]
1.86	0.519	[0.345, 0.694]	[0.355, 0.689]
1.91	0.410	[0.234, 0.587]	[0.287, 0.626]
2.05	0.328	[0.154, 0.502]	[0.184, 0.518]

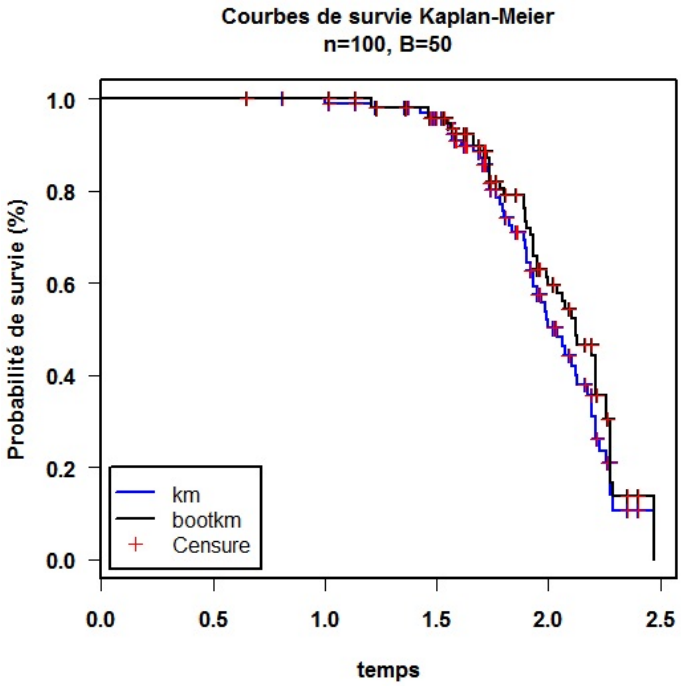


FIGURE 3.39 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 56%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=50 et comparaison avec la variance asymptotique.

Temps	1.47	1.57	1.73	1.80	1.89	1.92	1.99	2.04	2.13
$\widehat{\sigma}(\widehat{S}(t))$	0.0209	0.0285	0.0403	0.0496	0.0546	0.0585	0.0622	0.0638	0.0647
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0208	0.0281	0.0363	0.0455	0.0483	0.0527	0.0575	0.0592	0.0632

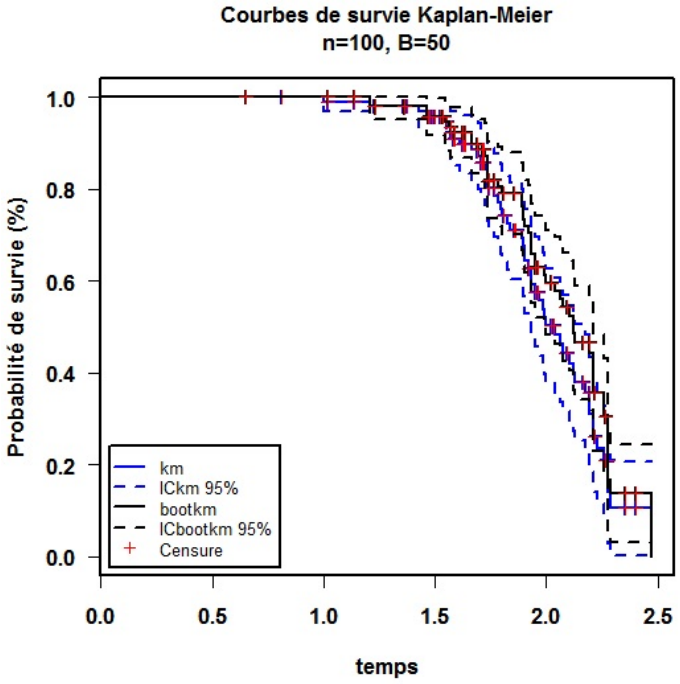


FIGURE 3.40 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.47	0.957	[0.916, 0.998]	[0.916, 0.998]
1.57	0.922	[0.865, 0.978]	[0.867, 0.978]
1.73	0.839	[0.764, 0.922]	[0.800, 0.943]
1.80	0.818	[0.658, 0.853]	[0.701, 0.880]
1.89	0.765	[0.585, 0.849]	[0.667, 0.857]
1.92	0.732	[0.599, 0.807]	[0.644, 0.799]
1.99	0.700	[0.416, 0.722]	[0.500, 0.726]
2.04	0.634	[0.356, 0.606]	[0.462, 0.604]
2.13	0.451	[0.341, 0.589]	[0.351, 0.505]

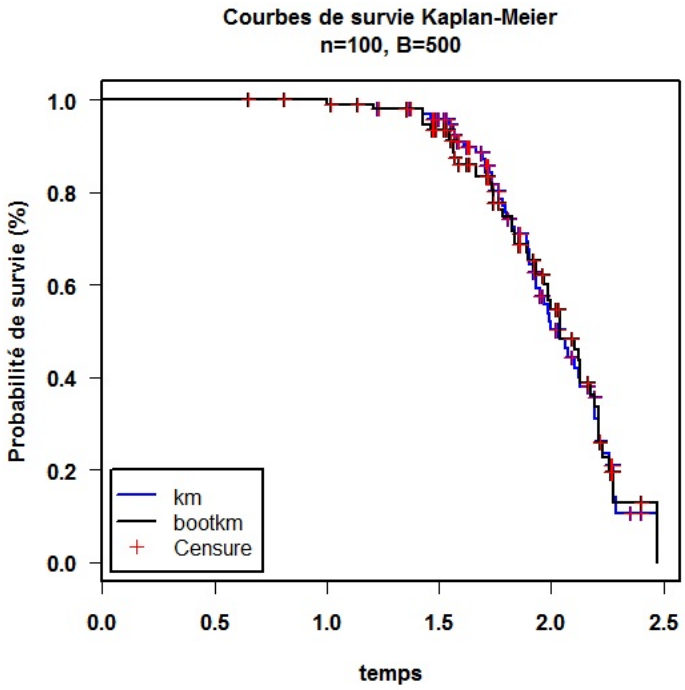


FIGURE 3.41 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 49%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=100, B=500 et comparaison avec la variance asymptotique.

Temps	1.47	1.57	1.73	1.80	1.89	1.92	1.99	2.04	2.13
$\widehat{\sigma}(\widehat{S}(t))$	0.0209	0.0285	0.0403	0.0496	0.0546	0.0585	0.0622	0.0638	0.0647
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0148	0.0257	0.0376	0.0484	0.0543	0.0573	0.0610	0.0617	0.0645

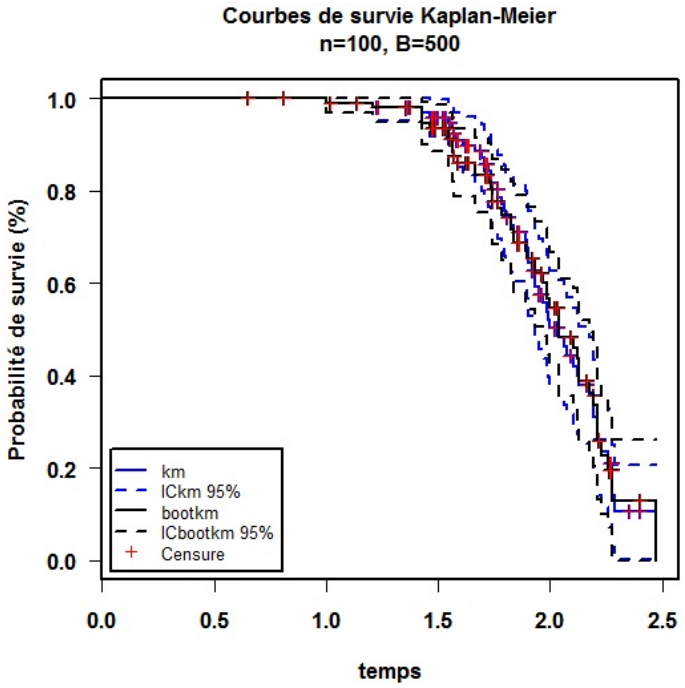


FIGURE 3.42 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.47	0.957	[0.916, 0.998]	[0.949, 1.000]
1.57	0.922	[0.865, 0.978]	[0.884, 0.985]
1.73	0.839	[0.764, 0.922]	[0.786, 0.934]
1.80	0.818	[0.658, 0.853]	[0.666, 0.856]
1.89	0.765	[0.585, 0.849]	[0.589, 0.847]
1.92	0.732	[0.599, 0.807]	[0.587, 0.799]
1.99	0.700	[0.416, 0.722]	[0.445, 0.726]
2.04	0.634	[0.356, 0.606]	[0.425, 0.667]
2.13	0.451	[0.341, 0.589]	[0.357, 0.603]

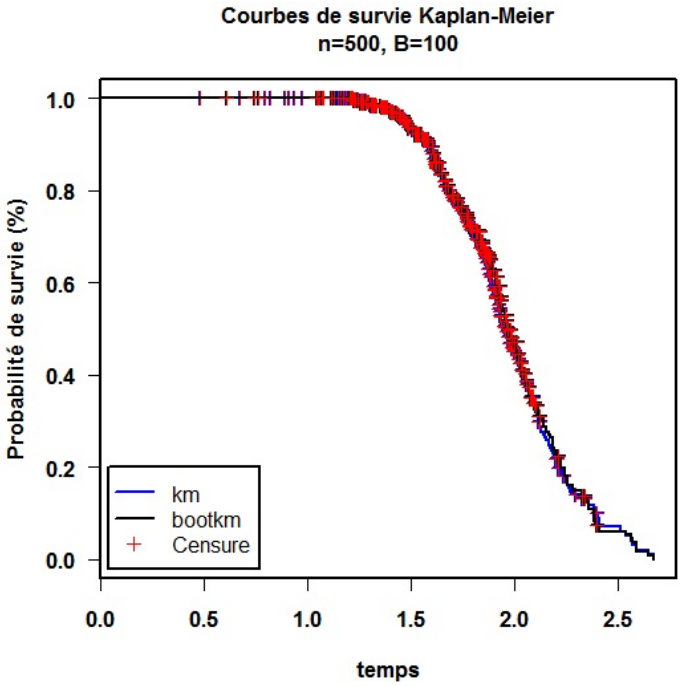


FIGURE 3.43 – Estimation de survie selon la méthode de Bootstrap

Le taux de censure est de 52%.

Le tableau ci-dessous affiche les estimations de la variance Bootstrap de la fonction de survie, n=500, B=100 et comparaison avec la variance asymptotique.

Temps	1.24	1.26	1.31	1.38	1.43	1.49	1.56	1.93	2.02
$\widehat{\sigma}(\widehat{S}(t))$	0.0030	0.0037	0.0057	0.0069	0.0085	0.0113	0.0125	0.0275	0.0289
$\widehat{\sigma}^*(\widehat{S}^*(t))$	0.0021	0.0029	0.0042	0.0052	0.0065	0.0103	0.0116	0.0265	0.0276

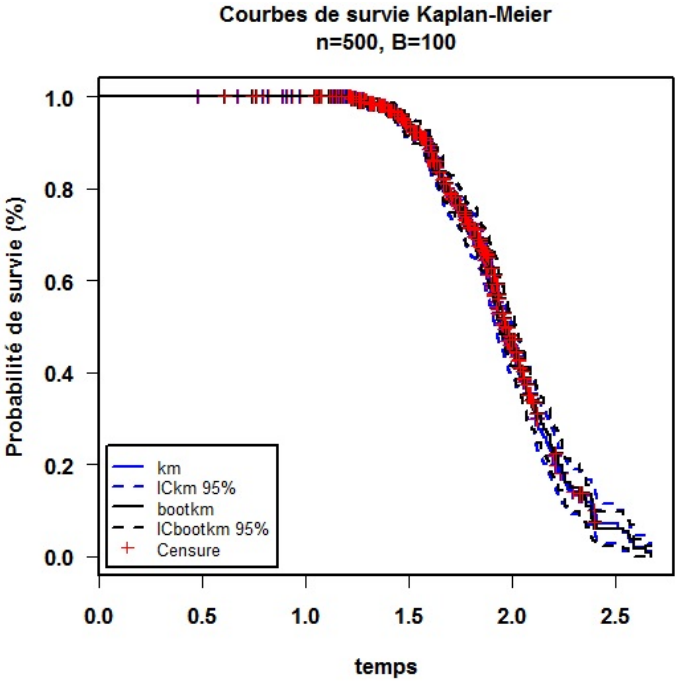


FIGURE 3.44 – Estimation d'I.C selon Bootstrap

On a le tableau suivant :

Temps	$\widehat{S}(t_i)$	$I.C.km95\%$	$I.C.bootkm95\%$
1.24	0.995	[0.989, 1.000]	[0.993, 1.000]
1.26	0.993	[0.986, 1.000]	[0.989, 1.000]
1.31	0.984	[0.973, 0.996]	[0.983, 0.999]
1.38	0.977	[0.977, 0.985]	[0.976, 0.983]
1.43	0.966	[0.949, 0.983]	[0.955, 0.977]
1.49	0.942	[0.921, 0.964]	[0.930, 0.970]
1.56	0.927	[0.918, 0.962]	[0.917, 0.959]
1.93	0.539	[0.485, 0.593]	[0.502, 0.606]
2.02	0.438	[0.381, 0.495]	[0.394, 0.502]

3.2.2 Conclusion

Dans les études de simulations présentées dans cette partie, nous remarquons que

- Les graphes obtenus pour les trois modèles montrent une bonne performance de l'estimateur de la fonction de survie et ce même pour les petites tailles de l'échantillon.

- La courbe de survie obtenue par bootstrap est très proche de la courbe de survie d'autant que la taille n de l'echantillon augmente et B aussi.

- La qualité de l'estimation s'améliore quand la taille de l'échantillon augmente, ce qui est tout à fait prévisible.

- Les intervalles de confiance sont meilleurs pour les grandes valeurs de n et B.

Conclusion et recommandations

Dans ce mémoire nous avons étudié l'analyse des données de survie et nous sommes intéressé au cas de censure aléatoire à droite avec l'hypothèse d'indépendance des durées de survie et les variables de censure.

Dans le chapitre un, nous avons étudié la méthode d'estimation de la fonction de survie en particulier l'estimateur de Kaplan-Meier.

Dans le chapitre deux, nous avons étudié l'estimateur de Kaplan Meier bootstrappé des durées de survie dans le cas de données censurées.

Dans le chapitre trois, nous avons consacré un temps considérable à la programmation à l'aide du logiciel R, afin de simuler des échantillons et calculer les estimateurs et les intervalles de confiance présentés aux chapitre 1 et 2. Nous avons présenté quelques résultats d'éxécusion de notre programme, d'autres résultats sont disponibles et d'autres modèles peuvent envisagés.

Ces résultats montrent la performance de la méthode du bootstrap pour l'analyse des durées de vie dans le cas i.i.d.

Des perspectives à notre étude se présentent à nous, nous citons en particulier :

1. L'étude par le boostrap, d'autres paramètres du modèle de censure comme le mode, la fonction de régression dans le cas de présence de covaiables.
2. L'analyse des mêmes paramètres et les mêmes modèles (exponentiel, Weibull, normale) dans le cas dépendant par exemple le cas où les X_i ne sont pas indépendants mais proviennent d'un processus ARMA.
3. Evaluer l'effet de la censure sur la qualité des estimateurs bootstrapés en mettant l'accent sur le taux de censure (le pourcentage d'observations censurées dans l'échatillon).

Annexe

Dans cette partie nous rappelons des résultats que nous avons utilisés dans ce mémoire ([10] chapitre 1).

Théorème 3.1. *(Th. 3.2.3 [10], p. 97). Si S(t) > 0 alors*

$$\frac{\widehat{S}(t)}{S(t)} = 1 - \int_0^s \frac{1 - \widehat{S}_n(s^-)}{S(s)} \left(\frac{d\overline{N}_n(s)}{\overline{Y}_n(s)} - dH(s) \right)$$

Théorème 3.2. *(Inégalité de Lenglart. cf [10], p. 113). Soit X un processus continu à droite et Y un processus prédictible tel que $Y(0) = 0$. Si pour tout temps d'arrêt borné T, $\mathbf{E}(|X(T)|) \leq \mathbf{E}(|Y(T)|)$ alors pour tout temps d'arrêt T et tout couple (ε, η) positifs*

$$\mathbf{P}(\sup_{t \leq T} |X(t)| \geq \varepsilon) \leq \frac{\eta}{\varepsilon} \mathbf{P}(Y(T) \geq \eta)$$

Corollaire(Coro. 3.4.1 de [10], p. 113). Soit$N = \{N(t), t \geq 0\}$ un processus ponctuel et $M = N - A$ la martingale locale de carré intégrable associée. Si H est un processus adapté continu à gauche possedant des limites à droite (ou plus généralement prédictible et localement borné), alors, pour tout temps d'arrêt T tel que $\mathbf{P}(T < \infty) = 1$ et tout $\varepsilon, \eta > 0$,

$$\mathbf{P}\left(\sup_{t \leq T} \{ \int_0^t H(s) dM(s) \}^2 \geq \varepsilon \right) \leq \frac{\eta}{\varepsilon} + \mathbf{P}\left(\int_0^T H^2(s) d < M > (s) \geq \varepsilon \right)$$

Les processus $\overline{N}_n, \overline{Y}_n$ et $\widehat{F}_n$ sont introduit en chapitre 1

Théorème 3.3. *(Th.3.4.2. de [10], p. 115). Si pour tout $t > 0$, $\overline{Y}_n(t) \to^P_{n \to \infty} \infty$, alors*

$$\sup_{0 \leq s \leq t} | \int_0^s \frac{d\overline{N}_n(u)}{\overline{Y}_n(u)} - H(u)| \to^P_{n \to \infty} 0$$

et

$$\sup_{0 \leq s \leq t} |\widehat{F}_n(s) - F(s)| \to^P_{n \to \infty} 0.$$

Définition 3.1. *(Convergence en distribution). On dit que la suite d'éléments aléatoires* $(X_n)_{n\geq 1}$ *converge en dstribution (ou en loi) vers* X *lorsque* $(P_{X_n})_{n\geq 1}$ *converge faiblement vers* P_X*. Ainsi, on notera*

$$X_n \underset{n\longrightarrow\infty}{\longrightarrow} X \quad si\ et\ seulement\ si \quad P_{X_n} \underset{n\longrightarrow\infty}{\longrightarrow} P_X.$$

Codes des fonctions R

```
##***********************************************
## Exemple 1. Données de Freireich.
##------------------------------
library(survival)
freireich <- read.table("Datafreireich.csv", sep = ";", header = T)
head(freireich)
data <- data.frame(freireich)
data
freireich.surv = Surv(freireich$temps, freireich$statut)
fit <- survfit(freireich.surv ~ groupe, data = freireich)
summary(fit)
g6MP = freireich$groupe == "6MP"
f6MP = freireich[g6MP, ]
gplacebo = freireich$groupe == "Placebo"
fplacebo = freireich[gplacebo, ]
freireich= Surv(f6MP$temps, f6MP$statut)
fit6MP <- survfit(freireich ~ 1, conf.int = 0.95, conf.type = "plain", data = f6MP)
summary(fit6MP)
t.u <- summary(fit6MP , f6MP$temps[f6MP$statut==0])$time
surv.u <- summary(fit6MP , f6MP$temps[f6MP$statut==0])$surv
```

```
# FIGURE 1.1
##------------------------------
par (fg="black",lwd=2)
par(cex.lab=1.1)
par(cex.axis=1)
par(cex.main=1.5)
plot(fit,pch=3,col="black",conf.int=F,lwd=2,xlab="Durée de rémission (semaines)",
font.axis =2,las=1,font.lab =2,col.main = "black",ylab="Probabilité de survie (%)") ;
par(new=TRUE)
plot(fit6MP,pch=3,col="blue",conf.int=F,lwd=2,las=1)
points(t.u, surv.u, col='red', pch=3,lwd=1)
legend(17,1.04, c("Groupe traité par 6-MP","Groupe traité par Placebo","censure"),
col=c("blue","black","red"),lwd=c(2,2,1), lty = c(1,1,-1), pch = c(-1,-1,3),cex=1)
```

```
# FIGURE 1.3
##------------------------------
par (fg="black",lwd=2)
par(cex.lab=1.1)
par(cex.axis=1)
par(cex.main=1.5)
plot(fit6MP,pch=2,col="blue",lwd=2,
xlab="Durée de rémission (semaines) ",font.axis =2,las=1,
col.axis = "black", font.lab =2,font.main = 2, ylab="Probabilité de survie (%)")
points(t.u, surv.u, col='red', pch=3,lwd=1)
points(23,0.5,pch=19,col="green",cex =2)
abline(h=0.5,lty=2,lwd=1)
abline(v=23,lty=2,col="red",lwd=1)
text(23,0,"23",col="black",cex=1.12)
text(1,0.5,"0.5",col="black",cex=1.12)
legend(0.4,0.14, c("Groupe traité par 6-MP","IC 95%","censure"),
col=c("blue","blue","red"), lwd=c(2,2,1), lty = c(1,2,-1), pch = c(-1,-1,3),cex=0.72)
```

```
##***********************************************
## Exemple 2. Données des ventilateurs.
##-----------------------------------
time<-c(4.5, 4.6, 11.5, 11.5, 15.6, 16.0, 16.6, 18.5, 18.5, 18.5, 18.5, 18.5, 20.3, 20.3, 20.3,
20.7, 20.7, 20.8, 22.0, 30.0, 30.0, 30.0, 30.0, 31.0, 32.0, 34.5, 37.5, 37.5, 41.5, 41.5, 41.5,
41.5, 43.0, 43.0, 43.0, 43.0, 46.0, 48.5, 48.5, 48.5, 48.5, 50.0, 50.0, 50.0, 61.0, 61.0, 61.0,
61.0, 63.0, 64.5, 64.5, 67.0, 74.5, 78.0, 78.0, 81.0, 81.0, 82.0, 85.0, 85.0, 85.0, 87.5, 87.5,
87.5, 94.0, 99.0, 101.0, 101.0, 101.0, 115.0)
cens<-c(1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0,0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,0, 0, 0, 0, 0,0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0)
data <- data.frame(time,cens)
data
my.surv<-Surv(time,cens)
km<- survfit(my.surv ~ 1, conf.int = 0.95, conf.type = "plain", data)
summary(km)
t.u <- summary(km, time[cens==0])$time
surv.u <- summary(km, time[cens==0])$surv

# FIGURE 1.2
##-----------------------------------
par (fg="black",lwd=2)
par(cex.lab=1.1)
par(cex.axis=1)
par(cex.main=1.1)
plot(km,pch=2,col="blue",conf.int=F,lwd=2,csi=10,
xlab="Durées de fonctionnement",font.axis =2,las=1,font.lab =2,
col.main = "black", ylab="Probabilité de survie") ;
points(t.u, surv.u, col='red', pch=3,lwd=1)
```

Bibliographie

[1] D. Bosq, J. Lecoutre. Théorie de l'estimation fonctionnelle. Economica, Paris (1987).

[2] G. Colletaz. Modèles de survie. Notes de cours. Master 2, ESA (Novembre 2012). http ://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Survie_Sas.pdf

[3] J. J. Droesbeke, B. Fichet, P. Tassi. Analyse statistique des durées de vie, Modélisation des données censurées. Economica, Paris (1989).

[4] T. Duchesne. Analyse des durées de vie. New York (2004).

[5] B. Efron. Censored data and the bootstrap. Technical report. Public health service grant 5 R01 GM21215-05 Stanford, California (February 1980).

[6] B. Efron. Censored data and the bootstrap. Journal of the American Statistical Association, 76, 312(319).(1981).

[7] B. Efron. Better bootstrap confidence intervals (with discussion). Journal of the American Statistical Association, 82, 171(200).(1987).

[8] B .Efron, R. J. Tibshirani. An Introduction to the Bootstrap. Chapitres 10-11, Chapman and Hall (1993).

[9] J. D. Fermanian. Modèles de durées. Cours ENSAE. http ://www.crest.fr/ckfinder/userfiles/files/Pageperso/fermania/JDF-duree3.pdf

[10] T. R. Fleming, D. P. Harrington. Counting processes and survival analysis. Wiley Series in Probability and Statistics. John Wiley and Sons. (1991).

[11] A. Guilloux. Bootstrap. (2007-2008). http ://www.lsta.upmc.fr/guilloux.php

[12] C. Huber. Une methode de rééchantillonnage : le bootstrap. (Septembre 2006). www.biomedicale.univ-paris5.fr/survie/enseign/cours_Bootstrap_C_Huber_web.pdf

[13] A.D. Hutson. Analytical Bootstrap Methods for Censored Data. Journal of applied mathematics and decision sciences 6(2), 129-141. USA (2002). http ://www.kurims.kyoto-u.ac.jp/EMIS/journals/HOA/JAMDS/Volume6_2/141.pdf

[14] S. Kankoé. Méthode actuarielle d'estimation des courbes de survie : principe, différences avec la méthode de Kaplan-Meier.

[15] E.L. Kaplan, P. Meier. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457(481). (1958).

[16] J. Kim.Two-sample Inference for Quantiles Based on the Bootstrap for Censored Survival Data, Journal of the Korean Statistical Society, 22, 159-169. (1993).

[17] S. Lemler. Modèles de durée, analyse de survie. ENSIIE (2012-2013).
http ://stat.genopole.cnrs.fr/_media/members/slemler/courssurvie.pdf

[18] J. C. Massé. Statistique computationnelle. Notes de cours (2009).
http ://www2.mat.ulaval.ca/fileadmin/Cours/STT-7320/Cours.pdf

[19] R. Miller, G. Gong. Survival Analysis. Stanford, California (1980).
http ://statistics.stanford.edu/c̃kirby/techreports/BIO/BIO%2058.pdf

[20] F. Planchet. Statistique des modèles non paramétriques. (2012-2013).
http ://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430AD6748CE3AFFC1256F130067B88E/FILE/Seance3.pdf?OpenElement

[21] P. Saint Pierre. Introduction à l'analyse des durées de survie. (Octobre 2011).
http ://www.lsta.upmc.fr/psp/Cours_Survie_1.pdf

[22] P. Saint Pierre. Processus de comptage et analyse de survie. (Avril 2012).
http ://www.lsta.upmc.fr/psp/Cours_Survie_2.pdf

[23] F. Utzet. Á. Sánchez. Some Applications of the Bootstrap to Survival Analysis. Anuario de Psicologia , no 55, 155-167 (1992).

[24] L. Samartzis. Survival and censored data. (2005-2006).
http ://infoscience.epfl.ch/record/112202/files/lafteris.pdf

[25] S. Sawyer. The Greenwood and Confidence Intervals in Survival Analysis. (2003).
http ://www.math.wustl.edu/s̃awyer/handouts/greenwood.pdf

[26] M. Tableman, J. S. Kim. Survival Analysis Using S, Analysis of Time-to-Event Data.Chapman and Hall (2004).
http ://www.amazon.com/Survival-Analysis-Using-Time-Event/dp/1584884088