UNIVERSITÉ ASSANE SECK DE ZIGUINCHOR

UFR SCIENCES ET TECHNOLOGIES
DÉPARTEMENT DE MATHÉMATIQUES

MÉMOIRE DE MASTER

Domaine :	Sciences et Technologies
Mention :	Mathématiques et Applications
Spécialité :	Mathématiques Appliquées
Option :	statistique

Présenté par :

FATOU DIENG

TITRE

Modèle Gamma bidimensionnel dans l'estimation paramétrique des variables de durée

Sous la direction de : `Dr Emmanuel Nicolas CABRAL`
Sous la supervision de : `Pr Alassane DIEDHIOU`

Soutenu publiquement le 30 avril devant le jury composé de :

M. Clément MANGA	Professeur Assimilé	Président	UASZ
M. Alassane DIEDHIOU	Professeur Titulaire	Superviseur	UASZ
M. Diène NGOM	Professeur Assimilé	Examinateur	UASZ
M. Emmanuel Nicolas CABRAL	Maître de Conférences Titulaire	Directeur	UASZ

Remerciements

Tout d'abord, je remercie Allah le tout puissant pour toute la volonté et le courage qu'Il m'a donné pour l'achèvement de ce travail.

Je voudrais exprimer toute ma reconnaissance à mon directeur de mémoire, Mr Emmanuel Nicolas CABRAL, avec qui j'ai eu un énorme plaisir à travailler, en plus de ses conseils et de ses suggestions, j'ai toujours bénéficié de ses encouragements et de sa disponibilité. Qu'il trouve ici, le témoignage de ma parfaite et profonde gratitude.

Je tiens à remercier vivement le professeur Clément MANGA, qui m'a fait l'honneur de présider le jury de ce mémoire.
Je remercie sincèrement le professeur Alassane DIEDHIOU d'avoir accepté de superviser ce travail. Je remercie également le professeur Diène NGOM qui a bien accepté d'examiner ce travail.

Mes remerciements s'adressent aussi à toutes les personnes qui ont contribué, de près ou de loin, à l'élaboration de ce travail notamment mes amis, mes collègues de travail.
Mes plus vifs remerciements à mon Papa et ma Maman, mes frères et sœurs, mes cousins et cousines pour le soutien qu'ils m'ont apporté durant toutes mes années d'études, ainsi que toute ma famille.

Mes remerciements s'adressent spécialement aux docteurs Marcel Sihintoé BADIANE, Souhaibou SAMBOU et Abdoulaye DIOUF pour leur aide et encouragement. Merci pour le temps que vous m'avez accordé.

J'exprime ma reconnaissance envers l'Université Assane SECK de Ziguinchor, l'UFR des Sciences et Technologies, le Département de Mathématiques.

Je remercie également mes amis plus que frères Sidy SALL, Makhfou DIOP, Moussa DIATTA, Babacar SY, Cheikh SECK, Alioune BA...

Enfin, je souhaiterais remercier mes camarades de promotion, plus particulièrement à Diéynaba SAMB, Awa BARRY, Arame GUEYE, Souadou DIALLO, Marie FAYE...

Dédicaces

Je dédie ce travail
A mes parents.
A mes frères.
A mes soeurs.
A mon grand père Cheikh Nah Mamadou SECK.
A ma tante.
A ma cousine Seynabou SECK.
A ma famille.
A toute la famille de Cheikh Déthialaw Seck.
A mes amis.
A toutes mes voisines de la résidence DIATTA, plus particulièrement à Fatou NDIAYE, Antoinette TINE, Khadidiatou DRAMÉ, Marème THIAM, Mariama BA pour tous les bons moments partagés.
A Mame Bousso THIAM, Mathurin SECK, Fatou DIÉYE, Mariama CISSOKHO, Diatou DIEDHIOU, Souhadou DIONE...
A l'amicale des étudiants ressortissants du lycée de Guéoul à Ziguinchor.
A tous ceux qui nous ont aidés, soutenus.

Résumé

Dans ce travail, nous nous sommes intéressés essentiellement à un modèle de durée de survie : le modèle Gamma à deux paramètres β et λ.
Nous abordons quelques outils nécessaires à l'étude de l'analyse de survie tels que la fonction densité, la fonction de répartition, la fonction de survie, les fonctions de hasard et de hasard cumulé ainsi que les différents types de données censurées.
Nous nous baserons sur l'estimation paramétrique pour pouvoir estimer les paramètres β et λ dans le cas des données censurées et non censurées pour les durées de survie.
Nous verrons aussi la fonction de la log-vraisemblance dans les deux cas.
Nous procéderons à l'estimation paramétrique de β et λ par deux approches : la méthode du maximum de vraisemblance et la méthode des moments et nous donnerons aussi leurs intervalles de confiance.
Enfin nous vérifierons, grâce à des simulations avec le logiciel R , l'efficacité de ces deux méthodes et les propriétés d'estimateurs.

Table des matières

Introduction générale

Le terme durée de survie est employé de manière générale pour désigner le temps qui s'écoule jusqu'à l'arrivée d'un évènement particulier. Autrement dit, il représente le temps écoulé entre le début d'une observation et l'arrivée d'un évènement qui n'est pas forcément la mort, mais peut être la guérison, l'apparition d'une maladie ou de complications. Dans l'industrie, il peut s'agir d'un bris d'une machine. En économie, on étudie la durée passée au chômage, dans un emploi ou entre deux emplois (c'est-à-dire le temps écoulé pour qu'une personne trouve un travail). Ce temps est connu sous le nom de temps de survie.
L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet évènement (voir [18]).
Ainsi les variables aléatoires de durées de vie trouvent des applications dans pratiquement tous les domaines statistiques, la médecine, l'économie, la fiabilité, l'assurance...
L'analyse des données de survie possède deux particularités intrinsèques : d'une part , celle-ci concerne les variables de durées positives et d'autre part, la présence de données censurées. La variable représentative est notée T.
Les données de survie sont souvent modélisées par des lois exponentielles ou des lois dérivées de la loi exponentielle telles que la loi Gamma, la loi de Weibull...
Nous nous intéressons particulièrement au cas du modèle Gamma bidimensionnel dans l'estimation paramétrique des variables de durée.
Ce travail est organisé comme suit :
Nous commençons d'abord par rappeler dans le premier chapitre quelques notions de probabilités et de statistiques.
Dans le deuxième chapitre, nous présenterons les fonctions de distribution de la durée de survie et leurs relations.
Dans le troisième chapitre, nous présentons deux méthodes d'estimation paramétrique appliquées à la loi Gamma bidimensionnelle.
Dans le quatrième chapitre, nous présentons les résultats numériques suivi de discussion.

Chapitre 1

Rappels d'outils probabilistes et statistiques

Nous allons commencer par énoncer quelques notions de probabilités et de statistiques utiles pour la suite du travail.

1.1 Outils probabilistes

La théorie des probabilités est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude. Le calcul des probabilités a commencé avec Blaise Pascal, Pierre Fermat, Christian Huygens et Jacques Bernoulli par l'analyse des jeux dits de hasard.
Nous allons présenter ici quelque définitions utiles pour notre travail.

Définition 1.
Une famille $\mathcal{E}$ de parties d'un ensemble Ω (appelé univers ou ensemble fondamental) est appelé ***tribu*** *ou* ***σ- algèbre*** *si elle vérifie les propriétés suivantes :*

- $\Omega \in \mathcal{E}$;
- *Si $(A_n)_n$ est une suite (éventuellement finie) d'éléments de $\mathcal{E}$, alors $\cup_{n\in\mathbb{N}} A_n \in \mathcal{E}$;*
- *Si A est un élément de $\mathcal{E}$, alors $\overline{A} \in \mathcal{E}$.*

Le couple $(\Omega; \mathcal{E})$ est un **espace probabilisable** et les éléments de $\mathcal{E}$ sont appelés **évènements**.

Exemple 1. *On se donne un ensemble E non vide. Les trois familles de sous-ensembles de E suivantes sont des tribus de E :*

- $\mathcal{M} = \mathcal{P}(E)$*, la famille de tous les sous-ensembles de E est appelée* ***tribu triviale****.*
- $\mathcal{M} = \{\emptyset, E\}$ *est la* ***tribu grossière****, c'est la plus petite tribu sur E.*
- *Si on fixe A un sous-ensemble de E alors $\mathcal{M} = \{\emptyset, E, A, \overline{A}\}$ est la* ***tribu engendrée par*** *A.*

Définition 2.
On appelle ***tribu borélienne*** *sur $\mathbb{R}$, notée $\mathcal{B}(\mathbb{R})$, la plus petite tribu, au sens de l'inclusion, contenant tous les intervalles de $\mathbb{R}$.*

On peut donc donner maintenant la définition d'un espace probabilisé :

Définition 3.
On appelle ***probabilité*** *sur $(\Omega; \mathcal{E})$ (ou mesure de probabilité) une application $\mathbb{P}$ de $\mathcal{E}$ dans $[0, 1]$ telle que :*

- $\mathbb{P}(\emptyset) = 0$ et $\mathbb{P}(\Omega) = 1$.
- *Pour toute suite $(A_n)_{n\geqslant 0}$ d'évènements, deux à deux incompatibles, :*

$$\mathbb{P}(\cup_{n=0}^{+\infty} A_n) = \sum_{n=0}^{+\infty} \mathbb{P}(A_n) \qquad \text{appelée la } \sigma\text{-additivité.}$$

On appelle ***espace probabilisé****, le triplet $(\Omega, \mathcal{E}, \mathbb{P})$.*

Définition 4.
Soient l'espace probabilisé $(\Omega, \mathcal{E}, \mathbb{P})$ et un évènement B tel que $\mathbb{P}(B) > 0$. On appelle ***probabilité conditionnelle*** *de A sachant B définie sur la tribu conditionnelle : $\mathcal{E}|B = \{A \cap B, A \in \mathcal{E}\}$ le rapport*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

La formule de définition de la probabilité conditionnelle peut aussi s'écrire, si $\mathbb{P}(A) > 0$ par : $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A/B) = \mathbb{P}(A)\mathbb{P}(B/A)$. Cette formule s'appelle parfois formule des probabilités composées.

Définition 5.
Soit $(\Omega, \mathcal{E}, \mathbb{P})$ un espace probabilisé. Toute application $X : \Omega \to \mathbb{R}$, $(\mathcal{E}, \mathcal{B}(\mathbb{R}))$-mesurable est appelée ***variable aléatoire*** *réelle.*

Suivant la nature de $X(\Omega)$, l'ensemble des valeurs prises par X, on peut distinguer deux types de variables aléatoires :

- Variable aléatoire discrète : $X(\Omega)$ est ensemble fini ou dénombrable.

- Variable aléatoire continue : c'est une variable qui peut prendre, du moins théoriquement, toute valeur de $\mathbb{R}$ ou d'un intervalle de $\mathbb{R}$.

Pour une variable aléatoire réelle, la fonction de répartition est définie par :

$$F(x) = \mathbb{P}(X \leq x), x \in \mathbb{R}.$$

Cette fonction de répartition vérifie certaines propriétés :

- $0 \leq F(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to +\infty} F(x) = 1$
- $F(x)$ est croissante et continue à droite
- Pour tout $x \in \mathbb{R}$, $\mathbb{P}(X = x) = F(x) - F(x^-)$
- L'ensemble des points de discontinuité de F est au plus dénombrable
- Si l'ensemble des points de discontinuité est vide, la variable aléatoire est continue. Autrement dit sa fonction de répartition est continue.
- Pour $a < b$: $\mathbb{P}(a < X < b) = F(b) - F(a)$.

Définition 6 (Loi absolument continue)**.**
La valeur moyenne de la probabilité d'un intervalle de longueur $h > 0$ est :

$$\frac{1}{h}\mathbb{P}(x \leq X < x + h) = \frac{F(x+h) - F(x)}{h}.$$

Si on fait tendre cette longueur vers 0, la limite, si elle existe, représentera la probabilité d'un intervalle de longueur infiniment petite Δx.
Ce sera le cas si F admet une dérivée f :

$$\lim_{h \to 0} \frac{1}{h}[F(x+h) - F(x)] = F'(x) = f(x).$$

Moment et variance d'une variable aléatoire

- Cas discret :
  Soit X une variable aléatoire réelle à valeurs dans $\{x_i, i \in I\}$, I dénombrable. On dira que X admet une espérance mathématique si la série $\sum_{i \in I} |x_i| \mathbb{P}(X = x_i)$ est convergente.
  Dans ce cas on appelle **espérance mathématique** de X le nombre réel noté $E(X)$ défini par :

$$E(X) = \sum_{i \in I} x_i \mathbb{P}(X = x_i).$$

On appelle **moment d'ordre** k de X, l'espérance mathématique de X^k(si elle existe) définie par

$$E(X^k) = \sum_{i \in I} x_i^k \mathbb{P}(X = x_i).$$

On appelle **moment centré** d'ordre $k, k \in \mathbb{N}^*$ de X, l'espérance mathématique de $(X - E(X))^k$ si elle existe.
On appelle **Variance** de X et on note $V(X)$ où $Var(X)$, le moment centré d'ordre 2 de X si elle existe, définie par :

$$V(X) = E[X - E(X)]^2.$$

Si X admet un moment d'ordre 2, alors

$$V(X) = E(X^2) - [E(X)]^2 \qquad \text{(formule de Koenig-Huyghens).}$$

- Cas de variable aléatoire à densité :
  Soit X une variable aléatoire réelle définie sur l'espace probabilisé $(\Omega, \mathcal{E}, \mathbb{P})$ admettant une densité f.
  On dit que X admet une espérance mathématique si l'intégrale généralisée $\int_{\mathbb{R}} xf(x)dx$ est absolument convergente. Ainsi l'espérance mathématique est définie par :

$$E(X) = \int_{\mathbb{R}} xf(x)dx.$$

  Comme dans le cas discret on a le moment centré d'ordre k suivant :

$$E(X^k) = \int_{\mathbb{R}} x^k f(x)dx.$$

  Ainsi la variance est donnée par cette formule :

$$V(X) = \int_{\mathbb{R}} [x - E(X)]^2 f(x)dx = E(X^2) - [E(X)]^2.$$

Proposition 1. *(Inégalité de Markov).*
Soit X une variable aléatoire positive. Alors,

$$\forall a > 0, \quad \mathbb{P}(X \geqslant a) \leqslant \frac{E(X)}{a}$$

Définition 7.
Nous appelons ***fonction gamma*** *la fonction définie par :*

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} \exp(-t)dt, \forall \alpha \in \mathbb{R}_+^* \tag{1.1}$$

La fonction gamma satisfait aux relations suivantes :

- $\forall n \in \mathbb{N}, \Gamma(n+1) = n!$
- $\forall \alpha \in \mathbb{R}_+^*, \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
- En particulier $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Elle est une généralisation directe de la notion de factorielle pour des nombres réels non entiers. On appelle fonction gamma tronquée (ou incomplète) la fonction définie par :

$$\gamma(a,x) = \frac{1}{\Gamma(a)} \int_0^x u^{a-1} \exp(-u)du$$

et on note

$$\Gamma(a,x) = \frac{1}{\Gamma(a)} \int_x^{+\infty} u^{a-1} \exp(-u)du.$$

1.2 Outils statistiques

Le terme statistique est apparu récemment, vers le milieu du $XVII^e$ siècle ; il vient du latin "statisticus" , relatif à l'état ("status"). Il s'agit d'un moyen scientifique d'analyse et de compréhension du phénomène étudié, s'appliquant très largement à l'économie et à toutes les sciences sociales et de la nature.
Dans cette section, nous allons donner la définition de quelques termes du vocabulaire utilisé.

Définition 8.
On appelle ***échantillon*** *de taille n d'une loi de probabilité $\mathbb{P}$, une suite $(X_1, ..., X_n)$ de variables aléatoires indépendantes et de même loi de probabilité $\mathbb{P}$.*

Soient $X_1, ..., X_n$ n variables aléatoires de même loi que X.
Le moment empirique d'ordre k, $k \in \mathbb{N}^*$ de l'échantillon notée $\mu_k(X)$ est définie par :

$$\mu_k(X) = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

$\mu_1(X)$, notée $\overline{X}_n$ est la moyenne empirique.
La variance empirique de l'échantillon notée S_n^2 est définie par :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

La variance empirique corrigée de l'échantillon $S_n'^2$ est définie par :

$$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}S_n^2.$$

Définition 9.
La fonction quantile d'ordre ω liée à la fonction de répartition F est la fonction inverse généralisée (ou pseudo-inverse) de F notée F^{-1} définie par :

$$q_\omega = \inf_{t\in\mathbb{R}}\{t : F(t) > \omega\}, 0 < \omega < 1.$$

Si F est continue et strictement croissante alors :

$$q_\omega = F^{-1}(\omega).$$

La **médiane** d'une variable aléatoire X est le quantile d'ordre $\frac{1}{2}$. Elle est notée $Me(X)$ et définie par :

$$Me(X) = q_{0.5}.$$

La **vraisemblance** (**likelihood** en anglais) de l'échantillon $(X_1, ..., X_n)$ est la loi de probabilité de ce n-uplet, notée $L(x_1, ..., x_n; \theta)$ et définie par :

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n}\mathbb{P}(X_i = x_i|\theta)$$

si X est une variable aléatoire discrète, et par :

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

si X est une variable aléatoire continue de densité $f(x_i; \theta)$.
Sous certaines conditions, la **quantité d'information de Fisher** sur θ fournie par l'échantillon est le réel positif noté $I_n(\theta)$, défini par :

$$I_n(\theta) = E_\theta\left[\left(\frac{\partial \ln L(X_1,...,X_n;\theta)}{\partial\theta}\right)^2\right].$$

On appelle **score** de l'échantillon, noté par $S_n(\theta)$, la dérivée de la log-vraisemblance définie par :

$$S_n(\theta) = \frac{1}{L(X_1, ..., X_n; \theta)} \frac{\partial}{\partial \theta} L(X_1, ..., X_n; \theta).$$

En posant $\ell(x, \theta) = \ln L(x, \theta)$, on a :

$$S_n(\theta) = \frac{\partial}{\partial \theta} \ell(X_1, ..., X_n; \theta).$$

Propriétés de l'information de Fisher (cf [14] pour les preuves)

- Si le domaine de définition de $f(x, \theta)$ est indépendant de θ alors on a :

$$I_n(\theta) = E\left(-\frac{\partial^2}{\partial \theta^2} \ln L(X_1, ..., X_n; \theta)\right).$$

- Si le domaine de définition de $f(x, \theta)$ est indépendant de θ, chaque observation apporte la même information,

$$I_n(\theta) = nI_1(\theta).$$

Ainsi pour le cas p-dimensionnel (c'est-à-dire $\theta = (\theta_1, ..., \theta_p)$), le score est un vecteur défini par :

$$S_n(\theta) = \left(\frac{\partial}{\partial \theta} l(X_1, ..., X_n; \theta_1), ..., \frac{\partial}{\partial \theta} l(X_1, ..., X_N; \theta_p)\right).$$

Il est centré et sa matrice de variance covariance qu'on appelle **matrice d'information de Fisher** est définie par

$$I_n(\theta) = (I^n_{i,j})_{1 \leq i,j \leq p}.$$

Cette matrice est définie positive et de terme générale :

$$I^n_{i,j} = cov\left(\frac{\partial}{\partial \theta_i} l(X_1, ..., X_n; \theta), ..., \frac{\partial}{\partial \theta_j} l(X_1, ..., X_n; \theta)\right).$$

Définition 10.
Soit X une variable aléatoire de loi dépendant d'un paramètre θ et $X_1, ..., X_n$ un n- échantillon extrait de X. On appelle estimateur de θ toute statistique T_n fonction de $(X_1, ..., X_n)$, c'est-à-dire $T_n = \varphi(X_1, ..., X_n)$. Sa valeur est notée par :

$$\widehat{\theta} = T_n(x_1, ..., x_n).$$

Le **biais** d'un estimateur est l'écart entre sa moyenne et la vraie valeur du paramètre. Il est noté souvent par $b(T_n, \theta)$ *ou par* $b_n(\theta)$.

$$b(T_n, \theta) = E(T_n) - \theta.$$

Un estimateur T_n de θ est dit **sans biais** si pour tout $\theta \in \Theta$ (espace des paramètres) et tout entier n :

$$b(T_n, \theta) = 0.$$

Un estimateur T_n *de* θ est dit **asymptotiquement sans biais** si pour tout θ :

$$\lim_{n\to\infty} b(T_n, \theta) = 0 \Leftrightarrow \lim_{n\to\infty} E(T_n) = \theta.$$

Définition 11.
Un estimateur T_n, de θ est ***consistant*** *(ou convergent) si :*

$$\forall \epsilon > 0, \lim_{n\to+\infty} \mathbb{P}(|T_n - \theta| > \epsilon) = 0.$$

Tout estimateur sans biais ou asymptotiquement sans biais dont la variance tend vers 0 est convergent en moyenne quadratique donc consistant.
Un estimateur T_n, de θ dont l'espérance mathématique tend vers θ et dont la variance tend vers 0 lorsque $n \to +\infty$, est un estimateur consistant :

$$E(T_n) = \theta \;\; et \;\; V(T_n) \to 0 \;\; quand \;\; n \to +\infty$$

La moyenne empirique $\bar{X}_n$ est un estimateur sans biais et convergent de l'espérance de X. En effet pour le cas de la loi gamma à deux paramètres β et h, on a :

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= E(X) = \frac{\beta}{h} \end{aligned}$$

Donc $\bar{X}_n$ est un estimateur sans biais de $E(X)$.
La variance de $\bar{X}_n$ est :

$$\begin{aligned} V(\bar{X}_n) &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n}V(X) \\ &= \frac{\beta}{h^2 n} \end{aligned}$$

$Var(\bar{X}_n) \to 0$ *quand* $n \to +\infty$. Par conséquent : la moyenne empirique $\bar{X}_n$ est un estimateur sans biais et convergent en moyenne quadratique de $E(X)$. De la même manière on montre que la variance empirique n'est pas un estimateur sans biais de la variance de X.
En effet on a :

$$\begin{aligned} E(S_n^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}_n^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\bar{X}_n^2) = E(X^2) - E(\bar{X}_n^2) \\ &= Var(X) + (E(X))^2 - Var(\bar{X}_n) - E((\bar{X}_n))^2 \\ &= \frac{n-1}{n}Var(X). \end{aligned}$$

Ainsi la variance empirique n'est pas un estimateur sans biais mais il est asymptotiquement sans biais. Il est aussi un estimateur convergent.
S_n^2 permet de déterminer un estimateur sans biais de la variance en posant $S_n'^2 = \frac{n}{n-1}S_n^2$.
En effet on a :

$$\begin{aligned} E(S_n'^2) &= \frac{n}{n-1}E(S_n^2) \\ &= \frac{n}{n-1} \times \frac{n-1}{n}Var(X) = Var(X) \end{aligned}$$

L'**erreur quadratique moyenne** notée par (EQM) d'un estimateur $\hat{\theta}_n$ de θ est la quantité :

$$EQM(T_n) = E\left[(T_n - \theta)^2\right].$$

On l'appelle aussi risque quadratique. Ainsi on distingue deux cas :

- Cas d'un estimateur sans biais

$$\begin{aligned} EQM(T_n) = EQM(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E(\hat{\theta}^2 - 2(\hat{\theta}\theta) + \theta^2) \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta}\theta) + E(\theta^2) \\ &= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + E(\theta^2) = E(\hat{\theta}^2) - \theta^2 \\ &= Var(\hat{\theta}) \end{aligned}$$

- Cas d'un estimateur quelconque
  Rappelons d'abord que le biais $(b(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta)$ et $E(\hat{\theta})$ sont des constantes , ce qui permet d'utiliser la linéarité de l'espérance :

$$E(c_1X + c_2) = c_1E(X) + c_2$$

$$
\begin{aligned}
EQM(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right] &= E\left[\left(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right)^2\right] \\
&= E\left[\left(\hat{\theta}-E(\hat{\theta})+b(\hat{\theta},\theta)\right)^2\right] \\
&= E\left[\left(\hat{\theta}-E(\hat{\theta})\right)^2+2\left(\hat{\theta}-E(\hat{\theta})\right)b(\hat{\theta},\theta)+b(\hat{\theta},\theta)^2\right] \\
&= E\left[\left(\hat{\theta}-E(\hat{\theta})\right)^2\right]+2E\left(\hat{\theta}-E(\hat{\theta})\right)b(\hat{\theta},\theta)+b(\hat{\theta},\theta)^2 \\
&= Var(\hat{\theta})+2\left(E(\hat{\theta})-E(\hat{\theta})\right)b(\hat{\theta},\theta)+b(\hat{\theta},\theta)^2 \\
&= Var(\hat{\theta})+b(\hat{\theta},\theta)^2
\end{aligned}
$$

Remarque. Si $T_n^{(1)}$ *et* $T_n^{(2)}$ sont deux estimateurs de θ, on dit que $T_n^{(1)}$ est meilleur que $T_n^{(2)}$ si $EQM(T_n^{(1)}) \leq EQM(T_n^{(2)})$.

Conséquence 1.
Si deux estimateurs sont sans biais le meilleur est celui qui est de variance minimale.

Si $\hat{\theta}_n$ est un estimateur sans biais de θ, $\hat{\theta}_n$ est dit **efficace** si $Var(\hat{\theta}_n) = \frac{1}{I_n}(\theta)$

Remarque. Si $\hat{\theta}_n$ est un estimateur biaisé de θ alors $\hat{\theta}_n$ est dit **efficace** si $Var(\hat{\theta}_n) = \frac{g'(\theta)}{I_n}(\theta)$, où $g'(\theta) = E(\hat{\theta}_n)$.

L'efficacité relative de deux estimateurs, $\hat{\theta}_n$ et $\tilde{\theta}_n$, est donnée par :

$$eff(\hat{\theta}_n, \tilde{\theta}_n) = \frac{EQM(\tilde{\theta}_n)}{EQM(\hat{\theta}_n)}.$$

Définition 12.
Soit $(X_1, ..., X_n)$ un échantillon extrait d'une variable aléatoire dont la loi dépend d'un paramètre θ inconnu et $\alpha \in]0,1[$ fixé.
On appelle ***intervalle de confiance*** *de paramètre θ, de niveau de confiance $1-\alpha$, tout intervalle de la forme $[a_n, b_n]$, avec a_n, b_n deux statistiques dépendant de $(X_1, ..., X_n)$ tel que :*

$$\mathbb{P}(a_n \leq \theta \leq b_n) = 1 - \alpha.$$

Chapitre 2

Les distributions de la durée de survie (variables de durée)

2.1 Notion de censure

Pour analyser les données de survie, on est souvent confronté à des données incomplètes à cause de deux phénomènes différents : la censure (voir [4], [8], [10], [12], [20]) et la troncature. Dans ce mémoire, on étudiera seulement le phénomène de la censure.
Les données censurées sont des observations pour lesquelles la valeur exacte n'est pas toujours connue. Il existe trois types de censure que l'on nomme : censure à droite, censure à gauche et censure par intervalle.
Formellement, soit X_i le temps de survie pour l'individu i, C_i son temps de censure et T_i la durée réelle observée.

1) **Censure à droite** :
La durée de vie est censurée à droite si l'individu n'a pas subi l'évènement à sa dernière observation. C'est-à-dire si $X_i \geqslant C_i$. La censure à droite est de trois types :

   a) Censure à droite de type I (Censure non aléatoire)
   Dans ce cas, les variables X_i avec $i = 1, ..., n$, ne sont pas toutes observées. On observe que les X_i qui sont inférieures ou égales à une durée de censure fixée C avec $C_i = C$, sinon on sait uniquement que $X_i > C$.
   On utilise la notation suivante : $T_i = X_i \wedge C = \min(X_i, C)$.
   Par exemple, on peut tester la durée de vie de n objets identiques (ampoules par exemple) sur un intervalle d'observation fixé $[0, u]$.
   En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps u sont sacrifiées).

b) Censure à droite de type II
Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables X_i et T_i. La date de censure est donc $X_{(k)}$ et on observe les variables suivantes

$$\begin{aligned} T_{(1)} &= X_{(1)} \\ &\vdots \\ T_{(k)} &= X_{(k)} \\ T_{(k+1)} &= X_{(k)} \\ &\vdots \\ T_{(n)} &= X_{(k)} \end{aligned}$$

c) Censure à droite de type III (Censure aléatoire)
Supposons que les C_i sont indépendantes et identiquement distribuées (i.i.d.). Dans ce cas, on observe que les variables $T_i = X_i \wedge C_i$. En présence de censure à droite de type III, l'information disponible pour chaque individu est $\{T_i, d_i\}$ avec

$$\begin{cases} T_i = X_i \wedge C_i = \min(X_i, C_i) \\ d_i = \begin{cases} 1 & Si \quad X_i \leq C_i \\ 0 & Sinon \end{cases} \end{cases}$$

Elle est la plus courante.
La censure aléatoire nous donne les informations qui peuvent être résumées comme suit :
∗ perte de vue ;
∗ arrêt de l'expérience à cause de faits indésirables inattendus ;
∗ fin d'étude.

2) **Censure à gauche** :
La durée de vie est censurée à gauche si la valeur exacte n'est pas connue avant C_i. Autrement dit, la durée de vie est censurée à gauche si l'individu a subi l'évènement avant qu'il soit observé. En présence de censure à gauche on a : $\begin{cases} T_i = X_i \vee C_i = \max(X_i, C_i) \\ d_i = \begin{cases} 1 & si \quad X_i \geq C_i \\ 0 & si \quad X_i \leq C_i \end{cases} \end{cases}$

3) **Censure par intervalle** : On parle de censure par intervalle si l'évènement n'est pas observé, mais il survient entre deux dates d'observation. La seule information disponible est qu'il a eu lieu entre deux dates connues.

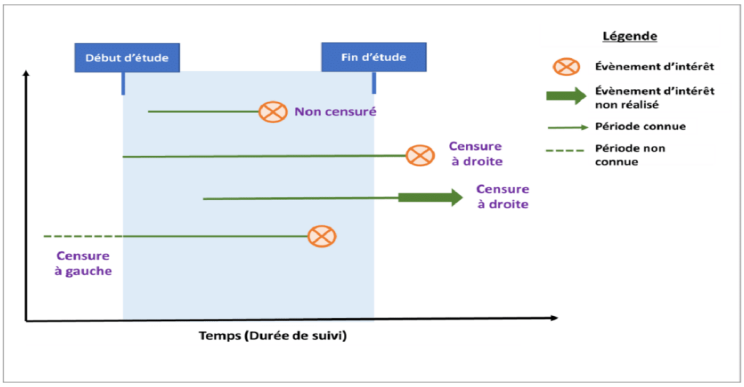

FIGURE 2.1 – Exemple de censure (à droite et à gauche)

2.2 Distribution de la durée de survie

Soit un espace probabilisé $(\Omega, \mathcal{E}, \mathbb{P})$, on définit une variable aléatoire de durée T continue, à valeurs réelles positives.
Il existe plusieurs fonctions qui caractérisent la loi de probabilité de la variable durée de survie T (cf.[6], [10], [12]). On peut en citer quelques-unes :

2.2.1 Fonction de répartition

Soit T une variable de durée de survie positive, absolument continue .
La fonction de répartition est donnée par :

$$F_T(t) = \mathbb{P}(T \leq t), t \in \mathbb{R}_+^* \tag{2.1}$$

Elle définit la probabilité que cette durée soit inférieure à une valeur donnée t et l'on a :

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

F est à valeurs dans $[0, 1]$, ce qui est évident car c'est la probabilité d'un évènement $(T \leq t)$.
La fonction de répartition est croissante.
En effet, soient t et t_1 tel que $t < t_1$, montrons que $F_T(t) \leq F_T(t_1)$. On a

$$\begin{aligned}]0, t] \cup]t, t_1] &=]0, t_1] \\ T^{-1}(]0, t_1]) &= T^{-1}(]0, t]) \cup T^{-1}(]t, t_1]) \end{aligned}$$

c'est-à-dire $(T \leq t_1) = (T \leq t) \cup (T \leq T \leq t_1)$.
Donc $\mathbb{P}(T \leq t_1) = \mathbb{P}(T \leq t) + \mathbb{P}(T \leq T \leq t_1)$.
D'où $F_T(t_1) = F(t) + \mathbb{P}(t \leq T \leq t_1) \geq F_T(t)$. Ainsi $F_T(t) \leq F_T(t_1)$.

Exemple 2. *Si T désigne la durée du chômage en mois, $F_T(t)$ représente la probabilité de rester au plus t mois au chômage à partir de la date d'inscription.*

Exemple 3. *Si T désigne la longueur d'un appel téléphonique en secondes, $F_T(t)$ représente la probabilité qu'un appel dure moins de t secondes.*

2.2.2 Densité de probabilité

La densité de la durée de probabilité de T est notée par f. Elle est définie sur $]0, +\infty[$ telle que pour tout $t \geq 0$

$$f_T(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t \leq T \leq t + \Delta t). \tag{2.2}$$

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

$$f_T(t) = \frac{dF_T(t)}{dt} \geq 0 \tag{2.3}$$

puisque f est la dérivée d'une fonction croissante.
$f_T(t)\Delta t$ s'interprète comme la probabilité que l'évènement survienne dans le petit intervalle de temps $[t, t + \Delta t]$.
Ainsi, la densité est intégrable sur $]0, +\infty[$ et d'intégrale égale à un :

$$\int_0^{+\infty} f_T(t)dt = 1$$

ceci provient du fait que $F(+\infty) = 1$.
La fonction densité est en rapport avec la fonction de répartition. On a :

$$F_T(t) = \int_0^t f_T(x)dx \tag{2.4}$$

En reprenant l'exemple 2, la quantité $f_T(t)\Delta t$ représente la probabilité que la durée du chômage soit comprise entre t et $t + \Delta t$, c'est-à-dire qu'elle soit à peu près égale à t mois.

2.2.3 Fonction de survie

La fonction de survie est définie comme :

$$S_T(t) = \mathbb{P}(T > t) = 1 - F_T(t) \tag{2.5}$$

$S_T(t)$ donne la probabilité que le temps de survie d'un d'individu dépasse t. C'est-à-dire la probabilité que l'individu soit toujours vivant après t unités de temps.
On a :

- $S_T(t)$ est positive et continue à droite, c'est-à-dire $S_T(t^+) = S_T(t)$, pour tout $t > 0$.

- $S_T(t)$ est décroissante et satisfait les conditions aux limites :

$$S_T(t) = \begin{cases} 1 & \text{si } t = 0 \\ 0 & \text{si } t = +\infty. \end{cases} \iff \begin{cases} \lim\limits_{t \to 0} S_T(t) = 1 \\ \lim\limits_{t \to +\infty} S_T(t) = 0 \end{cases}$$

Ainsi toute la population est en vie en $t = 0$ et plus personne n'est en vie en $t = +\infty$. En reprenant l'exemple 3, la quantité $S_T(t)$ représente la probabilité qu'un appel dure plus de t secondes.

Remarque.
Il est arbitraire de décider que $\mathbb{P}(T > t) = \mathbb{P}(T \geq t)$ ou $\mathbb{P}(T \leq t) = \mathbb{P}(T < t)$; lorsque la loi de T est continue.

NB : L'économétrie des durées utilise plus souvent la fonction de survie $S_T(t)$ au concept de la fonction de répartition.

2.2.4 Fonction de hasard ou fonction de risque

La fonction de hasard est aussi importante car :

- elle mesure l'intensité instantanée ;

- elle peut être utilisée pour identifier la forme spécifique d'un modèle (exponentielle, weibull...).

C'est la base des mathématiciens pour la modélisation des données de survie. La fonction de hasard, notée $h_T(t)$, est définie, pour tout t de $\mathbb{R}^+$ par :

$$\begin{aligned} h_T(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leqslant t + \Delta t \mid T > t) && (2.6) \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}((t < T \leqslant t + \Delta t) \bigcap (T > t))}{\Delta t \mathbb{P}(T > t)} && (2.7) \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leqslant t + \Delta t)}{\Delta t \mathbb{P}(T > t)} && (2.8) \\ &= \frac{1}{\mathbb{P}(T > t)} \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t) && (2.9) \\ &= \frac{f_T(t)}{S_T(t)}. && (2.10) \end{aligned}$$

C'est la probabilité la probabilité conditionnelle que le phénomène se termine après une durée t sachant que l'on a atteint cette durée. h_T peut désigner

par exemple le taux de sortie de chômage ou le taux de réemploi, le taux de perte d'emploi ou encore le taux de guérison...
Les caractéristiques de la fonction de hasard sont les :

- elle est toujours non négative (positive ou nulle) ;

- elle ne possède pas de limite supérieure.

Comme on a supposé que T est une variable aléatoire continue, le hasard s'exprime en fonction de la survie :

$$h_T(t) = \frac{f_T(t)}{S_T(t)} = -\frac{1}{S_T(t)}\frac{\partial}{\partial t}S_T(t) \tag{2.11}$$

$$= -\frac{\partial}{\partial t}\ln(S_T(t)). \tag{2.12}$$

Exemple 4. *Si T est la durée d'un handicap, la quantité $h_T(t)\Delta t$ représente la probabilité de guérison après t années.*

2.2.5 Fonction de hasard cumulée

C'est l'intégrale de la fonction de hasard h sur l'intervalle $[0, t]$, notée $H_T(t)$. Elle est définie par :

$$H_T(t) = \int_0^t h_T(x)dx = \int_0^t \frac{f_T(x)}{S_T(x)}dx = \int_0^t \frac{dS_T(x)}{S_T(x)}dx \tag{2.13}$$

$$= -\ln(S_T(t)). \tag{2.14}$$

Ainsi $H_T(t)$ vaut $+\infty$ quand $t \to 0$.

Remarque. A partir de la fonction de hasard cumulée, on peut déduire la fonction de survie grâce à la relation suivante :

$$H_T(t) = -\ln(S_T(t)) \Leftrightarrow S(t) = \exp(-H_T(t)). \tag{2.15}$$

D'où

$$S_T(t) = \exp(-\int_0^t h_T(x)dx). \tag{2.16}$$

La fonction de hasard et la fonction de hasard cumulée font partie des cinq fonctions qui caractérisent la loi de probabilité de la variable "durée de vie".

2.2.6 Relations entre les définitions

Nous pouvons définir la distribution de probabilité de la variable aléatoire de durée T à partir de sa fonction de répartition, sa densité, sa fonction de survie, sa fonction de hasard ou sa fonction de hasard cumulée.
Par manipulation des définitions précédentes, on peut en déduire les relations suivantes :

• Pour la fonction de répartition $F_T(t)$, on a :

$$\begin{aligned} F_T(t) &= P(T \leq t) && (2.17)\\ &= 1 - S_T(t) && (2.18)\\ &= \int_0^t f_T(t) && (2.19)\\ &= 1 + \exp(H_T(t)) && (2.20)\\ &= 1 + \exp(\int_0^t h_T(x)dx). && (2.21) \end{aligned}$$

• Pour la fonction de densité, on a :

$$\begin{aligned} f_T(t) &= \frac{dF_T(t)}{dt} && (2.22)\\ &= -\frac{dS_T(t)}{dt}. && (2.23) \end{aligned}$$

• Pour la fonction de survie, on a :

$$\begin{aligned} S_T(t) &= 1 - F_T(t) && (2.24)\\ &= 1 - \int_0^t f_T(x)dx && (2.25)\\ &= \exp(-H_T(t)) && (2.26)\\ &= \exp(-\int_0^t h_T(x)dx). && (2.27) \end{aligned}$$

• Pour la fonction de hasard, on a :

$$\begin{aligned} h_T(t) &= \frac{f_T(t)}{1 - F_T(t)} && (2.28)\\ &= -\frac{\partial}{\partial t}\ln(S_T(t)) && (2.29)\\ &= \frac{dH_T(t)}{dt}. && (2.30) \end{aligned}$$

• Pour la fonction de hasard cumulée, on a :

$$H_T(t) = \int_0^t h(x)dx \tag{2.31}$$
$$= -\ln(S_T(t)). \tag{2.32}$$

Ainsi nous pouvons calculer l'espérance, la variance, la médiane et le quantile de la durée de survie (voir [10]) :

- Pour l'espérance ou le temps moyen de survie, on a :

$$E(T) = \int_0^{+\infty} t f_T(t)dt.$$

Comme $f_T(t) = -\frac{dS_T(t)}{dt}$, on a :

$$E(T) = -\int_0^{+\infty} t\left(\frac{dS_T(t)}{dt}\right)dt$$
$$= -\lim_{u\to+\infty}\int_0^u t\left(\frac{dS_T(t)}{dt}\right)dt. \quad (*)$$

En faisant une intégration par parties de $\int_0^u t\left(\frac{dS_T(t)}{dt}\right)dt$ on obtient :

$$\int_0^u t\left(\frac{dS_T(t)}{dt}\right)dt = [tS_T(t)]_0^u - \int_0^u S_T(t)dt$$
$$= uS_T(u) - \int_0^u S_T(t)dt.$$

D'après l'inégalité de Markov on a :

$$tS_T(t) \leq E(T).$$

Donc le terme $uS_T(u)$ est borné. On en déduit que $\int_0^{+\infty} S_T(t)dt$ converge, ce qui implique $\lim_{t\to+\infty} tS_T(t) = 0$.
$(*) \Rightarrow E(T) = \int_0^{+\infty} S_T(t)dt$.

- Pour la variance de la durée de survie, on a :

$$V(T) = E(T^2) - (E(T))^2.$$

En faisant la même démarche que l'espérance, on obtient :

$$E(T^2) = 2\int_0^{+\infty} tS_T(t)dt.$$

Donc

$$\begin{aligned} V(T) &= 2\int_0^{+\infty} tS_T(t)dt - (E(T))^2 \\ &= 2\int_0^{+\infty} tS_T(t)dt - \left(\int_0^{+\infty} tf_T(t)dt\right)^2. \end{aligned}$$

- Pour le quantile, on a :

$$\begin{aligned} q_\omega &= \inf(t : F_T(t) \geqslant \omega) \\ &= \inf(t : S_T(t) \leq 1-\omega). \end{aligned}$$

Si de plus $F_T(t)$ est continue et strictement croissante, on a :

$$\begin{aligned} q_\omega &= F_T^{-1}(\omega) \\ &= S_T^{-1}(1-\omega). \end{aligned}$$

- Pour la médiane, on a :
On a :

$$M_e(T) = q_{\frac{1}{2}} = \inf\left(t : S_T(t) \leq \frac{1}{2}\right).$$

Si de plus $F_T(t)$ est continue et strictement croissante, on a :

$$M_e(T) = S_T^{-1}\left(\frac{1}{2}\right).$$

Donc il s'en suit que la durée de vie médiane pour une variable aléatoire continue T est la valeur $q_{0.5}$ de sorte que

$$S_T(q_{0.5}) = 0.5.$$

Conclusion

En conclusion, la fonction de densité, la fonction de répartition, la fonction de survie, la fonction de hasard et la fonction de hasard cumulée permettent de caractériser la loi de T. La fonction de hasard permet souvent d'interpréter un modèle pour des données de survie.
On voit que l'espérance et la variance de la durée de survie peuvent être calculées à partir de ces cinq fonctions.

Chapitre 3

Quelques modèles usuels de durée de survie

Les distributions les plus couramment utilisées sont : la loi exponentielle, la loi de Weibull, la loi gamma, la loi de Gompertz, la loi log-logistique, la loi de Pareto (voir [2], [3], [6], [10], [20]).

3.1 Distribution exponentielle

C'est la la distribution la plus importante et la plus simple dans les études de survie.

Définition 13.
Une variable aléatoire T suit une loi exponentielle de paramètre $\lambda > 0$ si c'est une variable aléatoire positive presque sûrement dont la densité est donnée par :

$$f_T(t) = \begin{cases} \lambda \exp(-\lambda t) & si \quad t \geqslant 0 \\ 0 & sinon. \end{cases} \tag{3.1}$$

Si $\lambda = 1$, la distribution est appelée exponentielle standard. Ainsi, on a les distributions de la durée de vie T suivantes :

$$F_T(t) = 1 - \exp(-\lambda t) \tag{3.2}$$
$$S_T(t) = \exp(-\lambda t) \tag{3.3}$$
$$h_T(t) = \lambda \tag{3.4}$$
$$H_T(t) = \lambda t. \tag{3.5}$$

On en déduit que la loi exponentielle est caractérisée par une fonction de hasard (taux de mortalité) constante λ qui est strictement positive.

Remarque. Un appareil dont la durée de vie est exponentielle garde les mêmes propriétés probabilistes quel que soit son âge : il ne vieillit pas.

Proposition 2.
Soit T est une variable aléatoire positive représentant une durée de vie. On dit que T suit une loi exponentielle si et seulement si elle vérifie la propriété de non-vieillissement

$$\mathbb{P}(T > t + h | T > t) = \mathbb{P}(T > h) \tag{3.6}$$

pour tout $t > 0$ et $h > 0$.

Pour démontrer la proposition 2, nous aurons besoin du lemme suivant :

Lemme 1.
Soit $S_T : \mathbb{R}_+ \to \mathbb{R}_+$ une fonction multiplicative, décroissante et vérifiant $\lim_{t\to 0} S_T(t) = 1$ et $\lim_{t\to+\infty} S_T(t) = 0$. Alors il existe $\lambda > 0$ tel que

$$S_T(t) = \exp(-\lambda t) \quad \textit{pour tout} \quad t \geq 0.$$

Preuve. Soit n un entier positif. On a : $S(n) = S(1+...+1) = S(1)^n$ d'après la propriété de multiplicativité.
Soit maintenant n un entier strictement positif, alors on a :

$$S(1) = S\left(\tfrac{1}{n} + ... + \tfrac{1}{n}\right) = S\left(\tfrac{1}{n}\right)^n.$$

Donc $S\left(\frac{1}{n}\right) = S(1)^{\frac{1}{n}}$. On en déduit que pour tout nombre rationnel positif $r = \frac{p}{q}$, $p \in \mathbb{N}$ et $q \in \mathbb{N}^*$, on a : $S(r) = S\left(\frac{p}{q}\right) = S(1)^{\frac{p}{q}}$.
Soit t un réel positif quelconque. Il existe une suite croissante $(r_n)_{n\geq 0}$ et une suite décroissante $(t_n)_{n\geq 0}$ de rationnels convergeant vers t telles que $r_n \leq t \leq t_n$ pour tout $n \in \mathbb{N}$. On alors pour tout n, $S(r_n) \leq S(t) \leq S(t_n)$. Comme r_n et t_n sont rationnels, on déduit que

$$S(1)^{r_n} \leq S(t) \leq S(1)^{t_n}. \qquad (*).$$

Finalement comme S est monotone, par passage à la limite dans $(*)$ on a : $S(t) = S(1)^t$ pour tout réel positif.
Comme $\lim_{t\to 0} S_T(t) = 1$ et $\lim_{t\to+\infty} S_T(t) = 0$, on a $0 < S(1) < 1$.
On pose $\lambda = -\log(S(1)) > 0$ et on a le résultat que $S(t) = \exp(-\lambda t)$ pour $t > 0$. □

Preuve. *(De la Proposition 2)*
Supposons que T suit une loi exponentielle de paramètre λ. On a :

$$\begin{aligned}\mathbb{P}(T > t+h | T > t) &= \frac{\mathbb{P}((T > t+h) \cap (T > t))}{\mathbb{P}(T > t)} = \frac{\mathbb{P}(T > t+h)}{\mathbb{P}(T > t)} \\ &= \frac{\exp(-\lambda(t+h))}{\exp(-\lambda t)} = \exp(-\lambda h) \\ &= \mathbb{P}(T > h).\end{aligned}$$

Réciproquement, supposons que T vérifie la propriété de non vieillissement. Posons $S_T(t) = \mathbb{P}(T > t)$ pour tout $t \geq 0$. On sait que $S_T(t)$ est décroissante sur $\mathbb{R}_+$ et vérifie $\lim\limits_{t\to 0} S_T(t) = 1$ et $\lim\limits_{t\to+\infty} S_T(t) = 0$. On a :

$$P(T > t+h/T > t) = \frac{\mathbb{P}((T > t+h) \cap (T > t)}{\mathbb{P}(T > t)}. \qquad (*)$$

Puisque $h > 0, (T > t+h) \subset (T > t) \Rightarrow ((T > t+h) \cap (T > t)) = T > t+h.$

$$(*) \Rightarrow \mathbb{P}(T > t+h | T > t) = \frac{\mathbb{P}(T > t+h)}{\mathbb{P}(T > t)} = \frac{S_T(t+h)}{S_T(t)} \qquad (i).$$

D'autre part, on a : $P(T > h) = S_T(h) \qquad (ii)$

$$(i) \text{ et } (ii) \Rightarrow \frac{S_T(t+h)}{S_T(t)} = S_T(h)$$

d'après la supposition.
Donc on a : $S_T(t+h) = S_T(t)S_T(h)$. On conclut à l'aide du lemme 1 que $S_T(t) = \exp(-\lambda t)$. □

Cependant, cette loi décrit un processus "sans mémoire" (ou de non-vieillissement ou bien sans usure) d'après la proposition 2 car le temps d'attente jusqu'à l'occurrence de l'évènement d'intérêt ne dépend pas du passé de l'individu. Les quantités associées à cette distribution sont :

1) Espérance et variance
   On a :

$$\begin{aligned}E(T^k) &= \int_0^{+\infty} t^k \lambda \exp(-\lambda t) dt, k \in \mathbb{N}^* \\ &= \frac{\Gamma(k+1)}{\lambda^k}.\end{aligned}$$

   Si $k = 1$, on en déduit que $E(T) = \frac{1}{\lambda}$ et avec $k = 2$, $V(T) = \frac{1}{\lambda^2}$.

2) Le quantile et la médiane
Soit $x \in]0,1[$.

$$x = F_T(t) \Leftrightarrow x = 1 - \exp(-\lambda t) \Leftrightarrow t = -\frac{\ln(1-x)}{\lambda}.$$

D'où $F^{-1}(t) = -\frac{\ln(1-t)}{\lambda}$. Ainsi le quantile d'ordre ω vaut : $q_\omega = -\frac{\ln(1-\omega)}{\lambda}$.
La médiane est donnée par : $Me(T) = q_{0.5} = \frac{\ln 2}{\lambda}$.

On voit que la moyenne (l'espérance) est toujours supérieure à la médiane avec cette distribution.
Les lois exponentielles sont des cas particuliers de deux types de lois classiques plus généraux, couramment utilisées pour modéliser des durées de survie. Il s'agit des lois gamma et les lois de Weibull.

3.2 Distribution de Weibull

C'est la loi la plus populaire et la plus couramment utilisée pour la modélisation des données de fiabilité.
La loi de weibull est très largement utilisée dans les domaines industriel (par exemple la fiabilité) et biomédical (analyse des durées de vie).

Définition 14. *Soit T une variable aléatoire positive. Elle suit la loi de Weibull de paramètres λ et α strictement positifs, si elle a pour densité de support $\mathbb{R}_+^*$ définie par :*

$$f_T(t) = \lambda\alpha t^{\alpha-1} \exp(-\lambda t^\alpha). \tag{3.7}$$

λ est le paramètre d'échelle qui est sans dimension et α le paramètre de forme. Le paramètre de forme est lié au vieillissement et le paramètre d'échelle à la durée de vie médiane.
Ainsi, si $\alpha = 1$ on a la loi exponentielle de paramètre λ.
Grâce à l'expression (3.7), on obtient les distributions suivantes pour $t \in \mathbb{R}_+^*$:

$$F_T(t) = 1 - \exp(-\lambda t^\alpha) \tag{3.8}$$
$$S_T(t) = \exp(-\lambda t^\alpha) \tag{3.9}$$
$$h_T(t) = \lambda\alpha t^{\alpha-1} \tag{3.10}$$
$$H_T(t) = \lambda t^\alpha. \tag{3.11}$$

D'après l'expression (3.10), on en déduit :

- ▶ si $\alpha = 1$, on retrouve une fonction de hasard constante (loi exponentielle) ;

- ▶ si $\alpha > 1$, une fonction de hasard est croissante ;

- ▶ si $\alpha < 1$, une fonction de hasard est décroissante.

On peut ainsi déduire les trois étapes de la vie d'un composant (population) :

- pour $\alpha = 1$, la vie est utile ;

- pour $\alpha > 1$, on a le vieillissement ;

- pour $\alpha < 1$, on a le phénomène de rajeunissement.

Les quantités associées à cette distribution sont :

1) Espérance et variance :
   On a : $E(T^k) = \left(\frac{1}{\lambda} \right)^{\frac{k}{\alpha}} \Gamma \left(\frac{k}{\alpha} + 1 \right)$.
   Ainsi pour $k = 1$, on a : $E(T) = \frac{\Gamma\left(\frac{1}{\alpha} + 1 \right)}{\lambda^{\frac{1}{\alpha}}}$ et avec $k = 2$, on a :

$$V(T) = \frac{\Gamma \left(\frac{2}{\alpha} + 1 \right) - \Gamma^2 \left(1 + \frac{1}{\alpha} \right)}{\lambda^{\frac{2}{\alpha}}}.$$

2) Le quantile et la médiane :
   Soit $x \in]0, 1[$. $x = F_T(t) \Leftrightarrow t = \left(-\frac{\ln(1-x)}{\lambda} \right)^{\frac{1}{\alpha}}$.
   Donc $F^{-1}(t) = \left(-\frac{\ln(1-t)}{\lambda} \right)^{\frac{1}{\alpha}}$.
   On a alors le quantile d'ordre ω qui est égal à : $q_\omega = \left(-\frac{\ln(1-\omega)}{\lambda} \right)^{\frac{1}{\alpha}}$.
   Sa médiane est alors $Me(T) = \left(\frac{\ln 2}{\lambda} \right)^{\frac{1}{\alpha}}$.

L'autre famille classique des modèles pour des durées de survie est la famille des lois Gamma.

3.3 Loi Gamma à un paramètre

T suit une loi Gamma à un paramètre $\beta > 0$ si c'est une variable aléatoire positive dont la densité de support $\mathbb{R}_+^*$ est donnée par :

$$f_T(t) = \frac{t^{\beta-1}\exp(-t)}{\Gamma(\beta)}. \tag{3.12}$$

Cette loi est notée par $G(\beta, 1)$.
D'après (3.12), on en déduit les distribution de survie correspondantes :

$$F_T(t) = \frac{1}{\Gamma(\beta)}\int_0^t x^{\beta-1}\exp(-x) = \gamma(\beta, t) \tag{3.13}$$

$$S_T(t) = 1 - \gamma(\beta, t) \tag{3.14}$$

$$h_T(t) = \frac{t^{\beta-1}\exp(-t)}{\Gamma(\beta)(1-\gamma(\beta, t))}. \tag{3.15}$$

Les quantités associées à cette distribution sont :

1) Espérance et variance :
   On a :

$$E(T^k) = \frac{\Gamma(k+\beta)}{\Gamma(\beta)}.$$

   Ainsi si $k = 1$, on a : $E(T) = \beta$ et avec $k = 2$, on a la variance : $V(T) = \beta$.

Pour cette distribution, l'espérance mathématique et la variance sont ainsi égales.

2) Le quantile et la médiane :
   Ainsi, pour calculer F^{-1}, on utilise la relation entre la loi Gamma et la loi normale. Car la loi Gamma est une généralisation de la loi de khi-deux, qui elle même est reliée à la loi normale.
   Soient $T_1, ..., T_k$, k variables aléatoires i.i.d de loi normale centrée réduite alors on peut vérifier que $\sum_{i=0}^{k} T_i^2$ suit une loi de khi-deux à k degrés de liberté, notée $\chi^2(k)$ ou χ_k^2.
   Sa densité de support $\mathbb{R}_+^*$ est donnée par :

$$f(t; k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} t^{\frac{k}{2}-1}\exp\left(-\frac{t}{2}\right).$$

Sa fonction de répartition est : $F_{\chi^2_k}(t) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^t x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) dx$.
En faisant un changement de variable (poser $u = \frac{x}{2}$), on obtient

$$F_{\chi^2_k}(t) = \gamma\left(\frac{k}{2}, \frac{t}{2}\right).$$

Par conséquent, $F_{\chi^2_{2a}}(2t) = \gamma(a,t)$.
On en déduit alors que $F^{-1}_{\chi^2_k}(t) = \frac{1}{2} F^{-1}_{\chi^2_{2\beta}}(t)$.
Donc le quantile d'ordre ω est donné par $q_\omega = \frac{1}{2} F^{-1}_{\chi^2_{2\beta}}(\omega)$ où $F^{-1}_{\chi^2_{2\beta}}$ est la fonction quantile du χ^2 à 2β degrés de liberté.
La médiane est : $Me(T) = \frac{1}{2} F^{-1}_{\chi^2_{2\beta}}\left(\frac{1}{2}\right)$.

3.4 Loi Gamma à deux paramètres

Définition 15.
Une variable aléatoire T suit une loi Gamma de paramètre β et λ strictement positifs si c'est une variable aléatoire positive dont la densité de support $\mathbb{R}_+^*$ est donnée par :

$$f_T(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}. \tag{3.16}$$

où β est un paramètre de forme (shape) et λ est un paramètre d'échelle (rate).

Cette loi, notée $G(\beta, \lambda)$, généralise la loi exponentielle et la loi gamma à un paramètre. En effet, on retrouve la densité de la loi exponentielle de paramètre λ en posant $\beta = 1$ et celle de la loi gamma à un paramètre pour $\lambda = 1$.
Sa fonction de répartition est donnée par :

$$F_T(t) = \frac{1}{\Gamma(\beta)} \int_0^{\lambda t} (u)^{\beta-1} \exp(-u) du = \gamma(\beta, \lambda t). \tag{3.17}$$

La fonction de survie est $S_T(t) = 1 - \gamma(\beta, \lambda t)$. La fonction de hasard est $h_T(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)(1-\gamma(\beta,\lambda t))}$.
Et la fonction de hasard cumulé est $H_T(t) = -\ln S_T(t) = -\ln\left(1 - \gamma(\beta, \lambda t)\right)$.
On voit alors que la fonction de répartition, la fonction de survie, la fonction de hasard et celle de hasard cumulée s'expriment à l'aide de la fonction gamma incomplète. Cette fonction est calculable numériquement.
Les quantités associées à cette distribution sont :

1) Espérance et variance :

$$E(T^k) = \frac{1}{\Gamma(\beta)\lambda^k} \int_0^{+\infty} u^{k+\beta-1} \exp(-u) du = \frac{\Gamma(\beta+k)}{\Gamma(\beta)\lambda^k}.$$

Pour $k = 1$, on a la moyenne de vie qui est égale à :

$$E(T) = \frac{\beta}{\lambda}. \tag{3.18}$$

et avec $k = 2$, la variance de vie est égale à :

$$V(T) = \frac{\beta}{\lambda^2}. \tag{3.19}$$

2) Le quantile et la médiane :
En se basant sur la loi Gamma à un paramètre, on a le quantile d'ordre ω suivant : $q_\omega = \frac{1}{2\lambda} F^{-1}_{\chi^2_{2\beta}}(\omega)$ et la médiane qui est donnée par :

$$Me(T) = \frac{1}{2\lambda} F^{-1}_{\chi^2_{2\beta}}(\frac{1}{2}).$$

Remarque. D'après l'étude de ces lois, on voit en général que les fonctions de hasard cumulé ont typiquement une "courbe en baignoire".

Conclusion

D'après ces distributions, on voit que la loi exponentielle est la seule caractérisée par une fonction de hasard constante qui est son unique paramètre. La loi de Weibull est très utilisée car elle a d'une part une fonction de hasard croissante et d'autre part une fonction de hasard décroissante.

Chapitre 4

Estimation paramétrique

Il existe de nombreuses méthodes pour estimer des paramètres, parmi lesquelles on peut citer l'estimation paramétrique, l'estimation semi-paramétrique et l'estimation non paramétrique. Notre travail porte sur l'estimation paramétrique.
En ayant retenu une forme de distribution de durée de vie de base spécifiée (ici la loi gamma à deux paramètres), nous cherchons à estimer les paramètres associés à cette distribution par des méthodes classiques. Nous passons en revue dans cette étude trois méthodes courantes d'estimation des paramètres à savoir la méthode du maximum de vraisemblance, la méthode des moments et la méthode par intervalle de confiance (voir [15], [16]).
La première est employée généralement pour ses propriétés asymptotiques intéressantes, la seconde pour sa simplicité et la troisième contourne l'inconvénient majeur des deux autres à donner une estimation ponctuelle (c'est-à-dire qu'elle contourne le fait qu'on est presque sûre de ne pas "tomber" sur la valeur théorique que l'on cherche à estimer).
Dans ce qui suit nous travaillerons sur la loi gamma à deux paramètres qui est très souvent utilisée pour modéliser la distribution de durée de vie.

4.1 Méthode du maximum de vraisemblance (EMV)

Quelques cas particuliers de l'EMV ont été connus depuis le XVIIIème siècle, mais sa définition générale et l'argumentation de son rôle fondamental en Statistique sont dues aux statisticiens Ronald Aylmer Fisher (en 1922). L'EMV est une méthode d'estimation ponctuelle puisqu'elle cherche à trouver une valeur estimée pour un paramètre θ inconnu à partir d'un ensemble donné. Ainsi, nous distinguons deux cas : celui d'une observation complète et celui

des données censurées (cf.[2], [6], [10], [20]).

4.1.1 EMV dans le cas complet

On considère une variable aléatoire de durée T dont la loi dépend du paramètre θ qui est inconnu.

Définition 16.
*Soit $T_i, i = 1, ..., n$ un n-échantillon de réalisations extrait de T et $t_1, ..., t_n$ les temps de survie. Ses réalisations sont i.i.d. selon une loi de densité $f_T(t, \theta)$. On appelle **vraisemblance** de l'échantillon l'application L qui représente l'intensité d'occurrence de l'échantillon $t = (t_1, ..., t_n)$, définie par :*

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

Soit T_i une variable aléatoire de durée indépendante et identiquement distribuée extrait de T avec une fonction de densité $f_T(t_i; \theta)$ et une fonction de survie $S_T(t_i; \theta)$, pour $i = 1, ..., n$ *et* $\theta = (\theta_1, ..., \theta_n)$. Soit E l'ensemble des individus non censurés aux temps t_i. Pour ces individus, la fonction de vraisemblance est donnée par :

$$L(t_i; \theta) = \prod_{i \in E} f_T(t_i; \theta).$$

Définition 17. *(EMV)*
*Soit $T_i, i = 1, ..., n$ un n-échantillon de T. On appelle **EMV** du paramètre θ, noté $\hat{\theta}$, un estimateur qui maximise la vraisemblance $L(t_1, ..., t_n; \theta)$. Elle est définie par :*

$$L(t_1, ..., t_n; \hat{\theta}) = \arg\max_{\theta} L(t_1, ..., t_n; \theta).$$

On suppose que ce maximum est unique.

Remarque. Cette unicité est due à la notion d'exhaustivité. L'EMV peut ne pas exister en général (cf.[16]).

On trouve ce maximum par les conditions du premier ordre et comme il est plus facile de dériver une somme qu'un produit, on utilise en général le logarithme de la vraisemblance (dit log-vraisemblance) plutôt que la vraisemblance. Elle est définie par :

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

Si $L(t_1, ..., t_n; \theta)$ est la vraisemblance du paramètre θ, alors l'estimateur du maximum de vraisemblance vérifie :

$$\begin{cases} \frac{\partial}{\partial \theta} \ln L(t_1, ..., t_n; \theta) = 0 \\ \frac{\partial^2}{\partial \theta^2} \ln L(t_1, ..., t_n; \theta) < 0 \end{cases}$$

On note souvent $\ln L(t_1, ..., t_n; \theta)$ par $\ell(t_1, ..., t_n; \theta)$.
Des fois toutes les données de durées de vie peuvent ne pas être complètes, dans ce cas on parle de données censurées (c'est-à-dire données incomplètes). Nous allons donc définir cette méthode d'estimation dans le cas des données censurées.

4.1.2 EMV en présence d'une censure

Soit un n-échantillon d'individus dont m individus sont censurés. Soit $T_1, ..., T_n$ des variables aléatoires de durée i.i.d extrait de T avec respectivement une fonction de densité $f_T(t_i; \theta)$ et une fonction de survie $S_T(t_i; \theta)$, $i = 1, ..., n$ et $\theta = (\theta_1, ..., \theta_n)$.
Avant de s'intéresser au cas de la censure à droite, nous introduisons tout d'abord la vraisemblance dans le cas de la censure à gauche puis celle de la censure par intervalle.

Fonction de vraisemblance pour une censure à gauche

Pour une observation censurée à gauche, où l'on sait seulement que l'évènement a eu lieu avant la durée t_i, la vraisemblance est associée à la probabilité correspondante

$$\mathbb{P}(T \leq t_i) = F_T(t_i) = 1 - S_T(t_i).$$

On a alors la fonction de vraisemblance :

$$L(t_1, ..., t_n; \theta) = \prod_{i \in G} F_T(t_i; \theta)$$

où G représente l'ensemble des individus censurés à gauche aux temps t_i.

Fonction de vraisemblance pour une censure par intervalle

Pour un individu censurée sur un intervalle, pour lequel on sait seulement que l'évènement s'est réalisé entre deux temps t_{i_1} *et* t_{i_2}, avec $t_{i_1} < t_{i_2}$, la vraisemblance associée sera :

$$\mathbb{P}(t_{i_1} < T < t_{i_2} = F_T(t_{i_2}) - F_T(t_{i_1}) = S_T(t_{i_1}) - S_T(t_{i_2}).$$

Donc

$$L(t_1, ..., t_n; \theta) = \prod_{i \in I} \left(S_T(t_{i_1}) - S_T(t_{i_2}) \right),$$

où I représente l'ensemble des individus censurés par intervalle aux temps t_{i_1} *et* t_{i_2}.

Fonction de vraisemblance pour une censure à droite

Nous allons maintenant présenter l'écriture de la vraisemblance pour le cas de la censure à droite. Les observations correspondantes à une censure à droite sont celles pour lesquelles la durée d'évènement est supérieure à la durée de censure. Leur vraisemblance est donc simplement la probabilité associée, soit, pour une durée T_i censurée à droite, $\mathbb{P}(T > t_i) = S_T(t_i)$. La vraisemblance est donnée par :

$$L(t_1, ..., t_n; \theta) = \prod_{i \in D} S_T(t_i; \theta),$$

où D représente l'ensemble des individus censurés à droites aux temps t_i.
En combinant les trois cas précédent d'une censure et le cas non censuré, on en déduit la fonction de vraisemblance suivante :

$$L(t_1, ..., t_n; \theta) = \prod_{i \in E} f_T(t_i; \theta) \prod_{i \in D} S_T(t_i; \theta) \prod_{i \in G} F_T(t_i; \theta) \prod_{i \in I} (S_T(t_{i_1}) - S_T(t_{i_2})).$$

Nous allons à présent évoquer les trois types de la censure à droite comme mentionnés dans la section 2.1.

a) Cas de la censure aléatoire et non aléatoire
Soit d_i une fonction indicatrice d'observation complète de la variable de durée T_i, avec T_i une durée de vie supposée indépendante et identiquement distribuée. Ainsi, les données impliquant une censure à droite peuvent être représentées par un couple de variable (T_i, d_i), où

$$d_i = \begin{cases} 1 & si \quad X_i \leq C_i \\ 0 & si \quad X_i > C_i \end{cases}$$

où X_i est la vraie durée passée dans l'état étudié par l'individu i (c'est-à-dire si la durée est observée $T_i = X_i$) et C_i la durée maximale observable pour l'individu i du fait de la censure (si la durée est censurée $T_i = C_i$). On observe donc une durée

$$T_i = min(X_i, C_i).$$

Soit $f(t_i, \theta)$, la densité des t_i et $S(t_i, \theta)$, la fonction de survie associée.

Proposition 3.
Supposons que le délai de la censure C_i de l'individu i est une variable aléatoire indépendante de la durée de vie X_i, pour $i = 1, ..., n$. La vraisemblance s'écrit alors :

$$L((t_1, d_1), ..., (t_n, d_n); \theta) = \prod_{i=1}^{n} (f(t_i, \theta))^{d_i} (S(t_i, \theta))^{1-d_i}.$$

Preuve.
On rappelle que :

- ▷ La durée de vie exacte sera connue si et seulement si T_i est inférieure ou égale C_i.
- ▷ La durée de vie est supérieure à C_i, l'individu est survivant et son temps d'évènement est censuré C_i.

Pour $d_i = 0$, on voit que :

$$\begin{aligned}\mathbb{P}(T_i = C_i, d_i = 0) &= \mathbb{P}(T_i = C_i | d_i = 0)\mathbb{P}(d_i = 0) = \mathbb{P}(d_i = 0) \\ &= P(X_i > C_i) = S(C_i)\end{aligned}$$

également pour $d_i = 1$,

$$\begin{aligned}\mathbb{P}(T_i = X_i, d_i = 1) &= \mathbb{P}(T_i = X_i | d_i = 1)\mathbb{P}(d_i = 1) \\ &= \mathbb{P}(X_i = T_i | X_i \leq C_i)\mathbb{P}(X_i \leq C_i) \\ &= \frac{\mathbb{P}(X_i = T_i, X_i \leq C_i)}{\mathbb{P}(X_i \leq C_i)}\mathbb{P}(X_i \leq C_i) \\ &= \left(\frac{f(t_i)}{1 - S(C_i)}\right)(1 - S(C_i)) = f(t_i).\end{aligned}$$

On peut combiner les deux expressions en une seule

$$\mathbb{P}((t_1, d_1), ..., (t_n, d_n)) = (f(t_1, ..., t_n))^{d_i} (S(t_1, ..., t_n))^{1-d_i}.$$

Ainsi, pour un couple variables aléatoires $(t_i, d_i), i = 1, ..., n$, la fonction de vraisemblance est donnée par :

$$L((t_1, d_1), ..., (t_n, d_n); \theta) = \prod_{i=1}^{n} \mathbb{P}[(t_i, d_i)] = \prod_{i=1}^{n} (f(t_i, \theta))^{d_i} (S(t_i, \theta))^{1-d_i}.$$

□

En connaissant la vraisemblance, on peut en déduire la log-vraisemblance d'une observation. Elle est définie par :

$$\begin{aligned}\ln L((t_1,d_1),...,(t_n,d_n);\theta) &= \ln \prod_{i=1}^{n} (f(t_i,\theta))^{d_i} (S(t_i,\theta))^{1-d_i} \\ &= \sum_{i=1}^{n} d_i \ln f(t_i,\theta) + \sum_{i=1}^{n} (1-d_i) \ln S(t_i,\theta).\end{aligned}$$

Mais on peut également écrire cette définition en utilisant :

$$\begin{aligned}h(t_i,\theta) &= \frac{f(t_i,\theta)}{S(t_i,\theta)} \\ \Leftrightarrow \ln f(t_i,\theta) &= \ln h(t_i,\theta) + \ln S(t_i,\theta),\end{aligned}$$

de sorte que

$$\ell((t_1,d_1),...,(t_n,d_n);\theta) = \sum_{i=1}^{n} d_i \ln h(t_i,\theta) + \sum_{i=1}^{n} \ln S(t_i,\theta),$$

avec $\ell((t_1,d_1),...,(t_n,d_n);\theta) = \ln L((t_1,d_1),...,(t_n,d_n);\theta)$.
D'autre part aussi, on a :

$$\ln S(t_i,\theta) = -H(t_i,\theta).$$

Alors, on peut écrire la log-vraisemblance en fonction du fonction de hasard cumulée :

$$\ell((t_1,d_1),...,(t_n,d_n);\theta) = \sum_{i=1}^{n} d_i \ln h(t_i,\theta) - \sum_{i=1}^{n} H(t_i,\theta).$$

Ainsi, la log-vraisemblance pour les données censurées peut être déterminée de trois manières équivalentes :

$$\ell((t_1,d_1),...,(t_n,d_n);\theta) = \sum_{i=1}^{n} d_i \ln f(t_i,\theta) + \sum_{i=1}^{n} (1-d_i) \ln S(t_i,\theta) \tag{4.1}$$

$$= \sum_{i=1}^{n} d_i \ln h(t_i,\theta) + \sum_{i=1}^{n} \ln S(t_i,\theta) \tag{4.2}$$

$$= \sum_{i=1}^{n} d_i \ln h(t_i,\theta) - \sum_{i=1}^{n} H(t_i,\theta). \tag{4.3}$$

On choisit entre (4.1), (4.2), (4.3) l'expression la plus pratique selon la distribution.
Par analogie au cas complet, on définit de la même manière l'estimateur du maximum de vraisemblance.

b) Cas de la censure de type II
Pour ce type de censure, nous avons (cf.[10]) :

$$L(t_{(1)},...,t_{(n)};\theta) = \frac{n!}{(n-m)!}\left[\prod_{i=1}^{m}(t_{(i)},\theta)\right]\left[S_T(t_{(m)})^{n-m}\right].$$

Application à la loi gamma à deux paramètres

D'après la section 3.4, nous avons :

$$f_T(t) = \frac{\lambda^\beta t^{\beta-1}\exp(-\lambda t)}{\Gamma(\beta)} \quad et \quad S_T(t) = 1-\gamma(\beta,\lambda t)$$

Nous estimons $\theta = (\beta,\lambda)$ (les paramètres inconnus) dans le cas non censuré et dans le cas censuré.

1) Cas des données non censurées
La vraisemblance est définie, pour $\theta = (\beta,\lambda)$, par :

$$\begin{aligned} L(t_1,...,t_n;\theta) &= \prod_{i=1}^{n} f_T(t_i;\theta) = \prod_{i=1}^{n}\frac{\lambda^\beta t_i^{\beta-1}\exp(-\lambda t_i)}{\Gamma(\beta)} \\ &= \frac{\lambda^{n\beta}\prod_{i=1}^{n} t_i^{\beta-1}\exp\left(-\lambda\sum_{i=1}^{n} t_i\right)}{(\Gamma(\beta))^n}. \end{aligned}$$

Par passage au logarithme on a :

$$\ln L(t_1,...,t_n;\theta) = n\beta\ln\lambda + (\beta-1)\sum_{i=1}^{n}\ln(t_i) - \lambda\sum_{i=1}^{n} t_i - n\ln(\Gamma(\beta)).$$

L'EMV $\widehat{\theta} = (\widehat{\beta},\widehat{\lambda})$ vérifie les conditions du premier et second ordre suivant :

$$\begin{cases} \frac{\partial}{\partial\beta}\ln L(t_1,...,t_n;\widehat{\theta}) = 0 \\ \frac{\partial}{\partial\lambda}\ln L(t_1,...,t_n;\widehat{\theta}) = 0 \\ \frac{\partial^2}{\partial\beta^2}\ln L(t_1,...,t_n;\widehat{\theta}) < 0 \\ \frac{\partial^2}{\partial\lambda^2}\ln L(t_1,...,t_n;\widehat{\theta}) < 0 \end{cases} \tag{4.4}$$

Les dérivées première et seconde par rapport à λ donnent respectivement :

$$\frac{n\beta}{\lambda} - \sum_{i=1}^{n} t_i \qquad \text{et} \qquad -\frac{n\beta}{\lambda^2} < 0.$$

Ainsi, d'après la deuxième équation du système (4.4), $\widehat{\lambda}$ est solution de l'équation $\frac{n\beta}{\lambda} - \sum_{i=1}^{n} t_i = 0$, soit

$$\widehat{\lambda} = \frac{\widehat{\beta}}{\overline{t}}, \quad \text{avec} \quad \overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

.

Pour le calcul de l'EMV de $\widehat{\beta}$, on a :

$$\begin{aligned}\frac{\partial \ln L(t_1, ..., t_n; \theta)}{\partial \beta} &= n \ln \lambda + \sum_{i=1}^{n} t_i - n\frac{\partial \ln(\Gamma(\beta)}{\partial \beta} \\ &= n \ln \lambda + \sum_{i=1}^{n} t_i - \frac{\partial \Gamma(\beta)}{\partial \beta}/\Gamma(\beta).\end{aligned}$$

Or, d'après le théorème de dérivation sous le signe intégral on a :

$$\frac{\partial \Gamma(\beta)}{\partial \beta} = \int_0^{+\infty} (\ln x) x^{\beta-1} \exp(-x) dx$$

et

$$\frac{\partial^2 \Gamma(\beta)}{\partial \beta^2} = \int_0^{+\infty} (\ln x)^2 x^{\beta-1} \exp(-x) dx.$$

On voit alors que le calcul exact de $\widehat{\beta}$ n'est pas possible. Donc nous pouvons seulement exprimer $\widehat{\lambda}$ en fonction $\widehat{\beta}$. Pour contourner cette difficulté on peut trouver, par exemple dans [6], une méthode pour exprimer ces estimateurs dite méthode de Newton-Raphson.

2) Cas des données censurées

Pour estimer des données censurées par l'EMV, nous considérons ici la censure à droite de type I.

Soit $t_1, t_2, ..., t_m$ les individus non censurés et $t_{m+1}^+, ..., t_n^+$ les individus censurés à droite .

Supposons que $f_T(t; \theta)$ et $S_T(t; \theta)$ la fonction densité des individus non censurés et la fonction de survie des individus censurés à droite respectivement. La fonction de vraisemblance ici notée $L_1(t_i; \theta)$ est donnée

par :

$$L_1(t_i;\theta) = \prod_{i=1}^{m} f_T(t_i;\theta) \prod_{i=m+1}^{n} S_T(t_i^+;\theta)$$
$$= \prod_{i=1}^{m} \frac{\lambda^\beta t_i^{\beta-1}\exp(-\lambda t_i)}{\Gamma(\beta)} \prod_{i=m+1}^{n} (1-\gamma(\beta,\lambda t_i^+))$$

Par passage au logarithme on a :

$$\ln L_1(t_i;\theta) = m\beta\ln\lambda + (\beta-1)\sum_{i=1}^{m}\ln(t_i) - \lambda\sum_{i=1}^{m} t_i - m\ln(\Gamma(\beta))$$
$$+ \sum_{i=m+1}^{n} \ln(1-F_{\chi^2_{2\beta}}(2\lambda t_i^+)).$$

D'autre part la log-vraisemblance peut être écrite d'une autre manière. En effet, on a :

$$S_{T_i}(t_i^+;\theta) = 1-\gamma(\beta,\lambda t_i^+) = 1 - \frac{1}{\Gamma(\beta)}\int_0^{\lambda t_i^+} u^{\beta-1}\exp(-u)du$$
$$= 1 - \frac{\lambda^\beta}{\Gamma(\beta)}\int_0^{t_i^+} x^{\beta-1}\exp(-\lambda x)dx.$$

Or

$$\frac{\lambda^\beta}{\Gamma(\beta)}\int_0^{+\infty} x^{\beta-1}e^{-\lambda x}dx = \frac{\lambda^\beta}{\Gamma(\beta)}\int_0^{t_i^+} x^{\beta-1}e^{-\lambda x}dx + \frac{\lambda^\beta}{\Gamma(\beta)}\int_{t_i^+}^{+\infty} x^{\beta-1}e^{-\lambda x}dx.$$

On en déduit alors que $1-\gamma(\beta,\lambda t_i^+) = \frac{\lambda^\beta}{\Gamma(\beta)}\int_{t_i^+}^{+\infty} x^{\beta-1}\exp(-\lambda x)dx$.
D'où la log-vraisemblance est donnée par :

$$\ln L_1(t_i;\theta) = m\beta\ln\lambda + (\beta-1)\sum_{i=1}^{m}\ln(t_i) - \lambda\sum_{i=1}^{m} t_i - m\ln(\Gamma(\beta))$$
$$+ \sum_{i=m+1}^{n} \ln\left(\frac{\lambda^\beta}{\Gamma(\beta)}\int_{t_i^+}^{+\infty} x^{\beta-1}\exp(-\lambda x)dx\right).$$

Or

$$\sum_{i=m+1}^{n} \ln\left(\frac{\lambda^\beta}{\Gamma(\beta)}\int_{t_i^+}^{+\infty} x^{\beta-1}e^{-\lambda x}dx\right) = \beta(n-m)\ln\lambda - (n-m)\ln\Gamma(\beta)$$
$$+ \sum_{i=m+1}^{n}\left(\ln\int_{t_i^+}^{+\infty} x^{\beta-1}e^{-\lambda}dx\right).$$

Donc

$$\ln L_1(t_i;\theta) = n\beta \ln \lambda - n \ln \Gamma(\beta) + (\beta - 1) \sum_{i=1}^{m} \ln t_i \\ - \lambda \sum_{i=1}^{m} t_i + \sum_{i=m+1}^{n} \ln \left(\int_{t_i^+}^{+\infty} x^{\beta-1} \exp(-\lambda x) dx \right).$$

Ainsi, pour $\theta = (\beta, \lambda)$, en dérivant $\ln L_1(t_i;\theta)$ par rapport à λ et β à on a :

$$\frac{n\beta}{\lambda} - \sum_{i=1}^{m} t_i - \sum_{i=m+1}^{n} \left(\frac{\int_{t_i^+}^{+\infty} x^{\beta} \exp(-\lambda x) dx}{\int_{t_i^+}^{+\infty} x^{\beta-1} \exp(-\lambda x) dx} \right) \qquad (*)$$

et

$$n \ln \lambda - n \frac{\Gamma'(\beta)}{\Gamma(\beta)} + \sum_{i=1}^{m} \ln t_i + \sum_{i=m+1}^{n} \left(\frac{\int_{t_i^+}^{+\infty} x^{\beta-1} \exp(-\lambda x) \ln(x) dx}{\int_{t_i^+}^{+\infty} x^{\beta-1} \exp(-\lambda x) dx} \right) \qquad (**)$$

respectivement. Pour obtenir l'EMV, il faut trouver les valeurs λ et β qui annulent $(*)$ et $(**)$. Cette résolution peut se faire à l'aide de l'algorithme de Newton-Raphson.

► Algorithme de Newton-Raphson (cf [6], [10]) :

C'est l'une des méthodes d'optimisation les plus utilisées en Statistique.
C'est une méthode numérique pour résoudre les racines d'une fonction. Il s'agit d'un algorithme itératif basé sur l'équation suivante :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Et la relation de récurrence est :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

On a un exemple de graphique suivant :

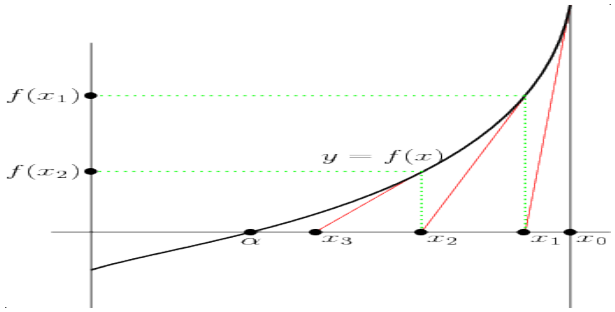

FIGURE 4.1 – Newton-Raphson appliquée à une fonction f.

Comme nous travaillons sur des durées de survie, la fonction f précédente est en fait la dérivée de la log-vraisemblance par rapport au paramètre $\theta = (\beta, \lambda)$. On aura donc pour $(\beta_k, \lambda_k) = \theta_k$,

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln L(t_1, ..., t_n; \theta_k) \right]^{-1} \frac{\partial}{\partial\theta} \ln L(t_1, ..., t_N; \theta_k)$$

$$\theta_{k+1} = \theta_k - \frac{\ell'(t_1, ..., t_n; \theta_k)}{\ell''(t_1, ..., t_n; \theta_k)}$$

$$\theta_{k+1} = \theta_k - \frac{\nabla\ell(t_1, ..., t_n; \theta_k)}{H(t_1, ..., t_n; \theta_k)}$$

où $\nabla\ell(t_1, ..., t_n; \theta_k)$ est le gradient de $\ell(t_1, ..., t_n; \theta_k)$ et H désigne la matrice Hessienne de $\ell(t_1, ..., t_n; \theta_k)$.
La procédure est exécutée jusqu'à ce qu'il n'y ait pas de différence significative entre θ_{k+1} et θ_k. $\theta_{k+1} = \theta_k$ est équivalent à $\ell'(t_1, ..., t_n; \theta_k) = 0$. Cela démontre que lorsque l'algorithme a convergé, nous avons atteint un point stationnaire de $\ell(t_1, ..., t_n; \theta_k)$.
Cela pourrait être un maximum, un minimum ou un point selle. Cependant si $\ell''(t_1, ..., t_n; \theta_k) < 0$, le point est un point maximal.
Ainsi, d'après les calculs, pour $\theta_k = (\beta_k, \lambda_k)$ et $(.)^T$ l'opérateur transposé, on a :

$$\begin{aligned}\nabla\ell(t_1, ..., t_n; \theta) &= \left(\frac{\partial}{\partial\beta}\ell(t_1, ..., t_n; \theta), \frac{\partial}{\partial\lambda}\ell(t_1, ..., t_n; \theta) \right)^T \\ &= \left(n\ln\lambda - n(\ln\Gamma(\beta))' + \sum_{i=1}^n \ln t_i, \frac{n\beta}{\lambda} - \sum_{i=1}^n t_i \right)^T\end{aligned}$$

et

$$\begin{aligned}H(t_1, ..., t_n; \theta_k) &= \begin{pmatrix} \frac{\partial^2}{\partial^2\beta}\ell(t_1, ..., t_n; \theta_k) & \frac{\partial^2}{\partial\beta\partial\lambda}\ell(t_1, ..., t_n; \theta_k) \\ \frac{\partial^2}{\partial\lambda\partial\beta}\ell(t_1, ..., t_n; \theta_k) & \frac{\partial^2}{\partial^2 h}\ell(t_1, ..., t_n) \end{pmatrix} \\ &= \begin{pmatrix} -n(\ln\Gamma(\beta))'' & \frac{n}{\lambda} \\ \frac{n}{\lambda} & -\frac{n\beta}{\lambda^2} \end{pmatrix}.\end{aligned}$$

D'où

$$[H(t_1,...,t_n;\theta_k)]^{-1} = \frac{1}{n(1-\beta(\ln\Gamma(\beta))'')}\begin{pmatrix} \beta & \lambda \\ \lambda & \lambda^2\ln(\Gamma(\beta))'' \end{pmatrix}.$$

Les étapes de l'algorithme sont les suivantes :

Etape 1 Fixer une valeur initiale θ_0.

Etape 2 Calcul du gradient et du hessien numériques, respectivement

$$\nabla\ell(t_1,...,t_N;\theta_k) \quad et \quad H(t_1,...,t_n;\theta_k).$$

Etape 3 Calcul de la nouvelle valeur du paramètre

(a) On calcule d'abord

$$\theta_{k+1} = \theta_k - \delta_k H^{-1}(t_1,...,t_n;\theta_k)\nabla\ell(t_1,...,t_n;\theta_k),$$

$k \geq 0$ avec $\delta_k = 1$.

(b) On vérifie ensuite la condition $\ell(t_i;\theta_{k+1}) \geq \ell(t_i;\theta_k)$: l'algorithme doit être croissant pour $i = 1,...,n$.

(c) Si cette condition est remplie, on passe à l'itération suivante (Etape 2), sinon on pose $\delta_k = 1/2, (1/2)^2, \ldots, (1/2)^c$ jusqu'à ce qu'elle soit remplie.

Etape 4 Arrêt des calculs dès que la log-vraisemblance est stable. Plus précisément, on arrête dès que

$$\left|\frac{\ell(t_1,...,t_n;\theta_{k+1}) - \ell(t_1,...,t_n;\theta_k)}{\ell(t_1,...,t_n;\theta_{k+1})}\right| < \varepsilon,$$

avec par exemple $\varepsilon = 10^{-6}$.

Etape 5 Affichage des résultats.

Remarque.

L'algorithme converge rapidement lorsque les valeurs de départ sont proches de la racine, mais peut ne pas converger lorsque les valeurs de départ sont mal choisies.
On peut aussi choisir comme valeurs initiales les estimateurs par la méthode des moments.

Ainsi dans le cas de la loi Gamma, les dérivées premières et secondes de la fonction gamma peuvent être approximées d'une part. Numériquement

celles-ci peuvent être approximées en utilisant le développement limité de Taylor-Young par :

$$\Gamma'(\beta) \approx \frac{\Gamma(\beta+\lambda)-\Gamma(\beta)}{\lambda}$$
$$\Gamma''(\beta) \approx \frac{\Gamma'(\beta+\lambda)-\Gamma'(\beta)}{\lambda} \approx \frac{\Gamma(\beta+2\lambda)-2\Gamma(\beta+\lambda)+\Gamma(\beta)}{\lambda^2}$$

Un autre estimateur des paramètres β et λ dans ce modèle est donné par l'estimateur par la méthode des moments (EMM).

4.2 Méthode des moments (EMM)

La méthode des moments à été introduite en 1894 par Karl Pearson (mathématicien britanique). Il est connu pour avoir développé le coefficient de corrélation et le test du χ^2. Soit un n-échantillon T_i extrait de T dont la densité est $f(t;\theta)$ où $\theta = (\theta_1, ..., \theta_m)$ sont les paramètres inconnus. La méthode des moments consiste à égaler les m premiers moments théoriques aux m premiers moments empiriques. Autrement dit les estimateurs $\theta_1, ..., \theta_m$ sont les solutions des équations :

$$E_\theta(T^k) = \mu_k(T).$$

Estimation des paramètres de la loi Gamma :
La loi Gamma possède deux paramètres à estimer, $\theta = (\theta_1 = \beta, \theta_2 = \lambda)$. L'estimation des deux premiers moments se fait grâce à la moyenne empirique pour l'espérance et la variance empirique pour la variance théorique. En effet, d'après la section 3.4, on a : $\begin{cases} E(T) = \frac{\beta}{\lambda} \\ V(T) = \frac{\beta}{\lambda^2}. \end{cases}$
Donc on peut exprimer β et λ en fonction de $E(T)$ et $V(T)$:

$$\begin{cases} \beta = \frac{E^2(T)}{V(T)} \\ \lambda = \frac{E(T)}{V(T)}. \end{cases}$$

Comme on sait que la moyenne empirique et la variance empirique sont des estimateurs convergents de $E(T)$ *et* $V(T)$ respectivement, nous obtenons pour estimateurs :

$$\begin{cases} \tilde{\beta} = \frac{\overline{T}_n^2}{S_n^2} \\ \tilde{\lambda} = \frac{\overline{T}_n}{S_n^2} \end{cases}$$

avec

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_i \; et \; S_n^2 = \frac{1}{n}\sum_{i=1}^{n} T_i^2 - \overline{T}_n^2.$$

4.3 Estimation par intervalle de confiance

L'estimation par intervalle de confiance consiste à déterminer un intervalle contenant la vraie valeur du paramètre. Le principe de cette méthode est de proposer un encadrement d'un paramètre inconnu d'une population dont la loi est connue avec un seuil α donné (voir [2], [6]).
Soit $T_1, ..., T_n$ un n-échantillon extrait d'une variable aléatoire T dont la loi dépend du paramètre θ inconnu et $f(t, \theta)$ sa densité. Soit $\alpha \in [0, 1]$. S'il existe des variables aléatoires réelles $a_n(T_1, ..., T_n)$ et $b_n(T_1, ..., T_n)$ telles que :

$$\mathbb{P}\left(\theta \in [a_n(T_1, ..., T_n), b_n(T_1, ..., T_n)]\right) = 1 - \alpha.$$

On dit alors que $[a_n(T_1, ..., T_n), b_n(T_1, ..., T_n)]$ est un intervalle de confiance pour le paramètre θ avec une certaine probabilité $1 - \alpha$. On le note souvent par $IC_{1-\alpha}(\theta)$.

4.3.1 Intervalle de confiance construit à partir des EMV

Il existe trois approches pour calculer les intervalles de confiance :

- la méthode exacte (basée sur la solution analytique de la distribution d'échantillonnage) ;
- l'approximation basée sur la théorie des grands échantillons (c'est-à-dire par approximation normale) ;
- la méthode de rééchantillonnage ou bootstrap.

Ainsi pour nos EMV, nous pouvons utiliser n'importe laquelle des trois approches. Mais généralement la distribution exacte est difficile à résoudre donc nous utiliserons la normalité asymptotique des EMV. Pour cela nous avons besoin du théorème suivant :

Théorème 1.
Soit un n-échantillon T_i i.i.d. de densité $f(t, \theta)$ et $\hat{\theta}_n$ est l'EMV de θ. Sous des conditions de régularité, $\hat{\theta}_n$ est consistant, et si $I_1(\theta)$ est inversible, alors pour tout $\theta \in \Theta$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, I_1(\theta)^{-1}\right)$$

où $I_1(\theta) = -E_\theta\left(\frac{\partial^2}{\partial \theta^2} \ln f(T_1, \theta)\right)$.

Preuve.
Soit $\ell(t; \theta) = \ln L(t; \theta)$, avec $L(t; \theta) = \prod_{i=1}^{n} f(t; \theta)$.

On a :

$$\ell(t;\theta) = \ln L(t;\theta) = \sum_{i=1}^{n} f(t_i;\theta)$$

$\ell'(t;\hat{\theta}_n) = \left(\frac{\partial}{\partial\theta_k}\ell_n(t;\hat{\theta}_n)\right)_{k=1,\ldots,dim(\theta)}$ et $\ell''(t;\theta) = H(t;\theta)$

Comme $\hat{\theta}_n$ est un maximum, $H(t;\theta)$ est définie négative autour de $\hat{\theta}_n$.
Par définition de l'EMV, on a : $\ell'(t,\theta) = 0$. Donc $\ell'(\hat{\theta}_n) = 0$.
En appliquant le développement limité de Taylor à $\ell'(\hat{\theta}_n)$ au point θ on a :

$$\ell'(\hat{\theta}_n) \approx \ell'(\theta) + (\hat{\theta}_n - \theta)\ell''(\theta)$$
$$(\hat{\theta}_n - \theta) \approx -\frac{\ell'(\theta)}{\ell''(\theta)}$$

En multipliant par $\sqrt{n}$ on a :

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta) &\approx -\sqrt{n}\frac{\ell'(\theta)}{\ell''(\theta)} \\ &\approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \end{aligned} \tag{4.5}$$

Or $\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln(f(T_i;\theta))$. Comme les T_i sont des variables indépendantes et identiquement distribuées, d'après le théorème central limite, on a :
$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln(f(T_i;\theta)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(E\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right], V\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right]\right)$
avec V la variance. Calculons l'espérance et la variance :

$$\begin{aligned} E\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right] &= \int\left[\frac{\partial}{\partial\theta}\ln(f(t;\theta))\right]f(t,\theta)dt \\ &= \int\left(\left[\frac{\partial}{\partial\theta}f(t;\theta)\right]/f(t;\theta)\right)f(t;\theta)dt \\ &= \int\frac{\partial}{\partial\theta}(f(t;\theta))dt \\ &= \frac{\partial}{\partial\theta}\int f(t;\theta)dt \\ &= 0 \end{aligned}$$

car $\int f(t;\theta)dt = 1$ puisque f est une densité.

$$\begin{aligned} Var\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right] &= E\left(\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right]^2\right) - \left(E\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right]\right)^2 \\ &= E\left(\left[\frac{\partial}{\partial\theta}\ln(f(T_1;\theta))\right]^2\right) \\ &= I_1(\theta) \end{aligned}$$

Donc

$$\frac{1}{\sqrt{n}}\ell'(\theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_1(\theta)).$$

Ainsi pour le dénominateur $-\frac{1}{n}\ell''(\theta)$ dans 4.5 on a :

$$-\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln(f(t_i;\theta)) \xrightarrow{\mathcal{LGN}} E\left(-\frac{\partial^2}{\partial\theta^2}\ln(f(T;\theta))\right) = I_1(\theta)$$

D'où

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} = \frac{1}{I_1(\theta)}\left(\frac{1}{\sqrt{n}}\ell'(\theta)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{I_1(\theta)}{I_1(\theta)^2}\right).$$

Car si $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$ alors $aX + b \rightsquigarrow \mathcal{N}(am, a^2\sigma^2)$.
Donc $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{I_1(\theta)}\right)$. □

Remarque.
Pour n assez grand la loi de la variable $Z = \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) = \sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)$ peut être assimilée à une loi normale centrée réduite.

Donc nous pouvons toujours construire un intervalle de confiance asymptotique à partir de ce résultat :

$$\mathbb{P}\left(q_{\frac{\alpha}{2}} \leq \sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

avec $q_{\frac{\alpha}{2}}$ et $q_{1-\frac{\alpha}{2}}$ les quantiles de la loi normale d'ordre $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ respectivement. Comme $Z \rightsquigarrow \mathcal{N}(0,1)$ alors $q_{\frac{\alpha}{2}} = -q_{1-\frac{\alpha}{2}}$.
Ainsi nous aurons alors

$$\mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \left(\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}}\right), \hat{\theta}_n + \left(\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}}\right)\right]\right) = 1 - \alpha.$$

L'intervalle de confiance au seuil α est donné par :

$$IC_{1-\alpha}(\theta) = \left(\left[\hat{\theta}_n - \left(\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}}\right), \hat{\theta}_n + \left(\frac{q_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}}\right)\right]\right).$$

On peut aussi remplacer $I_n(\theta)$ par son estimateur $I_n(\hat{\theta})$.

4.3.2 Intervalle de confiance construit à partir d'une fonction pivotale

Nous pouvons construire un intervalle de confiance en utilisant une fonction pivotale dépendant des observations qui suivrait une loi de khi-deux. Ainsi nous aurons besoin des résultats du Théorème 2 et du Théorème 3 suivants. Supposons que $2\beta \in \mathbb{N}^*$.

Théorème 2.
Si $T \rightsquigarrow G(\beta, \lambda)$ *et* $Y \rightsquigarrow G(h, \lambda)$ *sont deux variables aléatoires indépendantes alors* $T + Y \rightsquigarrow G(\beta + h, \lambda)$.

Preuve.
$T \rightsquigarrow G(\beta, \lambda) \Rightarrow f(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}$
$Y \rightsquigarrow G(h, \lambda) \Rightarrow f(t) = \frac{\lambda^h y^{h-1} \exp(-\lambda y)}{\Gamma(h)}$. Posons $Z = T + Y$ et $g(z)$ sa densité.

$$\begin{aligned} g(z) &= \int_0^z \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t}{\Gamma(\beta)} \frac{\lambda^h (z-t)^{h-1} \exp(-\lambda(z-t))}{\Gamma(h)} dt \\ &= \int_0^z \frac{\lambda^{\beta+h}}{\Gamma(\beta)\Gamma(h)} \exp(-\lambda z)(z-t)^{h-1} t^{\beta-1} dt. \end{aligned}$$

En posant $t = xz$, on a :

$$\begin{aligned} g(z) &= \frac{\lambda^{\beta+h}}{\Gamma(\beta)\Gamma(h)} \exp(-\lambda z) \int_0^1 (xz)^{\beta-1} (z - xz)^{h-1} z dx \\ &= \frac{\lambda^{\beta+h}}{\Gamma(\beta)\Gamma(h)} \exp(-\lambda z) \int_0^1 x^{\beta-1} z^{\beta+h-1} (1-x)^{h-1} dx \\ &= \frac{\lambda^{\beta+h}}{\Gamma(\beta)\Gamma(h)} \exp(-\lambda z) z^{\beta+h-1} \int_0^1 x^{\beta-1} (1-x)^{h-1} dx \\ &= \lambda^{\beta+h} z^{\beta+h-1} \exp(-\lambda z) \frac{1}{\Gamma(\beta)\Gamma(h)} \int_0^1 x^{\beta-1} (1-x)^{h-1} dx \end{aligned}$$

or

$$\int_0^1 x^{\beta-1} (1-x)^{h-1} dx = \frac{\Gamma(\beta)\Gamma(h)}{\Gamma(\beta+h)}.$$

D'où

$$g(z) = \frac{\exp(-\lambda z) \lambda^{\beta+h} z^{\beta+h-1}}{\Gamma(\beta+h)}$$

On en déduit que si T et Y sont indépendantes , $T + Y \rightsquigarrow G(\beta + h, \lambda)$. □

Dans notre cas, la variable aléatoire $\sum_{i=1}^{n} T_i \rightsquigarrow G(n\beta, \lambda)$.

Comme la loi $G(\frac{n}{2}, \frac{1}{2})$ est aussi appelée loi du khi-deux à n degrés de liberté, le but est alors de se ramener à une loi gamma qui aurait pour paramètre $\mathbb{N}$ *et* $\frac{1}{2}$ c'est-à-dire une loi $G(\mathbb{N}, \frac{1}{2})$. Donc il est intéressant d'énoncer le :

Théorème 3.
Si $T \rightsquigarrow G(\beta, \lambda)$ *et* h *un réel strictement positif, alors* $hT \rightsquigarrow G\left(\beta, \frac{\lambda}{h}\right)$.

Preuve.
$T \rightsquigarrow G(\beta, \lambda) \Rightarrow f(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}$.
Posons $Z = hT$, on a :

$$\begin{aligned}\mathbb{P}(Z \leq z) = \mathbb{P}(hT \leq z) = P\left(T \leq \frac{z}{h}\right) = F_T\left(T \leq \frac{z}{h}\right) = \frac{1}{h} f_T\left(T \leq \frac{z}{h}\right) \\ = \frac{1}{h} \frac{\lambda^\beta \left(\frac{z}{h}\right)^{\beta-1} \exp\left(-\lambda \frac{z}{h}\right)}{\Gamma(\beta)} = \frac{\left(\frac{\lambda}{h}\right)^\beta z^{\beta-1} \left(-\frac{\lambda}{h} z\right)}{\Gamma(\beta)}.\end{aligned}$$

D'où $Z \rightsquigarrow G\left(\beta, \frac{\lambda}{h}\right)$. □

Construction de la fonction pivotale
Si $T_1, ..., T_n$ est un n-échantillon extrait de $T \rightsquigarrow G(\beta, \lambda)$ alors

$$\sum_{i=1}^{n} T_i \rightsquigarrow G(n\beta, \lambda).$$

Ainsi on a :

$$2\lambda \sum_{i=1}^{n} T_i \rightsquigarrow G\left(n\beta, \frac{\lambda}{2\lambda}\right) = G\left(n\beta, \frac{1}{2}\right) = G\left(\frac{2n\beta}{2}, \frac{1}{2}\right).$$

Donc $2\lambda \sum_{i=1}^{n} T_i \rightsquigarrow \chi^2_{2\beta n}$. La fonction pivotale recherchée est :

$$Y = 2\lambda \sum_{i=1}^{n} T_i.$$

Donc nous pouvons construire un intervalle de confiance asymptotique pour λ à partir de cette fonction pivotale :

$$\mathbb{P}\left(\chi^2_{2\beta n}\left(\frac{\alpha}{2}\right) \leq 2\lambda \sum_{i=1}^{n} T_i \leq \chi^2_{2\beta n}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

avec $\chi^2_{2\beta n}\left(\frac{\alpha}{2}\right)$ et $\chi^2_{2\beta n}\left(1-\frac{\alpha}{2}\right)$ les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ respectivement pour la loi du khi-deux à $2\beta n$ degrés de liberté.
En posant $c_1 = \chi^2_{2\beta n}\left(\frac{\alpha}{2}\right)$ et $c_2 = \chi^2_{2\beta n}\left(1-\frac{\alpha}{2}\right)$, on a :

$$\mathbb{P}\left(c_1 \leq 2\lambda \sum_{i=1}^{n} T_i \leq c_2\right) = 1-\alpha \Leftrightarrow \mathbb{P}\left(\frac{c_1}{2\sum_{i=1}^{n} T_i} \leq \lambda \leq \frac{c_2}{2\sum_{i=1}^{n} T_i}\right) = 1-\alpha.$$

On en déduit que

$$IC_{1-\alpha}(\lambda) = \left[\frac{c_1}{2\sum_{i=1}^{n} T_i}, \frac{c_2}{2\sum_{i=1}^{n} T_i}\right].$$

Ainsi pour β donné, on aura l'intervalle de confiance pour λ.

Conclusion

Nous avons utilisé l'EMV et l'EMM pour estimer les paramètres de la loi Gamma. Du point de vue théorique, nous n'avons pas pu trouver une expression explicite des EMV. C'est la raison pour laquelle nous avons utilisé la méthode itérative de Newton Raphson pour notre étude.
L'intervalle de confiance a été obtenu grâce à l'approche d'une fonction pivotale et à celle des EMV.

Chapitre 5

Application numérique

Dans cet chapitre, nous allons d'abord tracer les fonctions associées aux lois utilisées dans le chapitre 2.
Nous allons ensuite de façon spécifique estimer le paramètre θ de la loi Gamma, voir ses propriétés et faire une étude comparative de la performance des EMV et EMM. L'application numérique se fera avec le logiciel R et se basera sur des données simulées et sur des données réelles.

5.1 Exemples de loi

Dans cette section, nous présentons les graphes des fonctions associées à quelques lois usuelles utilisées en analyse de survie : la densité de probabilité, la fonction de survie, la fonction de hasard et la fonction de hasard cumulée.

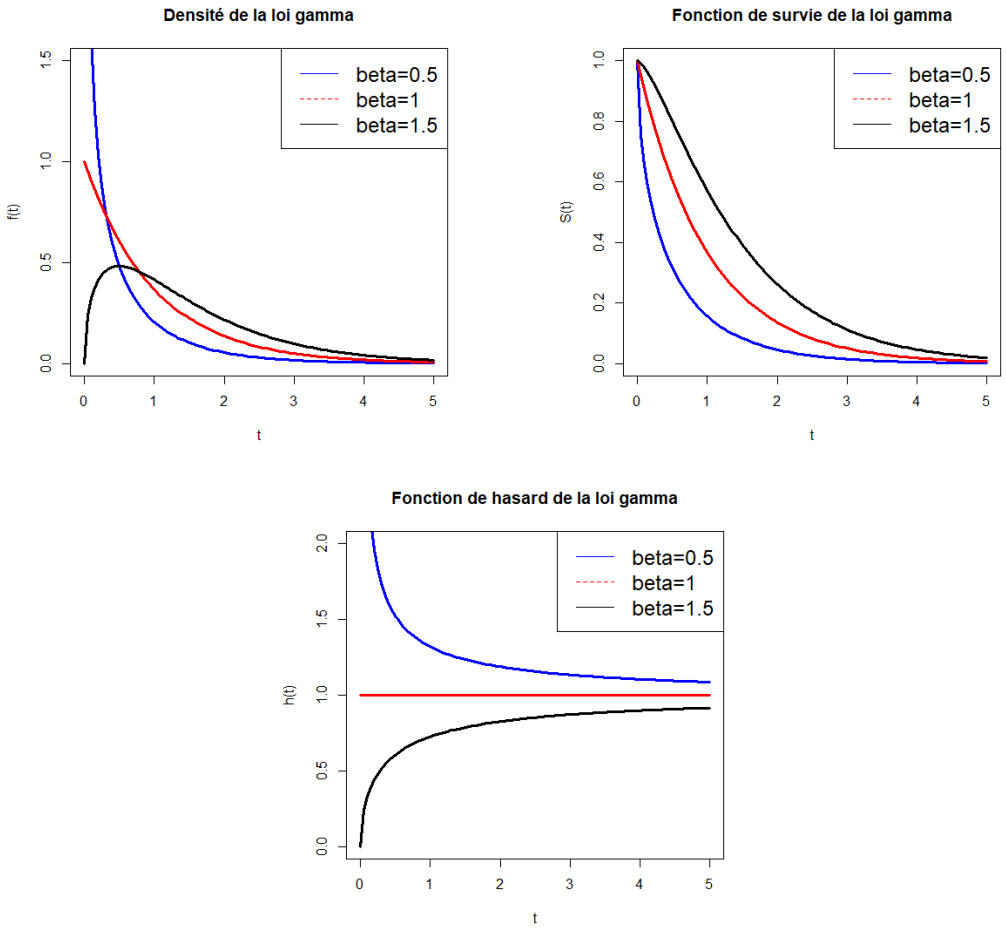

FIGURE 5.1 – Densité, fonction de survie et fonction de hasard de loi Gamma $G(\beta, 1)$ pour $\beta = 0.5, 1 \quad et \quad 1.5$.

La figure 5.1 illustre, pour différentes valeurs de β, les fonctions densité de probabilité, les fonctions de survie et les fonctions de hasard de la loi Gamma.

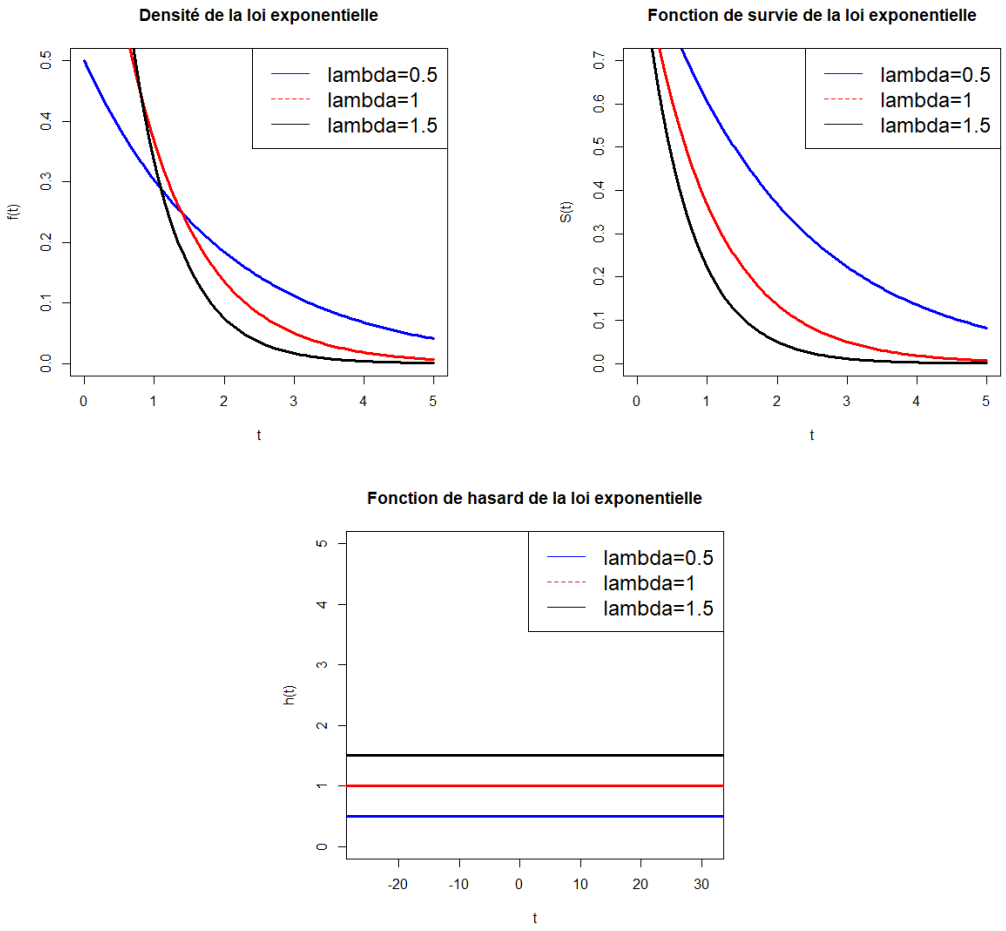

FIGURE 5.2 – Densité, fonction de survie et fonction de hasard de la loi exponentielle $\mathcal{E}(\lambda)$ pour $\lambda = 0.5, 1$ *et* 1.5.

La figure 5.2 représente les densités de probabilité, les fonctions de survie et les fonctions de hasard de plusieurs lois exponentielle.

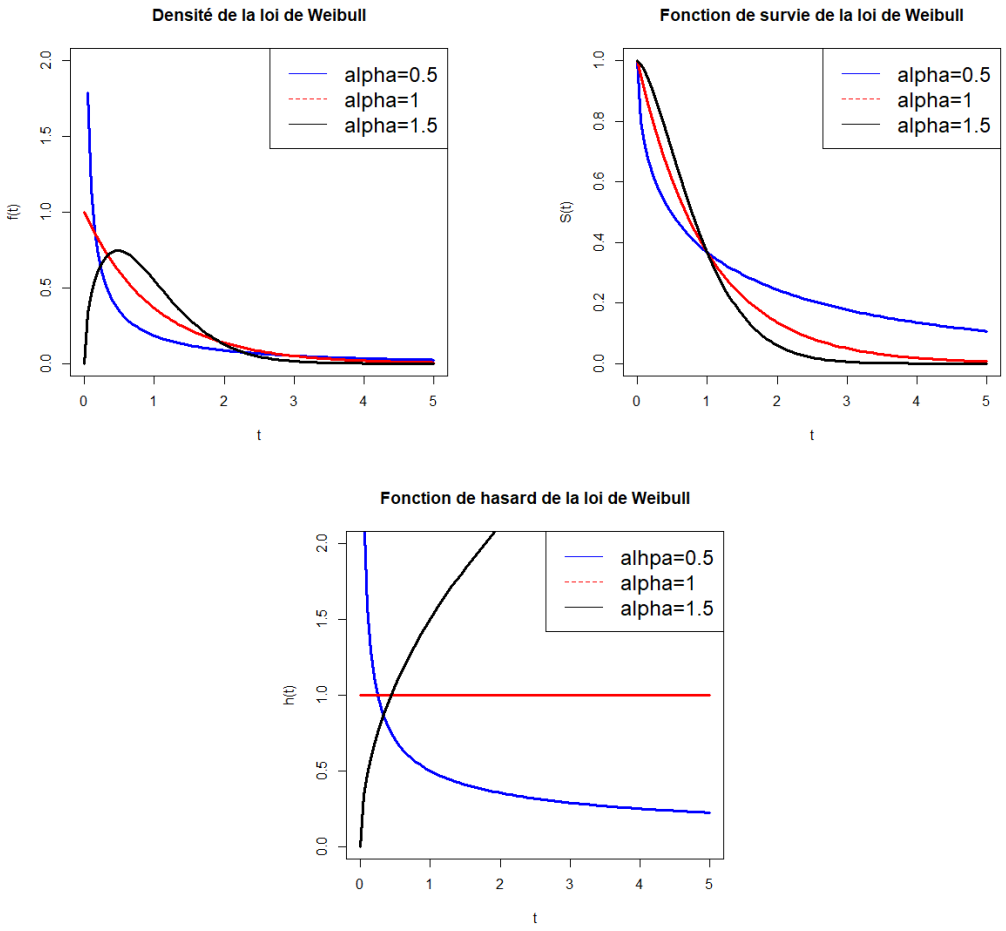

FIGURE 5.3 – Densité, fonction de survie et fonction de hasard de la loi de Weibull $W(\alpha, 1)$ pour $\alpha = 0.5, 1 \ \ et \ \ 1.5$.

La figure 5.3 illustre, pour un paramètre d'échelle $\lambda = 1$ fixé et différentes valeurs de α, les fonctions densité de probabilité, les fonctions de survie et les fonctions de hasard de la loi de Weibull.

5.2 Estimation paramétrique dans le cas non censuré avec des données simulées

L'estimation du paramètre $\theta = (\beta, \lambda)$ est obtenue de la manière suivante : nous générons un échantillon de taille $n = 100$ de loi Gamma de paramètres $\beta = 2$ et $\lambda = 4$ et nous estimons θ par la méthode du maximum de vraisem-

blance et par la méthode des moments.
Soient $\tilde{\theta} = (\tilde{\beta}, \tilde{\lambda})$ l'EMM de θ et $\hat{\theta} = (\hat{\beta}, \hat{\lambda})$ l'EMV de θ.

5.2.1 Méthode du maximum de vraisemblance

Avec une taille d'échantillon $n = 100$, pour déterminer l'EMV $\hat{\theta_n}$, nous utilisons trois approches à savoir :

- approche 1 : La méthode de Newton Raphson,
- approche 2 : La fonction **optim** de R [1],
- approche 3 : La commande **mle** de R [2].

	$\hat{\theta}_n$	
	$\hat{\beta}_n$	$\hat{\lambda}_n$
Approche 1	2.347	5.191
Approche 2	2.035	3.875
Approche 3	1.975	3.593

TABLE 5.1 – Les EMV pour les paramètres d'une loi gamma pour $n = 100$ avec des approches différentes dans le cas non censuré.

5.2.2 Méthode des moments

Pour la méthode des moments on obtient les estimateurs suivants pour $n = 100$:

$\tilde{\beta}_n$	$\tilde{\lambda}_n$
1.665	3.151

TABLE 5.2 – Les EMM pour les paramètres d'une loi Gamma pour $n = 100$ dans le cas non censuré.

1. On peut maximiser la log-vraisemblance grâce à la fonction **optim de R**. Il faut cependant faire attention au fait que la fonction **optim** minimise la solution, on entrera donc l'opposé de la log-vraisemblance pour maximiser les valeurs et obtenir les estimateurs.
2. Elle permet de calculer les estimateurs du maximum de vraisemblance dans un modèle paramétrique spécifié et aussi de déterminer les intervalles de confiance. Pour obtenir l'EMV à l'aide de la commande mle, il faut d'abord charger le package **stats4**.

	EMV			
Erreur commise	approche 1	approche 2	approche 3	EMM
|Estimateur−2|	0.347	0.035	0.025	0.335
|Estimateur−4|	1.191	0.125	0.407	0.849

TABLE 5.3 – Erreur absolue des EMV et EMM.

D'après les tableaux 5.1, 5.2 et 5.3, nous remarquons que les estimations obtenues sont proches des vraies valeurs. Le modèle Gamma est donc satisfaisant pour notre échantillon de taille n.

5.2.3 Propriétés d'estimateurs des paramètres de la loi gamma

Dans cette section, les propriétés d'estimateurs abordés sont le biais et l'erreur quadratique moyenne.

Calcul du biais pour les EMV

Ici, nous avons simulé m échantillons de taille n.
Pour différentes valeurs de m et de n, on a les résultats suivants du biais de $\hat{\beta}_n$:

Biais de $\hat{\beta}_n$	n=10	n=50	n=200	n=500	n=1000
m=10	0.64	-0.016	0.034	-0.026	0.013
m=50	0.33	0.129	0.063	0.0006	0.0038
m=200	0.59	0.084	0.036	0.021	0.015
m=500	0.78	0.10	0.037	0.011	0.004
m=1000	0.81	0.10	0.025	0.017	0.010

TABLE 5.4 – Biais de l'EMV $\hat{\beta}_n$ de la loi Gamma sur différents échantillons de taille différentes de la loi G(2,4) dans le cas non censuré.

Le tableau 5.4 illustre que les paramètres n et m n'ont pas la même influence car n agit sur la convergence asymptotique du biais vers 0 tandis que m a une influence sur le fait d'avoir des estimateurs de β et λ proches entre différentes simulations d'échantillons.
De plus l'EMV $\hat{\beta}$ est un estimateur asymptotiquement sans biais de β.

Calcul de l'erreur quadratique moyenne de l'EMV

Nous obtenons les résultats suivants de l'EQM de $\hat{\beta}_n$:

EQM de $\hat{\beta}_n$	n=10	n=50	n=200	n=500	n=1000
m=10	3.12	0.070	0.041	0.010	0.006
m=50	1.32	0.209	0.039	0.012	0.0063
m=200	2.74	0.179	0.035	0.014	0.0083
m=500	2.65	0.174	0.037	0.013	0.0070
m=1000	3.15	0.173	0.038	0.013	0.0070

TABLE 5.5 – L'EQM de l'EMV $\hat{\beta}$ de loi Gamma pour m différents échantillons de taille n différentes dans le cas non censuré.

Dans le tableau 5.5, nous remarquons que l'EQM de $\hat{\beta}_n$ est une fonction décroissante de n mais pas de m. De plus, l'EQM tend vers 0 quand n tend vers l'infini, donc $\hat{\beta}_n$ est un estimateur convergent de β.

Vérification empirique des propriétés de l'EMV

Dans cette section, nous vérifions que l'EMV est asymptotiquement gaussien. Pour ce faire, il suffit juste de vérifier qu'il suit une loi normale. Donc pour chaque $\hat{\beta}_n$ et $\hat{\lambda}_n$, on va tracer le graphe de probabilité pour voir si les points du nuage sont approximativement alignés (c'est-à-dire qu'ils s'ajustent correctement à la droite de Henry).
Pour $n = 100$, $m = 1000$, $\beta = 2$ et $\lambda = 4$, on obtient les graphes suivants :

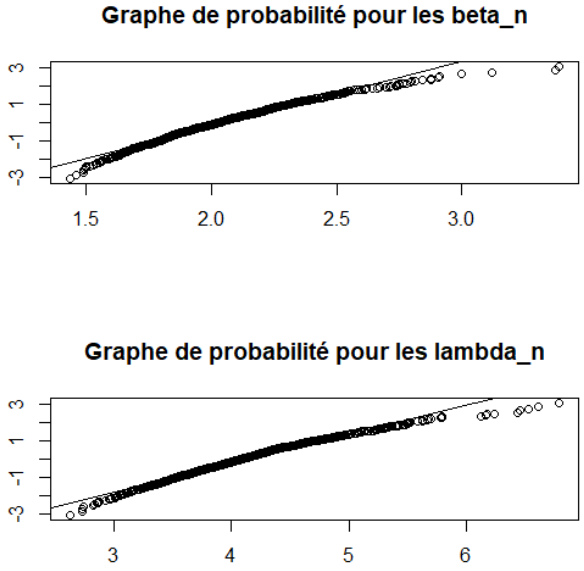

FIGURE 5.4 – Graphes de probabilité des EMV $\hat{\beta}_n$ et de $\hat{\lambda}_n$ d'une loi Gamma. Estimations sur 1000 échantillons de taille 100 de la loi G(2,4).

Dans la figure 5.8, nous voyons que les points nuages sont quasiment alignés à la droite de Henry, ce qui signifie alors que l'EMV est effectivement asymptotiquement gaussien.

Calcul du biais pour les EMM

Pour l'EMM, on fera la même procédure que les EMV.
On obtient les résultats suivants :

Biais de $\tilde{\beta}_n$	n=10	n=50	n=200	n=500	n=1000
m=10	0.97	0.28	0.0021	-0.015	-0.017
m=50	1.11	0.24	0.0921	0.030	-0.009
m=200	1.07	0.127	0.032	0.022	-0.0038
m=500	1.007	0.227	0.063	0.029	0.015
m=1000	1.05	0.19	0.036	0.015	0.0082

TABLE 5.6 – Biais de l'estimateur $\tilde{\beta}$ par la méthode des moments d'une loi Gamma pour différents échantillons m de taille n différentes dans le cas non censuré.

Nous pouvons en déduire, d'après le tableau 5.6, que l'EMM est un estimateur asymptotiquement sans biais de β. En effet le biais tend vers 0 quand $n \to \infty$.

Calcul de l'erreur quadratique moyenne de l'EMM

Nous avons les erreurs quadratiques moyennes de $\tilde{\beta}_n$ suivantes :

EQM de $\tilde{\beta}_n$	n=10	n=50	n=200	n=500	n=1000
m=10	0.67	0.27	0.025	0.028	0.010
m=50	3.47	0.39	0.056	0.021	0.015
m=200	3.799	0.35	0.065	0.024	0.012
m=500	4.11	0.27	0.066	0.026	0.0117
m=1000	3.90	0.295	0.061	0.0229	0.0113

TABLE 5.7 – L'EQM de l'estimateur $\tilde{\beta}_n$ par la méthode des moments d'une loi Gamma pour différents échantillons m de taille n différentes dans le cas non censuré.

D'après le tableau 5.7, l'EMM $\tilde{\beta}$ est un estimateur convergent de β.

Vérification empirique des propriétés de l'EMM

En faisant la même démarche que dans la cas des EMV, nous obtenons les graphes suivants :

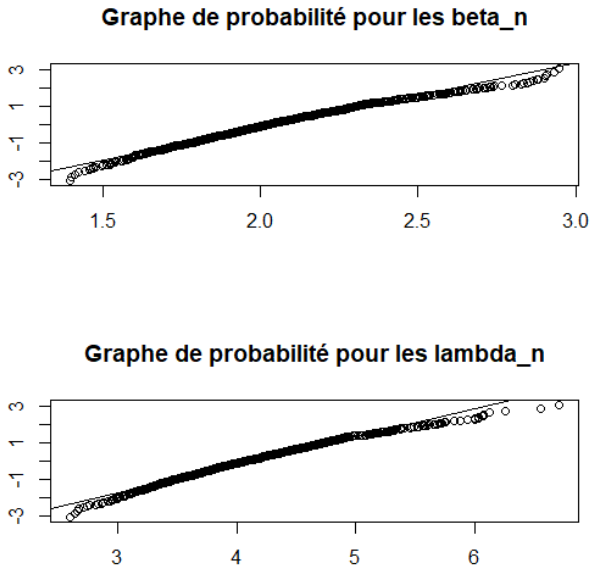

FIGURE 5.5 – Graphe de probabilite des estimateurs par la méthode des moments des paramètres $\tilde{\beta}$ et de $\tilde{\lambda}$ d'une loi Gamma pour 1000 échantillons de taille 100 de la loi G(2,4).

D'après le tracé de la figure 5.5, nous pouvons en déduire que les nuages de point s'ajustent correctement à la droite de Henry. Donc l'EMM $\tilde{\beta}$ est bien asymptotiquement gaussien.

Comparaison des résultats des deux estimateurs

Dans cette section, nous étudions la performance des estimateurs obtenus par l'EMV et l'EMM. Pour faire la comparaison des estimateurs étudiés, on peut recourir à l'EQM , le biais de l'estimateur, la variance...
Nous avons fixé les valeurs de β et λ ($\beta = 2$ et $\lambda = 4$) et nous avons tiré 1000 échantillons de taille 100 de la loi $G(\beta, \lambda)$. Nous calculons pour chaque échantillon les quatre estimations $\hat{\lambda}, \hat{\beta}, \tilde{\lambda}$ *et* $\tilde{\beta}$. Les résultats sont donnés dans le tableau 5.8.

Estimateurs	$\hat{\beta}$	$\hat{\lambda}$	$\tilde{\beta}$	$\tilde{\lambda}$
Biais	0.052	0.12	0.084	0.19
Variance	0.068	0.35	0.13	0.61
EQM	0.079	0.39	0.13	0.64

TABLE 5.8 – Biais et variances des EMV et des EMM pour les paramètres d'une loi Gamma. Estimation sur 1000 échantillons de taille 100 de la loi $G(2, 4)$.

D'après le tableau 5.8, le biais de $\hat{\beta}$ (respectivement $\hat{\lambda}$) est inférieur au biais de $\tilde{\beta}$ (respectivement $\tilde{\lambda}$) et de même pour les variances et l'EQM. Donc nous pouvons conclure que l'EMV est meilleur que l'EMM dans ce cas.

5.2.4 Intervalle de confiance

Pour obtenir les intervalles de confiance des paramètres β et λ de la loi gamma numériquement pour un échantillon de taille n à 95%, on utilise deux approches en simulant un échantillon de taille n de la loi Gamma de paramètres $\beta = 2$ et $\lambda = 4$:

- Approche 1 : La commande **mle** de R
- Approche 2 : La fonction pivotale.
  On obtient les résultats suivants :

Approche 1	$n = 100$	$n = 10000$
$\beta \in$	$[1.64, 2.76]$	$[1.98, 2.09]$
$\lambda \in$	$[2.94, 5.27]$	$[3.98, 4.22]$
Approche 2	$n = 100$	$n = 1000$
$\lambda \in$	$[3.251, 4.29]$	$3.95, 4.06]$

TABLE 5.9 – Les intervalles de confiance des paramètres β et λ simulés pour $n = 100$ et $n = 10000$ d'une loi Gamma $G(2, 4)$.

D'après le tableau 5.9, nous voyons que plus la taille de l'échantillon est grande plus l'intervalle de confiance se rétrécit (c'est-à-dire plus la taille de l'échantillon augmente mieux l'amplitude est petite).
Nous avons le graphe suivant de l'intervalle de confiance pour $n = 100$ d'une loi Gamma $G(\beta = 2, \lambda = 4)$:

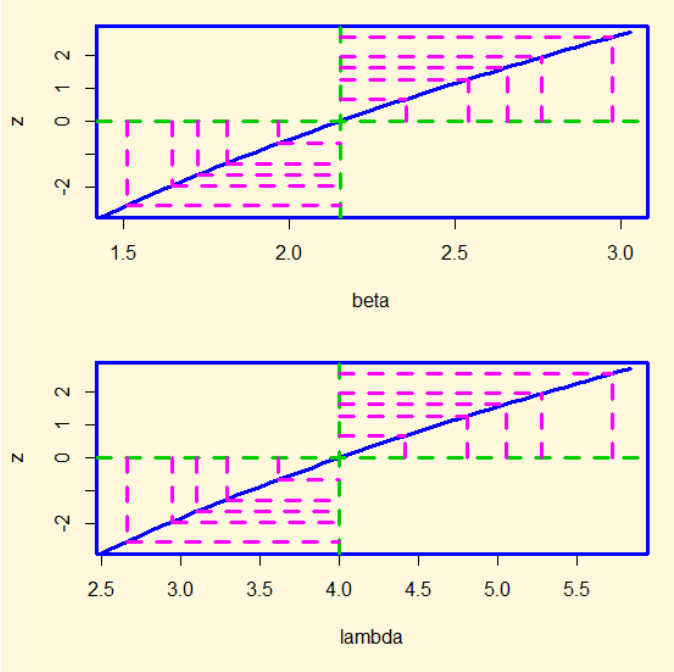

FIGURE 5.6 – Graphe de l'intervalle de confiance des paramètres β et λ pour $n = 100$ dans le cas des données non censurées.

La figure 5.6 illustre, l'intervalle de confiance des paramètres $\beta = 2$ et $\lambda = 4$.

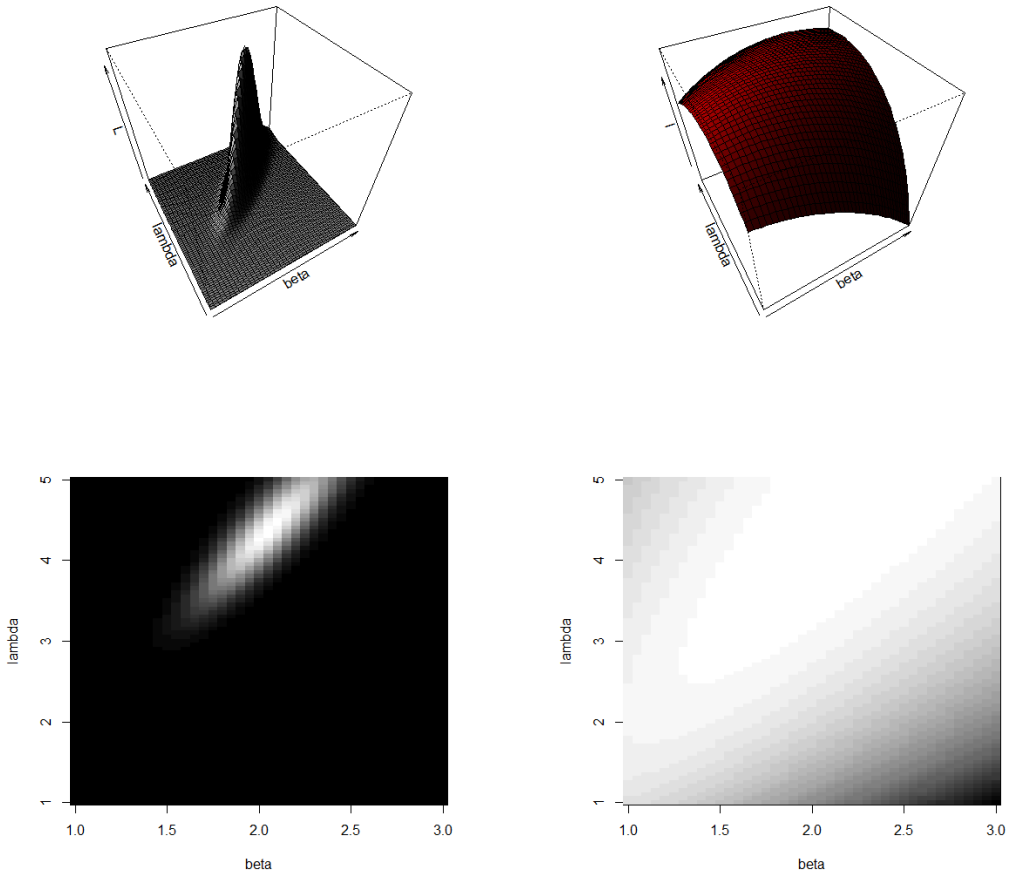

FIGURE 5.7 – Tracés de la fonction vraisemblance avec son image (à gauche) et de la log-vraisemblance avec son image (à droite) d'une loi gamma $G(2, 4)$ pour $n = 100$ dans le cas non censuré.

5.3 Estimation paramétrique dans le cas censuré avec des données simulées

Dans cette partie, nous estimons le paramètre θ numériquement pour un échantillon qui est censuré à droite de type I.
Nous générons un échantillon de taille $n = 100$ dont certaines observations sont censurées à droite : (T_i, D_i) avec T_i qui suit une loi Gamma de paramètres $\beta = 2$ et $\lambda = 4$. On suppose que t_{97}, t_{98}, t_{99} et t_{100} sont censurées.

5.3.1 Méthode du maximum de vraisemblance

Pour déterminer l'estimateur $\hat{\theta}$, nous utilisons deux approches :

- approche 1 : La fonction **optim** de R,
- approche 2 : la commande **mle**.

	$\hat{\theta}_{100}$	
	$\hat{\beta}_{100}$	$\hat{\lambda}_{100}$
Approche 1	1.93	3.65
Approche 2	2.05	3.90

TABLE 5.10 – Les estimateurs de $\hat{\theta}$ pour $n = 100$ en cas de censure.

5.3.2 Méthode des moments

Pour la méthode des moments, nous avons les estimateurs suivants pour $n = 100$ en cas de la censure :

$\tilde{\beta}_{100}$	$\tilde{\lambda}_{100}$
2.026	3.66

TABLE 5.11 – Les estimateurs de $\tilde{\theta}$ pour $n = 100$ dans le cas de la censure

	EMV		
Erreur commise	approche 1	approche 2	EMM
|Estimateur−2|	0.07	0.05	0.026
|Estimateur−4|	0.35	0.1	0.34

TABLE 5.12 – Erreur absolue des EMV et EMM

Nous avons la même remarque que dans le cas des données non censurées avec données simulées.

5.3.3 Propriétés d'estimateurs des paramètres de la loi gamma

Dans cette partie, nous avons tiré 1000 échantillons de taille n.

Calcul du biais et de l'erreur quadratique des EMV et EMM

Pour $m = 1000$ échantillons de taille $n = 100$, nous avons :

Estimateurs	$\hat{\beta}$	$\hat{\lambda}$	$\tilde{\beta}$	$\tilde{\lambda}$
Biais	0.0539	0.0092	0.0928	0.22
EQM	0.093	0.44	0.13	0.62

TABLE 5.13 – Biais et variances des EMV et des EMM pour les paramètres d'une loi Gamma. Estimation sur 1000 échantillons de taille 100 en cas de censure.

D'après le tableau 5.13, le biais de $\hat{\beta}$ (respectivement $\hat{\lambda}$) est inférieur au biais de $\tilde{\beta}$ (respectivement $\tilde{\lambda}$) et de même pour les variances et l'EQM. Donc nous pouvons conclure que l'EMV est meilleur que l'EMM en cas d'un échantillon censuré.
De plus nous constatons que les estimateurs $\hat{\theta} = (\hat{\beta}, \hat{\lambda})$ et $\tilde{\theta} = (\tilde{\beta}, \tilde{\lambda})$ sont asymptotiquement sans biais et convergents.

Vérification empirique des propriétés de l'EMV et de l'EMM

En présence de censure, nous avons les graphes de probabilité suivants :

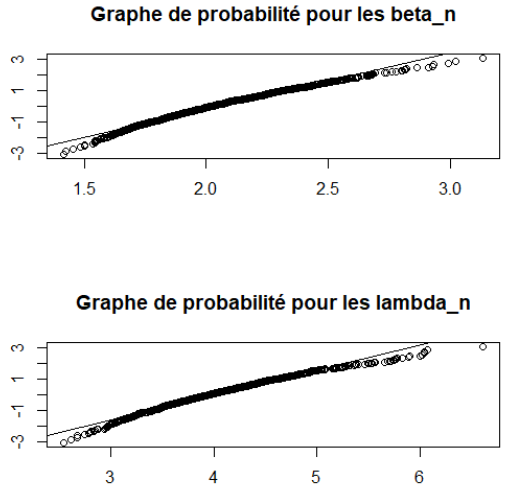

FIGURE 5.8 – Graphes de probabilité des EMV $\hat{\beta}_n$ et de $\hat{\lambda}_n$ d'une loi Gamma. Estimations sur 1000 échantillons de taille 100 de la loi.

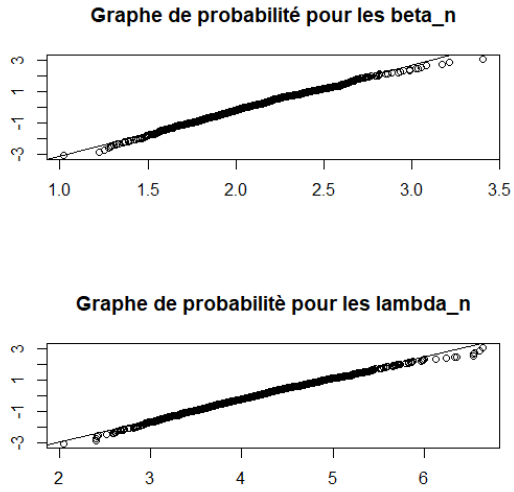

FIGURE 5.9 – Graphes de probabilité des estimateurs par la méthode des moments des paramètres $\tilde{\beta}$ et de $\tilde{\lambda}$ d'une loi Gamma pour 1000 échantillons de taille 100.

D'après les figures 5.8 et 5.9, nous pouvons en déduire que les nuages de point s'ajustent correctement à la droite de henry. Donc les EMV et EMM sont bien asymptotiquement gaussien.

5.3.4 Intervalle de confiance

A 95%, on obtient les intervalles de confiance des paramètres de la loi Gamma pour $n = 100$, $\beta = 2$ et $\lambda = 4$:

La commande **mle**	$n = 100$
$\beta \in$	$[1.569, 2.644]$
$\lambda \in$	$[2.846, 5.179]$

TABLE 5.14 – Les intervalles de confiance pour $n = 100$ de la loi Gamma avec $\beta = 2$ et $\lambda = 4$ en cas de censure.

Ainsi, pour $n = 100$, $\beta = 2$ et $\lambda = 4$ on a le graphe suivant :

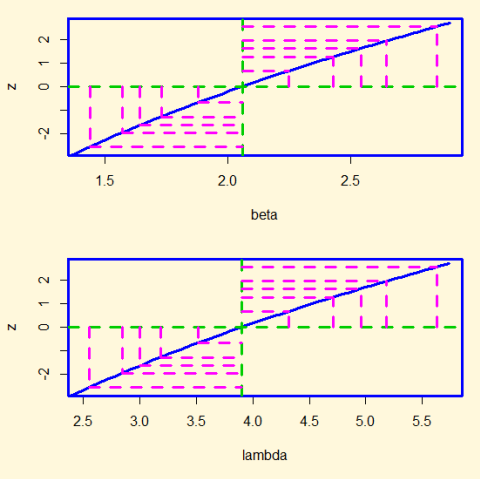

FIGURE 5.10 – Graphe de l'intervalle de confiance pour les EMV avec des données censurées

5.4 Estimation paramétrique avec des données réelles non censurées

Dans le cas des données non censurées, nous utilisons les durées de rémission en mois de dix patients atteints de mélanome qui ont obtenu une rémission après une chirugie et une thérapie.
Les données t_i obtenues sont les suivantes : $5, 8, 10, 11, 15, 20, 21, 23, 30, 40$.
On a le résumé de ces données ci-après :

Minimum	Maximum	Moyenne	Variance
5	40	18.30	105.61

TABLE 5.15 – Statistiques des durées de rémission.

D'après les relations (3.18) et (3.19), les paramètres β et λ de la loi sont donnés respectivement par : $\beta = 3.171006$ et $\lambda = 0.173279$. Pour vérifier que ces données suivent une loi Gamma, nous utilisons la méthode de l'approximation de la densité.[3]
Nous avons le graphique suivant :

3. On utilise une estimation de la densité inconnue. Puis on superpose les valeurs de la "densité" associée à la loi gamma en estimant éventuellement les paramètres inconnus de celle-ci.

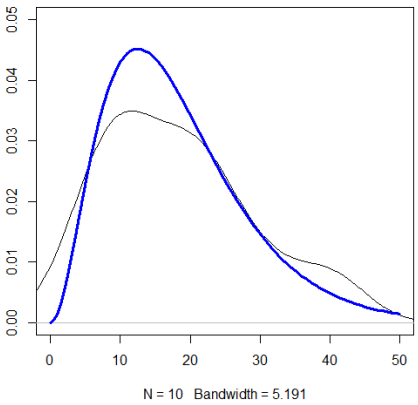

FIGURE 5.11 – Méthode de l'approximation de la densité des durées de rémission.

La figure 5.11 représente la méthode de l'approximation de la densité (la courbe en noir représente la densité des t_i et la courbe en blue la densité des t_i associée à la loi Gamma de paramètres $\beta = 3.171006$ et $\lambda = 0.173279$). Ainsi, les différences observées laissent penser que T suit bien une loi gamma.

5.4.1 Méthode du maximum de vraisemblance

Pour obtenir les EMV, nous utiliserons trois approches :
- ▶ approche 1 : méthode de Newton-Raphson ;
- ▶ approche 2 : la fonction optim de R ;
- ▶ approche 3 : la commande mle de R.

	$\hat{\beta}$	$\hat{\lambda}$
Approche 1	3.087846	0.1687348
Approche 2	3.0885171	0.1687814
Approche 3	3.0880761	0.1687474

TABLE 5.16 – Les EMV pour les paramètres d'une loi Gamma avec des approches différentes des durées de rémission.

5.4.2 Méthode des moments

Nous obtenons les estimateurs de β et de λ donnés dans la table 5.17 :

$\tilde{\beta}$	$\tilde{\lambda}$
3.171007	0.173279

TABLE 5.17 – Les EMM pour les paramètres d'une loi Gamma des durées de rémission.

	EMV			
Erreur commise	approche 1	approche 2	approche 3	EMM
$\lvert$Estimateur$-\beta\rvert$	0.083	0.082	0.082	10^{-6}
$\lvert$Estimateur$-\lambda\rvert$	0.004	0.004	0.004	0

TABLE 5.18 – Erreur absolue des EMV et EMM

D'après les tableaux 5.21, 5.17 et 5.18, nous remarquons que les estimateurs trouvés par la méthode des moments et la méthode du maximum de vraisemblance sont proches des vraies valeurs.

Calcul du biais et de l'erreur quadratique moyenne

Estimateurs	$\hat{\beta}$	$\hat{\lambda}$	$\tilde{\beta}$	$\tilde{\lambda}$
Biais	-0.08248889	-0.004497603	5.33472210^{-7}	4.55449310^{-8}
EQM	0.006804417	2.02284310^{-6}	2.84592610^{-13}	2.07434110^{-15}

TABLE 5.19 – Biais et EQM des EMV et EMM pour les paramètres d'une loi Gamma des durées de rémission.

D'après le tableau 5.19, nous voyons que l'EQM de l'EMV est plus grande que l'EQM de l'EMM. Ainsi, l'EMM est meilleur que l'EMV pour les données réelles non censurées.

5.4.3 Intervalle de confiance

Les intervalles de confiance de β et λ à 95% pour les données réelles non censurées sont donnés respectivement par :

$$IC_\beta = [1.195, 6.459] \quad \text{et} \quad IC_\lambda = [0.056, 0.369].$$

La représentation graphique est :

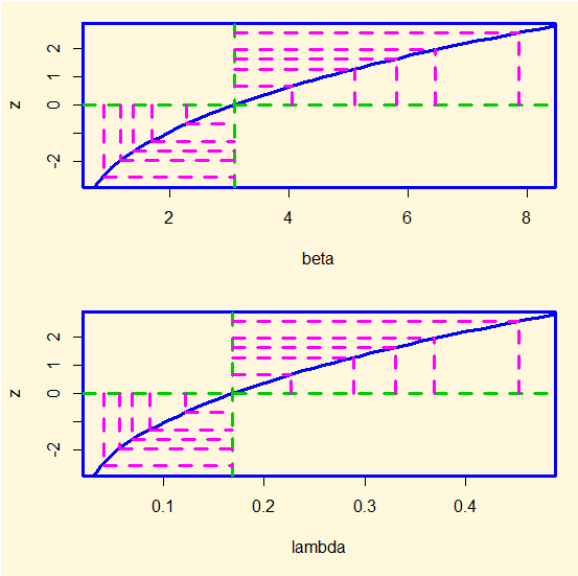

FIGURE 5.12 – Graphe de l'intervalle de confiance pour les EMV des données réelles non censurées.

5.5 Estimation paramétrique avec des données réelles censurées

Considérons une expérience avec $n = 34$ animaux. Les données suivantes sont les durées de vie t_i en semaines de 34 animaux (cf.[6]) : 3, 4, 5, 6, 6, 7, 8, 8, 9, 9, 9, 10, 10, 11, 11, 11, 13, 13, 13, 13, 13, 17, 17, 19, 19, 25, 29, 33, 42, 42, 52, 52^+, 52^+, 52^+.
Le signe "+" indique que la donnée est censurée. L'étude s'achève lorsque 31 animaux sont morts et les 3 autres sont sacrifiés.
Le tableau suivant illustre le résumé de ces données :

Minimum	Maximum	Moyenne	Variance
3	52	18.91	234.9628

TABLE 5.20 – Statistiques des durées de vie t_i en semaines de 34 animaux.

Les durées de vie t_i en semaines de 34 animaux suivent une loi Gamma de

paramètres $\beta = 1.521892$ et $\lambda = 0.08048083$.
En utilisant la méthode de l'approximation de la densité, on a le graphique suivant :

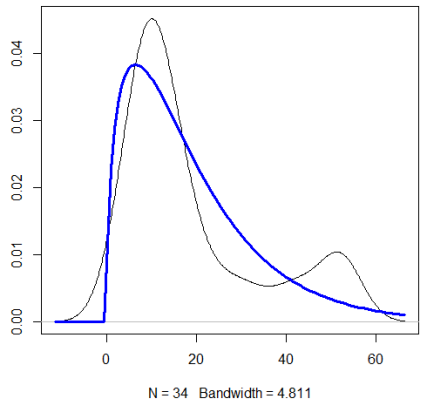

FIGURE 5.13 – Méthode de l'approximation de la densité des durées de vie t_i en semaines de 34 animaux.

D'après la figure 5.13, les différences observées laissent penser que T suit bien une loi Gamma.

5.5.1 Méthode du maximum de vraisemblance

Pour obtenir les EMV, nous utilisons deux approches :
► approche 1 : la fonction optim de R ;
► approche 2 : la commande mle.

	$\hat{\beta}$	$\hat{\lambda}$
Approche 1	1.62342911	0.08062225
Approche 2	1.62428791	0.08068949

TABLE 5.21 – Les EMV pour les paramètres d'une loi Gamma avec des approches différentes d'une expérience de 34 animaux.

5.5.2 Méthode des moments

On obtient les estimateurs de β et de λ donnés dans la table 5.17 :

$\tilde{\beta}$	$\tilde{\lambda}$
1.52217645	0.08048833

TABLE 5.22 – Les EMM pour les paramètres de la loi Gamma d'une expérience de 34 animaux.

	EMV		
Erreur commise	approche 1	approche 2	EMM
|Estimateur$-\beta$|	0.101	0.102	0.00028
|Estimateur$-\lambda$|	0.00014	0.0002	7.510^{-6}

TABLE 5.23 – Erreur absolue des EMV et EMM

D'après les tableaux 5.21, 5.22 et 5.23, nous voyons que les deux estimateurs sont presque sûrement égaux aux vraies valeurs.

Biais et EQM

Estimateurs	$\hat{\beta}$	$\hat{\lambda}$	$\tilde{\beta}$	$\tilde{\lambda}$
Biais	0.1015371	0.0001414235	0.000284447	7.510^{-6}
EQM	0.01030978	2.10^{-8}	8.0910^{-8}	5.63210^{-11}

TABLE 5.24 – Biais et EQM des EMV et EMV pour les paramètres de la loi Gamma d'une expérience de 34 animaux.

D'après la table 5.24, nous remarquons que l'EQM de l'EMV est plus grande que l'EQM de l'EMM. Donc l'EMM est meilleur que l'EMV dans le cas des données réelles censurées.

5.5.3 Intervalle de confiance

Avec la commande **mle**, les intervalles de confiance de β et de λ sont :

$$IC_\beta = [0.996, 2.500] \quad \text{et} \quad IC_\lambda = [0.043, 0.133].$$

Nous avons le graphe suivant :

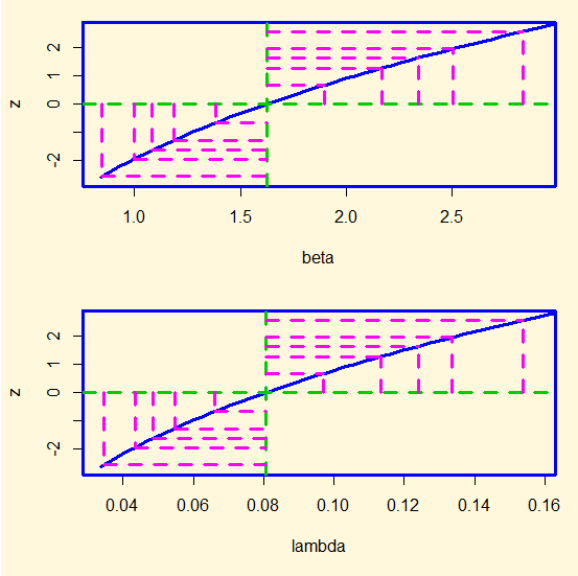

FIGURE 5.14 – Graphe de l'intervalle de confiance pour les EMV des données réelles censurées.

Conclusion

D'après les études de simulations, nous remarquons que les EMV et les EMM présentent des comportements satisfaisants à savoir la convergence, le biais, l'erreur quadratique moyenne pour les données censurées et non censurées. Une étude comparative a été faite en calculant le biais et l'erreur quadratique moyenne de l'estimateur. De ce fait, on a pu conclure que l'EMV est meilleur que l'EMM pour notre échantillon de taille n et l'EMM est meilleur que l'EMV pour des données réelles.

Conclusion générale

Dans ce travail, nous nous sommes intéressés à l'estimation paramétrique des variables de durée suivant la loi Gamma bidimensionnelle.
L'estimation des paramètres du modèle Gamma a été effectuée par la méthode du maximum de vraisemblance et la méthode des moments pour un échantillon de taille n avec des données simulées et des réelles.
Des données simulées de la loi Gamma ont montré que l'EMV est meilleur que l'EMM en termes d'erreur quadratique moyenne. Et pour les données réelles, l'EMM est meilleur que l'EMV.
En perspectives, nous pouvons élargir ce travail sur l'estimation non paramétrique et semi-paramétrique des variables de durée.

Bibliographie

[1] Catherine Huber. Modèles pour des durées de survie, disponible sur **https ://docplayer.fr/423140-Modeles-pour-des-durees-de-survie.html**.

[2] Cox D.R., Oakes D.(1984). *Analysis of survival data.* London, Edition Chapman and Hall.

[3] David G. Kleinbaum, Mitchel Klein. (2012). *Survival Analysis : A Self-Learning Text*, Third Edition, Springer.

[4] Diouf A. (2019). *Modélisation des effets de traitements fongiques sur la dynamique de population d'organismes sentinelles*, Thèse de doctorat en Mathématiques et Applications, Université Assane Seck de Ziguinchor.

[5] Droesbeke J.J., Fichet B., Tassi P. (1989). *Analyse statistique des durées de vie.* Paris : Economica.

[6] Elisa T. Lee, John Wenyu Wang. (2013). *Statistical Methods for Survival Data Analysis*, Third Edition, Wiley.

[7] Hill, C., Com-Nougue C., Kramar A., Moreau T., O'Quigley J., Senoussi R., Chastang C. (1996) . Analyse statistique des données de survie. *Collection : Statistique en biologie et en médecine.* Flammarion Sciences 1996 ; 3ème édition 2000.

[8] Hougaard P.(2000).*Analysis of Multivariate Survival Data.* New York : Springer-Verlag.

[9] Jean-David Fermanian, Modèles de durée, cours ensae troisième année, crest.fr/ckfinder/userfiles/files/Pageperso/fermania/JDF_duree3.pdf.

[10] Klein P., Melvin L. M. (2005) *Survival Analysis. Statistics for Biology and Health.* Springer.

[11] Li S. (1996). Survival analysis. *Marketing Research*, 7(4), 17-23.

[12] Li Jialiang, Ma Shuangge. (2013). *Survival Analysis in Medicine and Genetics*, Chapman and Hall/CRC.

[13] Mai Zhou, *Use software are R to survival analysis and simulation*, disponible sur **http ://www.ms.uky.edu/ mai//Rsurv.pdf**.

[14] Morris H.DeGroot (2012). *Probability and Statistics*, Fourth Edition.

[15] Michel Lejeune, Statistique (2010), *La théorie et ses applications*. Springer, Deuxième édition.

[16] O.Wintenberger, *Statistique Mathématique*, disponible sur wintenberger.fr/cours/L3MAS/StatMathPoly2013.pdf.

[17] Partrat C., Besson J.L. (2004). *Assurance non-vie - Modélisation, simulation*. Paris : Economica.

[18] Planchet F., Thérond P.E. (2006). *Modèles de durée. Applications actuarielles*. Economica.

[19] Planchet F., Thérond P.E. (2011). *Modélisation statistique des phénomènes de durée - applications actuarielles*. Paris : Economica.

[20] Saint Pierre Philippe (2015). *Introduction à l'analyse de survie (analyse des durées de vie)*. Support de cours (perso.math.univ-toulouse.fr/psaintpi/files/2021/04/Cours_Survie_1.pdf).