وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR -ANNABA
UNIVERSITY
UNIVERSITE BADJI MOKHTAR
ANNABA

جامعة باجي مختار
- عنابة -

Faculté des Sciences

Département de Mathématiques

Année : 2017

THÈSE

Présentée en vue de l'obtention du diplôme de **DOCTORAT**

SUR LES TESTS D'AJUSTEMENT

Option
Statistiques

Par
Mme Tilbi Djahida

DIRECTRICE DE THÈSE : Mme Seddik-Ameur Nacira Prof. U.B.M. ANNABA

Devant le jury

PRESIDENT:	Mr Haiour Mohamed	Prof	U.B.M. ANNABA
EXAMINATRICE:	Mme Chadli Assia	Prof	U.B.M. ANNABA
EXAMINATRICE:	Mme Sadki Ourida	Prof	USTHB ALGER
EXAMINATRICE:	Mme Laskri Yamina	Prof	E.S.I.I. ANNABA
EXAMINATEUR :	Mr Maouni Messaoud	M.C.A	Université SKIKDA

Table des matières

Dédicace

A mon papa et ma maman,
A mon mari et ma fille,
A mes frères et ma sœur.

Remerciements

Je suis très reconnaissante à ma directrice de thèse, Madame ***Seddik-Ameur Nacira*** *Professeur à l'université Badji-Mokhtar d'Annaba et membre du laboratoire LaPS, pour m'avoir fait confiance en me proposant ce travail de thèse. Je la remercie sincèrement pour son optimisme, sa gentillesse, ses conseils avisés et son soutien continu. Qu'elle trouve ici mon admiration et mes profonds respects.*

Je tiens à exprimer toute ma gratitude à Monsieur ***Haiour Mohamed*** *Professeur à l'université Badji-Mokhtar d'Annaba, d'avoir accepté de juger mon travail et de présider le jury de ma thèse.*

Je suis très honorée par la présence au sein du jury de ma thèse, en tant qu'examinateurs, de :

- Madame ***Sadki Ourida****, Professeur à l'université des Sciences et de la Technologie* Houari Boumediène *Alger,*

- Madame ***Chadli Assia****, Professeur à l'université Badji-Mokhtar Annaba,*

- Madame ***Laskri Yamina****, Professeur à l'université Badji-Mokhtar Annaba,*

- Monsieur ***Maouni Messaoud****, Maître de conférences A à l'université du 20 Août 1955 Skikda.*

Qu'ils soient vivement remerciés.

Je remercie vivement tous mes collègues d'e l'université du 20 Août 1955 Skikda, pour leur aide et leurs encouragements lors de préparation de ce travail ainsi que les membres du département de mathématiques de l'université Badji-Mokhtar Annaba, en particulier ceux du laboratoire de probabilités et statistique LaPS.

Je suis très reconnaissante à mon mari ***Reda*** *et ma fille* ***Assil****. Merci pour tous ses encouragements et leurs amours précieux qui ont été d'un grand réconfort à mon cœur. Je ne les remercierai jamais assez pour leur gentillesse*

et leur générosité.

Enfin et surtout, je remercie ***mon père****,* ***ma mère****,* ***mes frères****,* ***ma sœur*** *et* ***mes neveux*** *pour leur amour, leur soutien indéfectible, leurs encouragements bienveillants et pour m'avoir toujours poussée à aller de l'avant. Merci à toute ma famille qui m'a toujours soutenue et aidée.*

Je remercie très fort mes beaux parents, mes beaux-frères et mes belles sœurs.

Résumé

Ce travail de recherche est consacré à la construction et la mise en œuvre de tests d'ajustement du type du chi-deux modifié pour la distribution généralisée de Rayleigh et de modèles associés à celle-ci tels que le modèle à durée de vie accélérée dont la distribution de base est une Rayleigh généralisée et le modèle mélange de deux distributions de Rayleigh. Dans le cas des données complètes, on utilise la statistique NRR (Nikulin-Rao-Robson) qui est une modification de la statistique du chi-deux de Pearson. Pour les données censurées à droite, on se base sur l'approche introduite récemment par Bagdonavičius et Nikulin (2011). L'estimation des paramètres inconnus des modèles étudiés est basée sur les données initiales non groupées permettant ainsi de recouvrir toute l'information apportée par l'échantillon. Une importante étude par simulations numériques confirme les résultats théoriques obtenus.

Mots-clés : Données censurées, Test du chi-deux, Distribution de Rayleigh généralisée, Estimation du maximum de vraisemblance, Statistique NRR, Modèle AFT, Modèle de mélange.

Abstract

This research is devoted to the construction and implementation of modified chi-squared goodness-of- fit tests for the generalized Rayleigh distribution and the associated models such as the accelerated failure time model with generalized Rayleigh distribution as the baseline and the Rayleigh mixture model. For complete data, we use the NRR (Nikulin-Rao-Robson) statistic, which is a modification of the Pearson chi-squared statistic. In right censored data case, we use the approach introduced recently by Bagdonavičius and Nikulin (2011). The estimation of the unknown parameters of the studied models is based on the non-grouped initial data, which recover all the information provided by the sample. An important study by numerical simulations confirms the theoretical results obtained.

Keywords : Censored data, Chi-square test, Generalized Rayleigh distribution, Maximum likelihood estimation, NRR statistic, AFT model, Mixture model.

ملخص

هذا العمل البحثي يكرس في مجال بناء تنفيذ اختبارات الضبط المعدل χ^2 لتوزيع رايليغ المعمم و النماذج المرتبطة به مثل نموذج الحياة المتسارعة الذي توزيعه القاعدي هو توزيع رايليغ المعمم ونموذج مزيج من توزيعين لرايليغ. في حالة البيانات الكاملة، نستخدم احصائية NRR (Nikulin-Rao-Robson) التي هي تعديل لإحصائية χ^2 ل Pearson. من أجل البيانات المراقبة على اليمين، نرتكز على النهج الذي وضعه Nikulin و Bagdonvičius (2011). تقدير البارامترات المجهولة للنماذج المدروسة يستند على بيانات أولية غير مجمعة تسمح بالتالي لتغطية جميع المعلومات المقدمة من العينة. دراسة هامة عن طريق المحاكاة العددية تؤكد النتائج النظرية.

الكلمات الأساسية:

البيانات المراقبة، اختبار χ^2، توزيع رايليغ المعمم، تقدير الحد الاقصى من المصداقية، إحصائية NRR، نموذج AFT، نموذج المزيج.

Introduction générale

Dans la modélisation de la propagation des ondes radioélectriques, la distribution de Rayleigh (1980) compte parmi les distributions importantes pour définir certains paramètres physiques décrivant l'atmosphère, le comportement des signaux et les processus d'interaction qui lient ces paramètres entre eux. Elle peut être utilisée pour décrire la hauteur et la profondeur des vagues en océanographie, en théorie de la communication pour décrire les signaux radio reçus et en imagerie. Aussi, de nombreux problèmes de dynamique aboutissent à une équation différentielle linéaire d'ordre 2, où la solution est fonction d'une distribution de Rayleigh. Les domaines d'applications de ce modèle se sont généralisés aux études de fiabilité et de l'analyse de survie, ce qui a conduit les chercheurs à proposer de nouvelles généralisations de cette distribution.

La distribution généralisée de Rayleigh a été introduite par Surles et Padgett (2001). A l'origine, Mudholkar et Srivastava (1993), Mudholkar et al. (1995) ont proposé plusieurs distributions appelées les distributions Burr, dont la distribution de Rayleigh généralisée (GR) est un cas particulier de celles de Burr Type X. Selon les valeurs des paramètres, Kundu et Raqab (2005) ont montré que la densité de probabilité de cette distribution a différentes formes lui permettant de décrire beaucoup plus de données réelles. Ils ont utilisé différentes méthodes d'estimation sur données simples alors qu'Al-Khedhairi et al. (2007) ont calculé les estimateurs sur données groupées et données censurées. Fathipour et al. (2013) et Rao (2014) se sont intéressés à l'estimation de la fiabilité des composants décrits par des distributions de Rayleigh généralisées. Ces auteurs ont utilisé le test classique de Kolmogorov-Smirnov et celui du rapport de vraisemblance LR (maximum likelihood Ratio) pour ajuster ce modèle à des données observées. Récemment, Abd-Elfattah (2011) a fourni pour cette distribution, les tables des valeurs critiques de la statistique d'Anderson-Darling dans le cas de données com-

plètes. Toutefois en présence de censure, le problème reste ouvert. De plus les tests d'ajustement pour ce modèle n'ont pas encore été investis suffisamment.

Ce travail de recherche est consacré à la construction de tests d'ajustement du type du chi-deux modifié pour la distribution généralisée de Rayleigh et de modèles associés à celle-ci tels que le modèle à durée de vie accélérée dont la distribution de base est une Rayleigh généralisée et le modèle mélange de deux distributions de Rayleigh.

La validation des modèles choisis pour n'importe quelle analyse est nécessaire si nous voulons obtenir des résultats fiables. C'est pourquoi les méthodes et les techniques de tests d'ajustement sont en perpétuel développement. Quand la distribution est bien spécifiée, on peut utiliser n'importe quel test classique comme celui du chi-deux de Pearson, ou la statistique de Kolmogorov-Smirnov, la statistique d'Anderson-Darling, la statistique de Cramer-Von Mises et d'autres statistiques. On utilise aussi le test du rapport de vraisemblance ou des critères d'information comme le critère d'information d'Akaike (AIC) pour mesurer la qualité d'un modèle parmi deux modèles possibles. Cependant pour valider une hypothèse composite quand les paramètres sont inconnus et doivent être estimés à partir de l'échantillon et si en plus les données sont censurées, les tests classiques ne sont plus adaptés et les distributions des statistiques de test dépendent de la méthode d'estimation utilisée et du modèle proposé.

Parmi les nouvelles approches utilisées dans le cas de données complètes, nous considérons la statistique NRR (Nikulin-Rao-Robson) proposée séparément par Nikulin (1973) et Rao et Robson (1974). Ce test est une modification de la statistique du chi-deux de Pearson. Basée sur l'estimation du maximum de vraisemblance sur les données d'échantillon non regroupées, cette statistique suit à la limite une distribution du chi-deux [Drost (1988) et Van Der Vaart (1998)].

Pour les données censurées à droite, récemment Bagdonavicius et Nikulin (2011), Bagdonavicius et al. (2013) ont introduit un test du type du chi-deux modifié pour les distributions paramétriques continues. Cette statistique est basée sur la différence entre le nombre de pannes observées et le nombre de pannes théoriques. L'estimation des paramètres inconnus du modèle se fait sur les données initiales non groupées permettant ainsi de recouvrir toute l'information apportée par l'échantillon. Le critère de test est distribué selon un chi-deux.

Le manuscrit est structuré de la manière suivante :

Après la présentation des caractéristiques statistiques de la distribution de

Rayleigh généralisée, on étudie les estimateurs du maximum de vraisemblance des paramètres inconnus pour les données complètes et les données censurées à droite.

Dans un deuxième chapitre, nous construisons un test d'ajustement du type du chi-deux modifié basé sur la NRR statistique pour le modèle de Rayleigh généralisé pour données complètes. Nous déterminons les matrices d'information de Fisher pour données non groupées et pour données groupées ainsi que le critère de la statistique de test.

Quant au troisième chapitre, il est consacré à la construction et à la mise en place d'un test d'ajustement pour le modèle suscité en présence de censure aléatoire droite. Nous utilisons la statistique du test introduite par Bagdonavičius et Nikulin (2011) qui est basée aussi sur les estimateurs du maximum de vraisemblance sur les données initiales et suit une distribution du chi-deux. Tous les éléments constituant la statistique de test sont obtenus de manière explicite.

Un modèle à durée de vie accélérée AFT (accelerated failure time) dont la distribution de base est une distribution de Rayleigh généralisée est introduit dans le quatrième chapitre. Après l'étude des propriétés statistiques du modèle, on calcule les estimateurs du maximum de vraisemblance des paramètres inconnus de la distribution de base et des coefficients de régression représentant les variables explicatives. En utilisant la même approche, nous construisons pour ce modèle, un test d'ajustement qui nous permettra de vérifier si une suite de données censurées à droite est distribuée selon un modèle AFT-GR.

Dans le chapitre cinq, on présente un modèle mélange de deux distributions de Rayleigh. Ce type de modèle s'avère très utile pour décrire une variable provenant d'une population contenant des sous-populations ayant des paramètres différents. Pour déterminer les estimateurs du maximum de vraisemblance de ces paramètres, on utilise une méthode itérative basée sur l'algorithme espérance-maximisation (EM). Ensuite, on calcule le critère de test d'ajustement pour ce modèle dans le cas de données complètes.

Nous terminons par une étude par simulations numériques. Des dizaines de milliers d'échantillons avec différentes tailles et différentes valeurs de paramètres ont été simulés et ceci pour les différents modèles étudiés. Les estimateurs du maximum de vraisemblance des paramètres ainsi que leurs erreurs quadratiques moyennes sont déterminées. Les valeurs des critères de tests pour chacun des modèles ont été calculées selon plusieurs niveaux de confiance. Les résultats obtenus dans cette étude ont été appliqués à plusieurs

jeux de données provenant des études de fiabilité et d'analyse de survie.

Chapitre 1

Pésentation du modèle Rayleigh généralisé

1.1 Introduction

La distribution de Rayleigh a été proposée à l'origine dans les domaines de l'acoustique et de l'optique par Lord Rayleigh (1880). Ensuite, elle a trouvé ses applications en océanographie pour décrire la hauteur et la profondeur des vagues, la variation de la vitesse du vent en météorologie, la fiabilité des composants des systèmes et en théorie de la communication pour décrire le pic instantané de la puissance des signaux radio reçus. En 2001, Surles et Padgett ont introduit une généralisation de la distribution de Rayleigh en ajoutant un nouveau paramètre qui permet à cette nouvelle distribution d'être appliquée dans beaucoup plus de domaines tels que la fiabilité et l'analyse de survie. A l'origine, Mudholkar et Srivastava (1993), Mudholkar et al. (1995) ont proposé plusieurs distributions appelées les distributions Burr, dont la distribution de Rayleigh généralisée (GR) est un cas particulier de celles de Burr Type X. Selon les valeurs des paramètres, Kundu et Raqab (2005) ont montré que cette densité de probabilité a différentes formes lui permettant de décrire beaucoup plus de données réelles.

Dans ce chapitre, nous présentons la distribution de Rayleigh généralisée et nous déterminons les estimateurs des moments et les estimateurs du maximum de vraisemblance des paramètres inconnus pour les données complètes et pour les données censurées à droite.

1.2 La distribution de Rayleigh

On dit qu'une variable aléatoire T suit la distribution de Rayleigh $R(\sigma)$ avec le paramètre $\sigma(\sigma > 0)$, si sa fonction de densité est :

$$f(t,\sigma) = \frac{t}{\sigma^2} e^{-\frac{t^2}{2\sigma^2}}$$

La fonction de répartition d'une distribution de Rayleigh est donnée par :

$$F(t,\sigma) = 1 - \exp(\frac{-t^2}{2\sigma^2}) \quad \text{pour} \quad t \geq 0$$

L'espérance mathématique et la variance d'une variable aléatoire de Rayleigh T sont :

$$E(T) = \sigma\sqrt{\frac{\pi}{2}} \text{ et } V(T) = \frac{4-\pi}{2}\sigma^2$$

La distribution de Rayleigh s'applique à une variable continue positive non limitée et souvent utilisée seulement au voisinage de l'origine, c'est à dire pour les faibles valeurs de t. En particulier, la distribution de Rayleigh intervient dans les phénomènes de diffusion (comme le comportement dynamique des signaux utiles et des signaux brouilleurs). Cette distribution est liée avec plusieurs distributions continues, comme par exemple :

- Si $T \rightsquigarrow R(\sigma)$, alors $T = \sqrt{Y^2 + Z^2}$, où $Y \rightsquigarrow N(0,\sigma^2)$ et $Z \rightsquigarrow N(0,\sigma^2)$ sont deux variables gaussiennes indépendantes.

- Si $T \rightsquigarrow R(\sigma)$, alors $T = \sigma\sqrt{-2\ln(U)}$, où U est une variable uniforme sur l'intervalle $[0,1]$.

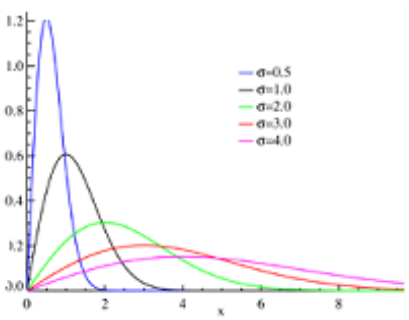
- Si $T \rightsquigarrow R(1)$, alors T^2 suit la loi du χ^2 avec deux degrés de liberté :

$$T^2 \rightsquigarrow \chi_2^2, \text{qui est une loi exponentielle de paramètre } \frac{1}{2}.$$

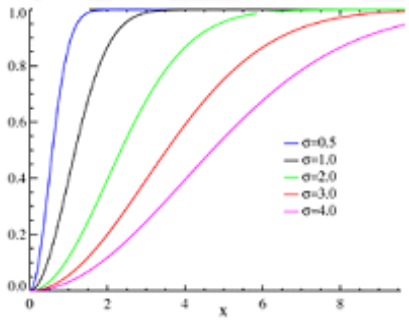
- Si Y suit une loi exponentielle de paramètre λ, alors

$$T = \sqrt{2Y\sigma^2\lambda} \rightsquigarrow R(\sigma).$$

- La loi de Rice est une généralisation de la loi de Rayleigh.



Densité de probabilité



Fonction de répartition

1.3 Modélisation par la distribution de Rayleigh

De nombreux problèmes de dynamique basés sur la deuxième loi de Newton (cette loi décrit ce qui se passe lorsqu'une force est exercée sur un objet et elle présente une formule mathématique permettant de calculer l'intensité de cette force en physique) aboutissent à une équation du deuxième ordre que l'on peut mettre sous la forme :

$$M\ \ddot{x} + g(x, \dot{x}) = f(t),$$

où $\dot{x} = \dfrac{dx}{dt}$, $\ddot{x} = \dfrac{d^2x}{dt^2}$, M est la masse et l'excitation $f(t)$ est centrée tandis que la fonction g est impaire par rapport à ses deux paramètres position et vitesse.

Ce type d'équation ne possède, sauf exception, de solution explicite que lorsqu'elle est linéaire (voir Équation différentielle linéaire d'ordre deux) :

$$M\ \ddot{x} + B\dot{x} + Kx = f(t)$$

La raideur K est le coefficient d'amortissement linéaire, B étant a priori inconnus, l'idée consiste à exprimer l'erreur comme la différence entre les termes des deux équations

$$\varepsilon\,(t) = B\dot{x} + Kx - g(x, \dot{x}),$$

On minimise alors le carré de sa moyenne quadratique $\overline{\varepsilon\,(t)^2}$ en annulant ses dérivées par rapport à B et K, ce qui fournit deux équations en B et K.

Ces équations permettent d'exprimer les deux inconnues en fonction des caractéristiques de la réponse. Par ailleurs, en utilisant l'équation linéaire, le calcul de cette réponse en fonction de B et K permet d'obtenir finalement une équation algébrique qui donne une valeur généralement approchée de la solution.

Cette méthode s'utilise dans deux cas particuliers, celui de l'excitation sinusoïdale et celui de l'excitation par un processus de Gauss.

Excitation sinusoïdale

À l'équation

$$M\ \ddot{x} + g(x, \dot{x}) = F \cos \omega t,$$

on substitue l'équation linéaire

$$M\ \ddot{x} + B\dot{x} + Kx = F \cos \omega t,$$

dont la solution est de la forme

$$\begin{aligned} x &= X \cos(\omega t + \varphi) \\ \dot{x} &= -\omega X \sin(\omega t + \varphi). \end{aligned}$$

Excitation gaussienne

Dans ce cas, le second membre est un processus aléatoire$F(t)$ représentant une infinité de signaux. La solution de l'équation est elle-même un processus $X(t)$constitué par une infinité de signaux :

$$M\ \ddot{X} + g(X, \dot{X}) = F\,(t)$$

Si le processus$F(t)$ est gaussien, l'équation linéarité

$$M\ \ddot{X} + B\dot{X} + KX = F\,(t)$$

a pour solution un processus également gaussien de densité de probabilité

$$P_{X\dot{X}}\,(x, \dot{x}) = \frac{1}{2\pi\sigma_X\sigma_{\dot{X}}} e^{-\frac{x^2}{2\sigma_X^2} - \frac{\dot{x}^2}{2\sigma_{\dot{X}}^2}}$$

Exemple de représentation de Rice

La représentation (voir Lin, 1976) d'un processus gaussien centré sous la forme

$$X(t) = A(t)\cos(\omega t + \phi(t))$$

ω : constante

$A(t)$: processus aléatoire à valeurs positives ou nulles

$\phi(t)$: processus aléatoire sur un intervalle$]0, 2\pi]$

Pour cela il faut également considérer le processus

$$Y(t) = A(t)\sin(\omega t + \phi(t))$$

Les deux processus gaussiens ont même variance σ^2 et sont indépendants. Leur densité de probabilité jointe se réduit donc à

$$P_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Le changement de variables défini par les deux formules du début conduit à une densité de probabilité jointe qui montre que les deux nouveaux processus sont indépendants :

$$P_A(a) = \frac{1}{2\pi\sigma^2} e^{-\frac{a^2}{2\sigma^2}} : \text{processus de Rayleigh}$$

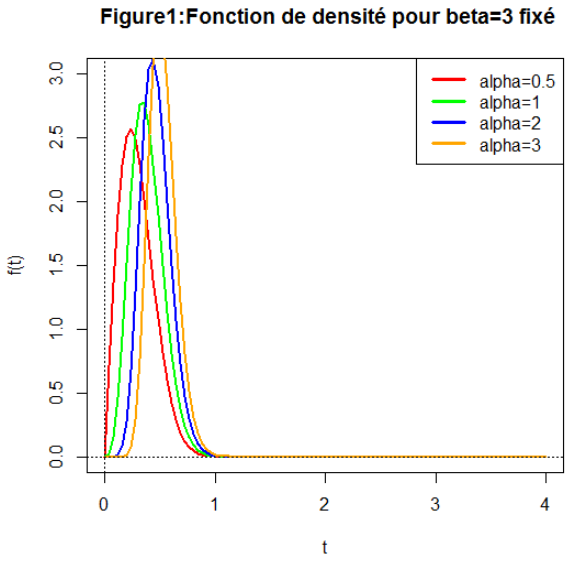
$$P\phi(\varphi) = \frac{1}{2\pi} : \text{processus uniforme}$$

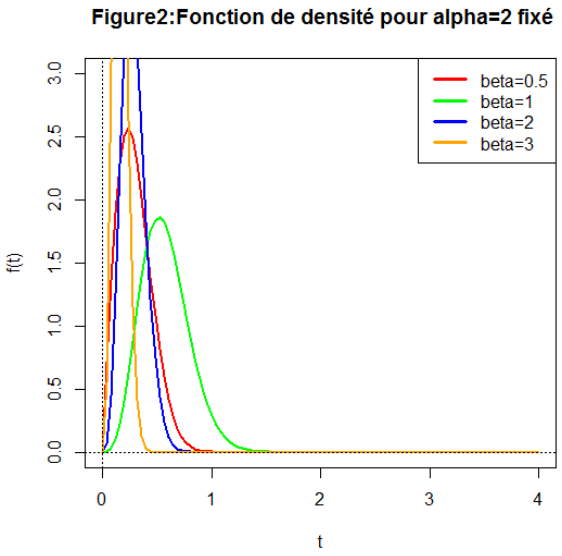
Le processus gaussien est donc décrit par un processus vaguement sinusoïdal $\cos(\omega t + \phi(t))$modulé par un processus enveloppe $A(t)$.

1.4 Distribution Rayleigh généralisée

Soit $T = (T_1, T_2, ...T_n)^T$ un échantillon de n variables aléatoires indépendantes. On dit que $T_i (i = 1, ..., n)$ suit la distribution de Rayleigh généralisée, on note $T_i \rightsquigarrow GR(\alpha, \beta)$ si sa densité de probabilité est définie par :

$$f(t, \alpha, \beta) = 2\alpha\beta^2 t e^{-(\beta t)^2} \left(1 - e^{-(\beta t)^2}\right)^{\alpha-1}, \qquad t > 0, \qquad \alpha > 0, \qquad \beta > 0. \tag{1.1}$$





où β est le paramètre d'échelle et α est le paramètre de forme.
Et sa fonction de répartition est donnée par :

$$F(t,\alpha,\beta) = \left(1 - e^{-(\beta t)^2}\right)^{\alpha}. \tag{1.2}$$

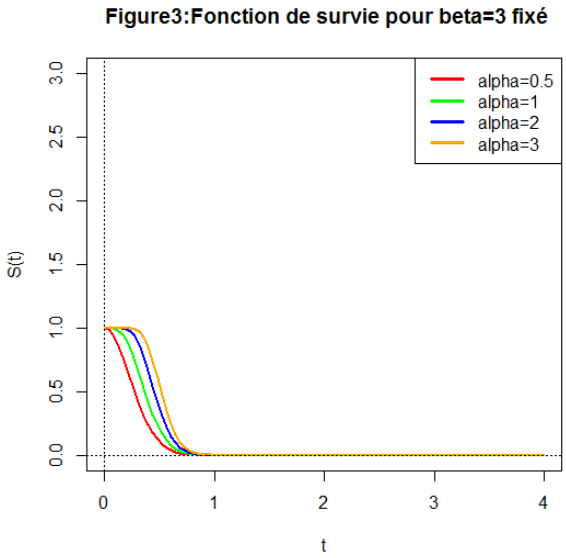
Dans la figure 1 et 2 nous représentons les densités de probabilité de la distribution de Rayleigh généralisée, en fixant un paramètre et en variant l'autre.

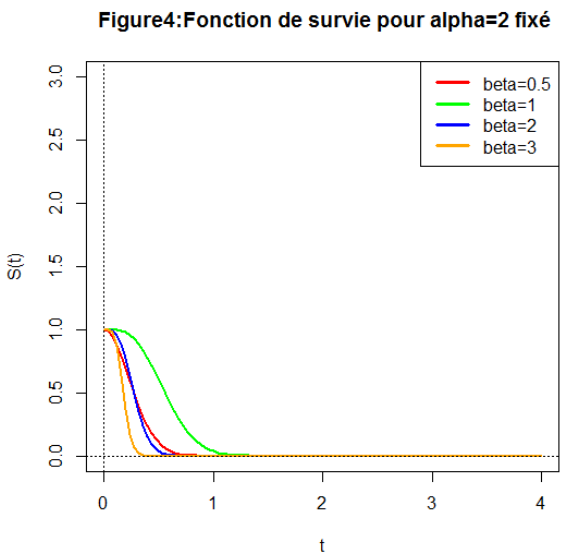
1.4.1 La fonction de survie

La fonction de survie de la distribution de Rayleigh généralisée, est définie par :

$$S(t,\alpha,\beta) = 1 - \left(1 - e^{-(\beta t)^2}\right)^{\alpha},$$

Les représentations graphiques de cette fonction sont illustrées dans les figures 3 et 4.





1.4.2 La fonction de hasard

La fonction de hasard (le taux de défaillance, de panne ou de risque) de la distribution de Rayleigh généralisée est définie par :

$$h\left(t,\alpha,\beta\right)=\frac{f\left(t,\alpha,\beta\right)}{S(t,\alpha,\beta)}=\frac{2\alpha\beta^{2}te^{-(\beta t)^{2}}\left(1-e^{-(\beta t)^{2}}\right)^{\alpha-1}}{1-\left(1-e^{-(\beta t)^{2}}\right)^{\alpha}}.$$

Sa présentation graphique se trouve dans les figures 5 et 6.

1.4.3 La fonction de hasard cumulé

La fonction de hasard cumulé de la distribution de Rayleigh généralisée est définie comme suit :

$$\Lambda(t)=-\ln[1-\left(1-e^{-(\beta\ t)^{2}}\right)^{\alpha}],$$

et sa présentation graphique est illustrée dans les figures 7 et 8.

Figure5:Fonction de hasard pour alpha=2 fixé

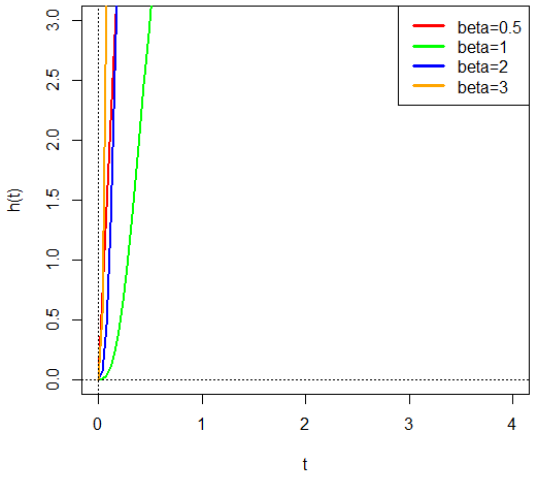


Figure6:Fonction de hasard pour beta=3 fixé

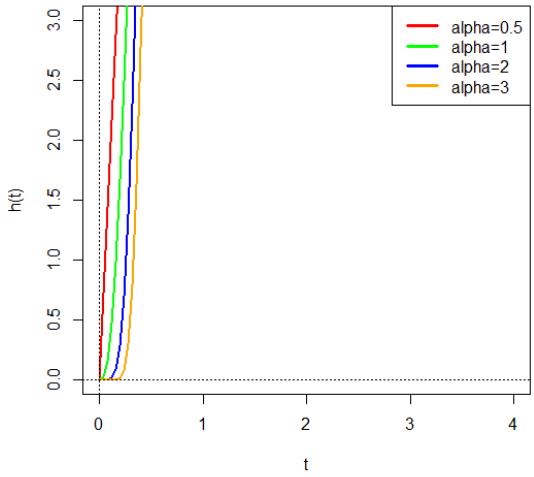


Figure7:Fonction de hasard cumulé pour alpha=2 fixé

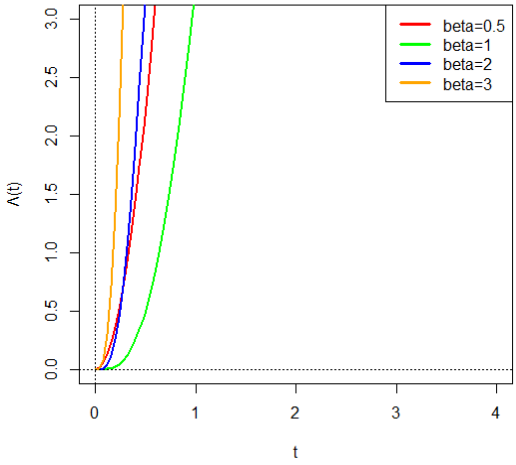


Figure8:Fonction de hasard cumulé pour beta=3 fixé

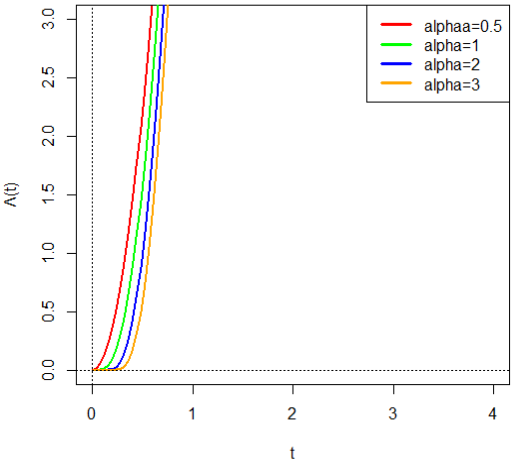
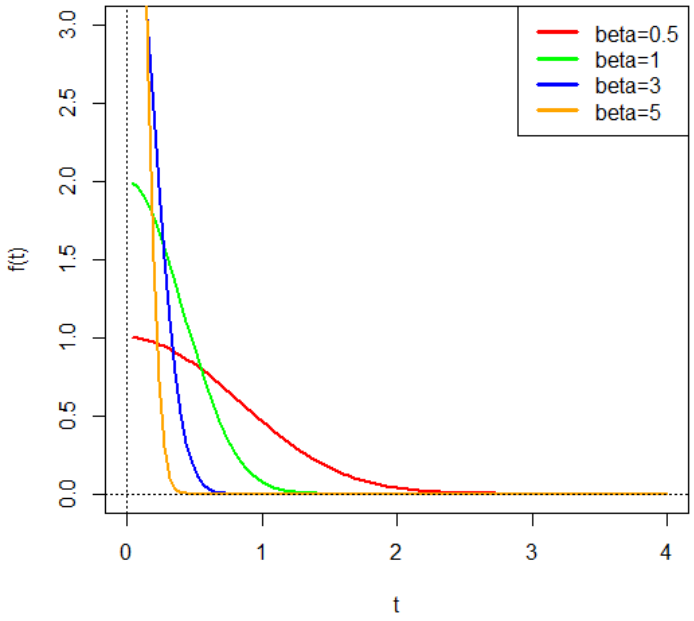


Figure 1.1: Densité pour alpha=0.5 fixé



Remarque :

1- Raqab et Kundu (2006) ont montré que la fonction de densité (pdf) de Rayleigh généralisée est une fonction décroissante si $\alpha \leq 0.5$ (voir Figure 1.1) . Et si $\alpha > 0.5$ c'est une fonction unimodale asymétrique à droite (voir la figure 9).

2- Il a également été montré par Raqab et Kundu (2006) que la fonction de taux de risque $h(t)$ est une fonction croissante pour $\alpha > 0.5$ (voir la figure 10).

3- Si $\alpha = 1$ et $\beta = \dfrac{1}{\sqrt{2}\sigma}$, on dit que T suit la distribution de Rayleigh à un paramètre $R(\sigma)$.

Figure9:Fonction de densité pour beta=1 fixé

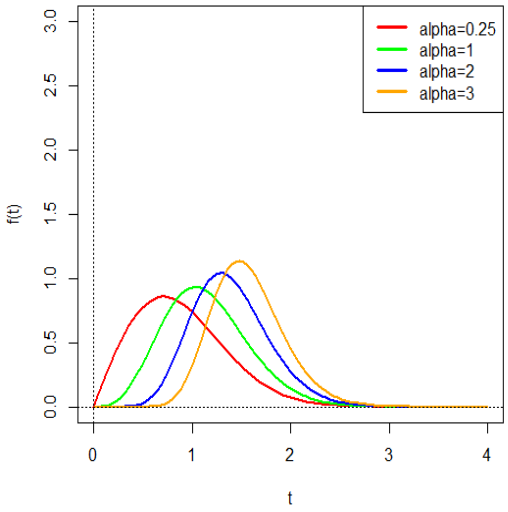
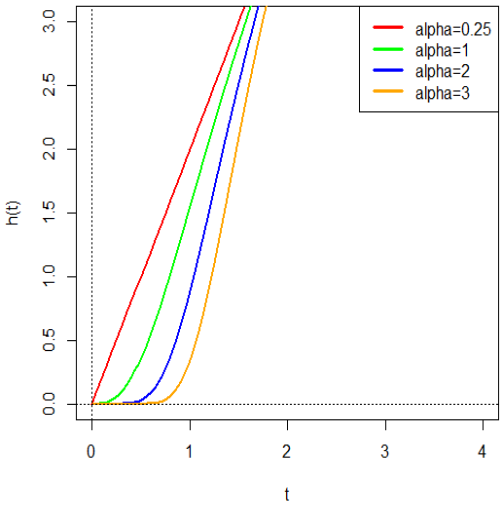


Figure10:Taux de risque pour beta=1 fixé



1.5 Moyenne du temps de bon fonctionnement (MTBF)

La moyenne du temps de bon fonctionnement d'une variable aléatoire (MTBF) est la moyenne de la variable aléatoire T définie par :

$$MTBF \ = E(T) = \int_0^{+\infty} t \ f \ (t) \ dt = \lim_{x \to +\infty} \int_0^{x} t \ f \ (t) \ dt.$$

Elle représente l'espérance de vie du dispositif.
Pour la distribution de Rayleigh généralisée, on obtient

$$MTBF \ = \int_0^{+\infty} 2t^2 \alpha \beta^2 e^{-(\beta \ t)^2} \left(1 - e^{-(\beta t)^2}\right)^{\alpha - 1} dt \tag{1.3}$$

On utilise le changement de variable

$$X = e^{-(\beta \ t)^2}.$$

Donc, on obtient l'intégrale suivante :

$$MTBF = \frac{\alpha}{\beta}\int_0^1 \sqrt{\log(\frac{1}{X})}\,(1-X)^{\alpha-1}\,dX$$

·Si $\alpha = 1 : MTBF = \dfrac{0.8862269255}{\beta}$.

· Si $\alpha = 2 : MTBF = \dfrac{1.145796782}{\beta}$.

· Si $\alpha = 3 : MTBF = \dfrac{1.290372924}{\beta}$.

1.6 Moments

L'espérance de T pour quelques valeurs du paramètre α :

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

où $S(t) = 1 - (1 - \exp(-(\beta t)^2)^\alpha$.

· Si $\alpha = 1 : E(T) = \lim\limits_{t\to+\infty}\left(\dfrac{\sqrt{\pi}\,\mathrm{erf}(\beta t)}{2\beta}\right)$, où $\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt$.

· Si $\alpha = 2 : E(T) = \lim\limits_{t\to+\infty}\left(-\dfrac{\sqrt{\pi}(-4\,\mathrm{erf}(\beta t) + \sqrt{2}\,\mathrm{erf}(\sqrt{2}\beta t))}{4\beta}\right)$.

· Si $\alpha = 3 : E(T) = \lim\limits_{t\to+\infty}\left(\dfrac{\sqrt{\pi}(18\,\mathrm{erf}(\beta t) - 9\sqrt{2}\,\mathrm{erf}(\sqrt{2}\beta t) + 2\sqrt{3}\,\mathrm{erf}(\sqrt{3}\beta t))}{12\beta}\right)$

1.6.1 Fonction génératrice des moments

Soit $X \rightsquigarrow GR\,(\alpha, \beta)$, la fonction génératrice des moments $M_X\,(t)$ de X est donnée par :

$$M_{_X}(t) = E(e^{tX}) = \int_0^{+\infty} e^{tx} f(x,\alpha,\beta)\, dx$$

$$M_{_X}(t) = \int_0^{\infty} e^{tx} 2\alpha\beta^2 x e^{-\beta^2 x^2} (1 - e^{-\beta^2 x^2})^{\alpha-1} dx$$

La résolution de l'intégrale précédente par changement de variable donne :

$$M_u(t) = \int_0^1 \alpha e^{\frac{t}{\beta}\sqrt{-\log(1-u)}} u^{\alpha-1} du$$

on obtient :

$$M_y(t) = \int_0^1 \alpha y^{-\frac{t}{\beta}} (1-y)^{\alpha-1} dy = \alpha \int_0^1 y^{-\frac{t}{\beta}} (1-y)^{\alpha-1} dy$$

$$= \alpha Beta(1 - \frac{t}{\beta}, \alpha) = \alpha \frac{\Gamma\left(1 - \frac{t}{\beta}\right)\Gamma(\alpha)}{\Gamma\left(1 - \frac{t}{\beta} + \alpha\right)} = \frac{\Gamma\left(1 - \frac{t}{\beta}\right)\Gamma(\alpha+1)}{\Gamma\left(1 - \frac{t}{\beta} + \alpha\right)}$$

où $\Gamma(\alpha)$ est la fonction gamma ($\alpha\Gamma(\alpha) = \Gamma(\alpha+1)$, $\Gamma(1) = 1$) et $Beta(1 - \frac{t}{\beta}, \alpha)$ est la fonction bêta à deux variables $1 - \frac{t}{\beta}$ et α.

La fonction gamma est définie par :

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} \exp(t) dt,$$

et

$$Beta(\beta,\alpha) = \int_0^1 t^{\beta-1} (1-t)^{\alpha-1} dt.$$

Donc

$$M_{_X}(t) = \frac{\Gamma\left(1 - \frac{t}{\beta}\right)\Gamma(\alpha+1)}{\Gamma\left(1 - \frac{t}{\beta} + \alpha\right)} \tag{1.4}$$

Le calcul de la dérivée première de $M_{X}(t)$ est

$$M_{X}{}'(t)=\frac{d}{dt}M_{X}(t)=-\frac{\left(\Psi\left(\frac{\beta-t}{\beta}\right)-\Psi\left(\frac{(\beta-t+\alpha\beta)}{\beta}\right)\right)\Gamma(\alpha+1)\Gamma\left(\frac{\beta-t}{\beta}\right)}{\beta\Gamma\left(\frac{\beta-t+\alpha\beta}{\beta}\right)},$$

et de sa dérivée deuxième, est

$$M_{X}{}'(t) = \frac{d}{dt}M_{X}{}'(t)= \frac{\left(\Psi\left(1,\frac{\beta-t}{\beta}\right)+\Psi^{2}\left(\frac{\beta-t}{\beta}\right)-2\Psi\left(\frac{\beta-t}{\beta}\right)\Psi\left(\frac{\beta-t+\alpha\beta}{\beta}\right)+\Psi^{2}\left(\frac{\beta-t+\alpha\beta}{\beta}\right)-\Psi\left(1,\frac{\beta-t+\alpha\beta}{\beta}\right)\right)\Gamma(\alpha+1)\Gamma\left(\frac{\beta-t}{\beta}\right)}{\beta^{2}\Gamma\left(\frac{\beta-t+\alpha\beta}{\beta}\right)}$$

où $\Psi(\alpha)$ est la fonction digamma est définie comme la dérivée logarithmique de la fonction gamma

$$\Psi(\alpha)=\frac{\Gamma'(\alpha)}{\Gamma(\alpha)},$$

et $\Psi(n,\alpha)$ est la fonction polygamma qui est la dérivée nième de la fonction digamma

$$\Psi(n,\alpha)=\Psi^{(n)}(\alpha)=(-1)^{(n+1)}\int_{0}^{+\infty}\frac{t^{n}\exp(-\alpha t)}{1-\exp(-t)}dt.$$

1.6.2 L'espérance mathématique et la variance

Soit $X \rightsquigarrow GR(\alpha,\beta)$, alors

$$\begin{aligned}
E(X) &= M_{X}{}'(0)=-\frac{\Psi(1)-\Psi(1+\alpha)}{\beta}=\frac{\gamma+\Psi(1+\alpha)}{\beta}\\
V(X) &= E(X^{2})-E^{2}(X)\\
E(X^{2}) &= M_{X}{}'(0)=\frac{\Psi(1,1)+\Psi^{2}(1)-2\Psi(1)\Psi(1+\alpha)+\Psi^{2}(1+\alpha)-\Psi(1,1+\alpha)}{\beta^{2}}\\
&= \frac{\frac{1}{6}\pi^{2}+\gamma^{2}+2\gamma\Psi(1+\alpha)+\Psi^{2}(1+\alpha)-\Psi(1,\alpha)-\frac{1}{\alpha^{2}}}{\beta^{2}}
\end{aligned}$$

où $\Psi(1)=-\gamma=-\,0.5772156649$ et $\Psi(1,1)=\frac{1}{6}\pi^{2}=1.644934068$
On a aussi comme $\alpha>0$,

$$\Psi(1+\alpha)=\Psi(1)+\sum_{n=1}^{\infty}\frac{\alpha}{n(n+\alpha)},$$

et

$$\Psi(1, 1+\alpha) = \Psi(1, \alpha) - \frac{1}{\alpha^2} = \sum_{n=0}^{\infty} \frac{1}{(n+\alpha)^2} - \frac{1}{\alpha^2}.$$

1.6.3 Estimateurs par la méthode des moments

La méthode des moments (MM) est un outil d'estimation intuitif et elle peut être utilisée à la place de la méthode du maximum de vraisemblance. Elle consiste à estimer les paramètres recherchés en égalisant certains moments théoriques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. Cette méthode peut être résumée comme suit.

On suppose que l'échantillon $X_1, \ldots, X_n$ est un échantillon iid (identiquement et indépendamment distribué) selon une famille de lois paramétriques, paramétrée par θ. Toute fonction des données de l'échantillon est donc une fonction $F(\theta)$.

On sélectionne alors s moments qui définit un vecteur $s \times 1$. Il existe donc une fonction G telle que $G(\theta) = [m_1(\theta), m_2(\theta), ..., m_s(\theta)]$. L'équivalent

empirique du vecteur G est le vecteur composé des s moments d'échantillon, noté $\widehat{G}$. Cela signifie que l'on remplace le i-ème moment théorique, à savoir $E_{\theta}(X^i)$, par la quantité :

$$\widehat{m}_i = \frac{1}{n} \sum_{k=1}^{n} x_k^i$$

L'estimateur de θ par la méthode des moments, noté $\widehat{\theta}$, consiste à résoudre l'équation vectorielle :

$$\widehat{G} = G\left(\widehat{\theta}\right)$$

Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires iid selon la loi de Rayleigh généralisée. On cherche à estimer le vecteur des paramètres

$$\theta = [\alpha, \beta].$$

On détermine d'abord les moments théoriques. Le premier moment, l'espérance, est donné par

$$E(X) = m_1 = \frac{\gamma + \Psi(1+\alpha)}{\beta}$$

et le second moment, l'espérance du carré de la variable, est

$$E(X^2) = m_2 = \frac{\frac{1}{6}\pi^2 + \gamma^2 + 2\gamma\Psi(1+\alpha) + \Psi^2(1+\alpha) - \Psi(1,\alpha) + \frac{1}{\alpha^2}}{\beta^2}$$

On exprime ensuite la relation entre les paramètres et les moments théoriques :

$$\begin{cases} \frac{\gamma+\Psi(1+\alpha)}{\beta} = m_1 & (1) \\ \frac{\frac{1}{6}\pi^2+\gamma^2+2\gamma\Psi(1+\alpha)+\Psi^2(1+\alpha)-\Psi(1,\alpha)+\frac{1}{\alpha^2}}{\beta^2} = m_2 & (2) \end{cases}$$

d'après l'équation (1), l'estimateur de β est

$$\widehat{\beta} = \frac{\gamma+\Psi(1+\widehat{\alpha})}{m_1} \quad (3)$$

Pour évaluer l'estimateur $\widehat{\beta}$, il faut trouver l'estimateur $\widehat{\alpha}$ de α qui est difficile à résoudre. Dans ce cas-là, on utilise la méthode suivante :
la méthode des moments consiste à utiliser les moments empiriques

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X} = m_1 \quad (4)$$

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = m_2 \quad (5)$$

d'après les relations précédentes, on obtient l'équation suivante :

$$m_2 - \frac{\left(\frac{1}{6}\pi^2 + \gamma^2 + 2\gamma\Psi(1+\widehat{\alpha}) + \Psi^2(1+\widehat{\alpha}) - \Psi(1,\widehat{\alpha}) + \frac{1}{\widehat{\alpha}^2}\right)\overline{X}^2}{\left(\gamma + \Psi(1+\widehat{\alpha})\right)^2} = 0 \quad (6)$$

pour résoudre l'équation (6), on utilise la méthode de Newton Raphson sous forme :

$$\widehat{\alpha}_{i+1} = \widehat{\alpha}_i - \frac{g(\widehat{\alpha}_i)}{g'(\widehat{\alpha}_i)}$$

où

$$g(\widehat{\alpha}_i) = m_2 - \frac{\left(\frac{1}{6}\pi^2 + \gamma^2 + 2\gamma\Psi(1+\widehat{\alpha}_i) + \Psi^2(1+\widehat{\alpha}_i) - \Psi(1,\widehat{\alpha}_i) + \frac{1}{\widehat{\alpha}_i^2}\right)\overline{X}^2}{\left(\gamma + \Psi(1+\widehat{\alpha}_i)\right)^2},$$

alors

$$g'(\widehat{\alpha}_i) = \frac{(A(\widehat{\alpha}_i, \gamma) + B(\widehat{\alpha}_i, \pi))\,\overline{X}^2}{3\widehat{\alpha}_i^4\,(\gamma + \Psi\,(1 + \widehat{\alpha}_i))^3}$$

où

$$A(\widehat{\alpha}_i, \gamma) = 3\Psi(2, \widehat{\alpha}_i)\widehat{\alpha}_i^4\gamma + 3\Psi(2, \widehat{\alpha}_i)\widehat{\alpha}_i^4\Psi\,(1 + \widehat{\alpha}_i) + 6\gamma\widehat{\alpha}_i,$$

et

$$B(\widehat{\alpha}_i, \pi) = 6\Psi\,(1 + \widehat{\alpha}_i)\,\widehat{\alpha}_i + \pi^2\widehat{\alpha}_i^4\Psi\,(1, \widehat{\alpha}_i) - \pi^2\widehat{\alpha}_i^2 - 6\Psi^2\,(1, \widehat{\alpha}_i)\,\widehat{\alpha}_i^4 + 12\Psi\,(1, \widehat{\alpha}_i)\,\widehat{\alpha}_i^2 - 6,$$

donc à chaque fois on trouve l'estimateur $\widehat{\alpha}$, on le substitue dans (3) pour obtenir $\widehat{\beta}$ (pour les calculs, on utilise Maple 13).

1.7 Estimation du maximum de vraisemblance en cas de données complètes

On considère les estimateurs du maximum de vraisemblance (MLE) des paramètres inconnus. Soit $x_1, ..., x_n$ un échantillon aléatoire de taille n de GR (α, β), la fonction de vraisemblance est donnée par l'équation suivante :

$$\begin{aligned} L\,(x, \theta) &= \prod_{i=1}^{n} f\ \ (x_i, \theta) \\ &= \prod_{i=1}^{n} 2\alpha\beta^2 x_i e^{-(\beta x_i)^2}\left(1 - e^{-(\beta x_i)^2}\right)^{\alpha-1} \\ &= 2^n\alpha^n\beta^{2n}\prod_{i=1}^{n} x_i e^{-(\beta x_i)^2}\left(1 - e^{-(\beta x_i)^2}\right)^{\alpha-1} \end{aligned}$$

Notons $\ell(\theta)$ la log-vraisemblance telle que :

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \ln f\ \ (x_i, \theta) \\ &= n\ln(2) + n\ln(\alpha) + 2n\ln(\beta) + \sum_{i=1}^{n}\ln(x_i) - \\ &\quad \beta^2\sum_{i=1}^{n} x_i^2 + (\alpha - 1)\sum_{i=1}^{n}\ln\left(1 - e^{-(\beta x_i)^2}\right) \end{aligned} \tag{1.5}$$

alors l'estimateur MLE vérifie l'équation :

$$\dot{\ell}(\widehat{\theta}) = 0$$

et $\dot{\ell}$ est le vecteur de score

$$\dot{\ell}(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \left(\frac{\partial}{\partial\alpha}\ell(\theta), \frac{\partial}{\partial\beta}\ell(\theta)\right)^T$$

Les fonctions de score sont obtenues comme suit :

$$\begin{cases} \dfrac{\partial}{\partial\alpha}\ell(\theta) = \dfrac{n}{\alpha} + \sum\limits_{i=1}^{n} \ln\left(1 - e^{-(\beta x_i)^2}\right) = 0 & (1) \\ \dfrac{\partial}{\partial\beta}\ell(\theta) = \dfrac{2n}{\beta} - 2\beta\sum\limits_{i=1}^{n} x_i^2 + 2\beta\,(\alpha-1)\sum\limits_{i=1}^{n} \frac{x_i^2 e^{-(\beta x_i)^2}}{1-e^{-(\beta x_i)^2}} = 0 & (2) \end{cases}$$

D'après (1), nous obtenons l'estimateur du maximum de vraisemblance de α comme suit :

$$\hat{\alpha} = -\frac{n}{\sum\limits_{i=1}^{n} \ln\left(1 - e^{-\left(\hat{\beta} x_i\right)^2}\right)} \tag{3}$$

et celui de β comme :

$$\hat{\beta}^2 = \frac{n}{\sum\limits_{i=1}^{n} x_i^2 - (\hat{\alpha}-1)\sum\limits_{i=1}^{n}\left(\frac{x_i^2 e^{-(\hat{\beta}x_i)^2}}{1-e^{-(\hat{\beta}x_i)^2}}\right)} \tag{4}$$

Les équations (3) et (4) forment un système non linéaire. La résolution de ce système peut être donnée par des méthodes numériques (par exemple, la méthode de Newton Raphson, l'algorithme BB, l'algorithme maxLik). Pour plus de détails, on peut consulter Kundu et Raqab (2005).

1.8 Estimation du maximum de vraisemblance en cas de données censurées

Soit $T_1, T_2, ..., T_n$ les taux de défaillance et $C_1, C_2, ..., C_n$ les taux de censure à droite. Les données observées sont (X_1, δ_1), (X_2, δ_2), ..., (X_n, δ_n),

où $X_i = \min(T_i, C_i)$ et la fonction indicatrice définie comme suit :

$$\delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \\ 0 & \text{sinon} \end{cases}$$

Soit T_i une variable aléatoire distribuée avec le vecteur de paramètres $\theta = (\alpha, \beta)^T$, où T_i et C_i sont des variables aléatoires indépendantes.

La fonction de vraisemblance est :

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) = \prod_{i=1}^{n} h^{\delta_i}(X_i, \theta) S(X_i, \theta)$$

où $f(X_i, \theta)$ est la fonction de densité de la distribution Rayleigh généralisée, $h(X_i, \theta)$ est la fonction de hasard et $S(X_i, \theta)$ est la fonction de survie.

Alors

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{2\alpha\beta^2 X_i e^{-(\beta\ X_i)^2} \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha-1}}{1 - \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha}} \right]^{\delta_i} \left[1 - \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha}\right],$$

et la fonction de log-vraisemblance est :

$$\ell(\theta) = \ln(L(\theta)) = \sum_{i=1}^{n} \delta_i \ln h(X_i, \theta) + \sum_{i=1}^{n} \ln S(X_i, \theta),$$

alors

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \delta_i [\ln(2) + \ln(\alpha) + 2\ln(\beta) + \ln(X_i) - \beta^2 X_i^2 + (\alpha - 1)\ln(1 - e^{-(\beta\ X_i)^2})] + \\ &\quad \sum_{i=1}^{n} (1 - \delta_i) \ln\left(1 - \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha}\right); \\ \ell(\theta) &= \sum_{i=1}^{n} \delta_i \left(\ln(2) + \ln(\alpha) + 2\ln(\beta)\right) + \sum_{i \in P} \ln(X_i) - \beta^2 \sum_{i \in P} X_i^2 + \\ &\quad (\alpha - 1) \sum_{i \in P} \ln\left(1 - e^{-(\beta\ X_i)^2}\right) - \sum_{i \in P} \ln\left(1 - \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha}\right) \\ &\quad + \sum_{i \in C} \ln\left(1 - \left(1 - e^{-(\beta\ X_i)^2}\right)^{\alpha}\right), \end{aligned} \tag{1.6}$$

où P et C désignent les ensembles d'observations non censurées et censurées, respectivement.

Les fonctions de score pour les paramètres α et β sont données par :

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \alpha} &= \sum_{i=1}^{n} \delta_i \left[\frac{1}{\alpha} + \frac{\ln(1-e^{-(\beta\ x_i)^2})}{1-\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha}} \right] - \sum_{i=1}^{n} \frac{\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha} \ln\left(1-e^{-(\beta\ x_i)^2}\right)}{1-\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha}} \\
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{i=1}^{n} \delta_i \left[\frac{2}{\beta} + \frac{2\beta x_i^2 \left(-1+\alpha e^{-(\beta\ x_i)^2}\right)}{1-e^{-(\beta\ x_i)^2}} + \frac{2\alpha\beta x_i^2 e^{-(\beta\ x_i)^2}\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha-1}}{1-\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha}} \right] - \\
&\quad \sum_{i=1}^{n} \frac{2\alpha\beta x_i^2 e^{-(\beta\ x_i)^2}\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha-1}}{1-\left(1-e^{-(\beta\ x_i)^2}\right)^{\alpha}}.
\end{aligned}$$

Pour résoudre le système de fonctions de score, assez compliquées, on fait appel aux méthodes numériques, telles que la méthode de Newton Raphson, l'algorithme BB, l'algorithme maxLix et autres.

1.9 Estimateurs du maximum de vraisemblance avec l'algorithme EM

L'algorithme espérance-maximisation (expectation-maximization algorithm, EM), proposé par Dempster et al. (1977), est un algorithme itératif qui permet de trouver les paramètres du maximum de vraisemblance d'un modèle probabiliste lorsque ce dernier dépend de variables non observables dites variables cachées. Depuis et vu leur utilité, de nombreuses variantes ont été proposées dans la littérature, formant une classe entière d'algorithmes.

Définition 1 :

Les situations où l'algorithme EM est utile peuvent être résumées ainsi. Avec les données dont y, nous disposons réellement, l'estimation est difficile. Avec les données plus fines, t, l'estimation est facile, mais nous n'en disposons pas. Dans un contexte général, soient donc y_{obs} les données observées et soit θ le paramètre que nous souhaitons estimer.

La densité des données est $f_Y(y|\theta)$ et la fonction de vraisemblance est $V_Y(\theta) = f_Y(y_{obs}|\theta)$. L'estimateur $\widehat{\theta}$ du maximum de vraisemblance est tel que $\ell_Y\left(\widehat{\theta}\right) = \ln\left(V_Y\left(\widehat{\theta}\right)\right) \geq \ell_Y(\theta)$ pour tout θ. En ajoutant une composante

d'information suplémentaire Z aux données Y la vraisemblance se simplifie. Tout en augmentant la complexité

$$\begin{array}{ccc} Y & \rightarrow & (Y,Z)=T \\ \text{données} & \longrightarrow & \text{données augmentées} \end{array}$$

La loi de $f_T(x|\theta)$ devient plus facile à analyser que $f_Y(y|\theta)$. Parce qu'on ne dispose pas de la valeur t_{obs}, on procède à une estimation en remplaçant

$$\ell_T(\theta) = \ln(V_T(\theta)) = \ln(f_T(t_{obs}|\theta))$$

par son espérance mathématique

$$Q(\theta|\eta) = E[\ln(f_T(t|\theta))|Y = y_{obs}, \theta = \eta] = E[\ln(\ell_T(\theta))|Y = y_{obs}, \theta = \eta].$$

L'espérance est calculée par rapport à la densité conditionnelle de T sous

condition $Y = y_{obs}$.

Le but consiste à optimiser $\ell_Y(\theta)$ et l'algorithme EM y arrive en s'appuyant sur $Q(\theta|\eta)$. La démarche est telle que l'on calcule une suite $\theta_0, \theta_1, \theta_2, ...$d'approximation de $\widehat{\theta}$.

Définition 2 :

L'algorithme EM est une sorte de log-vraisemblance approximative Q avec l'optimisation de cette fonction. Au départ, on choisit une valeur initiale du paramètre, θ_0.

Ensuite, on utilise θ_0 pour calculer l'espérance des statistiques dont on a besoin pour déterminer $Q(\theta|\theta_0)$. Puis on trouve la valeur de θ qui maximise cette fonction Q. Cette valeur nous donne θ_1, et ainsi de suite.

En général, le schéma de l'algorithme EM est le suivant :

[**EM0**] choisir une valeur initiale θ_0 et poser i=0.

[**EM1**] Calculer $Q(\theta|\theta_i) = E[\ln(V_T(\theta))|\theta = \theta_i, y_{obs}]$, où l'espérance est par rapport à la densité conditionnelle $f_T|_Y(T|\theta_i, y_{obs})$.

[**EM2**] maximiser $Q(\theta|\theta_i)$ par rapport à θ et poser $\theta_{i+1} = \arg\max Q(\theta|\theta_i)$.

[**EM3**] Tester la convergence $(\theta_{i+1} - \theta_i \approx 0)$. Soit on s'arrête, soit on pose $i = i + 1$ et

on reprend avec [**EM1**].

Chapitre 2

Tests d'ajustement pour données complètes

2.1 Introduction

Quand la distribution est bien spécifiée, on peut utiliser n'importe quel test classique comme celui du chi-deux de Pearson, la statistique de Kolmogorov-Smirnov, la statistique d'Anderson-Darling, la statistique de Cramer-Von Mises et d'autres statistiques pour valider le choix du modèle utilisé dans l'analyse. Le test du rapport de vraisemblance ou des critères d'information comme le critère d'information d'Akaike (AIC) sont utilisés pour mesurer la qualité d'un modèle parmi deux modèles possibles. Cependant pour valider une hypothèse composite quand les paramètres sont inconnus et doivent être estimés à partir de l'échantillon, les tests classiques ne sont plus adaptés et les distributions des statistiques de test dépendent de la méthode d'estimation utilisée et du modèle proposé. Récemment, en utilisant la méthode de Monte Carlo, Abd Elfattah (2011) a calculé les valeurs critiques de la statistique d'Anderson-Darling pour la distribution de Rayleigh généralisée dans le cas où les paramètres sont inconnus.

Dans ce chapitre, nous donnons un aperçu sur le test du chi-deux de Pearson et les statistiques classiques d'Anderson-Darling et Kolmogorov-Smirnov. Ensuite nous construisons un test du type du chi-deux modifié pour la distribution de Rayleigh généralisé, basé sur la statistique de Nikulin-Rao-Robson (NRR). La NRR statistique introduite par Nikulin (1973) et Rao et Robson (1974) est une modification de la statistique de Pearson qui est basée

sur l'estimation du maximum de vraisemblance sur les données initiales Sa distribution est une loi du chi-deux à $r-1$ degrés de libertés où r représente le nombre de classes choisies pour le groupement des données. Nous procurons la formule explicite du critère de test permettant de vérifier si une série d'observations est distribuée selon un modèle de Rayleigh généralisé.

2.2 La théorie de test du chi-deux de Pearson

Soit un échantillon $T_1, T_2, \cdots, T_n$ de T suivant une distribution paramétrique $F(t,\theta)$,c-à-d :

$P\{T_i \leq t \setminus H_0\} = F(t,\theta)$, où $\theta = (\theta_1, \theta_2, ...\theta_s)^T \in \Theta \subset \mathbb{R}^s$ et $t \in \mathbb{R}^1$.

où Θ est un ensemble ouvert.

On considère de tester l'hypothèse H_0 selon laquelle la distribution de $T = (T_1, T_2, ...T_n)^T$ vérifie :

$$H_0 : P(T_i \leq t) = F(t,\theta) \tag{2.1}$$

où $F(t,\theta)$est la fonction de répartition de t_i.

On partage la droite réelle en r sous-intervalles :$I_1, I_2, \cdots, I_r$ mutuellement disjoints, par les points :

$$-\infty = a_0 < a_1 < \cdots < a_{r-1} < a_r = +\infty,\ I_j =]a_{j-1}, a_j];\ I_i \cap I_j = \phi,\ i \neq j;\ \bigcup_{j=1}^{r} I_j = R^1.$$

Soit $\nu = (\nu_1, \nu_2, \cdots, \nu_r)^T$ le vecteur des fréquences obtenu en regroupant les T_i dans des intervalles I_j (r classes), avec

$$\nu_j = card\{i,\ T_i \in I_j,\ i = 1, 2, ..., n\}.$$

La matrice d'information de Fisher pour les donnée groupées est $nJ = nB^T B$,où

$$J = J(\theta) = \left[\sum_{l=1}^{r} \frac{1}{\sqrt{p_l(\theta)}} \frac{\partial p_l(\theta)}{\partial \theta_i} \frac{\partial p_l(\theta)}{\partial \theta_j}\right]_{s\times s} = B^T(\theta) B(\theta),\ \text{rang}(J) = s,$$

et

$$B(\theta) = \left[\frac{1}{\sqrt{p_l(\theta)}} \frac{\partial p_l(\theta)}{\partial \theta_j} \right]_{r \times s}.$$

La statistique du chi-deux de Pearson est :

$$X_n^2(\theta) = X_n^T(\theta) X_n(\theta) = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)} \tag{2.2}$$

où

$$X_n(\theta) = \left(\frac{\nu_1 - np_1(\theta)}{\sqrt{np_1(\theta)}}, \frac{\nu_2 - np_2(\theta)}{\sqrt{np_2(\theta)}}, ..., \frac{\nu_r - np_r(\theta)}{\sqrt{np_r(\theta)}} \right)^T, \tag{2.3}$$

et que

$$p_j(\theta) = \int_{a_{j-1}}^{a_j} f(t,\theta)\, dt = F(a_j) - F(a_{j-1}),\ j = 1, 2, ...r.$$

sous H_0, si θ est connu, Karl. Pearson a montré que pour n assez grand, on a :

Théorème 2.1.(Karl Pearson 1900)

$$\lim_{n \to \infty} P\left(X_n^2(\theta) \geq t | H_0\right) = P\left(\chi_{r-1}^2 \geq t\right) \tag{2.4}$$

où $\chi_{r-1}^2(\theta)$ est la distribution de chi-deux à $r-1$ degrés de liberté.
L'hypothèse H_0 est rejetée au seuil α, si :

$$X_n^2(\theta) \geq C_\alpha; \qquad C_\alpha = \chi_{r-1,\alpha}^2$$

C_α est la valeur critique du test de Pearson.

Si θ est inconnu, il faut l'estimer à l'aide des données. Par conséquent, la limite de distribution de la statistique de Pearson $X_n^2(\theta)$ ne suit plus asymptotiquement une loi du chi-deux χ_{r-1}^2 et dépend de la méthode d'estimation utilisée.

2.3 Tests d'ajustement du chi-deux modifié

Si les paramètres du modèle à tester sont inconnus, la NRR statistique Y_n^2 introduite par Nikulin (1973) et Rao et Robson (1974) se base sur les estimateurs du maximum de vraisemblance sur les données non regroupées recouvrant ainsi toute l'information apportée par l'échantillon.

La statistique de test s'écrit

$$Y_n^2\left(\hat{\theta}_n\right) = X_n^T\left(\hat{\theta}_n\right) W^-\left(\hat{\theta}_n\right) X_n\left(\hat{\theta}_n\right) \tag{2.5}$$

où $W^-(\theta)$ est la matrice inverse généralisée de$W(\theta)$, et

$$W(\theta) = I_r - q(\theta)q^T(\theta) - B(\theta)\, I^{-1}(\theta)\, B^T(\theta), \quad \text{rang } (W) = r-1,$$

où

$$q(\theta) = \left(\sqrt{p_1(\theta)}, \sqrt{p_2(\theta)}, ..., \sqrt{p_r(\theta)}\right)^T,$$

et I_r est la matrice unité d'ordre r.

En utilisant l'équation (2.5), Nikulin (1973) a montré que la matrice inverse généralisée W^- de W peut s'écrire :

$$W^-(\theta) = I_r + B(\theta)\left(I(\theta) - J(\theta)\right)^{-1} B^T(\theta),$$

et donc l'équation (2.5) devient :

$$Y_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + X_n^T\left(\hat{\theta}_n\right) B(\hat{\theta}_n)\left(I\left(\hat{\theta}_n\right) - J(\hat{\theta}_n)\right)^{-1} B^T(\hat{\theta}_n) X_n\left(\hat{\theta}_n\right).$$

Pour calculer la statistique NRR, on utilise souvent la formule quadratique simple suivante :

$$Y_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + Q_n\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + \frac{1}{n} L^T\left(\hat{\theta}_n\right)\left(I\left(\hat{\theta}_n\right) - J(\hat{\theta}_n)\right)^{-1} L\left(\hat{\theta}_n\right), \tag{2.6}$$

où $L(\theta) = \left(L_1(\theta),\ L_2(\theta), ..., L_s(\theta)\right)^T$, $s = \overline{1,\ j}$ et

$$L_j(\theta) = \sum_{i=1}^{r} \frac{\nu_i}{p_i} \frac{\partial p_i(\theta)}{\partial \theta_j}.$$

Le comportement asymptotique de la statistique $Y_n^2\left(\hat{\theta}_n\right)$ est donné par le théorème suivant.

Théorème 2.2. [Nikulin (1973a)] Pour n suffisamment grand, on a

$$\lim_{n\to\infty} \mathbf{P}(Y_n^2\left(\hat{\theta}_n\right) \geq t) = \mathbf{P}\left(\chi_{r-1}^2 \geq t\right). \tag{2.7}$$

2.4 Test d'ajustement du chi-deux modifié pour le modèle de Rayleigh généralisé

Nous voulons tester l'hypothèse composite H_0 selon laquelle la distribution de l'échantillon $T = (T_1, T_2, ...T_n)^T$ vérifie :

$$H_0 : P(T_i \leq t) = F_{GR}(t, \theta),\ t \geq 0, \qquad \theta = (\alpha, \beta)^T\ , \tag{2.8}$$

où $F_{GR}(t, \theta)$ est la fonction de répartition de la distribution de Rayleigh généralisée définie par l'équation (1.2) et $f_{GR}(t, \theta)$ sa densité de probabilité. Nous allons adapter la statistique de Nikulin-Rao-Robson. Pour cela, on partage l'ensemble $\mathbb{R}$ en r sous-intervalles $I_1, I_2, \cdots, I_r$ mutuellement disjoints, par les points :

$$-\infty = a_0 < a_1 < \cdots < a_{r-1} < a_r = +\infty,$$

où

$$I_j =]a_{j-1}, a_j];\ I_i \cap I_j = \phi,\ i \neq j;\ \bigcup_{j=1}^{r} I_j = R^1.$$

Soit ν le vecteur de fréquences $\nu = (\nu_1, \nu_2, \cdots, \nu_r)^T$ obtenu en regroupant l'échantillon $T_1, T_2, ..., T_n$ dans les sous-intervalles I_j.

Dans le cas d'équiprobabilité (i.e. $p_1 = p_2 = ... = p_r$), les a_j sont définies telles que :

$$\begin{aligned} a_j &= F_{GR}^{-1}(p_1 + p_2 + ... + p_r) = F_{GR}^{-1}\left(\frac{j}{r}\right);\ j = 1, ..., r-1, \\ a_j &= \frac{\sqrt{-\ln\left(1 - \left(\frac{j}{r}\right)^{\frac{1}{\alpha}}\right)}}{\beta}. \end{aligned}$$

Soit $p(\theta) = (p_1(\theta), p_2(\theta), ..., p_r(\theta))^T$ et

$$p_j(\theta) = \int_{I_j} f_{GR}(t, \theta)dt = F_{GR}(a_j) - F_{GR}(a_{j-1}), j = 1, 2, ...r.$$

Soit le vecteur $X_n(\theta)$

$$X_n(\theta) = \left(\frac{\nu_1 - np_1(\theta)}{\sqrt{np_1(\theta)}}, \frac{\nu_2 - np_2(\theta)}{\sqrt{np_2(\theta)}}, ..., \frac{\nu_r - np_r(\theta)}{\sqrt{np_r(\theta)}}\right)^T.$$

L'information de Fisher $I_n(\theta)$ pour les données non groupées est :

$$I_n(\theta) = nI(\theta) = n\left(I_{ij}\right)_{2\times 2} = n\begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\alpha^2} & -\frac{2n\beta t^2 \exp\left(-\beta^2 t^2\right)}{1-\exp\left(-\beta^2 t^2\right)} \\ -\frac{2n\beta t^2 \exp\left(-\beta^2 t^2\right)}{1-\exp\left(-\beta^2 t^2\right)} & \frac{2n}{\beta^2} + 2nt^2 - n(\alpha-1)\left[\frac{(2t^2-4\beta^2 t^4)\exp\left(-\beta^2 t^2\right)-2t^2\exp\left(-2\beta^2 t^2\right)}{\left(1-\exp\left(-\beta^2 t^2\right)\right)^2}\right] \end{pmatrix},$$

pour la calculer, nous prenons la relation suivante :

$$I(\theta) = -E\left(\frac{\partial^2 \ln(f_{GR}(t,\theta))}{\partial\theta^2}\right).$$

Pour tester l'hypothèse H_0, on utilise la statistique NRR donnée par l'équation (2.5) :

$$Y_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + \frac{1}{n}L^T\left(\hat{\theta}_n\right)\left(I\left(\hat{\theta}_n\right) - J\left(\hat{\theta}_n\right)\right)^{-1}L\left(\hat{\theta}_n\right)$$

où

$$L(\theta) = \left(L_1(\theta),\ L_2(\theta)\right)^T \text{ et } L_1(\theta) = \sum_{j=1}^{r}\frac{\nu_j}{p_j}\frac{\partial p_j(\theta)}{\partial\alpha};\ L_2(\theta) = \sum_{j=1}^{r}\frac{\nu_j}{p_j}\frac{\partial p_j(\theta)}{\partial\beta}$$

et la matrice d'information pour les données groupées $J(\theta)$ est :

$$J(\theta) = B(\theta)^T B(\theta), \qquad B(\theta) = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ . & . \\ . & . \\ b_{r1} & b_{r2} \end{pmatrix}$$

où

$$b_{j1}(\theta) = \frac{1}{\sqrt{p_j}}\frac{\partial p_j(\theta)}{\partial\alpha},\ b_{j2}(\theta) = \frac{1}{\sqrt{p_j}}\frac{\partial p_j(\theta)}{\partial\beta},\ j = 1, 2, ...r.$$

Pour la distribution de Rayleigh généralisée, on obtient

$$p_j(\theta) = \left(1 - e^{-\beta^2 a_j^2}\right)^{\alpha} - \left(1 - e^{-\beta^2 a_{j-1}^2}\right)^{\alpha}, \tag{2.9}$$

les dérivées partielles de $p_j(\theta)$, sont

$$\frac{\partial p_j(\theta)}{\partial \alpha} = \left(1 - e^{-\beta^2 a_j^2}\right)^{\alpha} \ln(1 - e^{-\beta^2 a_j^2}) - \left(1 - e^{-\beta^2 a_{j-1}^2}\right)^{\alpha} \ln\left(1 - e^{-\beta^2 a_{j-1}^2}\right);$$
$$\frac{\partial p_j(\theta)}{\partial \beta} = 2\alpha\beta \left(a_j^2 e^{-\beta^2 a_j^2} \left(1 - e^{-\beta^2 a_j^2}\right)^{\alpha} - a_{j-1}^2 e^{-\beta^2 a_{j-1}^2} \left(1 - e^{-\beta^2 a_{j-1}^2}\right)^{\alpha}\right),$$

et les éléments de la matrice d'information $J(\theta) = (J_{ij})_{2\times 2}$ sont :

$$J_{11} = \sum_{j=1}^{r} \frac{\left(A^{\alpha}(\beta, a_j) \ln A(\beta, a_j) - A^{\alpha}(\beta, a_{j-1}) \ln A(\beta, a_{j-1})\right)^2}{A^{\alpha}(\beta, a_j) - A^{\alpha}(\beta, a_{j-1})};$$

$$J_{12} = J_{21} = \left[\sum_{j=1}^{r} \frac{A^{\alpha}(\beta, a_j) \ln A(\beta, a_j) - A^{\alpha}(\beta, a_{j-1}) \ln A(\beta, a_j)}{A^{\alpha}(\beta, a_j) - A^{\alpha}(\beta, a_{j-1})}\right] \times \left[2\alpha\beta \sum_{j=1}^{r} \left(a_j^2 e^{-\beta^2 a_j^2} A^{(\alpha-1)}(\beta, a_j) - a_{j-1}^2 e^{-\beta^2 a_{j-1}^2} A^{(\alpha-1)}(\beta, a_{j-1})\right)\right];$$

$$J_{22} = 4\alpha^2\beta^2 \sum_{j=1}^{r} \frac{\left(a_j^2 e^{-\beta^2 a_j^2} A^{(\alpha-1)}(\beta, a_j) - a_{j-1}^2 e^{-\beta^2 a_{j-1}^2} A^{(\alpha-1)}(\beta, a_{j-1})\right)^2}{A^{\alpha}(\beta, a_j) - A^{\alpha}(\beta, a_{j-1})},$$

où

$$A(\beta, a_j) = 1 - e^{-\beta^2 a_j^2}.$$

Après un calcul laborieux, on obtient la forme explicite de la statistique NRR, pour les différentes situations, selon que les paramètres soient connus ou non, comme suit :

1. Si $\theta = (\alpha,\ \beta)^T$ est connu, la statistique Y_n^2 devient :

$$Y_n^2 = X_n^T(\theta) X_n(\theta) = X_n^2(\theta) = \sum_{j=1}^{r} \frac{(\nu j - n p_j)^2}{n p_j}.$$

2. a) Si β est connu et α est inconnu, on obtient :

$$Y_n^2 = X_n^2(\hat{\alpha}) + \frac{1}{n} \frac{\left(\sum_{j=1}^{r} \frac{\nu_j}{p_j} \frac{\partial p_j(\hat{\alpha})}{\partial \alpha}\right)^2}{\frac{n}{\hat{\alpha}^2} - \sum_{j=1}^{r} \frac{1}{p_j} \left(\frac{\partial p_j(\hat{\alpha})}{\partial \alpha}\right)^2}.$$

b) Si β est inconnu et α est connu, on a alors :

$$Y_n^2 = X_n^2\left(\hat{\beta}\right) + \frac{1}{n}\frac{\left(\sum_{j=1}^{r}\frac{\nu_j}{p_j}\frac{\partial p_j\left(\hat{\beta}\right)}{\partial\beta}\right)^2}{nI_{22} - \sum_{j=1}^{r}\frac{1}{p_j}\left(\frac{\partial p_j\left(\hat{\beta}\right)}{\partial\beta}\right)^2}.$$

3. Si $\theta = (\alpha,\ \beta)^T$ est inconnu, donc la statistique $Y_n^2\left(\hat{\theta}_n\right)$ devient :

$$Y_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + \frac{1}{n|M|}\left[\left(I_{22} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j2}^2\left(\hat{\theta}_n\right)\right) - \left(\sum_{j=1}^{r}\frac{\nu_j}{p_j}u_{j1}\left(\hat{\theta}_n\right)\right)^2\right] +$$
$$\frac{2}{n|M|}\left[\left(\sum_{j=1}^{r}\frac{1}{p_j}u_{j1}\left(\hat{\theta}_n\right)u_{j2}\left(\hat{\theta}_n\right) - I_{12}\right)\left(\sum_{j=1}^{r}\frac{\nu_j}{p_j}u_{j1}\left(\hat{\theta}_n\right)\right)\left(\sum_{j=1}^{r}\frac{\nu_j}{p_j}u_{j2}\left(\hat{\theta}_n\right)\right)\right] +$$
$$\frac{1}{n|M|}\left[\left(\frac{1}{\hat{\alpha}_n^2} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j1}^2\left(\hat{\theta}_n\right)\right)\left(\sum_{j=1}^{r}\frac{\nu_j}{p_j}u_{j2}\left(\hat{\theta}_n\right)\right)^2\right],$$

où

$$M = \begin{pmatrix} \frac{1}{\hat{\alpha}_n^2} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j1}^2\left(\hat{\theta}_n\right) & I_{12} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j1}\left(\hat{\theta}_n\right)u_{j2}\left(\hat{\theta}_n\right) \\ I_{12} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j1}\left(\hat{\theta}_n\right)u_{j2}\left(\hat{\theta}_n\right) & I_{22} - \sum_{j=1}^{r}\frac{1}{p_j}u_{j2}^2\left(\hat{\theta}_n\right) \end{pmatrix},$$

$$u_{j1}\left(\hat{\theta}_n\right) = \frac{\partial p_j\left(\hat{\theta}_n\right)}{\partial\alpha} \text{ et } u_{j2}\left(\hat{\theta}_n\right) = \frac{\partial p_j(\hat{\theta}_n)}{\partial\beta},$$

et $|M|$ est le déterminant de la matrice M.

Au seuil α, la valeur critique

$$C_\alpha = \chi^2_{r-1,1-\alpha},$$

et pour n suffisamment grand, la statistique $Y_n^2\left(\hat{\theta}_n\right)$ suit une distribution χ^2_{r-1} à $r-1$ degrés de liberté, l'hypothèse nulle H_0 est acceptée si

$$Y_n^2\left(\hat{\theta}_n\right) \leq C_\alpha.$$

Dans le cas contraire, H_0 est rejetée.

2.5 Autres tests d'ajustement

Soit $T = (T_1, T_2, ..., T_n)^T$ un n-échantillon, nous voulons tester l'hypothèse H_0 selon laquelle la distribution de l'échantillon T suit une distribution $F(t, \theta)$. Il existe d'autres statistiques, basées sur la différence entre les fonctions de répartition empiriques et les fonctions de répartition théoriques. Néanmoins ces tests ne peuvent être utilisés généralement que si les paramètres sont spécifiés.

2.5.1 Test de Kolmogorov-Smirnov

Le test d'ajustement de Kolmogorov-Smirnov est un test non paramétrique qui permet de tester l'hypothèse H_0 selon laquelle les données observées sont engendrées par une loi de probabilité théorique considérée comme étant un modèle convenable.

Soit la statistique :

$$D_n = \sup_{|n|<\infty} |F_n(t) - F(t, \theta)|, \tag{2.12}$$

où $F_n(t)$ est la fonction de répartition empirique associée à cet échantillon. En pratique, il est préférable d'utiliser le test basé sur n D_n avec la correction de Bolshev (1987), Bolshev et Smirnov (1983). Le problème de l'ajustement peut être énoncé comme suit

$$S_k = \frac{6nD_n + 1}{6\sqrt{n}},$$

où $D_n = \max(D_n^+, D_n^-)$,

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(t_i, \theta)\right\}, D_n^- = \max_{1\le i\le n}\left\{F(t_i, \theta) - \frac{i-1}{n}\right\}, \tag{2.13}$$

avec $t_1, t_2, ...t_n$ sont en ordre croissant.

Si θ est connu, la distribution de la statistique de bolshev obéit à la distribution de Kolmogorov -Smirnov (Bolshev et Smirnov, 1983).

2.5.2 Test d'Anderson-Darling

Avec la statistique d'ordre $T_{(1)} < T_{(2)} < .... < T_{(n)}$ générée par une distribution particulière, Anderson et Darling (1954) ont proposé une statistique

de test d'ajustement A_n^2 comme suit :

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left\{(2i-1)\ln F(t_{(i)},\theta) + (2n-2i+1)\ln(1-F(t_{(i)},\theta))\right\}. \tag{2.14}$$

Remarque 2.4 : Si θ est inconnu, la distribution de ces statistiques dépend d'un certain nombre de facteurs, la forme de la loi $F(t,\theta)$ correspondant à l'hypothèse H_0, la méthode d'estimation des paramètres et le nombre de paramètres estimés, (voir Lemeshko et Lemeshko (2009), Lemeshko et al. (2010a)).

Chapitre 3

Tests d'ajustement pour données censurées

3.1 Introduction

Lorsque les données sont censurées à droite, les tests d'ajustement classiques ne sont plus valables. Quelques techniques ont été suggérées comme le test d'Habib et Thomas (1986) basé sur l'estimateur de Kaplan Meyer. Récemment, Bagdonavičius et Nikulin (2011) ont proposé une modification de la statistique NRR basée sur l'estimation du maximum de vraisemblance sur des données non groupées et qui suit une distribution du chi-deux. Le principe de cette approche est que les nombres de défaillance théoriques doivent être égaux pour toutes les classes choisies pour le groupement des données. Ce test donne de bons résultats pour les modèles paramétriques, voir par exemple Bagdonavicius et al. (2013), Nikulin et Tran (2013), Goual et Seddik-Ameur (2014), Aidi et seddik-Ameur (2016), Chouia et Seddik-Ameur (2017). En utilisant cette approche, on propose la construction d'un test d'ajustement pour la distribution de Rayleigh généralisée. Nous calculons explicitement tous les éléments qui composent le critère de test pour ce modèle.

On peut consulter le livre de Voinov et al. (2013) qui donne une synthèse sur la construction et les applications des tests du type du chi-deux.

3.2 Test d'ajustement de Bagdonavičius et Nikulin pour données censurées à droite

Pour tester l'hypothèse nulle H_0 selon laquelle une variable aléatoire T suit une distribution paramétrique $F(t,\theta)$, Bagdonavičius et Nikulin (2011) procèdent comme suit.

Soit l'échantillon censuré à droite

$$(X_1,\delta_1),(X_2,\delta_2),.....,(X_n,\delta_n), \tag{3.1}$$

où $X_i=\min(T_i,C_i)$, C_1, C_2,.....,C_n sont les taux de censure et δ_i est la fonction indicatrice définie comme

$$\delta_i=\begin{cases}1 & \text{si } T_i\leq C_i\\ 0 & \text{ailleurs}\end{cases}.$$

Supposons que les taux de panne $T_1, T_2, ..., T_n$ sont n variables aléatoires indépendantes et identiquement distribuées. La fonction de densité de T_i appartient à une famille paramétrique $\{f(t,\theta),\theta\in\Theta\subset\mathbb{R}^m\}$ et dont le taux de hasard cumulé est noté par :

$$\Lambda(t,\theta)=-\ln S(t,\theta)=\int_0^t h(u,\theta)du.$$

La fonction de la log-vraisemblance est :

$$\ell(\theta)=\sum_{i=1}^{n}\delta_i\ln h(X_i,\theta)+\sum_{i=1}^{n}\ln S(X_i,\theta), \tag{3.2}$$

l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ vérifie :

$$\dot{\ell}(\hat{\theta})=0;$$

où $\dot{\ell}$ est le vecteur de score :

$$\dot{\ell}(\theta)=\frac{\partial}{\partial\theta}\ell(\theta)=\left(\frac{\partial}{\partial\theta_1}\ell(\theta),....,\frac{\partial}{\partial\theta_s}\ell(\theta)\right)^T. \tag{3.3}$$

La matrice d'information de Fisher est

$$I(\theta)=-E_\theta\ddot{\ell}(\theta), \tag{3.4}$$

où

$$\ddot{\ell}(\theta) = \sum_{i=1}^{n} \delta_i \frac{\partial^2}{\partial\theta^2} \ln h\left(X_i, \theta\right) - \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \Lambda\left(X_i, \theta\right). \qquad (3.5)$$

L'échantillon censuré (3.1) est représenté sous la forme de processus de comptage

$$\left(N_1\left(t\right), Y_1\left(t\right), t \geq 0\right), ..., \left(N_n\left(t\right), Y_n\left(t\right), t \geq 0\right), \qquad (3.6)$$

où

$$\begin{aligned} N_i\left(t\right) &= 1_{\{X_i \leq t, \delta_i = 1\}}, Y_i\left(t\right) = 1_{\{0 < t \leq X_i\}}, \\ N\left(t\right) &= \sum_{i=1}^{n} N_i\left(t\right) \text{ et } Y\left(t\right) = \sum_{i=1}^{n} Y_i\left(t\right). \end{aligned} \qquad (3.7)$$

les fonctions (3.2), (3.3), (3.4) et (3.5) seront écrites en termes des processus stochastiques N_i et Y_i. Les trajectoires de comptage $N_i\left(t\right)$ ont la forme :

$$N_i\left(t\right) = \begin{cases} 0, & 0 \leq t < X_i \\ 1, & t \geq X_i \end{cases}$$

Si $\delta_i = 1$ et $N_i\left(t\right) = 0$, si $\delta_i = 0, \delta_i = 0$, les trajectoires du processus de comptage Y_i ont la forme :

$$Y_i\left(t\right) = \begin{cases} 1, & 0 \leq t \leq X_i \\ 0, & t > X_i \end{cases}$$

L'utilisation de ces processus donne les deux relations :

$$\int_0^{\infty} \ln h\left(u, \theta\right) dN_i\left(u\right) = \delta_i \ln h\left(X_i, \theta\right), \qquad (3.8)$$

et

$$\int_0^{\infty} Y_i\left(u\right) h\left(u, \theta\right) du = \int_0^{X_i} h\left(u, \theta\right) du = -\ln S\left(X_i, \theta\right). \qquad (3.9)$$

Alors l'équation (3.2) s'écrit :

$$\ell(\theta) = \sum_{i=1}^{n} \int_0^{\infty} \{\ln h\left(u, \theta\right) dN_i\left(u\right) - Y_i\left(u\right) h\left(u, \theta\right)\} du.$$

La matrice d'information de Fisher est :

$$I(\theta) = -E_\theta \ddot{\ell}(\theta) = E_\theta \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial\theta} \ln h(u,\theta) \left(\frac{\partial}{\partial\theta} \ln h(u,\theta) \right)^T Y_i(u)\, h(u,\theta)\, du.$$

Les processus N_i et Y_i sont supposés être observés pendant un temps fini τ. L'intervalle $[0,\tau]$ est partagé en $r > s$ sous-intervalles où s est le nombre de paramètres

$$I_j = (a_{j-1}, a_j], \quad a_0 = 0, \quad a_r = \tau,$$

Soit :

$$U_j = N(a_j) - N(a_{j-1}) = \sum_{i:X_i \in I_j} \delta_i, \tag{3.10}$$

le nombre des défaillances observées dans le j^{ème} intervalle, $j = 1, 2, ..., r$.
On peut prévoir d'observer le nombre de défaillance

$$e_j = \int_{a_{j-1}}^{a_j} Y_i(u)\, h\left(u, \hat{\theta}\right) du = \sum_{i:X_i > a_{j-1}} \left(\Lambda\left(a_j \wedge X_i, \hat{\theta}\right) - \Lambda\left(a_{j-1}, \hat{\theta}\right) \right), \tag{3.11}$$

où $a \wedge b = \min(a,b)$; ici $\hat{\theta}$ est l'estimateur du maximum de vraisemblance du paramètre θ.

Pour tester l'hypothèse H_0, Bagdonavičius et Nikulin (2011) ont proposé la statistique :

$$\hat{Y}^2 = Z^T \widehat{V}^- Z, \tag{3.12}$$

basée sur le vecteur

$$Z = (Z_1, Z_2, ..., Z_r)^T, \; Z_j = \frac{1}{\sqrt{n}} (U_j - e_j), \tag{3.13}$$

où $\widehat{V}^-$ est l'inverse généralisée de la matrice $\widehat{V}$ et donnée par :

$$\widehat{V}^- = \widehat{A}^{-1} + \widehat{A}^{-1} \widehat{C}^T \widehat{G}^- \widehat{C} \widehat{A}^{-1}, \; \widehat{G} = \widehat{i} - \widehat{C} \widehat{A}^{-1} \widehat{C}^T.$$

Donc la statistique du test peut être écrite sous la forme :

$$\hat{Y}^2 = Z^T \widehat{A}^{-1} Z \; + Z^T \widehat{A}^{-1} \widehat{C}^T \widehat{G}^{-1} \widehat{C} \widehat{A}^{-1} Z = \sum_{j=1}^{r} \frac{(U_j - e_j)^2}{U_j} + Q, \tag{3.14}$$

où

$$Q = W^T \widehat{G}^- W,\ \widehat{W} = \widehat{C}\widehat{A}^{-1}Z = (\widehat{W}_1, ..., \widehat{W}_s)^T\ ,\ \widehat{W}_l = \sum_{j=1}^{r} \widehat{C}_{lj}\widehat{A}_j^{-1}Z_j,$$

$$\widehat{G} = [\widehat{g}_{ll'}]_{sxs},\ \ \widehat{g}_{ll'} = \widehat{i}_{ll'} - \sum_{j=1}^{r} \widehat{C}_{lj}\widehat{C}_{l'j}\widehat{A}_j^{-1},$$

$$\widehat{C}_{lj} = \frac{1}{n}\sum_{i:X_i\in I_j} \delta_i \frac{\partial}{\partial\theta}\ln h(X_i,\widehat{\theta}), \quad \widehat{i}_{ll'} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \ln h(X_i,\widehat{\theta})}{\partial\theta_l}\frac{\partial \ln h(X_i,\widehat{\theta})}{\partial\theta_{l'}},$$

$$\begin{aligned} \widehat{A}_j &= \frac{U_j}{n}, \qquad U_j = \sum_{i:X_i\in I_j} \delta_i, \\ i &= 1, ...n,\ \ j = 1, 2, ..., r \text{ et } \ l, l' = 1, ..., s. \end{aligned}$$

Choix de $\hat{a}_j$

Le principe de cette approche est que l'on doit s'assurer de l'égalité des effectifs théoriques dans les classes de groupement des données. Pour cela, les auteurs recommandent de prendre a_j , les limites des intervalles choisis pour le groupement des données, comme fonctions de données aléatoires. L'idée est de partager l'intervalle $[0, \tau]$ en r intervalles dont les nombres de défaillances prévus (qui ne sont pas nécessairement des entiers) sont tous égaux.

Soit

$$b_i = (n-i)\,\Lambda\left(X_{(i)}, \hat{\theta}\right) + \sum_{l=1}^{i} \Lambda\left(X_{(l)}, \hat{\theta}\right),$$

où $X_{(i)}$ est le i^{ème} élément dans les statistiques ordonnées $(X_{(1)}, X_{(2)}, ..., X_{(n)})$.

Si i est le plus petit nombre naturel qui vérifie que

$$E_j \in [b_{i-1}, b_i]\,, j = 1, 2, ..., r-1,$$

alors $\hat{a}_j$ est solution de l'équation :

$$(n-i+1)\,\Lambda\left(a_j, \hat{\theta}\right) + \sum_{l=1}^{i-1} \Lambda\left(X_{(l)}, \hat{\theta}\right) = E_j, \tag{3.15}$$

d'où

$$\hat{a}_j = \Lambda^{-1}\left(\left[E_j - \sum_{l=1}^{i-1}\Lambda\left(X_{(l)},\hat{\theta}\right)\right]/\left(n-i+1\right),\hat{\theta}\right), \qquad \hat{a}_r = \max\left(X_{(n)},\tau\right)$$

où Λ^{-1} est l'inverse de la fonction de hasard cumulé Λ.

Nous avons $0 < \hat{a}_1 < \hat{a}_2 < ..... < \hat{a}_r = \tau$, avec ce choix d'intervalles $e_j = \dfrac{E_r}{r}$ pour tout j où $E_r = \sum_{i=1}^{n}\Lambda(X_{(i)},\hat{\theta})$.

Bagdonavičius et al. (2010a) et Greenwood, Nikulin (1996) donnent des recommandations pour le choix des intervalles. S'il n'y a aucune autre hypothèse, le nombre d'intervalles r peut être pris comme $\dfrac{n}{r} > 5$.

Ainsi, en remplaçant a_j par $\hat{a}_j$ dans l'expression de la statistique $\hat{Y}^2$, la distribution limite de la statistique $\hat{Y}^2$ est encore un chi-deux, comme dans le cas de a_j fixe.

De la manière classique de la sélection des intervalles équiprobables nous fixons r et prenons $0 < P_1 < ... < P_r < 1$ de telle façon que

$$P_j = \frac{j}{r+1},\ j = 1,...,r$$

Par exemple, en prenant $r = 9$ nous avons

$$P_1 = 0,1, P_2 = 0,2, ..., P_9 = 0,9$$

$$a_j = F^{-1}(Pj,\hat{\theta}) = inf\{t : F\tau(t,\hat{\theta}) \geq P\}.$$

Sous H_0 la distribution limite de la statistique Y_n^2 est une distribution du chi-deux avec $k = rang(V^-) = Tr\left(V^-V\right)$ degrés de liberté. Si G est non dégénérée alors $k = r$.

3.3 Test du chi-deux modifié pour le modèle de Rayleigh généralisé avec données censurées

3.3.1 Choix des intervalles du groupement des données

1) Dans notre cas, le choix de $\widehat{a}_j$ est obtenu comme suit :

$$\hat{a}_j = \frac{1}{\hat{\beta}} \sqrt[)]{-\ln(1-(1-\exp(-M))^{\hat{\alpha}})}$$

où

$$M = \left[E_j - \sum_{l=1}^{i-1} \Lambda\left(X_{(l)}, \hat{\theta}\right) \right] / (n-i+1),$$

pour $j = 1, ..., k-1$ et $l = 1, ..., s$.Et

$$\begin{aligned} E_j &= \sum_{i:X_i > a_{j-1}} \ln \frac{S(a_{j-1}, \hat{\theta})}{S(a_j \wedge X_i, \hat{\theta})} \\ E_j &= \sum_{i:X_i > a_{j-1}} \ln \frac{1-(1-\exp(-\hat{\beta}^2 a_{j-1}^2)^{\hat{\alpha}}}{1-(1-\exp(-\hat{\beta}^2 (a_j \wedge X_i)^2)^{\hat{\alpha}}} \end{aligned}$$

2) Pour $r = 5$, nous avons

$$P_1 = \frac{1}{6}, P_2 = \frac{1}{3}, P_3 = \frac{1}{2}, P_4 = \frac{2}{3}, P_5 = \frac{5}{6},$$

les limites a_j des intervalles sont données par :

$$\begin{aligned} a_j &= F_{GR}^{-1}\left(P_j, \hat{\theta}\right), \ j = \overline{1,k}. \\ &= \frac{\sqrt{-\ln\left(1-(P_j)^{\frac{1}{\hat{\alpha}}}\right)}}{\hat{\beta}}. \end{aligned}$$

3.3.2 Calcul de la matrice $\widehat{W}$

Pour le calcul de la matrice $\widehat{W}$, on doit d'abord calculer la matrice $\widehat{C}$. Pour cela, on a besoin des dérivées partielles de la fonction $\ln h\left(X_i,\theta\right)$ par rapport α et β :

$$\begin{cases} \dfrac{\partial \ln h(X_i,\theta)}{\partial \alpha} = \dfrac{1}{\hat{\alpha}} + \dfrac{\ln\left(1-e^{-(\hat{\beta}X_i)^2}\right)}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}} \\ \dfrac{\partial \ln h(X_i,\theta)}{\partial \beta} = \dfrac{2}{\hat{\beta}} + \dfrac{2\hat{\beta}X_i^2\left(-1+\hat{\alpha}e^{-(\hat{\beta}X_i)^2}\right)}{1-e^{-(\hat{\beta}X_i)^2}} + \dfrac{2\hat{\alpha}\hat{\beta}X_i^2 e^{-(\hat{\beta}\ X_i)^2}\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}-1}}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}. \end{cases}$$

Après calcul, nous trouvons les éléments de la matrice $\widehat{C}$ qui sont donnés par :

$$\widehat{C}_{1j} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\left[\frac{1}{\hat{\alpha}} + \frac{\ln\left(1-e^{-(\hat{\beta}X_i)^2}\right)}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right]$$

$$\widehat{C}_{2j} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\left[\frac{2}{\hat{\beta}} + \frac{2\hat{\beta}X_i^2\left(-1+\hat{\alpha}e^{-(\hat{\beta}X_i)^2}\right)}{1-e^{-(\hat{\beta}X_i)^2}} + \frac{2\hat{\alpha}\hat{\beta}X_i^2 e^{-(\hat{\beta}\ X_i)^2}\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}-1}}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right]$$

3.3.3 Calcul de la matrice d'information de Fisher estimée

Dans ce cas, à partir de la fonction de la log-vraisemblance, nous pouvons calculer la matrice d'information de Fisher estimée $\widehat{i} = \left(\widehat{i}_{ll'}\right)_{2\times 2}$:

$$\widehat{i}_{ll'} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial \ln h(X_i,\widehat{\theta})}{\partial\theta_l}\frac{\partial \ln h(X_i,\widehat{\theta})}{\partial\theta_{l'}}, \quad l',l = 1,2.$$

Ils sont donnés comme suit :

$$\widehat{i}_{11} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \left[\frac{1}{\hat{\alpha}} + \frac{\ln\left(1-e^{-(\hat{\beta}X_i)^2}\right)}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right]^2;$$

$$\widehat{i}_{12} = \widehat{i}_{21} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \left[\frac{1}{\hat{\alpha}} + \frac{\ln\left(1-e^{-(\hat{\beta}X_i)^2}\right)}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right] \times$$
$$\left[\frac{2}{\hat{\beta}} + \frac{2\hat{\beta}X_i^2\left(-1+\hat{\alpha}e^{-(\hat{\beta}X_i)^2}\right)}{1-e^{-(\hat{\beta}X_i)^2}} + \frac{2\hat{\alpha}\hat{\beta}X_i^2 e^{-(\hat{\beta}\ X_i)^2}\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}-1}}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right];$$

$$\widehat{i}_{22} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \left[\frac{2}{\hat{\beta}} + \frac{2\hat{\beta}X_i^2\left(-1+\hat{\alpha}e^{-(\hat{\beta}X_i)^2}\right)}{1-e^{-(\hat{\beta}X_i)^2}} + \frac{2\hat{\alpha}\hat{\beta}X_i^2 e^{-(\hat{\beta}\ X_i)^2}\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}-1}}{1-\left(1-e^{-(\hat{\beta}\ X_i)^2}\right)^{\hat{\alpha}}}\right]^2.$$

Enfin, on obtient la forme explicite de la statistique$\hat{Y}^2$ du test du chi-deux modifié

$$\hat{Y}^2 = \sum_{j=1}^{r}\frac{(U_j - e_j)^2}{U_j} + Q,$$

pour la distribution de Rayleigh généralisée (GR), comme suit :

$$\hat{Y}^2 = \sum_{j=1}^{r}\frac{(U_j - e_j)^2}{U_j} + W^T\left[\widehat{i}_{ll'} - \sum_{j=1}^{r}\widehat{C}_{lj}\widehat{C}_{l'j}\hat{A}_j^{-1}\right]^{-1} W,\ l,l' = 1,2$$

$$\hat{Y}^2 = \sum_{j=1}^{r}\frac{(U_j - e_j)^2}{U_j} +$$
$$\sum_{j=1}^{r}\left[\frac{1}{n^{3/2}}\sum_{i=1}^{n}\delta_i\frac{\partial}{\partial\theta}\ln\left(h(X_i,\widehat{\theta})\right)\hat{A}_j^{-1}(U_j - e_j)\right]^T \times$$
$$[\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial}{\partial\theta_l}\ln\left(h(X_i,\widehat{\theta})\right)\frac{\partial}{\partial\theta_{l'}}\ln\left(h(X_i,\widehat{\theta})\right) -$$
$$\sum_{j=1}^{k}\widehat{C}_{lj}\widehat{C}_{l'j}\hat{A}_j^{-1})^{-1}] \times$$
$$\sum_{j=1}^{r}\left[\frac{1}{n^{3/2}}\sum_{i=1}^{n}\delta_i\frac{\partial}{\partial\theta}\ln\left(h(X_i,\widehat{\theta})\right)\hat{A}_j^{-1}(U_j - e_j)\right],$$

où $\widehat{\theta} = \left(\hat{\alpha}, \hat{\beta}\right)^T$, est l'estimateur du maximum de vraisemblance sur les données non regroupées du vecteur des paramètres inconnus θ.

Sous l'hypothèse H_0 et pour n suffisamment grand, la statistiqueY_n^2 suit asymptotiquement une loi de χ_k^2 à r degrés de liberté. L'hypothèse est rejetée avec niveau α si $\hat{Y}^2 > \chi_\alpha^2(r)$.

Chapitre 4

Le modèle AFT de la distribution de Rayleigh généralisée $(AFT - GR)$

4.1 Introduction

En fiabilité, pour mesurer les taux de défaillance en un temps court, on met le matériel sous des conditions appelées stress pour accélérer le processus provoquant les pannes, les modèles utilisés sont dits modèles à temps de vie accéléré AFT (accelerated failure time). Ces stress sont représentés par des covariables qui interviennent sur les fonctions d'intérêt du modèle. Ces modèles trouvent également des applications en analyse de survie, pour mesurer l'effet des différents schémas thérapeutiques ou les différentes caractéristiques des patients ayant des impacts directs sur la santé des patients.

Nous introduisons un modèle AFT dont la distribution de base est une distribution de Rayleigh généralisée. Après la présentation des caractéristiques du modèle $AFT - GR$ et l'étude des estimateurs du maximum de vraisemblance, on propose pour ce modèle, un test d'ajustement basé sur la statistique de Bagdonavičius et Nikulin (2011) dans le cas où les données sont censurées à droite et les paramètres inconnus.

Considérons E l'ensemble de tous les stress possibles, défini par :

$$E = \left\{ z\left(.\right) = (z_1(.), z_2(.), ..., z_m(.))^T, z : [0, \infty[\rightarrow R^m, z\left(.\right) \right\}$$

Nous écrivons z au lieu de $z(\cdot)$ si le stress est constant et on note $E_1 \subset E$ l'ensemble des stress constants. On suppose que le taux de défaillance (le

taux de panne) $T_{z(.)}$ sous $z(\cdot)$ est une variable aléatoire positive avec fonction de survie

$$S_{z(.)}(t) = P\left(T_{z(.)} \geq t\right),\ t > 0, z(\cdot) \in E.$$

Définition 4.1 : Le modèle AFT est défini sur E, si la fonction de survie sous le stress $z\,(.) \in E$ est

$$S_{z(.)}(t) = S_0\left(\int_0^t r(z(u))du\right), \qquad z(.) \in E. \tag{4.1}$$

La fonction $r\ (z)$ est souvent paramétrisée comme suit :

$$r(z) = e^{-\beta_0 - \beta_1 \psi(z)},$$

où $\psi\,(z)$ est une fonction donnée de x, cette fonction peut prendre plusieurs formes :

- $r(z) = e^{-\beta_0 - \beta_1 z}$, $\psi\,(z) = z$, c'est le modèle log-linéaire.
- $r(z) = e^{-\beta_0 - \beta_1 \ln z}$, $\psi\,(z) = \ln(z)$, c'est le modèle puissance.
- $r(z) = e^{-\beta_0 - \beta_1 / z}$, $\psi\,(z) = 1/z$, c'est le modèle d'Arrhenius.
- $r(z) = e^{-\beta_0 - \beta_1 \ln \frac{z}{1-z}}$, $0 < z < 1$, $\psi\,(z) = \ln \frac{z}{1-z}$, c'est le modèle de Meeker-Luvalle.

Généralement, nous prenons :

$$r(z) = \exp\{-\beta^T z\},$$

où $\beta = (\beta_0, \beta_1, ..., \beta_m)^T$ est le vecteur de régression inconnu. Dans ce cas, le modèle AFT paramétrique (4.1) est donné par la formule :

$$S_{z(.)}(t) = S_0\left(\int_0^t e^{-\beta^T z(u)}du\right), \qquad z(.) \in E, \tag{4.2}$$

sur l'ensemble des stress E_1 constants dans le temps, on a

$$S_{z(.)}(t) = S_0\left(e^{-\beta^T z}t\right), \quad z \in E_1, \tag{4.3}$$

et le logarithme du temps de défaillance T_z sous z peut s'écrire

$$\ln(T_z) = \beta^T z + \epsilon,$$

où la fonction ϵ ne dépend pas de z et est $S(t) = S_0\,(\ln t)$.

4.2 Construction du modèle $AFT - GR$

Pour un modèle AFT dont la distribution de base est une Rayleigh généralisée, de fonction de survie de base $S_0(t)$ donnée par :

$$S_0(t) = 1 - \left(1 - e^{-(\lambda\, t)^2}\right)^{\alpha},\ t > 0,\ \alpha > 0,\ \lambda > 0.$$

La fonction de distribution cumulative de $AFT - GR$ est déduite comme suit :

$$\begin{aligned} F(t) &= 1 - S_0(\exp\left\{-\beta^T z\right\} t) \\ &= \left(1 - \exp(-\lambda^2 \exp\left\{-2\beta^T z\right\} t^2)\right)^{\alpha}, \end{aligned} \tag{4.4}$$

où α et λ sont les paramètres de la distribution de Rayleigh généralisée GR, $\beta = (\beta_0, .., \beta_m)^T$ sont les paramètres de la régression du modèle AFT et $z = (1, z_1, ..., z_m)^T$ sont les covariables représentant les stress éventuels.

Ainsi, la fonction de survie pour le modèle $AFT - GR$ est obtenue comme suit :

$$\begin{aligned} S(t) &= S_0(\exp\left\{-\beta^T z\right\} t) \\ &= 1 - \left(1 - \exp(-\lambda^2 \exp\left\{-2\beta^T z\right\} t^2)\right)^{\alpha}, \end{aligned}$$

les fonctions risque et risque cumulé sont alors, respectivement :

$$\begin{aligned} h(t) &= -S^{'}(t)/S(t) = \frac{2\alpha\lambda^2 t \exp\left(-2\beta^T z - \lambda^2 t^2 e^{-2\beta^T z}\right)\left(1 - \exp(-\lambda^2 t^2 e^{-2\beta^T z})\right)^{\alpha-1}}{1 - \left(1 - \exp(-\lambda^2 t^2 \exp\left\{-2\beta^T z\right\})\right)^{\alpha}}, \\ \Lambda(t) &= -\ln S(t) = -\ln(1 - \left(1 - \exp(-\lambda^2 t^2 \exp\left\{-2\beta^T z\right\})\right)^{\alpha}). \end{aligned}$$

4.2.1 Estimation du maximum de vraisemblance en cas de données censurées

Soit T_i est une variable aléatoire selon une distribution $AFT - GR$, avec le vecteur de paramètres $\theta = (\alpha, \lambda, \beta_0, \beta_1)^T$. Supposons que les données se composent de n observations indépendantes

$$t_i = \min(T_i, C_i) \text{ pour } i = 1, ..., n.$$

La distribution de C_i ne dépend pas de paramètres inconnus de T_i.

La fonction de vraisemblance dans le cas de la censure

$$L(t,\theta) = \prod_{i=1}^{n} f(t_i,\theta) = \prod_{i=1}^{n} \lambda^{\delta_i}(t_i,\theta) S(t_i,\theta),\ \delta_i = 1_{\{T_i \leq C_i\}}. \tag{4.5}$$

Ce qui donne pour la distribution $AFT - GR$:

$$\begin{aligned} L(t,\theta) &= \prod_{i=1}^{n} \left[\frac{2\alpha\lambda^2 t_i \exp\left(-2\beta^T z_i - \lambda^2 t_i^2 e^{-2\beta^T z_i}\right)\left(1-\exp(-\lambda^2 e^{-2\beta^T z_i} t_i^2)\right)^{\alpha-1}}{1-\left(1-\exp(-\lambda^2 \exp\left\{-2\beta^T z_i\right\} t_i^2)\right)^{\alpha}} \right]^{\delta_i} \times \\ &\quad \left[1-\left(1-\exp(-\lambda^2 t_i^2 \exp\left\{-2\beta^T z_i\right\})\right)^{\alpha}\right]. \end{aligned}$$

La fonction de la log-vraisemblance est :

$$\begin{aligned} \ell(t,\theta) &= \sum_{i=1}^{n} \delta_i [\ln(2) + 2\ln(\lambda) + \ln(\alpha) + \ln(t_i) - 2\beta^T z_i - \lambda^2 t_i^2 \exp\left(-2\beta^T z_i\right) + \\ &\quad (\alpha-1)\ln(1-\exp(-\lambda^2 t_i^2 e^{-2\beta^T z_i}))] + \\ &\quad \sum_{i=1}^{n} (1-\delta_i)\ln(1-(1-\exp(-\lambda^2 t_i^2 e^{-2\beta^T z_i}))^{\alpha}). \end{aligned}$$

Après de longs calculs, nous obtenons les fonctions de score pour les paramètres $\alpha, \lambda, \beta_0, \beta_1, ..., \beta_m$:

$$\begin{aligned} \frac{\partial \ell(\alpha,\lambda,\beta)}{\partial \alpha} &= \sum_{i=1}^{n} \delta_i \left[\frac{1}{\alpha} + \ln(1-T(t_i,z_i)\right] - \\ &\quad \sum_{i=1}^{n} (1-\delta_i)\left[\frac{(1-T(t_i,z_i))^{\alpha}\ln(1-T(t_i,z_i))}{1-(1-T(t_i,z_i))^{\alpha}}\right]; \\ \frac{\partial \ell(\alpha,\lambda,\beta)}{\partial \lambda} &= \sum_{i=1}^{n} \delta_i \left[\frac{2}{\lambda} - 2\lambda t_i^2 e^{-2\beta^T z_i} + 2(\alpha-1)\frac{\lambda t_i^2 Z(t_i,z_i)}{1-T(t_i,z_i)}\right] - \\ &\quad \sum_{i=1}^{n} (1-\delta_i)\left[\frac{2\alpha\lambda t_i^2 Z(t_i,z_i)(1-T(t_i,z_i))^{\alpha-1}}{1-(1-T(t_i,z_i))^{\alpha}}\right]; \\ \frac{\partial \ell(\alpha,\lambda,\beta)}{\partial \beta_0} &= \sum_{i=1}^{n} \delta_i \left[-2 + 2\lambda^2 t_i^2 e^{-2\beta^T z_i} - 2(\alpha-1)\frac{\lambda^2 t_i^2 Z(t_i,z_i)}{1-T(t_i,z_i)}\right] + \end{aligned}$$

$$\sum_{i=1}^{n}(1-\delta_i)\left[\frac{2\alpha\lambda^2 t_i^2 Z(t_i,z_i)(1-T(t_i,z_i))^{\alpha-1}}{1-(1-T(t_i,z_i))^{\alpha}}\right]$$

$$\frac{\partial \ell(\alpha,\lambda,\beta)}{\partial \beta_1} = \sum_{i=1}^{n}\delta_i\left[-2z_{i1}+2\lambda^2 t_i^2 z_{i1}e^{-2\beta^T z_i}-2(\alpha-1)\frac{\lambda^2 t_i^2 z_{i1}Z(t_i,z_i)}{1-T(t_i,z_i)}\right]+$$

$$\sum_{i=1}^{n}(1-\delta_i)\left[\frac{2\alpha\lambda^2 t_i^2 z_{i1}Z(t_i,z_i)(1-T(t_i,z_i))^{\alpha-1}}{1-(1-T(t_i,z_i))^{\alpha}}\right];$$

$$\vdots$$

$$\frac{\partial \ell(\alpha,\lambda,\beta)}{\partial \beta_m} = \sum_{i=1}^{n}\delta_i\left[-2z_{im}+2\lambda^2 t_i^2 z_{im}e^{-2\beta^T z_i}-2(\alpha-1)\frac{\lambda^2 t_i^2 z_{im}Z(t_i,z_i)}{1-T(t_i,z_i)}\right]+$$

$$\sum_{i=1}^{n}(1-\delta_i)\left[\frac{2\alpha\lambda^2 t_i^2 z_{im}Z(t_i,z_i)(1-T(t_i,z_i))^{\alpha-1}}{1-(1-T(t_i,z_i))^{\alpha}}\right].$$

où $Z(t_i,z_i)=\exp\left(-2\beta^T z_i-\lambda^2 t_i^2 e^{-2\beta^T z_i}\right)$ et $T(t_i,z_i)=\exp(-\lambda^2 t_i^2 e^{-2\beta^T z_i})$.

4.2.2 Estimation et intervalle de confiance de la fonction de survie

Estimation de la fonction de survie

Si $\hat{\alpha},\hat{\lambda}$ et $\hat{\beta}$ sont les estimateurs de maximum de vraisemblance des paramètres du modèle $AFT-GR$, alors la fonction de survie sous un stress d'accélération constant $z \in E_0$ est :

$$\hat{S}_z(t) = 1-\left(1-\exp(-\hat{\lambda}^2\exp\left\{-2\hat{\beta}^T z\right\}t^2)\right)^{\hat{\alpha}}, \tag{4.6}$$

et la fonction de survie estimée sous un stress normal ou usuel quand $z=z^{(0)}$ peut être écrite comme suit :

$$\hat{S}_{z^{(0)}}(t) = 1-\left(1-\exp(-\hat{\lambda}^2\exp\left\{-2\hat{\beta}^T z^{(0)}\right\}t^2)\right)^{\hat{\alpha}}$$

Intervalle de confiance de la fonction de survie

En utilisant les propriétés des estimateurs du maximum du vraisemblance, sous le stress usuel, pour un risque α, l'intervalle de confiance pour la fonction

de survie $\hat{S}_{z^{(0)}}(t)$ est :

$$\left(1+\frac{1-\hat{S}_{z^{(0)}}(t)}{\hat{S}_{z^{(0)}}(t)}\exp\left[\pm\hat{\sigma}Q_{z^{(0)}}w_{1-\frac{\alpha}{2}}\right]\right)^{-1},$$

où $w_{1-\frac{\alpha}{2}}$ est le $\left(1-\frac{\alpha}{2}\right)$-quantile de $N(0,1)$ et

$$\hat{\sigma}^2Q_{z^{(0)}}=\frac{J_g^T\left(\hat{\theta}\right)I^{-1}\left(\hat{\theta}\right)J_g\left(\hat{\theta}\right)}{\left(\hat{S}_{z^{(0)}}(t)\right)^2\left(1-\hat{S}_{z^{(0)}}(t)\right)^2},$$

où

$$J_g^T\left(\hat{\theta}\right)I^{-1}\left(\hat{\theta}\right)J_g\left(\hat{\theta}\right)=\left(\frac{\partial S}{\partial\beta_0},...,\frac{\partial S}{\partial\beta_m},\frac{\partial S}{\partial\alpha},\frac{\partial S}{\partial\lambda}\right)\begin{pmatrix}I_{00}&.&I_{0(m+2)}\\I_{i0}&.&I_{i(m+2)}\\.&.&.\\I_{(m+2)0}&.&I_{(m+2)(m+2)}\end{pmatrix}\begin{pmatrix}\frac{\partial S}{\partial\beta_0}\\.\\\frac{\partial S}{\partial\beta m}\\\frac{\partial S}{\partial\alpha}\\\frac{\partial S}{\partial\lambda}\end{pmatrix}$$

$$=\left(\frac{\partial S}{\partial\beta_0},...,\frac{\partial S}{\partial\beta_m},\frac{\partial S}{\partial\alpha},\frac{\partial S}{\partial\lambda}\right)\begin{pmatrix}\sum_{j=0}^{m+2}I_{0j}\frac{\partial S}{\partial\beta j}\\\sum_{j=0}^{m+2}I_{1j}\frac{\partial S}{\partial\beta j}\\.\\\sum_{j=0}^{m+2}I_{(m+2)j}\frac{\partial S}{\partial\beta j}\end{pmatrix}$$

$$=\sum_{k=0}^{m+2}\sum_{j=0}^{m+2}\frac{\partial S}{\partial\beta_k}I_{kj}\frac{\partial S}{\partial\beta j}.$$

Ou il peut être écrit sous la forme simple :

$$\hat{\sigma}^2Q_{z^{(0)}}=\frac{1}{\left(\hat{S}_{z^{(0)}}(t)\right)^2\left(1-\hat{S}_{z^{(0)}}(t)\right)^2}\sum_{k=0}^{m+2}\sum_{j=0}^{m+2}\alpha_k\left(t,\hat{\alpha},\hat{\lambda},\hat{\beta}\right)I_{kj}\left(\hat{\alpha},\hat{\lambda},\hat{\beta}\right)\alpha_j\left(t,\hat{\alpha},\hat{\lambda},\hat{\beta}\right),$$

où $\alpha_k\left(t,\hat{\alpha},\hat{\lambda},\hat{\beta}\right)$ et $\alpha_j\left(t,\hat{\alpha},\hat{\lambda},\hat{\beta}\right)$ sont les dérivées partielles de la fonction de survie par rapport aux paramètres du modèle $AFT-GR$ tels que :

$$\frac{\partial S}{\partial\alpha}(t,\theta)=-\left(1-T(t,z)\right)^{\alpha}\ln(1-T(t,z));$$

$$\frac{\partial S}{\partial \lambda}(t,\theta) = -\frac{2\alpha\lambda t^2 Z(t,z)\left(1-T(t,z)\right)^{\alpha}}{1-T(t,z)};$$

$$\frac{\partial S}{\partial \beta_0}(t,\theta) = \frac{2\alpha\lambda^2 t^2 Z(t,z)\left(1-T(t,z)\right)^{\alpha}}{1-T(t,z)};$$

$$\frac{\partial S}{\partial \beta_1}(t,\theta) = \frac{2\alpha z_1\lambda^2 t^2 Z(t,z)\left(1-T(t,z)\right)^{\alpha}}{1-T(t,z)};$$

$$\vdots$$

$$\frac{\partial S}{\partial \beta_m}(t,\theta) = \frac{2\alpha z_m\lambda^2 t^2 Z(t,z)\left(1-T(t,z)\right)^{\alpha}}{1-T(t,z)}.$$

où $Z(t,z) = \exp\left(-2\beta^T z - \lambda^2 t^2 e^{-2\beta^T z}\right)$ et $T(t,z) = \exp(-\lambda^2 t^2 e^{-2\beta^T z})$

Et $I_{kj}\left(\hat{\alpha},\hat{\lambda},\hat{\beta}\right)$ sont les éléments de la matrice $I^{-1}\left(\hat{\alpha},\hat{\lambda},\hat{\beta}\right)$. La matrice de Fisher $I_{kj}(\alpha,\lambda,\beta)$ est estimée par :

$$I\left(\hat{\alpha},\hat{\lambda},\hat{\beta}\right) = -\frac{\partial^2 \ell\left(\hat{\alpha},\hat{\lambda},\hat{\beta}\right)}{\partial\theta_i \partial\theta_j}, \quad \theta = (\alpha,\lambda,\beta_0,\beta_1,.....,\beta_m)^T.$$

4.3 Test d'ajustement pour le modèle AFT-Rayleigh généralisé en cas de données censurées

En utilisant l'approche de Bagdonavičius et Nikulin (2011) et présentée plus haut, nous construisons un test du chi-deux modifié pour le modèle de durée de vie accélérée dont la distribution de base est une distribution de Rayleigh généralisée ($AFT-GR$).

Considérons l'hypothèse :

$$H_0 : F(t) \in F_0 = \{F_0(t,\theta),\ \theta \in \Theta \subset R^s\}, \qquad (4.7)$$

où $\theta = (\theta_1, ..., \theta_s)^T \in \Theta \subset R^s$ est le vecteur des paramètres inconnus de dimension s et F_0 est la fonction de répartition du modèle $AFT - GR$. On divise l'intervalle $[0, \tau]$ en $r > s$ intervalles $I_j = (a_{j-1}, a_j]$, $j = 1, 2, ..., r$ et $a_0 = 0$, $a_r = \tau$.

Le test est basé sur le vecteur $Z = (Z_1, ..., Z_r)^T$, $Z_j = \frac{1}{\sqrt{n}}(U_j - e_j)$, $j = 1, 2, ..., r$,

où U_j représente le nombre des défaillances observés dans ces intervalles.

Sous l'hypothèse H_0, la distribution limite de la statistique Y^2 (voir (3.14)) est chi-deux avec r degrés de liberté. L'hypothèse nulle H_0 est rejetée avec un niveau de signification approximatif α si $Y_n^2 > \chi_\alpha^2(r)$.

4.3.1 Le choix de $\hat{a}_j$

La valeur de $\widehat{a}_j$ estimée est donnée par :

$$\hat{a}_j = \Lambda^{-1}\left(\left[E_j - \sum_{l=1}^{i-1}\Lambda\left(t_{(l)}, \hat{\theta}\right)\right] / (n - i + 1), \hat{\theta}\right), \qquad \hat{a}_r = t_{(n)},$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance du paramètre θ, Λ^{-1} est l'inverse de la fonction de hasard cumulé Λ et

$$E_j = \frac{j}{r}E_r, \text{ où } E_r = \sum_{i=1}^{n}\Lambda(t_{(i)}, \widehat{\theta}).$$

Pour notre modèle, Le choix de $\hat{a}_j$ est obtenu comme suit :

$$\begin{aligned} \widehat{a}_j &= \frac{1}{\hat{\lambda}}\exp\left\{\beta^T z_i\right\}\left[-\ln\left(1 - \left[1 - \exp\left(-\frac{E_j - \sum_{l=1}^{i-1}\Lambda(t_l, \widehat{\theta})}{n - i + 1}\right)\right]^{1/\hat{\alpha}}\right)\right]^{1/2}, \\ j &= 1, ..., r-1, \widehat{a}_r = t_{(n)}. \end{aligned}$$

Pour un tel choix d'intervalles, nous avons $e_j = \frac{E_r}{r}$ pour tout j.

4.3.2 Calcul de la matrice $\widehat{W}$

Les éléments de la matrice $\widehat{W}$ qui est définie par :

$$\widehat{W_l} = \sum_{j=1}^{k} \widehat{C}_{lj} \widehat{A}_j^{-1} Z_j, \qquad l = 1, 2, ..., 4. \;\; j = 1, ..., r,$$

sont obtenus de la manière suivante :

$$\widehat{C}_{1j} = \frac{1}{n} \sum_{i:x_i \in I_j}^{n} \delta_i \left[\frac{\alpha \ln h(t_i,\theta) - \alpha h^{\alpha}(t_i) \ln h(t_i) + 1 - 2h^{\alpha}(t_i) + h^{2\alpha}(t_i)}{\alpha(1 - h^{\alpha}(t_i))^2} \right];$$

$$\begin{aligned} \widehat{C}_{2j} &= \frac{1}{n} \sum_{i:x_i \in I_j}^{n} \delta_i [2(\alpha - 1) \lambda t^2 h^{-1}(t_i) Z(t_i, z_i) + \\ &\quad \frac{2}{\lambda} - 2\lambda t_i^2 e^{-2\beta^T z_i} + 2(1 - h^{\alpha}(t_i))^{-1} \alpha \lambda t_i^2 h^{\alpha-1}(t_i) Z(t_i, z_i)]; \end{aligned}$$

$$\begin{aligned} \widehat{C}_{3j} &= \frac{1}{n} \sum_{i:x_i \in I_j}^{n} \delta_i \times \\ &\quad [-2(\alpha - 1) \lambda^2 t^2 h^{-1}(t_i) Z(t_i, z_i) + 2\lambda^2 t^2 e^{-2\beta^T z_i} - \\ &\quad 2 - 2\alpha \lambda^2 t_i^2 h^{\alpha-1}(t_i) (1 - h^{\alpha}(t_i))^{-1} Z(t_i, z_i)]; \end{aligned}$$

$$\begin{aligned} \widehat{C}_{4j} &= \frac{1}{n} \sum_{i:x_i \in I_j}^{n} \delta_i \times \\ &\quad [-2z_i(\alpha - 1) \lambda^2 t^2 h^{-1}(t_i) Z(t_i, z_i) + 2\lambda^2 t^2 z_i e^{-2\beta^T z_i} - \\ &\quad 2z_i - 2\alpha \lambda^2 t_i^2 z_i h^{\alpha-1}(t_i) (1 - h^{\alpha}(t_i))^{-1} Z(t_i, z_i)], \end{aligned}$$

où $h(t_i) = 1 - \exp(-\lambda^2 t_i^2 e^{-2\beta^T z_i})$.

4.3.3 Matrice d'information de Fisher $\widehat{i}_{[4\times4]}$

Les composantes de la matrice d'information $\widehat{i}_{[4\times4]}$ sont requises pour le calcul de la matrice G. Elles sont données comme suit :

$$\widehat{i}_{11} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\frac{\alpha\ln h(t_i,\theta)-\alpha h^{\alpha}(t_i)\ln h(t_i)+1-2h^{\alpha}(t_i)+h^{2\alpha}(t_i)}{\alpha(1-h^{\alpha}(t_i))^2}\right]^2;$$

$$\begin{aligned}\widehat{i}_{12} &= \widehat{i}_{21} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\frac{\alpha\ln h(t_i,\theta)-\alpha h^{\alpha}(t_i)\ln h(t_i)+1-2h^{\alpha}(t_i)+h^{2\alpha}(t_i)}{\alpha(1-h^{\alpha}(t_i))^2}\right]\times\\ &\quad[2(\alpha-1)\lambda t_i^2(h(t_i))^{-1}Z(t_i,z_i)+\frac{2}{\lambda}-2\lambda t_i^2e^{-2\beta^Tz_i}+\\ &\quad 2(1-(h(t_i))^{\alpha})^{-1}\alpha\lambda t_i^2(h(t_i))^{\alpha-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{13} &= \widehat{i}_{31} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\frac{\alpha\ln h(t_i,\theta)-\alpha h^{\alpha}(t_i)\ln h(t_i)+1-2h^{\alpha}(t_i)+h^{2\alpha}(t_i)}{\alpha(1-h^{\alpha}(t_i))^2}\right]\times\\ &\quad[-2(\alpha-1)\lambda^2t_i^2h^{-1}(t_i)Z(t_i,z_i)+2\lambda^2t_i^2e^{-2\beta^Tz_i}-2-\\ &\quad 2\alpha\lambda^2t_i^2h^{\alpha-1}(t_i)(1-h^{\alpha}(t_i))^{-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{14} &= \widehat{i}_{41} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\frac{\alpha\ln h(t_i,\theta)-\alpha h^{\alpha}(t_i)\ln h(t_i)+1-2h^{\alpha}(t_i)+h^{2\alpha}(t_i)}{\alpha(1-h^{\alpha}(t_i))^2}\right]\times\\ &\quad[-2z_i(\alpha-1)\lambda^2t_i^2h^{-1}(t_i)Z(t_i,z_i)+2\lambda^2t_i^2z_ie^{-2\beta^Tz_i}-2z_i-\\ &\quad 2\alpha\lambda^2t_i^2z_ih^{\alpha-1}(t_i)(1-h^{\alpha}(t_i))^{-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{22} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i[2(\alpha-1)\lambda t_i^2h^{-1}(t_i)Z(t_i,z_i)+\frac{2}{\lambda}-2\lambda t_i^2e^{-2\beta^Tz_i}+\\ &\quad 2(1-h^{\alpha}(t_i))^{-1}\alpha\lambda t_i^2h^{\alpha-1}(t_i)Z(t_i,z_i)]^2;\end{aligned}$$

$$\begin{aligned}\widehat{i}_{23} &= \widehat{i}_{32} = \frac{1}{n}\sum_{i=1}^{n}\delta_i[2\left(\alpha-1\right)\lambda t_i^2 h^{-1}(t_i)Z(t_i,z_i) + \\ &\frac{2}{\lambda} - 2\lambda t_i^2 e^{-2\beta^T z_i} + 2\left(1-h^{\alpha}(t_i)\right)^{-1}\alpha\lambda t_i^2 h^{\alpha-1}(t_i)Z(t_i,z_i)] \times \\ &[-2\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 e^{-2\beta^T z_i} - 2 - \\ &2\alpha\lambda^2 t_i^2 h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{24} &= \widehat{i}_{42} = \frac{1}{n}\sum_{i=1}^{n}\delta_i[2\left(\alpha-1\right)\lambda t_i^2 h^{-1}(t_i)Z(t_i,z_i) + \frac{2}{\lambda} - 2\lambda t_i^2 e^{-2\beta^T z_i} + \\ &2\left(1-h^{\alpha}(t_i)\right)^{-1}\alpha\lambda t_i^2 h^{\alpha-1}(t_i)Z(t_i,z_i)] \times \\ &[-2z_i\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 z_i e^{-2\beta^T z_i} - 2z_i - \\ &2\alpha\lambda^2 t_i^2 z_i h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{33} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i \times \\ &[-2\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 e^{-2\beta^T z_i} - 2 - \\ &2\alpha\lambda^2 t_i^2 h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)]^2;\end{aligned}$$

$$\begin{aligned}\widehat{i}_{34} &= \widehat{i}_{43} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \times \\ &[-2z_i\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 z_i e^{-2\beta^T z_i} - \\ &2z_i - 2\alpha\lambda^2 t_i^2 z_i h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)] \times \\ &[-2\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 e^{-2\beta^T z_i} - 2 - \\ &2\alpha\lambda^2 t_i^2 h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)];\end{aligned}$$

$$\begin{aligned}\widehat{i}_{44} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i \times \\ &[-2z_i\left(\alpha-1\right)\lambda^2 t_i^2 h^{-1}(t_i)Z(t_i,z_i) + 2\lambda^2 t_i^2 z_i e^{-2\beta^T z_i} - \\ &2z_i - 2\alpha\lambda^2 t_i^2 z_i h^{\alpha-1}(t_i)\left(1-h^{\alpha}(t_i)\right)^{-1}Z(t_i,z_i)]^2.\end{aligned}$$

Ensuite, nous obtenons la statistique Y^2

$$Y^2 = \sum_{j=1}^{r} \frac{(U_j - e_j)^2}{U_j} + \widehat{W}^T \left[\widehat{i}_{ll'} - \sum_{j=1}^{k} \widehat{C}_{lj} \widehat{C}_{l'j} \hat{A}_j^{-1} \right]^{-1} \widehat{W}, \quad l, l' = 1, 2, ..., 4$$

Chapitre 5

Le modèle mélange de deux distributions de Rayleigh généralisées

5.1 Introduction

Grâce à leur grande flexibilité, les modèles de mélange fini sont très utiles pour décrire une large variété de phénomènes aléatoires. Ils modélisent diverses situations réelles, où des mesures proviennent de différentes populations qui ont toutes des distributions différentes. De plus, les modèles de mélange définissent de nouvelles classes de distributions comme par exemple des densités asymétriques (un mélange de deux lois gaussiennes peut donner lieu à une densité asymétrique).

La fonction de répartition cumulative d'un modèle mélange (et la densité de probabilité si elle existe) peuvent être exprimées sous forme d'une combinaison convexe (par exemple une somme pondérée, avec des probabilités positives dont la somme est 1) d'autres fonctions de distribution et de fonctions de densité. Les répartitions individuelles qui sont combinées pour former la distribution du mélange sont appelées les composants du mélange, et les probabilités associées à chaque composant sont appelées les probabilités du mélange. On peut consulter les travaux d'Ahmad et Abdul Rahman (1994) pour le mélange de distributions de Weibull, et Liu et Shao (2003) pour le mélange de distributions normales.

Dans ce chapitre, on s'intéresse à l'estimation des paramètres d'un mo-

dèle mélange de deux distributions de Rayleigh généralisées. Nous utilisons l'algorithme EM et le logiciel R pour déterminer les estimateurs du maximum de vraisemblance des paramètres inconnus du modèle. Ensuite, nous construisons pour ce modèle, un test d'adéquation basé sur la statistique NRR.

5.2 Mélange fini de lois

Soit P l'ensemble de toutes les lois de probabilités définies sur un espace probabilisable (Ω, B). On appelle mélange fini de lois de probabilités p_i la probabilité p définie sur (Ω, B) par :

$$p = \sum_{i=1}^{k} \pi_i p_i,$$

où

$$\pi_i > 0, \ \sum_{i=1}^{r} \pi_i = 1, \ p_i \in P \text{ pour } i = 1, 2, ..., k.$$

Un certain nombre de résultats et de remarques découlent de cette définition.

Prenons $\Omega = \mathbb{R}$, ce qui ne restrient en rien la portée de ce qui suit :

a) $P\left(]-\infty, t[\right) = \sum_{i=1}^{k} \pi_i p_i \left(]-\infty, t[\right),$

et on a :

$$F(t) = \sum_{i=1}^{k} \pi_i F_i(t),$$

F et F_i, $i = 1, 2, ..., k$, étant les fonctions de répartitions de p et p_i.

b) Si les p_i admettent des densité f_i, par rapport à une même mesure $\sigma-$finie, p admet aussi une densité f qui vérifie, par simple dérivation du résultat précédent :

$$f(t) = \sum_{i=1}^{k} \pi_i f_i.$$

c) Si les probabilités p_i sont discrètes, c'est-à-dire définies sur un même

espace fini ou dénombrable, on a :

$$p\left(\{t\}\right)=\sum_{i=1}^{k}\pi_i p_i\left(\{t\}\right).$$

d) Il peut arriver que, les p_i étant connues, la loi p puisse être parfaitement déterminée. En effet, si φ_i est la fonction caractéristique de p_i la fonction φ de p s'écrit :

$$\varphi\left(x\right)=\int e^{ixt}dp=\sum_{i=1}^{k}\pi_i\int e^{ixt}dp=\sum_{i=1}^{k}\pi_i\varphi_i\left(x\right).$$

On sait alors que si p probabilité sur $\mathbb{R}$, admet une fonction φ intégrable par rapport à la mesure de Lebesgue, p admet une densité continue $f(t)$ définie par :

$$f(t)=\frac{1}{2\pi}\int e^{-ixt}\varphi\left(x\right)dx.$$

5.3 Cas d'un mélange de deux distributions de Rayleigh généralisées

Soit $T=(T_1,T_2,.....T_n)$ un échantillon d'observations issues d'un mélange de deux distributions Rayleigh généralisées (MGR) de paramètres $\lambda_i,\alpha_i,\beta_i$. La densité de X est donnée par

$$f(t,\lambda_i,\alpha_i,\beta_i)=\sum_{i=1}^{2}2\lambda_i\alpha_{1i}\beta_i^2te^{-(\beta_i x)^2}\left(1-e^{-(\beta_i t)^2}\right)^{\alpha_i-1},\qquad(5.1)$$

où $x>0$, $\lambda_i>0$, $\alpha_i>0$, $\beta_i>0$, telle que $\lambda_1=\lambda$, $\lambda_2=1-\lambda$. λ est le paramètre de mélange suivant une loi de Bernoulli $B(\lambda)$.

Sa fonction de répartition F_T est définie par la formule suivante :

$$F_T(t)=\lambda_1\left(1-e^{-(\beta_1 t)^2}\right)^{\alpha_1}+\lambda_2\left(1-e^{-(\beta_2 t)^2}\right)^{\alpha_2}.\qquad(5.2)$$

Et soit $Z=(Z_1,Z_2,.....Z_n)$ la donnée cachée où Z_i détermine la distribution dont est issue T_i :

$$L\left(T_i\mid Z_i=1\right)=GR(\alpha_1,\beta_1);\ L\left(T_i|Z_i=2\right)=GR(\alpha_2,\beta_2),$$

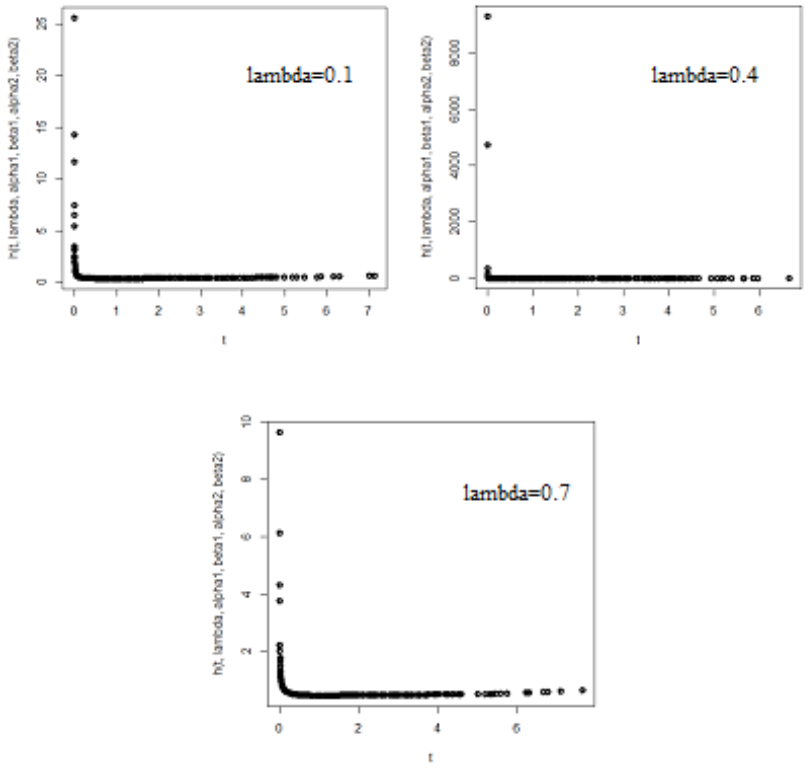


Fig 1.5: Le taux de hasard pour le modèle MGR avec alpha1=0.2, beta1=0.3, alpha2=0.4, beta2=0.5 et lambda=(0.1,0.4,0.7)

avec $P\{Z_i = 1\} = \lambda_1$ et $P\{Z_i = 2\} = \lambda_2 = 1 - \lambda_1$.

5.4 Estimation des paramètres inconnus d'un modèle mélange de distributions de Rayleigh généralisées

On cherche à estimer les cinq paramètres inconnus $\theta = (\lambda, \alpha_1, \beta_1, \alpha_2, \beta_2)$ du modèle mélange.

La vraisemblance des données complètes est définie par

$$L_n(T, Z\ |\theta) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2} 1_{\{Z_i=j\}} \lambda_j f_j(t_i) \right], \tag{5.3}$$

où $f_j(t) = 2\alpha_j \beta_j^2 t e^{-(\beta_j t)^2} \left(1 - e^{-(\beta_j t)^2}\right)^{\alpha_j - 1}$ est une densité de Rayleigh généralisée de paramètres α_j et β_j.

Passons à la log-vraisemblance des données complètes :

$$\begin{aligned} \log(L_n(T, Z\ |\theta)) &= \sum_{i=1}^{n} [\sum_{j=1}^{2} 1_{\{Z_i=j\}} \left(\log(\lambda_j) + \log(2) + \log(\alpha_j) + \log(\beta_j^2) + \log(t_i)\right) - \\ &\quad (\beta_j t_i)^2 + \log\left(1 - e^{-(\beta_j t_i)^2}\right)^{\alpha_j - 1}]. \end{aligned} \tag{5.4}$$

L'estimation des paramètres par les méthodes classiques telles que le maximum de vraisemblance s'avère difficile, c'est pourquoi on utilise des approches numériques. La méthode d'estimation via l'algorithme EM (Expectation-maximisation) très développée et largement utilisée ces dernières années nous donne des résultats très satisfaisants dans le cas de lois mélanges (voir Borman, 2004, Santos, 2010).

On procède comme suit :

- L'étape E : Calculons $E_{Z\ |T,\theta_m}[\log P(T, Z\ |\theta)]$.
- L'étape M : On maximise l'étape E

$$\theta_{m+1} = \arg\max_{\theta} \left\{ E_{Z\ |T,\theta_m}[\log P(T, Z\ |\theta)] \right\},$$

l'étape E nécessite de définir la distribution a posteriori de Z_j connaissant T_i et θ_m.

On définit

$$\tilde{p}_{i,j} = P\left\{Z_j = j \mid T_i = t_i, \theta_m\right\} = \frac{\lambda_j f_j\left(t_i\right)}{\lambda_1 f_1\left(t_i\right) + \lambda_2 f_2\left(t_i\right)},$$

la probabilité a posteriori pour que le point T_i soit issu de la distribution $f_j = GR(\alpha_j, \beta_j)$, connaissant θ_m.

Alors, on a :

$$\begin{aligned} E_{Z\mid T,\theta_m}\left[\log P\left(T, Z \mid \theta\right)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{2}\tilde{p}_{i,j}(\log(\lambda_j) + \log(2) + \log(\alpha_j) + \log(\beta_j^2) + \\ &\quad \log\left(t_i\right) - \left(\beta_j t_i\right)^2 + \log\left(1 - e^{-\left(\beta_j t_i\right)^2}\right)^{\alpha_j - 1}). \quad (5.5) \end{aligned}$$

La maximisation en θ de cette expression, bien qu'un peu lourde, ne présente aucune difficulté majeure et conduit aux estimateurs suivants (pour $j = 1$ ou 2)

$$\begin{cases} \lambda_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{i,j} \\ \alpha_j^{(m+1)} = -\frac{\sum_{i=1}^{n}\tilde{p}_{i,j}}{\sum_{i=1}^{n}\tilde{p}_{i,j}\log(1-e^{-\left(\beta_j t_i\right)^2})} \\ \beta_j^{(m+1)} = \frac{\sum_{i=1}^{n}\tilde{p}_{i,j}}{\sum_{i=1}^{n}\tilde{p}_{i,j}\beta_j t_i^2 + (1-\alpha_j)\sum_{i=1}^{n}\tilde{p}_{i,j}t_i^2 e^{-\left(\beta_j t_i\right)^2}(1-e^{-\left(\beta_j t_i\right)^2})^{-1}} \end{cases} \quad (5.6)$$

La résolution de ce système peut être donnée par des méthodes numériques (par exemple, la méthode de Newton Raphson, l'algorithme BB, etc...).

5.5 Test d'ajustement du chi-deux modifié pour le modèle mélange

Pour tester l'hypothèse H_0, nous pouvons utiliser la statistique de Nikulin-Rao-Robson donnée par l'équation :

$$\hat{Y}_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + \frac{1}{n}L^T\left(\hat{\theta}_n\right)\left(I\left(\hat{\theta}_n\right) - J\left(\hat{\theta}_n\right)\right)^{-1}L\left(\hat{\theta}_n\right).$$

où $L(\theta) = (L_1(\theta),\ L_2(\theta), L_3(\theta),\ L_4(\theta), L_5(\theta))^T$ et

$$L_k(\theta) = \sum_{j=1}^{r} \frac{\nu_j}{p_j} \frac{\partial p_j(\theta)}{\partial \theta_l}, l = 1, 2, 3, 4, 5,$$

et la matrice d'information pour les données groupées $J(\theta)$ est :

$$J(\theta) = B(\theta)^T B(\theta), \qquad B(\theta) = (b_{jl}(\theta))_{1 \le l \le 5,\ 1 \le j \le r},$$

où

$$b_{j1}(\theta) = \frac{1}{\sqrt{p_j}} \frac{\partial p_j(\theta)}{\partial \lambda},\ b_{j2}(\theta) = \frac{1}{\sqrt{p_j}} \frac{\partial p_j(\theta)}{\partial \alpha_1}, b_{j3}(\theta) = \frac{1}{\sqrt{p_j}} \frac{\partial p_j(\theta)}{\partial \beta_1},$$

$$b_{j4}(\theta) = \frac{1}{\sqrt{p_j}} \frac{\partial p_j(\theta)}{\partial \alpha_2},\ b_{j5}(\theta) = \frac{1}{\sqrt{p_j}} \frac{\partial p_j(\theta)}{\partial \beta_2},\ j = 1, 2, ...r.$$

Pour le modèle de mélange, on obtient

$$\begin{aligned} p_j(\theta) &= = \int_{a_{j-1}}^{a_j} f_{MGR}(t, \theta)\, dx = F_{MGR}(a_j) - F_{MGR}(a_{j-1}) \\ & \lambda \left[\left(1 - e^{-(\beta_1 a_j)^2}\right)^{\alpha_1} - \left(1 - e^{-(\beta_1 a_{j-1})^2}\right)^{\alpha_1} \right] + \\ & (1 - \lambda) \left[\left(1 - e^{-(\beta_2 a_j)^2}\right)^{\alpha_2} - \left(1 - e^{-(\beta_2 a_{j-1})^2}\right)^{\alpha_2} \right], \end{aligned}$$

et les dérivées partielles sont

$$\begin{aligned} \frac{\partial p_j(\theta)}{\partial \lambda} &= \left(1 - e^{-(\beta_1 a_j)^2}\right)^{\alpha_1} - \left(1 - e^{-(\beta_1 a_{j-1})^2}\right)^{\alpha_1} - \\ & \left(1 - e^{-(\beta_2 a_j)^2}\right)^{\alpha_2} + \left(1 - e^{-(\beta_2 a_{j-1})^2}\right)^{\alpha_2}; \end{aligned}$$

$$\begin{aligned} \frac{\partial p_j(\theta)}{\partial \alpha_1} &= \lambda[\left(1 - e^{-\beta_1^2 a_j^2}\right)^{\alpha_1} \ln(1 - e^{-\beta_1^2 a_j^2}) - \\ & (1 - e^{-\beta_1^2 a_{j-1}^2})^{\alpha_1} \ln(1 - e^{-\beta_1^2 a_{j-1}^2})]; \end{aligned}$$

$$\begin{aligned} \frac{\partial p_j(\theta)}{\partial \beta_1} &= 2\alpha_1 \beta_1 \lambda[a_j^2 e^{-\beta_1^2 a_j^2} \left(1 - e^{-\beta_1^2 a_j^2}\right)^{\alpha_1} - \\ & a_{j-1}^2 e^{-\beta_1^2 a_{j-1}^2} \left(1 - e^{-\beta_1^2 a_{j-1}^2}\right)^{\alpha_1}]; \end{aligned}$$

$$\begin{aligned}\frac{\partial p_j(\theta)}{\partial \alpha_2} &= (1-\lambda)[\left(1-e^{-\beta_2^2 a_j^2}\right)^{\alpha_2}\ln(1-e^{-\beta_2^2 a_j^2}) - \\ &\quad \left(1-e^{-\beta_2^2 a_{j-1}^2}\right)^{\alpha_2}\ln(1-e^{-\beta_2^2 a_{j-1}^2})];\end{aligned}$$

$$\begin{aligned}\frac{\partial p_j(\theta)}{\partial \beta_2} &= 2\alpha_2\beta_2(1-\lambda)[a_j^2 e^{-\beta_2^2 a_j^2}\left(1-e^{-\beta_2^2 a_j^2}\right)^{\alpha_2} - \\ &\quad a_{j-1}^2 e^{-\beta_2^2 a_{j-1}^2}\left(1-e^{-\beta_2^2 a_{j-1}^2}\right)^{\alpha_2}].\end{aligned}$$

5.6 Calcul de la matrice d'information de Fisher

La matrice d'information de Fisher pour un modèle mélange à cinq paramètres $(\theta = \lambda, \alpha_1, \beta_1, \alpha_2, \beta_2)$ est donnée par :

$$I_n(\theta) = nI(\theta)$$

où

$$I(\theta) = -\frac{\partial^2 \ln(f(t,\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2))}{\partial\theta^2}$$

où

$$\begin{aligned}f(t,\theta) &= f(t,\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2) \\ &= 2\lambda\alpha_1\beta_1^2 t e^{-(\beta_1 t)^2}\left(1-e^{-(\beta_1 t)^2}\right)^{\alpha_1-1} + 2(1-\lambda)\alpha_2\beta_2^2 t e^{-(\beta_2 t)^2}\left(1-e^{-(\beta_2 t)^2}\right)^{\alpha_2-1},\end{aligned}$$

posons $A_1 = \left(1-e^{-(\beta_1 t)^2}\right)$, $A_2 = \left(1-e^{-(\beta_2 t)^2}\right)$ et on obtient :

$$I_{11} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\lambda^2} = \frac{\left(2\alpha_1\beta_1^2 t e^{-(\beta_1 t)^2}A_1^{\alpha_1-1} - 2\alpha_2\beta_2^2 t e^{-(\beta_2 t)^2}A_2^{\alpha_2-1}\right)^2}{(f(t,\theta))^2};$$

$$\begin{aligned}I_{12} &= I_{21} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\lambda\partial\alpha_1} \\ &= -\frac{\beta_1^2\alpha_2\beta_2^2 A_1^{\alpha_1-1}A_2^{\alpha_2-1}\left(1-\alpha_1\beta_1^2 t^2 + \alpha_1\ln A_1\right)}{e^{-\left(\beta_1^2+\beta_2^2\right)t^2}\left[(\lambda-1)\alpha_2\beta_2^2 A_2^{\alpha_2-1} - \lambda\alpha_1\beta_1^2 A_1^{\alpha_1-1}\right]^2};\end{aligned}$$

$$I_{13} = I_{31} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\lambda\partial\beta_1} = -\frac{2\alpha_1\beta_1\alpha_2\beta_2^2 A_1^{\alpha_1-1}A_2^{\alpha_2-1}\left(1-2\beta_1^3t^4e^{-2(\beta_1 t)^2}(\alpha_1-1)\right)}{e^{-\left(\beta_1^2+\beta_2^2\right)t^2}\left[(\lambda-1)\,\alpha_2\beta_2^2A_2^{\alpha_2-1}-\lambda\alpha_1\beta_1^2A_1^{\alpha_1-1}\right]^2};$$

$$I_{14} = I_{41} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\lambda\partial\alpha_2} = \frac{\beta_1^2\alpha_1\beta_2^2A_1^{\alpha_1-1}A_2^{\alpha_2-1}\left(1-\alpha_2\beta_2^2t^2+\alpha_2\ln A_2\right)}{e^{-\left(\beta_1^2+\beta_2^2\right)t^2}\left[(\lambda-1)\,\alpha_2\beta_2^2A_2^{\alpha_2-1}-\lambda\alpha_1\beta_1^2A_1^{\alpha_1-1}\right]^2}$$

$$I_{15} = I_{51} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\lambda\partial\beta_2} = \frac{2\alpha_1\beta_1^2\alpha_2\beta_2A_1^{\alpha_1-1}A_2^{\alpha_2-1}\left(1-2\beta_2^3t^4e^{-2(\beta_2 t)^2}(\alpha_2-1)\right)}{e^{-\left(\beta_1^2+\beta_2^2\right)t^2}\left[(\lambda-1)\,\alpha_2\beta_2^2A_2^{\alpha_2-1}-\lambda\alpha_1\beta_1^2A_1^{\alpha_1-1}\right]^2};$$

$$I_{22} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\alpha_1^2} = -\frac{8\lambda\beta_1^2te^{-(\beta_1 t)^2}A_1^{\alpha_1-1}\left(-\beta_1^2t^2+\ln A_1\right)}{f(t,\theta)} + \frac{\left[2\lambda\beta_1^2te^{-(\beta_1 t)^2}A_1^{\alpha_1-1}\left[1+\alpha_1\left(-\beta_1^2t^2+\ln A_1\right)\right]\right]^2}{(f(t,\theta))^2};$$

$$I_{23} = I_{32} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial\alpha_1\partial\beta_1} = 2\lambda\beta_1e^{-(\beta_2 t)^2}e^{-\left(\beta_1^2+\beta_2^2\right)t^2}A_2 \times$$
$$[2\alpha_1^2\lambda\beta_1^5t^4A_2A_1^{2\alpha_1-1} + (\lambda-1)\,A_1^{\alpha_1+1}A_2^{\alpha_2}\alpha_2\beta_2^2 \times$$
$$(1-2\alpha_1^2t^4\beta_1^3e^{-2(\beta_1 t)^2}B_1(\alpha_1^2-\alpha_1-1)+\alpha_1B_1-4\alpha_1t^4\beta_1^3e^{-2(\beta_1 t)^2})] \times$$
$$\left(-\lambda\alpha_1\beta_1^2e^{-\left(\beta_1^2+\beta_2^2\right)t^2}A_1^{\alpha_1}A_2+(\lambda-1)\alpha_2\beta_2^2e^{-\left(\beta_1^2+\beta_2^2\right)t^2}A_2^{\alpha_2}A_1\right)^{-2};$$

où $B_1 = -\beta_1^2t^2+\ln A_1$ et posons $B_2 = -\beta_2^2t^2+\ln A_2$.

$$I_{24} = I_{42} = -\frac{\partial^2 \ln(f\,(t,\theta))}{\partial\alpha_1\partial\alpha_2} = \frac{\left[2\lambda t\beta_1^2 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}(1+\alpha_1 B_1)\right]\left[2(1-\lambda)t\beta_2^2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}(1+\alpha_2 B_2)\right]}{(f(t,\theta))^2};$$

$$\begin{aligned} I_{25} &= I_{52} = -\frac{\partial^2 \ln(f\,(t,\theta))}{\partial\alpha_1\partial\alpha_2} = (f(t,\theta))^{-2} \times \\ &\left[2\lambda t\beta_1^2 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left(1+\alpha_1 \ln A1\right)\right] \times \\ &\left[4\left(1-\lambda\right)\alpha_2 t\beta_2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}\left(1-t^2\beta_2^2+t^2\beta_2^2\left(\alpha_2-1\right)e^{-(\beta_2 t)^2}A_2^{-1}\right)\right]; \end{aligned}$$

$$\begin{aligned} I_{33} &= -\frac{\partial^2 \ln(f\,(t,\theta))}{\partial\beta_1^2} = -4\lambda\alpha_1 t e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left(f(t,\theta)\right)^{-2} \times \\ &[f^3(t,\theta) - \left(4\lambda\alpha_1\beta_1 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left(1-\beta_1^2t^2+\beta_1^2t^2e^{-(\beta_1 t)^2}A_1^{-1}\left(\alpha_1-1\right)\right)\right)^2]^{-1} - \\ &\lambda\alpha_1 t^3\beta_1^2 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left(f(t,\theta)\right)^{-2} \times [-20+\left(\alpha_1-1\right)e^{-(\beta_1 t)^2}A_1^{-1}+8t^2\beta_1^2 - \\ &24t^2\beta_1^2 e^{-(\beta_1 t)^2}A_1^{-1}\left(\alpha_1-1\right)+8t^2\beta_1^2\left(e^{-(\beta_1 t)^2}\right)^2 A_1^{-1}\left(\alpha_1^2-3\alpha_1+2\right)] \times \\ &[f^3(t,\theta) - \left(4\lambda\alpha_1\beta_1 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left(1-\beta_1^2t^2+\beta_1^2t^2e^{-(\beta_1 t)^2}A_1^{-1}\left(\alpha_1-1\right)\right)\right)^2]^{-1}; \end{aligned}$$

$$\begin{aligned} I_{34} &= I_{43} = -\frac{\partial^2 \ln(f\,(t,\theta))}{\partial\beta_1\partial\alpha_2} \\ &= \left[4\lambda\alpha_1 t\beta_1 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}(1-t^2\beta_1^2+t^2\beta_1^2 e^{-(\beta_1 t)^2}A_1^{-1}\left(\alpha_1-1\right))\right] \times \\ &\left(2\left(1-\lambda\right)t\beta_2^2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}\left[1+\alpha_2\ln A_2\right]\right)\left(f(t,\theta)\right)^{-2}; \end{aligned}$$

$$\begin{aligned} I_{35} &= I_{53} = -\frac{\partial^2 \ln(f\,(t,\theta))}{\partial\beta_1\partial\beta_2} = (f(t,\theta))^{-4} \times \\ &4\lambda\alpha_1 t\beta_1 e^{-(\beta_1 t)^2} A_1^{\alpha_1-1}\left[1-t^2\beta_1^2+t^2\beta_1^2 e^{-(\beta_1 t)^2}A_1^{-1}\left(\alpha_1-1\right)\right] \times \\ &4\left(1-\lambda\right)\alpha_2 t\beta_2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}\left[1-t^2\beta_2^2+t^2\beta_2^2 e^{-(\beta_2 t)^2}A_2^{-1}\left(\alpha_2-1\right)\right]; \end{aligned}$$

$$I_{44} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial \alpha_2^2} = -\frac{4(1-\lambda)t\beta_2^2 e^{-(\beta_2 t)^2} A_2^{\alpha_2 - 1} \ln A_2(\alpha_2+2)}{f(t,\theta)} + \frac{\left(2(1-\lambda)t\beta_2^2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}[1+\alpha_2 \ln A_2]\right)^2}{(f(t,\theta))^2};$$

$$\begin{aligned} I_{45} &= I_{54} = -\frac{\partial^2 \ln(f(t,\theta))}{\partial \alpha_2 \partial \beta_2} \\ &= -4\left(f(t,\theta)\right)^{-2}(1-\lambda)\, t\beta_2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1} \times \\ &\quad \left[1 - t^2\beta_2^2 + t^2\beta_2^2 e^{-(\beta_2 t)^2} A_2^{-1}\left(2\alpha_2 - 1 + \alpha_2(\alpha_2-1)\ln A_2\right) + \alpha_2 \ln A_2(1 - t^2\beta_2^2)\right] \times \\ &\quad [f(t,\theta) - \left(2(1-\lambda)\, t\beta_2^2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}[1+\alpha_2 \ln A_2]\right) \times \\ &\quad \left(4(1-\lambda)\,\alpha_2 t\beta_2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}\left[1 - t^2\beta_2^2 + t^2\beta_2^2 e^{-(\beta_2 t)^2} A_2^{-1}(\alpha_2 - 1)\right]\right)]^{-1} \end{aligned}$$

$$\begin{aligned} I_{55} &= -\frac{\partial^2 \ln(f(t,\theta))}{\partial \beta_2^2} = -\left(f(t,\theta)\right)^{-2}(1-\lambda)\,\alpha_2 t e^{-(\beta_2 t)^2} A_2^{\alpha_2-1} \times \\ &\quad [(4 - 20t^2\beta_2^2 + 20t^2\beta_2^2 e^{-(\beta_2 t)^2} A_2^{-1}(\alpha_2 - 1) + 8t^4\beta_2^4 - 24t^4\beta_2^4 e^{-(\beta_2 t)^2} A_2^{-1}(\alpha_2 - 1) + \\ &\quad 8^4\beta_2^4\left(e^{-(\beta_2 t)^2}\right)^2 A_2^{-1}\left(\alpha_2^2 - 3\alpha_2 + 2\right)) \times \\ &\quad \left(f(t,\theta) - \left(4(1-\lambda)\,\alpha_2 t\beta_2 e^{-(\beta_2 t)^2} A_2^{\alpha_2-1}\left[1 - t^2\beta_2^2 + t^2\beta_2^2 e^{-(\beta_2 t)^2} A_2^{-1}(\alpha_2-1)\right]\right)^2\right)^{-1}]. \end{aligned}$$

Tous les composants des matrices sont obtenus, donc nous obtenons la statistique NRR

$$\hat{Y}_n^2\left(\hat{\theta}_n\right) = X_n^2\left(\hat{\theta}_n\right) + \frac{1}{n} L^T\left(\hat{\theta}_n\right)\left(I\left(\hat{\theta}_n\right) - J\left(\hat{\theta}_n\right)\right)^{-1} L\left(\hat{\theta}_n\right);$$

où $\hat{\theta}_n = \left(\hat{\lambda}, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2\right)$ est l'estimateur du maximum de vraisemblance de $\theta = (\lambda, \alpha_1, \beta_1, \alpha_2, \beta_2)$ obtenu après la résolution de système (5.6).

Chapitre 6

Simulations et Applications

6.1 Simulations

Une importante étude par simulations numériques a été réalisée pour illustrer l'utilité et la maniabilité des tests proposés dans ce travail. De plus, nous avons appliqué ces tests à des données réelles issues de la fiabilité et de l'analyse de survie. Toutes les simulations ont été réalisées à l'aide du logiciel statistique R, l'algorithme EM pour le modèle de mélange et l'algorithme Barzilai-Borwein (BB).

6.1.1 Estimateurs du maximum de vraisemblance des paramètres, cas des données complètes et censurées

Nous considérons des échantillons d'une distribution de Rayleigh généralisée avec des paramètres $\alpha = 2$ et $\beta = 3$. Les données ont été simulées $N = 10.000$ fois. Les valeurs de la moyenne des estimateurs du maximum de vraisemblance MLEs $\hat{\alpha}$ et $\hat{\beta}$ et leurs erreurs quadratiques moyennes (SME) sont calculées et données dans la table 1 et la table 2 (taille des échantillons $n = 100, n = 300$, et $n = 500$)

$N = 10.000$	$n = 100$	$n = 300$	$n = 500$
$\widehat{\alpha}$	2.060411	2.083386	2.157434
SME	0.0895411	0.031215	0.018386
$\hat{\beta}$	2.793059	3.078906	3.096955
SME	0.0262139	0.009344	0.005561

Table 1 : MLEs des paramètres α et β de la distribution de Rayleigh généralisée dans le cas des données complètes

$N = 10.000$	$n = 100$	$n = 300$	$n = 500$
$\widehat{\alpha}$	2.02226	2.01668	2.00434
SME	0.01786	0.01351	0.00480
$\hat{\beta}$	3.01807	3.00732	3.00176
SME	0.01521	0.00693	0.00433

Table 2 :MLEs des paramètres α et β de la distribution de Rayleigh généralisée dans le cas des données censurées

6.1.2 Estimateurs du maximum de vraisemblance des paramètres, cas du modèle AFT-GR

Dans le cas du modèle $AFT - GR$, nous avons simulé $N = 1000$ échantillons avec les valeurs des paramètres $\alpha = 2$, $\lambda = 3, \beta_0 = 6$ et $\beta_1 = -0.8$ de tailles respectives $n = 100, n = 300$ et $n = 500$. La méthode du maximum de vraisemblance donne les valeurs des MLEs $\hat{\alpha}$,.$\hat{\lambda}$, $\hat{\beta}_0$et $\hat{\beta}_1$ et leurs valeurs d'erreurs quadratiques moyennes EQM. Les résultats sont résumés dans la table 3 suivante :

$N = 1000$	$n = 100$	$n = 300$	$n = 500$
$\widehat{\alpha}$	1.71063	2.195526	2.005487
EQM	0.00209182	0.0001940519	0.00023154
$\widehat{\lambda}$	3.742451	2.819077	3.014142
EQM	0.00740269	0.0001946275	0.00005478
$\widehat{\beta_0}$	6.187793	5.910105	5.996235
EQM	0.0007913436	$3.27343e-05$	$2.354612e-05$
$\widehat{\beta_1}$	-0.7908987	-0.802198	-0.8002223
EQM	$6.945995e-07$	$6.460948e-09$	$5.654897e-10$

Table 3 : MLEs des paramètres α, λ, β_0 et β_1 du modèle $AFT - GR$

6.1.3 Estimateurs du maximum de vraisemblance des paramètres, cas de modèle mélange de deux distributions de Rayleigh généralisées

Pour le modèle mélange, nous avons simulé $N = 1000$ échantillons d'un modèle mélange de deux distributions de Rayleigh généralisées de paramètres $\lambda = 0.7, \alpha_1 = 0.2, \beta_1 = 0.3, \alpha_2 = 0.4$ et $\beta_2 = 0.5$ de taille $n = 100, n = 300$ et $n = 500$ respectivement. En utilisant l'algorithme EM et le logiciel R, on a calculé les estimateurs MLEs $\hat{\lambda}, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2$ et $\hat{\beta}_2$ et les valeurs des erreurs quadratiques moyennes *EQM*. Les résultats sont résumés dans la table 4 suivante :

$N = 1000$	$n = 100$	$n = 300$	$n = 500$
$\widehat{\lambda}$	0.69451233	0.7171572	0.7098541
EQM	0.003755767	0.002187682	0.007990055
$\widehat{\alpha}_1$	0.2463238	0.2296811	0.2097716
EQM	0.00855037	0.010145451	0.003655481
$\hat{\beta}_1$	0.3275523	0.2816178	0.3017516
EQM	0.003163524	0.001315777	0.005501159
$\widehat{\alpha}_2$	0.41221431	0.3998054	0.4062134
EQM	0.00907314	0.020288942	0.002214715
$\hat{\beta}_2$	0.4820595	0.5052816	0.4981365
EQM	0.004844981	0.006900255	0.002457866

Table 4 : MLEs des paramètres $\lambda, \alpha_1, \beta_1, \alpha_2$ et β_2 pour le modèle mélange

6.1.4 La statistique NRR

NRR pour les données complètes et censurées à droite

Dans cette section, nous avons effectué une étude par simulations pour évaluer la performance des tests proposés, à la fois dans le cas de données complètes et de données censurées à droite. Pour cela, nous avons généré $N = 10.000$ échantillons de différentes tailles ($n = 25$, $n = 30$, $n = 60$, $n = 100$, $n = 300$ et $n = 500$) respectivement de la distribution généralisée Rayleigh GR ($\alpha = 2$, $\beta = 3$) avec des données complètes et $N = 10.000$ échantillons avec différents pourcentages de censure non informative à droite

(20%, 25% et 30%). Après calcul des MLEs $\hat{\alpha}$, $\hat{\beta}$, nous calculons les valeurs des statistiques de critères $(Y_n^2, \hat{Y}^2)$ pour chaque échantillon. Ensuite, on compte le nombre de rejets de l'hypothèse H_0, avec différents niveaux de signification $\alpha = 1\%$, $\alpha = 2\%$ et $\alpha = 5\%$. Les résultats des niveaux de signification empiriques par rapport à leurs valeurs théoriques sont résumés dans la table 5 et la table 6 :

$N = 10,000$	$\alpha = 1\,\%$	$\alpha = 2\,\%$	$\alpha = 5\,\%$
$n = 25$	0.0057	0.0144	0.0435
$n = 30$	0.0060	0.0146	0.0442
$n = 60$	0.0063	0.0150	0.0463
$n = 100$	0.0079	0.0180	0.0474
$n = 300$	0.0080	0.0198	0.0483
$n = 500$	0.0105	0.0204	0.0499

Table 5 :"Niveaux de signification simulés pour Y_n^2 pour les données complètes par rapport à leurs valeurs théoriques"

$N = 10,000$	$\alpha = 1\,\%$	$\alpha = 2\,\%$	$\alpha = 5\,\%$
$n = 25$	0.0044	0.0182	0.0477
$n = 30$	0.0084	0.0188	0.0479
$n = 60$	0.0087	0.0191	0.0482
$n = 100$	0.0098	0.0198	0.0499
$n = 300$	0.0122	0.0210	0.0513
$n = 500$	0.0110	0.0209	0.0508

Table 6 :"Niveaux de signification simulés de $\hat{Y}^2$ pour les données censurées à droite par rapport à leurs valeurs théoriques "

Nous observons que les niveaux de signification simulés pour les statistiques Y_n^2 et $\hat{Y}^2$ coïncident dans les niveaux théoriques correspondants des distributions du chi-deux. Ainsi, les tests proposés peuvent être utilisés pour ajuster des données à une distribution généralisée de Rayleigh dans le cas complet et le cas de censure aléatoire droite.

Deuxièmement, les histogrammes de la distribution de la statistique $\hat{Y}^2$, sous l'hypothèse nulle H_0, par rapport aux distributions du chi-deux avec r degrés de liberté, donnés dans la figure 1, pour différentes valeurs de paramètres et différentes tailles d'échantillon confirment que $\hat{Y}^2$ suit une distribution χ_r^2 à r degrés de libertés.

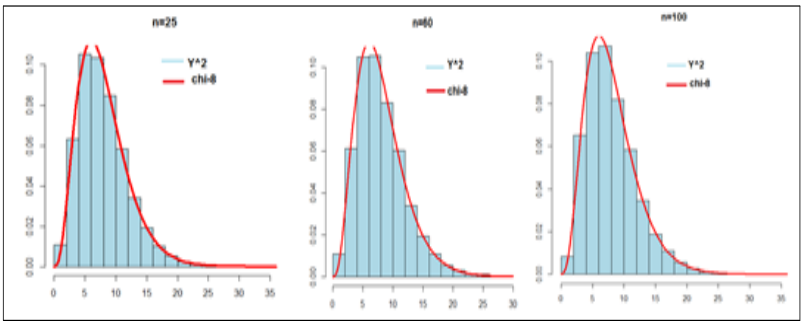


Figure 1: Y^2 sous l'hypothèse nulle H_0, ($\alpha = 2, \beta = 3$) versus la distribution du chi-deux avec 8 degrés de liberté ($r = 8, n = 25, 60, 100$), $N = 10.000$

NRR pour le modèle AFT-GR

De même, les valeurs du test statistique Y^2 sont calculées à partir des échantillons générés ($N = 1000$ fois) d'une distribution $AFT - GR$ de tailles respectives $n = 30, n = 60, n = 100$, $n = 300$ et $n = 500$. Nous calculons le nombre de cas d'hypothèses rejetées H_0 (le nombre de cas où $Y^2 < \chi^2_r$) respectivement ($\alpha = 2, \lambda = 3, \beta_0 = 6, \beta_1 = -0,8$) avec des données de différents pourcentages de censures non-informatives à droite ($20\%, 30\%$). Ensuite, on compte le nombre de rejets de l'hypothèse H_0, avec différents niveaux de signification alpha $\alpha = 1\,\%$, $\alpha = 2\,\%$ et $\alpha = 5\,\%$. Les résultats des niveaux de signification empiriques contre leurs valeurs théoriques sont résumés dans la table7.

$N = 1000$	$\alpha = 1\,\%$	$\alpha = 2\,\%$	$\alpha = 5\,\%$
$n = 30$	0.005	0.014	0.037
$n = 60$	0.005	0.016	0.044
$n = 100$	0.006	0.018	0.047
$n = 300$	0.008	0.021	0.052
$n = 500$	0.009	0.019	0.049

Table 7 :"Niveaux de signification simulés pour Y^2 pour les données censurées par rapport à leurs valeurs théoriques"

NRR pour le modèle de mélange MGR

La valeur du test statistique $\hat{Y}^2_n$ pour données complètes, est calculée à partir d'échantillons générés ($N = 1000$) à partir d'un modèle mélange de

deux distributions de Rayleigh généralisées de paramètres $\lambda = 0.7, \alpha_1 = 0.2, \beta_1 = 0.3, \alpha_2 = 0.4$, $\beta_2 = 0.5$, et de taille respectives $n = 30, n = 60, n = 100$, $n = 300$ et $n = 500$. Nous calculons le nombre de rejets de l'hypothèse nulle H_0 (le nombre de cas où $\hat{Y}_n^2 \succ \chi_\alpha^2$). Ensuite, on compte le nombre de rejets de l'hypothèse H_0, avec différents niveaux de signification $\alpha = 1\,\%$, $\alpha = 2\,\%$ et $\alpha = 5\,\%$. Les résultats des niveaux de signification empiriques contre leurs valeurs théoriques sont résumés dans la table 8.

$N = 1000$	$\alpha = 1\,\%$	$\alpha = 2\,\%$	$\alpha = 5\,\%$
$n = 30$	0.006	0.017	0.039
$n = 60$	0.007	0.018	0.046
$n = 100$	0.009	0.018	0.048
$n = 300$	0.009	0.019	0.051
$n = 500$	0.011	0.022	0.049

Table 8 :"Niveaux de signification simulés pour $\hat{Y}_n^2$ pour les données complètes par rapport à leurs valeurs théoriques"

6.2 Applications

Pour montrer l'utilité et les différentes applications des tests du chi-deux modifiés construits dans cette thèse, nous présentons quelques exemples illustratifs.

6.2.1 Distribution de Rayleigh généralisée

Cas de Données complètes

1) Les données (Pashardes et Christofides, 2008) portant sur la vitesse quotidienne du vent ont été extraites de l'index des données de l'Institut suédois de météorologie et d'hydrologie pendant 30 mois entre 2006 et 2008. Les 30 valeurs de la vitesse moyenne du vent sont les suivantes : 5.5, 4, 5.3, 5.7 , 4,1, 6,7, 5,4, 3,9, 2,8, 3,7, 2,9, 4,7, 3,8, 3,4, 2,5, 3,3, 3,5, 2,6, 4,1, 3,3, 6,9, 2,7, 2,0, 2,5, 2,8, 2,0, 3,2, 2,6, 3,8 et 4.0. On veut tester l'hypothèse H_0 selon laquelle ces observations sont modélisées par une distribution de Rayleigh généralisée. En utilisant le logiciel statistique R, nous trouvons les valeurs des estimateurs ML des paramètres :

$$\hat{\alpha} = 3.077557, \hat{\beta} = 0.342554.$$

Nous choisissons $r = 5$ intervalles. On obtient la valeur de la statistique NRR

$$Y_n^2 = 2.01468$$

Pour le niveau de signification $\alpha = 0.05$, $\chi^2(4) = 9.487729$, l'hypothèse nulle $H_0(Y_n^2 < \chi^2)$ que les données de la vitesse quotidienne du vent suivent une distribution de Rayleigh généralisée ne peut pas être rejetée.

Récemment, Abd-Elfattah (2011) a fourni les valeurs critiques de la statistique d'Anderson–Darling (AD) pour la distribution de Rayleigh généralisée et il a utilisé ces données pour les ajuster à ce modèle. Selon cette étude, la valeur correspondante de cette statistique est égale à $A_n^2 = 0.481$.Comme la valeur critique est de 2.113 pour le niveau de signification $\alpha = 0.05$, il accepte l'hypothèse selon laquelle ces données sont distribuées selon un modèle de Rayleigh généralisé. Ceci confirme notre résultat.

2) Dans un deuxième exemple, on veut tester si la hauteur des vagues peut être décrite par une distribution de Rayleigh généralisée. On utilise les données de prévisions de la hauteur des vagues à Alger plage pour la dernière mise à jour du 14/02/2017 qui sont les suivantes :

0.9 m, 1 m, 1.1 m, 1.1 m, 1 m, 0.9 m, 0.8 m, 0.7 m pour les heures 2h, 5h, 8h, 11h, 14h, 17h, 20h, 23h respectivement

(voir le site https ://www.meteoblue.com/fr/meteo/prevision/semaine/alger-plage_alg%C3%A9rie_2507474).

En utilisant le logiciel statistique R, nous trouvons les valeurs des estimateurs ML :

$$\hat{\alpha} = 26.59652, \hat{\beta} = 2.06841.$$

Après calcul de la statistique de NRR, nous trouvons $Y_n^2 = 2.015697$. Au seuil $\alpha = 0.05$, et pour $r = 5$, la valeur critique $\chi_{r-1}^2 = 9.487729$, d'où H_0 est loin d'être rejetée. On peut dire que la hauteur des vagues est modélisée par une distribution généralisée.

Cas de Données censurées

Pour montrer l'utilisation du test proposé pour le modèle de Rayleigh généralisé, on considère les données d'échantillon sur les cellules de réduction

d'aluminium de Whitmore (1983) représentant pour 20 cellules de réduction d'aluminium, le temps de rupture en unités de 1.000 jours :

$0.468-0.725-0.838-0.853-0.965-1.139-1.142-1.304-1.317-1.427-1.554-1.658-1.764-1.776-1.990-2.010-2.224-2.279^{*}-2.244^{*}-2.286^{*}$.
* Censure

Sous l'hypothèse nulle H_0, que ces données correspondent à la distribution de Rayleigh généralisée $GR(\alpha, \beta)$, et en utilisant le logiciel R, nous calculons d'abord les estimateurs du maximum de vraisemblance $\hat{\alpha} = 1.56743$, $\hat{\beta} = 0.67498$. Ensuite, nous donnons les valeurs des éléments des matrices $\hat{G}, \widehat{W}$, et $\hat{\imath}$. Si l'on prend par exemple $r = 5$ intervalles de groupement, les résultats sont donnés dans la table 9 :

j	1	2	3	4	5
$\hat{a}_j$	0.9277222	1.2496460	1.5397171	1.8501269	2.286
U_j	4	3	3	4	6
Z_j	0.13149031	-0.09211648	-0.09211648	0.13149031	0.57870391
e_j	3.411957	3.411957	3.411957	3.411957	3.411957

Table 9 : Les valeurs de $\hat{a}_j$, U_j, Z_j et e_j

La matrice estimée $\hat{G}$ est :

$$\hat{G} = \begin{bmatrix} 0.03156306 & -0.03201908 \\ -0.03201908 & 0.79903700 \end{bmatrix}$$

La matrice estimée $\widehat{W}$

$$\widehat{W} = \begin{bmatrix} -0.1139747 \\ 1.2496075 \end{bmatrix}$$

La matrice d'information estimée est :

$$\hat{\imath} = \begin{bmatrix} 0.3546345 & -1.60288 \\ -1.60288 & 11.56418 \end{bmatrix}$$

On obtient donc la valeur du test statistique $\hat{Y}^2 = 3.491465$. Pour un niveau de signification $\alpha = 0.05$, la valeur critique est $\chi^2_5 = 11.0705$, de sorte que les temps de rupture des cellules de réduction en aluminium peuvent être modélisés par une distribution généralisée de Rayleigh.

6.2.2 Modèle de Rayleigh généralisé à temps de vie accéléré $AFT - GR$

Nous étudions l'exemple des données de survie censurées, extraites du fichier "lupus" sur le site web : http ://lib.stat.cmu.edu/data sets/. Les données représentent le temps de survie (en mois) d'une biopsie rénale initiale pour 40 lupus de patients qui ont subi une biopsie rénale entre 1967 et 1983 et ont été suivis jusqu'au décès ou à la fin de 1990. En utilisant ces données, nous appliquerons le test chi-deux pour le modèle $AFT-GR$. Pour ce modèle nous prenons $z_{i1} = \ln z_{i1}$.

t_i	157*	268*	209	21	28	46	169	39	99	132	114	127	103	39	63
z_{i1}	1	10	2	0.1	7	24	2	22	0.1	0.1	1	4	11	57	1

t_i	159	75	32	148*	138	145	182	4	103	180
z_{i1}	0.1	1	1	44	10	0.1	14	106	1	7

t_i	54	188	19	118	48	13	165	86	78	25	4	65	62	89*	44
z_{i1}	1	0.1	0.1	10	1	6	0.1	8	0.1	15	34	0.1	8	0.1	77

En utilisant le logiciel statistique R, nous trouvons les valeurs des estimateurs ML : $\widehat{\alpha} = 0.5486051$, $\widehat{\lambda} = 18.92193$, $\widehat{\beta}_0 = 7.983122$, $\widehat{\beta}_1 = -0.01995037$.

Nous choisissons $r = 5$ intervalles. Le tableau suivant montre que les longueurs des classes aléatoires $(a_{j-1}, a_j]$ sont très différentes

$\widehat{a}_j$	33.00048	64.44906	99.90688	147.76466	268
U_j	8	8	5	8	7
Z_j	0.01885848	−0.07055504	−0.08886287	0.25060626	0.27840963
e_j	7.322908	7.322908	7.322908	7.322908	7.322908

$$\widehat{C} = \begin{bmatrix} -0.66414448 & -0.27642029 & -0.106444992 & -0.08248429 & -0.02926638 \\ 0.01258793 & 0.01423843 & 0.009648072 & 0.01705604 & 0.01630865 \\ -0.23818803 & -0.26941863 & -0.182560193 & -0.32273335 & -0.30859128 \\ -0.28754353 & -0.55696039 & 0.172728740 & -0.15107578 & 0.08718125 \end{bmatrix}$$

Par conséquent

$$\widehat{W} = \begin{bmatrix} -0.228893806, & 0.000896427, & -0.016962133, & -0.936919930 \end{bmatrix}^T.$$

Notons que dans le cas de la distribution $AFT - GR$, la matrice $\widehat{G}$ est dégénérée. Donc, le rang de la matrice V est $r-1$, et nous avons et la matrice d'information de Fisher

$$\widehat{I} = \begin{bmatrix} 3.07430936 & -0.078595052 & 1.4871704 & 2.05174286 \\ -0.07859505 & 0.005529806 & -0.1046346 & -0.04909417 \\ 1.48717045 & -0.104634625 & 1.9798896 & 0.92895674 \\ 2.05174286 & -0.049094173 & 0.9289567 & 9.30495798 \end{bmatrix}$$

Nous trouvons

$$Q = 0.9247978,\ X^2 = 1.522067$$

la statistique de test est alors

$$Y^2 = 2.446865, \text{ et } \chi^2_{0.5}(4) = 9.487729.$$

Nous concluons donc que l'hypothèse H_0 est acceptée et par conséquent ces données sont distribuées selon un modèle $AFT - GR$.

6.2.3 Modèle mélange de deux distributions de Rayleigh généralisées

Dans cet exemple, on considère les données de la durée de vie de 20 composants électroniques qui ont été prises de Murthy et al (2004). Les 20 valeurs sont les suivantes : 0.03, 0.12, 0.22, 0.35, 0.73, 0.79, 1.25, 1.41, 1.52, 1.79, 1.80, 1.94, 2.38, 2.4, 2.87, 2.99, 3.14, 3.17, 4.72, 5.09. On suppose que ces données peuvent provenir du mélange de deux distributions de Rayleigh généralisées de paramètres respectifs α_1, β_1 et α_2, β_2 et le paramètre de mélange λ.

En utilisant l'algorithme "BB" et l'algorithme EM, on calcule les valeurs des estimateurs du maximum de vraisemblance de 5 paramètres du modèle :

$$\hat{\lambda} = 0.6202931, \hat{\alpha}_1 = 0.3704238,\ \hat{\beta}_1 = 0.4069023, \hat{\alpha}_2 = 0.6899404 \text{ et } \hat{\beta}_2 = 0.3205562.$$

Si par exemple, on choisit $r = 5$ intervalles pour regrouper ces données, la valeur de la statistique du test NRR est :

$$\hat{Y}_n^2 = 6.978987$$

Pour le niveau de signification $\alpha = 0.05$, $\chi^2(4) = 9.487729$, l'hypothèse nulle $H_0(\hat{Y}_n^2 < \chi^2)$ n'est pas rejetée. On conclut que les données représentant la durée de vie de ces composants électroniques suivent une distribution mélange de deux distributions de Rayleigh généralisées.

Chapitre 7

Conclusion et perspectives

La validation des modèles choisis pour n'importe quelle analyse statistique est nécessaire si nous voulons obtenir des résultats fiables. C'est pourquoi les méthodes et les techniques de tests d'ajustement sont en perpétuel développement. Quand la distribution est bien spécifiée, on peut utiliser n'importe quel test classique, cependant pour valider une hypothèse composite quand les paramètres sont inconnus et doivent être estimés à partir de l'échantillon et si en plus les données sont censurées, ces tests ne sont plus adaptés et les distributions des statistiques de test dépendent de la méthode d'estimation utilisée et du modèle proposé. Par ce travail, on a proposé des tests d'ajustement pour le modèle de Rayleigh généralisé dans les cas de données complètes et censurées et les paramètres inconnus, pour le modèle de Rayleigh généralisée à temps de vie accéléré et pour un modèle mélange de deux distributions de Rayleigh généralisées. L'intérêt de ces modèles et la maniabilité des tests proposés dans ce travail sont mis en évidence dans les différentes applications. On espère que ces résultats seront bénéfiques pour les utilisateurs.

En perspective, on se propose de construire de nouveaux modèles plus flexibles généralisant la distribution de Rayleigh ainsi que des tests pour valider ceux-ci.

Bibliographie

[1] Abd-Elfattah, A. M., (2011). Goodness-of-fit test for the generalized Rayleigh distribution with unknown parameter. Journal of statistical computation and simulation,Vol.81, issue 3, pp. 357-366.

[2] Ahmad, K.E. and Abdul Rahman, A.M., (1994), Upgrading a Nonlinear Discriminate Function Estimated from a Mixture of two Weibull Distributions, Mathematics and Computer Modelling, 18, pp. 41-51.

[3] Aidi, K. and Seddik-Ameur, N., (2016), Chi-square tests for generalized exponential distributions with censored data, Electronique journal of applied statistical analysis, 9, 2, 371-384.

[4] Al Khedhairi, A., Sarhan, A. and Tadj, L., (2007). Estimation of the generalized Rayleigh distribution parameter. International journal of reliability and applications, vol.12, pp. 199-210.

[5] Bagdonavicius, V., Nikulin, M. (2011). Chi-squared tests for general composite hypotheses from censored samples. Comptes Rendus de l'Académie des Sciences de Paris, Mathématiques, V.349, N° 3-4, 219-223.

[6] Bagdonavicius, V., Levuliene, R.J., and Nikulin, M. (2013). Chi-squared goodness-of-fit tests for parametric accelerated failure time model. Communications in Statistics-Theory and Methods, Volume : 42 Issue : 15 Pages :2768-2785.

[7] Bagdonavicius, V., Kruopis, J., and Nikulin, M. (2010a). Nonparametric tests for Censored Data. ISTE and J. Wiley.

[8] Bolshev L.N. and Smirnov N.V. Tables of Mathematical Statistics. -M. : Science, 1983.

[9] Chernoff, H. and Lehmann, E. (1954). The use of maximum likelihood estimates in tests for goodness of fit. Ann. Math. Statist., 25 :579–586. 4, 23, 28.

[10] Chouia, S. and Seddik-Ameur N. (2017), A modified chi-square test for Bertholon model with censored data, communications in statistics -simulation

and computation, 46, 1, 593-602.

[11] Dempster, A.P. ; Laird, N.M. ; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) : 1–38. JSTOR 2984875. MR0501537.

[12] Drost, F. (1988) Asymptotics for Generalized Chi-squared Goodness-of-fit Tests, amsterdam : Centre for Mathematics and Computer Sciences, CWI Tracs, 48.

[13] Fathipour, P., Abolhasani, A. and Khamnei, H. J. (2013). Estimating $R = P(Y < X)$ in the generalized Rayleigh distribution with different scale parameters. Applied mathematical sciences, Vol. 7, N°2, 87-92.

[14] Greenwood, P., Nikulin, M.S. (1996). A Guide to Chi-squared Testing,Wiley : New York.

[15] Goual, H. and Seddik Ameur, N. (2014). Chi-squared type test for the AFT-generalized inverse Weibull distribution. Cummunication in Statistics-Theory and Method, 43, 13, 2605-2617.

[16] Habib, M.G., Thomas, D.R. (1986) Chi-squared Goodness-of-Fit Tests For Randomly Censored Data, Annals of Statistics, V.14, N 2, p 759-765.

[17] Hata M., (1980), " Empirical formula for propagation loss in land mobile radio service ", IEEE Transactions on Vehicular Technology vol. 29, p. 317-325.

[18] Hjort, N.L. (1990) Goodness of Fit Test in Models for Life History Data Based on Cumulative Hazard Rates, The annals of statistics, V.18, N3, p 1221-1258.

[19] Kundu, D. and Raqab, M.Z. (2005). Generalized Rayleigh distribution : different methods of estimation. Computational Statistics and Data Analysis, 49, 187-200.

[20] Lemeshko, B.Yu. (2009) Models for statistical distributions in nonparametric tting tests on composite hypotheses based on maximum-likelihood estimators. Part II / B.Yu. Lemeshko, S.B. Lemeshko // Measurement Techniques. Vol. 52, No. 8. - P.799-812.

[21] Lemeshko,B.Yu, Chimitova, E.V., Pleshkova, T. A. (2010a). Testing simple and composite goodness of fit hypotheses by censored samples. Nauchniy Vestnik NGTU 41 : 13–28.

[22] Liu X. et Shao Y. (2003). Asymptotics for the likelihood ratio test in a two-component normal mixture model. Annals of Statistics, vol. 31.

[23] Meeker, W.Q., Escobar, L.A., (1998) Statistical Methods for reliability Data. John Wiley and Sons, INC.

[24] Mudholkar, G.S.and Srivastava, D.K. (1993). "Exponentiated Weibull family for analyzing bathtub failure-rate data". IEEE Transactions on Reliability. 42, 299- 302.

[25] Mudholkar, G.S, Srivastava, D.K, and Lin, C.T. (1995). Some p-variate adaptations of the Shapiro-Wilk test of normality. Communications in Statistics-Theory and Methods. 24, 953-985.

[26] M. Z. Raqab and D. Kundu, (2003) "Burr Type X Distribution : Revisited,". http ://home.iitk.ac.in/ kundu/paper118.pdf.

[27] Murthy, D.N.P, Xie, M. and Jiang, R., 2004, Weibull Models, John Wiley & Sons, Inc., Hoboken, New Jersey.

[28] Nikulin, M.S. (1973a) Chi-square Test For Continuous Distributions with Shift and Scale Parameters, teor. Veroyatn. Primen., 18, No. 3, p 559-568.

[29] Nikulin, M. S. (1973b) On a Chi-square test for continuous distributions, Theory of Probability and its Applications, 18, p 638-639.

[30] Nikulin. M.s, (1973c). On a Chi-squared test for continuous distributions, Theory of Probability and its Aplications. vol.18, 3, p 638-639.

[31] Nikulin, M.S., Solev, V.N. (1999) Chi-squares Goodness-of- t Test for Doubly Censored Data with Applications in Survival Analysis and Reliability, In : Statistical and Probabilistic Models in Reliability, D.C.Ionescu, N.Limnios (eds), Birkhauser, Boston, p 101-112.

[32] Raqab, M.Z. and Kundu, D. (2006) Burr type X distribution : Revisited, Journal of Probability and Statistical Sciences, 4(2), 179-193.

[33] Rao, Gadde Srinivasa (2014). "Estimation of Reliability in Multicomponent Stress-Strength Based on Generalized Rayleigh Distribution". Journal of Modern Applied Statistical Methods : Vol. 13 : Iss. 1, Article 24. Available at : http ://digitalcommons.wayne.edu/jmasm/vol13/iss1/24.

[34] Rayleigh, L. (1880). "On the resultant of a large number vibrations of the same pitch and of arbitrary phase", Phil, Mag 10, pp 73–80.

[35] Rao, K.C. , Robson, D.S. (1974). A chi-square statistic for goodness-of-fit for tests with in the exponential family. Communications in Statistics, 3 , 1139-1153.

[36] Santos, F., (2010). L'algorithme EM : Une courte présentation ; université Bordeaux 1.

[37] Sean Borman (2004). The Expectation Maximization algorithm : A short tutorial. www.isi.edu/natural-language/teaching/cs562/.../B06.pdf.

[38] S. Pashardes and C. Christofides (2008). Statistical analysis of wind speed and direction in Cyprus, Sol. Energy 55(4), pp. 405–414.

[39] Surles, J.G. and Padgett, W.J. (2001). Inference for reliability and stress-strength for a scaled Burr type X distribution. Lifetime Data Analysis, 7, 187-200.

[40] Surles, J.G. and Padgett, W.J. (2004). Some Properties of a Scaled Burr X Distribution. Journal of Statistical Planning and Inference, 128, 271–280.

[41] Van der Vaart, A. W. (1998). Asymptotic statistics, vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[42] Voinov, V., Nikulin, M. and Balakrishnan, N. (2013). "Chi-Squared Goodness of Fit Tests with Applications". Academic Press, Elsevier.

[43] Vodă VG (1976b) procedures on a generalized Rayleigh variate, II. Appl Math 21 :413–419

[44] Whitmore, G.A. (1983). A regression method for censored inverse-gaussian data. Canadian Journal of Statistics, 11(4) :305-315, 14.

[45] Y. K. Lin, (1976), Probabilistic Theory of Structural Dynamics, New York, Robert E. Krieger Publishing Company, 368 p.