

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

=====

UNIVERSITE MENTOURI – CONSTANTINE
FACULTE DES SCIENCES EXACTES

=====

DEPARTEMENT DE MATHEMATIQUES

N° d'ordre :

N° de série :

MEMOIRE PRESENTE POUR L'OBTENTION
DU
DIPLOME DE MAGISTERE
EN
MATHEMATIQUES

« Tests d'hypothèses dans un modèle de censure. »

Par
Khaled Dib

OPTION
Probabilités et Statistique

Devant le jury :

Président	M.	Z. Mohdeb	Prof.	Université Mentouri
Rapporteur	M ^{me}	F. Messaci	Prof.	Université Mentouri
Examineur	M.	F. L. Rahmani	M. C.	Université Mentouri
Examineur	M ^{me}	N. Nemouchi	M. C.	Université Mentouri

Soutenu le :

Remerciements

Grâce à Dieu le tout puissant, j'ai pu terminer ce travail.

En tout premier lieu, je tiens à remercier chaleureusement Madame Messaci Fatiha qui m'a guidé et soutenu tout au long de ce travail de recherche aussi bien par ses précieux conseils que par sa rigueur et sa disponibilité à mon égard, qu'elle trouve ici l'expression de toute ma gratitude.

Je remercie vivement, Le professeur Zaher Mohdeb, Professeur à l'université Mentouri de Constantine, d'avoir accepté de présider mon jury.

Tous mes remerciements et respects les plus profonds vont à Monsieur Fouad Lazhar Rahmani, chef de département de Mathématiques de l'Université Mentouri, et à Madame Nemouchi Nahima, Maître de conférence au département de Mathématiques, pour avoir examiné mon travail et accepté d'être membres de mon jury, sans oublier tous mes enseignants.

إختبارات لفرضيات لنماذج بمراقبات.

ملخص :

هذا البحث يهتم بدراسة إختبارات غير وسيطية لفرضيات تهدف إلى مقارنة مجتمعات إحصائية وهذا في حالة وجود معطيات كاملة أو مراقبات من اليمين أو حتى معطيات بمراقبات مزدوجة. قدمنا كذلك أمثلة لشرح الطرق المدروسة لاستكمال الدراسة النظرية إضافة إلى إختبارات على معطيات مولدة عن طريق محاكات.

الكلمات الدلالية : إختبارات غير وسيطية لفرضيات،مراقبة من اليمين،المراقبة بالمجالات،المحاكات.

Tests d'hypothèses dans un modèle de censure.

Résumé: ce travail est consacré à l'étude de tests non paramétriques, de comparaison de populations, basés sur des données complètes ou censurées à droite ou même doublement censurées. Des exemples d'illustration des méthodes étudiées complètent l'étude théorique aussi bien pour des données simulées que pour des données réelles.

Mots-clés : tests non paramétriques, censure à droite, censure par intervalles, simulation.

Hypothesis Tests in a censored model.

Summary : This work is devoted to non-parametric test studies of population comparison based on complete, right censored data or categorically bouble censored ones. Illustration examples of studied methodes complete theoretical study of simulated data as well as real ones .

Keywords : non-parametric tests, right censorship , interval-censored data, simulation.

Table des matières

1	Rappels préliminaires	5
1.1	Tests d'hypothèses	5
1.1.1	Introduction	5
1.1.2	Tests paramétriques et non paramétriques : avantages et inconvénients	7
1.1.3	Quelques tests usuels	8
1.2	Introduction à l'analyse de survie	11
1.2.1	Définitions	11
1.2.2	Estimateur de Kaplan-Meier (produit limite)	15
1.2.3	Estimateur de Nelson-Aalen :	17
2	Tests non paramétriques dans le cas des données censurées à droite	19
2.1	Les tests de rang	19
2.1.1	Test de Gehan(1965)	21
2.1.2	Test de Log-Rang	21
2.1.3	Test de Tarone et Ware(1977)	22
2.1.4	Test de Peto et Peto	22
2.1.5	Tests de Fleming et Harrington (1981)	22
2.1.6	Application sur données réelles	23
2.2	Autres tests de comparaison	25
2.2.1	Le test de Renyi	25
2.2.2	Tests basés sur les tracés P-P et Q-Q	26
3	Tests basés sur des données censurées par intervalles	38
3.1	L'algorithme EM de Turnbull	38
3.2	Généralisation des tests de Wilcoxon et de Log-rang sur des données censurées par intervalles	39

3.2.1	Données censurées par intervalles	40
3.3	Tests non paramétriques avec données doublement censurées par intervalles	41
3.3.1	Tests de Jianguo Sun (2001)	41
3.4	Généralisations des tests de log-rang par Jinheum Kim et autres (2006)	44
3.5	Introduction	45
3.6	Tests de types de log-rang	45
4	Application	48
4.1	Données réelles	48
4.2	Données simulées	51

Introduction

La théorie des tests est l'une des deux branches de la statistique mathématique. Elle se subdivise en deux volets principaux, les tests paramétriques et les tests non-paramétriques. Ces derniers n'imposent aucune forme à la loi de probabilité des phénomènes étudiés contrairement au cas paramétrique qui requiert un modèle à fortes contraintes (comme la normalité des distributions, l'égalité des moyennes, ...).

Parmi les tests les plus usuels en statistique, on peut citer le test de normalité d'une population, les tests d'égalité des paramètres, les tests de corrélation, etc. La littérature statistique abonde de types de tests statistiques. Notre choix s'est porté sur des tests non-paramétriques et plus particulièrement sur les tests de comparaison de populations, vu leur importance dans la pratique (domaine médical, économique, social, ...). Parmi les tests les plus connus dans ce cadre, citons par exemple le test de Wilcoxon, le test de Mann-Whitney et celui de Komogorov-Smirnov.

Classiquement, les tests se font grâce à des statistiques calculées sur la base de données complètes, ce qui veut dire qu'elles sont de véritables réalisations des variables d'intérêt. Mais, dans la pratique de telles données ne sont pas toujours observables. Ainsi, à cause de divers facteurs, comme la fixation du temps de l'étude (très souvent obligatoire!) ou la disparition d'individus sous étude (migration, mort par accident ...), certaines observations ne donnent qu'une information partielle sur la vraie réalisation. Par exemple, cette dernière est inconnue mais on sait qu'elle est supérieure (respectivement inférieure) à l'observation recueillie qui est alors dite censurée à droite (respectivement à gauche); ou alors plus généralement on sait qu'elle appartient à un certain intervalle, auquel cas l'observation est dite censurée par intervalles.

C'est pourquoi, le but de notre travail est l'étude de tests, précédemment cités, en se basant sur des observations censurées.

Ce mémoire est divisé en quatre chapitres. Dans le premier, nous nous limitons à un bref rappel sur des notions et définitions de base de la théorie des tests et sur quelques éléments fondamentaux de l'analyse de survie (qui est la partie de la statistique qui s'intéresse à l'inférence dans le cas d'observations censurées), que nous jugeons utiles pour la suite de notre travail.

Dans le deuxième chapitre nous nous intéressons essentiellement aux tests non-paramétriques visés, dans le cas de données censurées à droite ; nous exposons l'idée des tests de rangs et nous étudions différents travaux dans ce contexte : Gehan (1965), Tarone et Ware (1977), Peto et Peto (1973) et puis la généralisation de Fleming et Harrington (1981). Ensuite nous explorons d'autres types de tests, le test de type 'Renyi' (c'est la généralisation du test de Kolmogorov-Smirnov au cas des données censurées à droite) et nous terminons le chapitre par une étude détaillée des travaux de P. Deheuvels et H. J. Einmahl (1992) sur une classe de tests basés sur les tracés P-P et Q-Q.

Au troisième chapitre, une généralisation des tests de rangs au cas de la censure par intervalles est présentée. Nous parlons aussi du cas des données doublement censurées par intervalles (ce qui veut dire que la variable d'intérêt s'écrit sous forme d'une différence de deux variables aléatoires censurées par intervalles) basé sur l'algorithme E-M (voir [40]), une étude détaillée de l'article de Kim et autres (2006) est présentée à la fin de ce chapitre, dans le cadre de la censure par intervalles.

Le quatrième chapitre est, quant à lui, consacré à l'application de quelques procédures présentées dans les chapitres antérieurs. Nous avons d'abord pratiqué plusieurs tests sur des données réelles censurées à droite, les unes prises de la littérature et les autres recueillies auprès d'éleveurs de bovins de la région de Teleghma, wilaya de Mila. Finalement les mêmes tests ont été appliqués à des données simulées.

Chapitre 1

Rappels préliminaires

1.1 Tests d'hypothèses

1.1.1 Introduction

Généralités

Dans la pratique on est souvent amené à prendre diverses décisions au sujet d'une population et ce à partir de l'information que donne un échantillon. On appelle de telles décisions des décisions statistiques.

On veut par exemple décider si un nouveau médicament est efficace, si une méthode pédagogique est meilleure qu'une autre, si une pièce de monnaie est bien équilibrée, etc...

Une hypothèse statistique, notée H , est une proposition logique contenant les caractéristiques d'une ou plusieurs populations données, (comme la forme éventuelle de la distribution ou l'égalité entre des lois) ou des valeurs pour des paramètres.

Un test statistique de l'hypothèse H_0 (dite hypothèse nulle) contre l'hypothèse H_1 (dite hypothèse alternative) est une démarche qui a pour but de fournir une règle de décision permettant de faire un choix entre les hypothèses H_0 et H_1 sur la base des réalisations de l'échantillon. Évidemment, il ne doit pas exister d'événement réalisant les hypothèses H_0 et H_1 simultanément. Ce qui s'écrit $H_0 \cap H_1 = \Phi$.

La région critique d'un test est l'ensemble des valeurs observées pour lesquelles l'hypothèse nulle H_0 est rejetée. les valeurs limites de cette région constituent les valeurs critiques. La région d'acceptation de H_0 est le com-

plément de la région critique, autrement dit elle est formée par l'ensemble des valeurs observées pour lesquelles l'hypothèse nulle H_0 est acceptée.

Que l'on rejette ou que l'on accepte une hypothèse nulle, donc quelle que soit la décision, on prend le risque de commettre l'une des erreurs suivantes :

1. L'hypothèse H_0 est rejetée à tort.
2. L'hypothèse H_0 est retenue de façon injustifiée.

Nous sommes donc face à deux erreurs possibles.

Le risque d'erreur de 1^{ère} espèce . Il représente la probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie, en d'autres termes, accepter H_1 alors qu'elle est fautive. Elle s'écrit $P(C/H_0)$ où C représente la région critique.

Le risque d'erreur de 2^{ème} espèce Il est noté β et représente la probabilité d'accepter H_0 alors qu'elle est fautive, c'est à dire qu'il s'écrit $P(\bar{C}/H_1)$. L'erreur β est généralement exprimée par son complément à 1, appelé puissance du test, celle ci exprime la capacité du test à éviter une hypothèse erronée ; c'est la probabilité de rejeter l'hypothèse nulle tout en ayant raison. Signalons le fait que, malheureusement, si une des deux erreurs diminue, l'autre augmente. La démarche à adopter dans la pratique est de contrôler le risque de 1^{ère} espèce en lui imposant de ne pas dépasser une valeur α (dite seuil ou niveau de signification) et de chercher alors à minimiser β . Remarquons cependant que le calcul de la puissance d'un test est une opération complexe. La difficulté tient essentiellement au fait que l'hypothèse alternative est vague. Ceci conduit souvent à ignorer l'existence même de β .

Selon que $\alpha = 5\%$ ou 1% ou $0,1\%$, le test est dit significatif ou hautement significatif ou très hautement significatif. Afin d'attirer l'attention sur les interprétations fausses concernant l'erreur α , en rejetant H_0 pour $\alpha = 5\%$, la bonne interprétation est qu'on a seulement 5 chances sur 100 de le faire par le simple fait de l'aléa.

La p-valeur d'un test statistique Après avoir calculé la valeur de la statistique T , nous avons une simple réponse binaire : on accepte H_0 ou bien on rejette H_0 pour une valeur de α fixé au début. Afin de mieux exploiter le résultat de la statistique, on cherche souvent le plus petit niveau de probabilité α pour lequel la valeur observée t de la statistique T conduit au rejet

de H_0 , c'est la p-valeur du test considéré. Autrement dit c'est le risque de première espèce minimal que l'on doit accepter de prendre en décidant de rejeter H_0 sur la base de l'échantillon observé. Une "faible" valeur de la p-valeur indique qu'il y a très peu de chances d'obtenir, pour la statistique de test sous H_0 , un tel ordre de grandeur et conduit donc au rejet de H_0 . L'introduction de la p-valeur a l'avantage de ne pas être obligé de fixer le seuil α au départ et c'est cette valeur qui est donnée dans les programmes de tests .

- Dans le cas d'un test unilatéral à droite (par exemple $\mu = \mu_0$ contre $\mu > \mu_0$ où μ est une moyenne inconnue qu'on veut comparer à la valeur fixée μ_0) la p-valeur d'une valeur t prise par une statistique T est :

$$p_t = 1 - F_{H_0}(t),$$

$F_{H_0}(t)$ étant la fonction de répartition de T supposée connue sous l'hypothèse H_0 .

- Pour un test unilatéral à gauche la p-valeur d'une valeur t prise par une statistique T est :

$$p_t = F_{H_0}(t)$$

- Pour le cas d'un test bilatéral la p-valeur d'une valeur t prise par la statistique de test T est :

$$p_t = \begin{cases} 2F_{H_0}(t) & \text{si } F_{H_0}(t) < 0.5, \\ 2(1 - F_{H_0}(t)) & \text{si } F_{H_0}(t) \geq 0.5. \end{cases}$$

1.1.2 Tests paramétriques et non paramétriques : avantages et inconvénients

Un test paramétrique requiert un modèle à fortes contraintes (normalité des distributions, égalité des variances etc...). Ces hypothèses sont d'autant plus difficiles à vérifier que les effectifs étudiés sont plus réduits.

Un test non paramétrique est un test dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Son emploi est possible même sans aucune connaissance sur les lois régissant les variables d'intérêt. Cependant les échantillons considérés doivent être aléatoires (tous les individus ont la même probabilité de faire partie de l'échantillon). Pour des échantillons de tailles très faibles (jusqu'à $N = 6$), la seule possibilité est l'utilisation d'un test non paramétrique,

sauf si la nature exacte de la distribution de la population est connue. Ceci permet la diminution du coût et du temps nécessaires à la collecte des informations. De plus des tests non paramétriques permettent de traiter des échantillons composés à partir d'observations provenant de différentes populations. Cependant, les tests paramétriques sont plus puissants que les tests non paramétriques, quand les conditions imposées sont remplies.

1.1.3 Quelques tests usuels

La littérature statistique abonde de types de tests statistiques. Notre choix s'est porté sur des tests non-paramétriques des plus connus. En particulier le test de comparaison de populations a été traité en vue de le généraliser au cas de données plus "complexes" dans les chapitres suivants.

Tests de Wilcoxon et Mann-Whitney [Tests de rang]

On se propose de tester l'effet d'un traitement sur un caractère donné (taux du glucose dans le sang par exemple). Un groupe témoin sans traitement correspond à un premier échantillon (X_1, \dots, X_{n_x}) de loi P_X . Sur un deuxième groupe, avec traitement, les valeurs mesurées sont celles de l'échantillon (Y_1, \dots, Y_{n_y}) de loi P_Y . Les deux lois P_X et P_Y sont inconnues. Si le traitement n'a aucun effet (hypothèse nulle), les deux lois sont identiques, d'où

$$H_0 : P_X = P_Y. \quad (1.1)$$

Test de Wilcoxon L'idée du test de Wilcoxon est comme suit : On rassemble les deux échantillons, et on range les valeurs dans l'ordre, alors sous H_0 l'alternance des X_i et des Y_j devrait être presque régulière. Si les Y_j ont tendance à être plus grands que les X_i , ou plus petits, ou très concentrés dans une certaine partie de valeurs dans la statistique d'ordre, on devrait rejeter l'hypothèse H_0 . On commence donc par écrire la statistique d'ordre de l'échantillon global (s'il y a des ex-æquo, on tire au hasard une permutation). On obtient une suite mélangée des X_i et des Y_j . On calcule ensuite la somme des rangs des X_i , notée W_X (c'est la statistique de Wilcoxon). Sous l'hypothèse H_0 , la loi de W_X se calcule facilement : sur un échantillon de taille $n_x + n_y$, il y a $(n_x + n_y)!$ ordres possibles. Le nombre de rangements possibles des X_i est $\binom{n_x + n_y}{n_x}$, et ils sont équiprobables. On a donc pour tout

m entier allant de $n_x(n_x + 1)/2$ à $n_x(n_y + (n_x + 1)/2)$

$$P_{H_0}[W_X = m] = \frac{k_m}{\binom{n_x+n_y}{n_x}} \quad (1.2)$$

où k_m désigne le nombre de n_x -uplets d'entiers r_1, \dots, r_{n_x} dont la somme vaut m et qui sont tels que : $1 \leq r_1 < r_2 < \dots < r_{n_x} \leq n_x + n_y$.

On rejette H_0 si la valeur de W_X est trop grande. Pour un seuil α , il faut trouver la valeur critique w_c telle que $\alpha = P[W_X \geq w_c]$. H_0 est rejetée si $W_X \geq w_c$, où w_c peut être lue sur la table de Wilcoxon pour les petites valeurs de n_x et de n_y ($\max(n_x, n_y) \leq 25$). Pour les grandes valeurs, on dispose du résultat d'approximation normale suivant (cf Vo Khac Khoan (1984).

Théorème 1 *Sous l'hypothèse H_0 , la loi de $\frac{W_X - n_x(n_x + n_y + 1)/2}{\sqrt{n_x n_y (n_x + n_y + 1)/12}}$, converge vers la loi normale $\mathcal{N}(0, 1)$.*

Voici un exemple de deux échantillons de taille 10 :

1.0364	0.5818	1.5273	0.7455	1.2545	0.9636	0.3091	0.5818	0.4545	1.3455
1.4727	1.0000	0.6182	1.4364	0.8364	0.2909	1.5455	1.2909	1.5818	1.0364

La statistiques d'ordre de l'échantillon regroupé de taille 20 s'écrit (les valeurs X_i du premier échantillon sont soulignées).

0.2909	<u>0.3091</u>	<u>0.4545</u>	<u>0.5818</u>	<u>0.5818</u>	0.6182	<u>0.7455</u>	0.8364	<u>0.9636</u>	1.0000
1.0364	<u>1.0364</u>	<u>1.2545</u>	1.2909	<u>1.3455</u>	1.4364	1.4727	<u>1.5273</u>	1.5455	1.5818.

La statistique W_X prend la valeur : $2+3+4+5+7+9+12+13+15+18 = 88$. Les valeurs du premier échantillon ont tendance à être plus petites que celles du second. On cherche à savoir si cette tendance est significative, on réalise donc un test unilatéral à gauche (rejet d'une valeur trop petite de W_x). La p -valeur correspondante est : $p(88) = 0.1088$. Donc on accepte l'hypothèse H_0 au seuil $\alpha = 0.05$ car $\alpha < p(88)$.

Test Mann-Whitney Le test de Mann-Whitney provient d'une autre approche mais il est équivalent au précédent. Dans l'exemple ci-dessus, nous voulions vérifier que les valeurs du premier échantillon étaient plus souvent plus petites que celles du second. On aurait pu pour cela compter le nombre de couples (X_i, Y_j) pour lesquels $X_i > Y_j$ (avec choix aléatoire en cas d'exæquo).

$$U = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} 1_{\{X_i > Y_j\}} \quad (1.3)$$

On vérifie que les deux statistiques U et W_X sont liées par la relation suivante :

$$U = W_x - n_x(n_x + 1)/2 \quad (1.4)$$

Les deux tests sont donc strictement équivalents. Pour l'exemple précédent, la statistique U prend la valeur $10 + 6 + 5 + 4 + 4 + 2 + 1 + 1 + 0 + 0 = 33 = 88 - 45$.

Le test de Kolmogorov-Smirnov soit l'échantillon (x_1, \dots, x_n) d'une loi inconnue P . L'hypothèse nulle est : H_0 la loi P a pour fonction de répartition F_0 où F_0 est la fonction de répartition d'une loi continue connue.

Si l'hypothèse H_0 est vraie, alors la fonction de répartition empirique F de l'échantillon doit être proche de F_0 . La fonction de répartition empirique est la fonction de \mathbb{R} dans $[0, 1]$, qui vaut :

$$\hat{F}(x) = \frac{i}{n} 1_{\{X_{(i)} \leq x < X_{(i+1)}\}} + 1_{\{x \geq X_{(n)}\}}, \quad (1.5)$$

où les $X_{(i)}$ sont les éléments de la statistique d'ordre de l'échantillon (valeurs de l'échantillon rangées par ordre croissant). En d'autres termes, $\hat{F}(x)$ est la fréquence des éléments de l'échantillon qui sont inférieurs ou égaux à x .

On mesure l'adéquation de la fonction de répartition empirique à la fonction F_0 par la distance de Kolmogorov-Smirnov, qui est la distance de la norme uniforme entre fonctions de répartitions. Pour la calculer, il suffit d'évaluer la différence entre \hat{F} et F_0 aux points $X_{(i)}$.

$$D_{KS}(F_0, \hat{F}) = \max_{i=1, \dots, n} \left\{ \left\| F_0(X_{(i)}) - \frac{i}{n} \right\|, \left\| F_0(X_{(i)}) - \frac{i-1}{n} \right\| \right\} \quad (1.6)$$

Sous l'hypothèse H_0 , la loi de la statistique $D_{KS}(F_0, \hat{F})$ est indépendante de loi de F_0 , car les images des X_i par F_0 sont des variables aléatoires de loi

$\mathcal{U}(0, 1)$.

Mais la fonction de répartition de $D_{KS}(F_0, \widehat{F})$ n'a pas d'expression explicite simple et doit être calculée numériquement. Pour des échantillons de taille suffisante, on utilise le résultat asymptotique suivant.

Proposition 1 *Sous l'hypothèse H_0 , on a, pour tout $t \geq 0$:*

$$\lim_{n \rightarrow \infty} P[\sqrt{n}D_{KS}(F_0, \widehat{F}) \leq t] = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k+1} \exp(-2k^2t^2). \quad (1.7)$$

La série converge très rapidement. En pratique, pour $t > 0.56$, la somme des trois premiers termes donne déjà une approximation avec une erreur inférieure à 10^{-4} .

Si l'hypothèse H_0 est fautive, $\sqrt{n}D_{KS}(F_0, \widehat{F})$ diverge vers $+\infty$ avec n . Le test est donc nécessairement unilatéral à droite (rejet des valeurs trop grandes).

Exemple 1 *Si on donne à $D_{KS}(F_0, \widehat{F})$ la valeur 0.05 pour un échantillon de taille $n = 1000$. La statistique $\sqrt{n}D_{KS}(F_0, \widehat{F})$ vaut 1.5811. La p -valeur correspondante est :*

$$p(t) \simeq 2 \sum_{k=1}^3 (-1)^{k+1} \exp(-2k^2t^2) = 0.0135.$$

Ce test étant indépendant de l'expression de la loi F_0 donc il peut s'étendre à la comparaison de deux fonctions de répartition empiriques, et permet alors de tester l'hypothèse que deux échantillons sont issus de la même loi ce qui nous intéresse dans ce mémoire.

1.2 Introduction à l'analyse de survie

1.2.1 Définitions

Lors d'essais cliniques (ou autres), l'analyse de survie, autrement dit la modélisation du temps de survenue d'un événement, apporte un outil principal d'évaluation théorique. L'analyse des données possède deux particularités intrinsèques, d'une part, celle-ci ne concerne que des variables aléatoires positives et d'autre part, la présence des observations censurées est possible comme nous l'expliquons ci-dessous, après l'introduction de notions fondamentales de l'analyse de survie.

Fonction de survie et de risque

Soit T une variable aléatoire positive définie sur un espace probabilisé (Ω, \mathcal{A}, P) et présentant le temps écoulé jusqu'à la survenue de l'événement d'intérêt (comme par exemple le décès, l'apparition de symptômes, la guérison, la panne d'un composant,...). Cette variable aléatoire est appelée durée de vie (ou de survie). Sa fonction de répartition $F(t) = P(T \leq t)$ représente la probabilité que l'événement d'intérêt se produise durant l'intervalle $[0, t]$. Notons que chaque individu est susceptible de subir une et une seule fois l'événement d'intérêt. L'interprétation des durées de survie a conduit à l'introduction des fonctions suivantes, permettant de mieux caractériser sa loi. La fonction de survie S d'une variable aléatoire de durée de vie T est la probabilité que l'événement d'intérêt survienne après l'instant t , elle est donc donnée par

$$S(t) = 1 - F(t) = P(T > t) \quad (1.8)$$

Par définition, S est décroissante, continue à droite et admet des limites à gauche avec :

$$\lim_{t \rightarrow 0} S(t) = 1 \quad \text{et} \quad \lim_{t \rightarrow +\infty} S(t) = 0 \quad (1.9)$$

Si la loi de T admet une densité de probabilité f , par rapport à la mesure de Lebesgue, nous pouvons définir sa fonction de risque instantané par

$$\lambda(t) = \frac{f(t)}{S(t)} \text{ si } S(t) \neq 0 \quad (1.10)$$

La fonction de risque instantané λ s'interprète comme la probabilité de subir l'événement à l'instant t sachant qu'on a pas encore subi l'événement à l'instant t_- (juste avant l'instant t), autrement dit si f est continue

$$\forall t > 0 (S(t) \neq 0, \lambda(t) = \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt / T \geq t)}{dt} \quad (1.11)$$

C'est ainsi que l'allure de cette fonction apporte une information immédiate sur les caractéristiques de l'objet étudié, les parties décroissantes de la courbe représentative de λ correspondent à un phénomène de rodage, les parties croissantes à un phénomène de vieillissement et les parties planes à un phénomène sans vieillissement. La fonction de risque cumulé est défini, pour tout $t \in \mathbb{R}^+$ par :

$$\Lambda(t) = \int_0^t \lambda(u) du \quad (1.12)$$

La donnée de la loi de probabilité de T repose de manière équivalente sur sa densité de probabilité, sa fonction de survie, sa fonction de risque instantané ou sa fonction de risque cumulé (sous réserve de leur existence). En effet, les relations suivantes entre les différentes fonctions définies peuvent être établies. Pour $S(t) > 0$;

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-d \log(S(t))}{dt}. \quad (1.13)$$

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log(S(t)). \quad (1.14)$$

$$S(t) = \exp(-\Lambda(t)). \quad (1.15)$$

Ces expressions se démontrent facilement mais elles exigent l'existence de f . Soulignons le fait que la relation 1.15 se généralise au cas d'une loi quelconque puisque la fonction de survie n'est que l'intégrale produit de la fonction de hasard cumulé, qui est défini dans le cas général par $\Lambda(t) = \int_0^t \frac{dF}{S(t-)}$ où $S(t-)$ est la limite de S à gauche du point t (cf Shorak et Wellner (1986) page 301).

Censure et troncature

Définition de la variable de censure Toute inférence statistique nécessite l'observation d'un échantillon, lequel est, dans les meilleurs cas, constitué par des vraies observations de la variable d'intérêt (données complètes). Or en analyse de survie, on ne se trouve justement pas dans ce cas de figure. Une variable de censure C peut empêcher l'observation de la vraie valeur d'intérêt et ne nous fournit alors qu'une information partielle sur elle, dont la nature nous conduit à différents types de censure.

Définition 1 (Censure à droite)

Il y a censure à droite lorsque nous observons la censure C (et non pas la durée de vie T) et que nous savons que $T > C$. Ce modèle est le plus fréquent en pratique, il est par exemple adapté au cas où l'événement d'intérêt est le temps de survie à une maladie et où la date de fin de l'étude est préalablement fixée ; les patients vivants à la fin de l'étude fournissent des données censurées à droite.

Définition 2 (Censure à gauche)

Il y a censure à gauche lorsque nous observons la censure C (et non pas la

durée de vie T) et que nous savons que $T < C$. Un phénomène symétrique au précédent se produit, le patient a déjà subi l'événement avant l'instant où on commence l'étude. Ce modèle est par exemple adapté au cas où l'on s'intéresse à l'âge auquel un individu commence à accomplir une tâche. Tout ce qu'on sait chez l'individu censuré est que le véritable âge est inférieur à la valeur observée (l'âge au moment de l'étude par exemple).

Définition 3 (*Censure par intervalles*)

Il y a censure par intervalles lorsque la censure à droite et la censure à gauche sont conjuguées. Dans ce cas l'information apportée par l'expérience se traduit par l'appartenance de la durée de vie à un intervalle de temps ($C_1 < T < C_2$). Ce modèle est adapté au cas de suivis périodiques de patients et généralise aussi bien le modèle de censure à droite que celui de censure à gauche.

Définition 4 (*Censure mixte*)

Il y a censure mixte lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Dans le modèle II décrit dans l'article de Patilea et Rolin (2006), au lieu d'observer un échantillon de la variable d'intérêt T , on observe un échantillon du couple (Z, A) avec $Z = \min(\max(T, L), R)$ et

$$A = \begin{cases} 0 & \text{si } L < T < R \\ 1 & \text{si } R < \max(T, L) \\ 2 & \text{si } T \leq L \leq R \end{cases}$$

où L et R sont des variables de censure et A est l'indicateur de censure. Un exemple de ce modèle est donné par un système formé par trois composants, dont deux sont placés en parallèle (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en série avec ce système en parallèle.

Les modèles de censure peuvent être affinés comme suit

La censure non-aléatoire (ou fixe) de type I Etant donné un nombre positif fixé c et un n -échantillon T_1, \dots, T_n , les observations consistent en (X_i, δ_i) , où $X_i = T_i \wedge c$ et $\delta_i = 1_{\{T_i < c\}}$. Ce modèle, souvent utilisé dans les études épidémiologiques, peut correspondre à l'observation de la durée de survie de n patients au cours d'une expérience de durée prédéterminée c .

La censure aléatoire de type I Etant donné un n-échantillon T_1, \dots, T_n , il existe une variable aléatoire n-dimensionnelle C_1, \dots, C_n de \mathbb{R} telle que les observations consistent en (X_i, δ_i) , où $X_i = T_i \wedge C_i$ et $\delta_i = 1_{\{T_i < C_i\}}$. Ce modèle est typiquement utilisé pour les essais thérapeutiques. Nous nous intéressons à une cause de décès ayant lieu au bout d'un temps T et nous désirons connaître la loi de T ; cependant, une autre cause aléatoire (décès dû à une autre cause aléatoire, abandon, ...) peut survenir auparavant et empêcher par conséquent l'observation de T .

La censure aléatoire de type II Etant donné un nombre positif fixé r et un n-échantillon T_1, \dots, T_n , les observations consistent en (X_i, δ_i) où $X_i = T_i \wedge T_{(r)}$ et $\delta_i = 1_{\{T_i < T_{(r)}\}}$ avec $T_{(1)} < T_{(2)} < \dots < T_{(n)}$ est la statistique d'ordre associée à T_1, \dots, T_n . Ce modèle, souvent utilisé dans les études de fiabilité, correspond à l'observation de la durée de fonctionnement de n machines tant que r d'entre elles ne sont pas tombées en panne

Troncature

Nous parlons de troncature à droite (respectivement à gauche) lorsque la variable d'intérêt n'est pas observable quand elle est supérieure (respectivement inférieure) à un seuil C fixé. Dans le cadre de la censure, la variable C est observée alors que dans le cas de la troncature à droite (respectivement à gauche) l'analyse porte uniquement sur la loi de T conditionnellement à l'événement $\{T < C\}$ (respectivement $\{T > C\}$) et une donnée tronquée ne peut faire partie de l'échantillon. Si une maison de retraite n'accepte que des personnes âgées d'au moins soixante ans, aucun individu décédé avant cet âge n'a la possibilité d'y avoir été admis et est de ce fait tronqué à gauche.

1.2.2 Estimateur de Kaplan-Meier (produit limite)

Soient T_1, T_2, \dots, T_n n variables aléatoires indépendantes et identiquement distribuées (i.i.d) et $F(t) = P(T_1 \leq t)$. Il est naturel et usuel d'estimer $F(t)$ par la fonction de répartition empirique donnée par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{\{T_i \leq t\}}. \quad (1.16)$$

d'autant que $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \rightarrow 0$ presque sûrement par le théorème de Glivenko-Cantelli.

Si T_i est censurée à droite par C_i , on observe $X_i = T_i \wedge C_i$ et $\delta_i = 1_{\{T_i < C_i\}}$.

Commençons par faire remarquer que pour $t'' < t' < t$, on a

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

et ainsi de suite, en faisant intervenir des instants de plus en plus antérieurs.

Par ailleurs, soient $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ la statistique d'ordre associée à (X_n) et $\delta_{(i)}$ l'indice de censure associée à $X_{(i)}$.

L'estimateur de Kaplan-Meier de S est encore appelé Produit limite car il s'obtient comme la limite d'un produit. En effet, il est basé sur l'idée intuitive qu'être en vie après l'instant t , c'est être vivant juste avant l'instant t et ne pas décéder à l'instant t . Cette idée se traduit en termes probabilistes, pour t_i désignant la valeur prise par $X_{(i)}$, comme suit

$$S(t_i) = P(T > t_i) = \prod_{j=1}^i P(T > t_j/T > t_{j-1}), \quad (1.17)$$

où $S(t_0) = 1$.

Notons par p_j la probabilité de survie pendant $[t_{j-1}, t_j[$, sachant que l'individu était vivant au début de cet intervalle.

On pose $q_j = 1 - p_j$; c'est la la probabilité de mourir durant l'intervalle $[t_{j-1}, t_j[$, sachant que l'individu était vivant en t_{j-1} . q_j est naturellement estimée par \hat{q}_j où

$$\hat{q}_j = \frac{M(X_{(j)})}{R(X_{(j)})} \quad (1.18)$$

où $M(X_{(j)}) = \sum_{i=1}^n \delta_i 1_{\{X_i = X_{(j)}\}}$ est le nombre de morts observées à l'instant $X_{(j)}$ et

$R(X_{(j)}) = \sum_{i=1}^n 1_{\{X_i \geq X_{(j)}\}}$ est le nombre des individus ni morts ni censurés juste avant $X_{(j)}$, dits à risque (de mourir).

On déduit alors que

$$\widehat{S}_n(t) = \prod_{X_{(j)} \leq t} \left[1 - \frac{M(X_{(j)})}{R(X_{(j)})} \right] \quad (1.19)$$

C'est l'estimateur de Kaplan-Meier, noté EKM; il s'écrit aussi

$$\widehat{S}_n(t) = \prod_{X_{(j)} \leq t} \left[\frac{n-j}{n-j+1} \right]^{\delta_{(j)}} \quad (1.20)$$

Pour les valeurs t supérieures à la plus grande observation t_{max} , cet estimateur n'est pas bien défini. Si t_{max} correspond à un véritable temps de mort, alors l'estimateur s'annule après ce point. Si t_{max} est censuré alors $\widehat{S}(t)$ ne tendra pas vers zéro à l'infini alors que c'est un estimateur d'une fonction de survie. Plusieurs suggestions sont proposées pour traiter cette ambiguïté. Efron(1967) à proposé d'estimer $S(t)$ par 0 pour $t > t_{max}$. Gill(1980) a, quant à lui, proposé d'estimer $S(t)$ par $\widehat{S}_n(t_{max})$ pour $t > t_{max}$. Ces deux estimateurs ont les mêmes propriétés asymptotiques et convergent vers la vraie fonction de survie. Une étude de simulation, menée par Klein (1991), a montré que la version de Gill est préférable.

Signalons enfin que la consistance forte de l'EKM est donnée par exemple dans Shorak et Wellner (1986) et que sa normalité asymptotique est montrée dans Breslow et Crowley (1974). Des lois du logarithme itéré existent aussi pour cet estimateur (voir par exemple Földes et Rejtö (1981)).

1.2.3 Estimateur de Nelson-Aalen :

L' EKM peut être utilisé pour estimer la fonction du hazard cumulé , en posant :

$$\widehat{\Lambda}(t) = -\ln[\widehat{S}_n(t)]$$

Mais la relation 1.14 n'est pas valable pour toute loi, un autre estimateur pour la fonction de hazard cumulé est proposé par Nelson-Aalen comme suit. Supposons que T_1, \dots, T_n est indépendant de C_1, \dots, C_n où C_1, \dots, C_n sont n v.a i.i.d (de censure), de fonction de répartition G avec $S = 1 - F$ et $\bar{G} = 1 - G$.

Posons $H(x) = P(X_i \leq t) = 1 - S(x)\bar{G}(x)$ et $H_1(x) = P(X_i \leq t, \delta_i = 1)$ la sous loi correspondant à une donnée complète, on a $H_1(x) = \int_0^x \bar{G}(t-)dF(t)$. Introduisons les fonctions empiriques suivantes :

$$N_n(t) = n^{-1} \sum_{i=1}^n 1_{\{X_i \leq t, \delta_i = 1\}},$$

et

$$Y_n(t) = n^{-1} \sum_{i=1}^n 1_{\{X_i \geq t\}}.$$

Leurs espérances sont données par :

$$EN_n(t) = P(X_i \leq t, \delta_i = 1) = H_1(x)$$

et

$$EY_n(t) = P(X_i \geq t) = 1 - H(t-)$$

Remarquons que $EN_n(t)$ et $EY_n(t)$ ne dépendent pas de n et que

$$\Lambda(x) = \int_0^x \frac{dEN_n(t)}{EY_n(t)}.$$

Puisque, par le théorème de Glivenko Cantelli, $\sup_t |N_n(t) - H_1(t)| \rightarrow 0$ p.s. et de même $\sup_t |Y_n(t) - (1 - H(t-))| \rightarrow 0$ p.s. cela suggère d'estimer $\Lambda(x)$ par

$$\widehat{\Lambda}_n(x) = \int_0^x \frac{dN_n(t)}{Y_n(t)} = \sum_{X_{(i)} \leq x} \frac{M(X_{(i)})}{R(X_{(i)})}, \quad (1.21)$$

qui est l'estimateur de Nelson Aalen. Il est fortement consistant et jouit de la propriété de normalité asymptotique (cf Shorak et Wellner (1986)). De plus il est facile de voir qu'il est lié à l'EKM par la relation

$$\widehat{S}_n(t) = \prod_{s \leq t} (1 - \Delta \widehat{\Lambda}_n(s)),$$

où $\Delta \widehat{\Lambda}_n(s) = \widehat{\Lambda}_n(s) - \widehat{\Lambda}_n(s-)$ est le saut de $\widehat{\Lambda}_n$ au point s .

Chapitre 2

Tests non paramétriques dans le cas des données censurées à droite

2.1 Les tests de rang

Nous allons traiter dans ce paragraphe le problème de comparaison $k(k \geq 2)$ populations en se basant sur leurs taux de hasard $\lambda_i(1 \leq i \leq k)$, pour cela testons l'hypothèse suivante :

$$H_0 : \lambda_1(t) = \lambda_2(t) = \dots = \lambda_k(t), \text{ pour } t \leq \tau,$$

contre

$$H_1 : \text{Au moins un des } \lambda_j(t) \text{ est différent pour un certain } t \leq \tau,$$

où τ est le temps maximum de l'étude.

Les données utilisées pour mener le test sont censurées à droite (et probablement tronquées à gauche) pour chacune des k populations.

Soient $t_1 < t_2 < \dots < t_D$ les différents temps de morts dans l'échantillon global. On observe d_{ij} événements dans le $j^{\text{ème}}$ échantillon au temps t_i parmi Y_{ij} individus à risque, $j = 1, \dots, k$ et $i = 1, \dots, D$.

Soient $d_i = \sum_{j=1}^k d_{ij}$ et $Y_i = \sum_{j=1}^k Y_{ij}$ respectivement le nombre de données complètes et le nombre d'individus à risque dans l'échantillon global au temps $t_i; i = 1, \dots, D$.

Le test H_0 est basé sur des comparaisons pondérées de l'estimateur du taux de hasard de la $j^{\text{ème}}$ population et utilise l'estimateur de Nelson-Aalan.

Si l'hypothèse nulle est vraie alors le taux de hasard espéré dans la $j^{\text{ème}}$ population est estimé par le taux de hasard dans l'échantillon global : $\frac{d_i}{Y_i}$.

En utilisant les données du $j^{\text{ème}}$ échantillon, l'estimateur du taux de hasard

est $\frac{d_{ij}}{Y_{ij}}$.

Soit, $W_j(t)$ une fonction positive des poids avec la propriété que $W_j(t_i) = 0$ lorsque $Y_{ij} = 0$ ($\frac{0}{0} := 0$).

Le test H_0 est basé sur les statistiques :

$$U_j(\tau) = \sum_{i=1}^D W_j(t_i) \left\{ \frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right\}, \quad j = 1, \dots, k \quad (2.1)$$

Si toutes les valeurs $U_j(\tau)$ sont proches de zéro alors il est peu probable que l'hypothèse nulle soit fautive, tandis que, si une des $U_j(\tau)$ est loin de zéro alors nous pouvons admettre que la j^{me} population a un taux de hasard différent de celui attendu sous H_0 .

Quoique la théorie permet de mener le test donné à la formule (2.1) pour différentes fonctions de poids, dans la pratique les poids les plus fréquemment utilisés sont :

$$W_j(t_i) = Y_{ij} W(t_i),$$

$W(t_i)$ est la contribution de chaque groupe, et avec ce choix il vient :

$$U_j(\tau) = \sum_{i=1}^D W(t_i) \left\{ d_{ij} - Y_{ij} \left(\frac{d_i}{Y_i} \right) \right\}, \quad j = 1, \dots, k \quad (2.2)$$

On note qu'avec cette fonction de poids, la statistique (2.2) est la somme des différences pondérées entre les nombres de morts observés et les nombres de morts espérés sous H_0 dans le $j^{\text{ème}}$ échantillon.

La variance de $U_j(\tau)$ dans (2.2) est donnée par :

$$\hat{\sigma}_{jj} = \sum_{i=1}^D W(t_i)^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad j = 1, \dots, k. \quad (2.3)$$

et la covariance de $(U_j(\tau), U_g(\tau))$ par :

$$\hat{\sigma}_{jg} = - \sum_{i=1}^D W(t_i)^2 \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad g \neq j. \quad (2.4)$$

où $\frac{Y_i - d_i}{Y_i - 1} = 1$ s'il n'y a pas d'exæquo.

Les termes $\frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) d_i$ et $-\frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} d_i$ proviennent de la variance et de la covariance d'une variable aléatoire multinomiale de paramètres $(d_i; p_j = \frac{Y_{ij}}{Y_i})$; $j =$

$1, \dots, k$.

Les composantes du vecteur $(U_1(\tau), \dots, U_k(\tau))$ sont linéairement dépendantes car $\sum_{j=1}^D U_j(\tau) = 0$, En effet :

$$\sum_{i=1}^D W(t_i) d_i \sum_{j=1}^D \{d_{ij} - Y_{ij}(\frac{d_i}{Y_i})\} = \sum_{i=1}^D W(t_i) [d_i - Y_i \frac{d_i}{Y_i}] = 0.$$

Le test statistique est construit en choisissant $k - 1$ parmi les U_j .

L'estimateur de la matrice des variances-covariances de ces statistiques est donné par la matrice carrée d'ordre $k - 1 : \Sigma_{(k-1) \times (k-1)}$ contenant les $\hat{\sigma}_{jg}$. La statistique du test est donnée par la forme quadratique :

$$\chi^2 = (U_1(\tau), \dots, U_{k-1}(\tau)) \Sigma^{-1} (U_1(\tau), \dots, U_{k-1}(\tau))^t. \quad (2.5)$$

Sous l'hypothèse nulle, cette statistique se distribue asymptotiquement selon la loi du χ^2 à $k - 1$ degrés de liberté.

Pour $k = 2$ la statistique du test peut s'écrire comme :

$$N = \frac{\sum_{i=1}^D W(t_i) [d_{i1} - Y_{i1}(\frac{d_i}{Y_i})]}{\sqrt{\sum_{i=1}^D W(t_i)^2 \frac{Y_{i1}}{Y_i} (1 - \frac{Y_{i1}}{Y_i}) (\frac{Y_i - d_i}{Y_i - 1}) d_i}} \quad (2.6)$$

qui suit asymptotiquement une loi normale standard sous H_0 .

Pour une valeur choisie de α nous obtenons les règles de décision selon l'hypothèse alternative H_1 considérée, comme suivant :

1. $H_1 : \lambda_1(t) > \lambda_2(t)$ pour un certain $t \leq \tau$, On rejette H_0 si $N \geq N_\alpha$.
2. $H_1 : \lambda_1(t) \neq \lambda_2(t)$ pour un certain $t \leq \tau$, On rejette H_0 si $|N| \geq N_{\alpha/2}$.

Les choix de la fonction poids permettent de définir différents tests parmi lesquels.

2.1.1 Test de Gehan(1965)

C'est une généralisation du test de Wilcoxon-Mann-Whitney au cas des données censurées à droite, il est donné par le choix de la fonction poids suivante $W(t_i) = Y_i$.

2.1.2 Test de Log-Rang

En prenant $W = 1$ nous obtenons le test de Log-Rang, c'est le test le plus utilisé dans les packages statistiques et se prête à être généralisé à différents cas de censure (voir chapitre suivant)

2.1.3 Test de Tarone et Ware(1977)

C'est une classe de tests proposée par Tarone et Ware (1977), qui ont choisi $W(t_i) = f(Y_i)$ où f est une fonction fixée par exemple ($f(y) = \sqrt{y}$). Cette classe de poids donne un poids relativement grand à la différence entre le nombre observé et le nombre espéré de morts dans l'échantillon j en un point où il y a plus de données.

2.1.4 Test de Peto et Peto

Ici le choix de la fonction de poids est $W(t_i) = \tilde{S}(t_i)$; où :

$$\tilde{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i + 1}\right) \quad (2.7)$$

Une modification du test de Peto et Peto est obtenue en choisissant :

$$W(t_i) = \frac{\tilde{S}(t_i)Y_i}{Y_i + 1}. \quad (2.8)$$

Chacun des poids définis dans (2.7) et (2.8) dépend de l'allure de la survie dans l'expérience; dans le cas où $W(t_i) = Y_i$ on remarque que le poids dépend beaucoup plus du temps d'événement et de la variable de censure alors que le poids de Gehan peut nous induire en erreur si la distribution de censure diffère d'un échantillon à un autre (voir exemple détaillé dans Prentice et Marek (1979) avec une étude de cas).

2.1.5 Tests de Fleming et Harrington (1981)

Fleming et Harrington ont proposé une large classe de tests de rang dont le test de Log-Rang et qui s'approche de celui de Peto et Peto (2.1.4), pour cela soit $\hat{S}(t)$ l'estimateur EKM basé sur l'échantillon global, alors la fonction des poids proposé par Fleming et Harrington est donnée par :

$$W_{p,q}(t_i) = \hat{S}(t_{i-1})^p [1 - \hat{S}(t_{i-1})]^q. \quad (2.9)$$

avec $p \geq 0$, $q \geq 0$.

Dans cette fonction de poids la fonction de survie au temps précédent de mort est utilisé comme un poids pour assurer que ces poids soient connus juste avant le temps où on a fait la comparaison.

- dans le cas où $p = q = 0$ nous avons le test de log-rang.
- Le cas $q = 0$ et $p > 0$ donne plus de poids aux résultats.
- Pour $p = 0$ et $q > 0$; Ces tests donnent plus de poids aux données qui apparaissent tard dans le temps.

En fonction des données, nous pouvons effectuer un choix judicieux de p et q .

2.1.6 Application sur données réelles

Nehman et al (1992) ont comparé le temps d'infection dû à deux emplacements différents du tube utilisé pour la dialyse chez des patients souffrant d'insuffisance rénale, dont voici les résultats :

Premier emplacement

- tempsinfect1 = [1.5 3.5 4.5 4.5 5.5 8.5 8.5 9.5 10.5 11.5 15.5 16.5 18.5 23.5 26.5];
- obsercensur1 = [2.5 2.5 3.5 3.5 3.5 4.5 5.5 6.5 6.5 7.5 7.5 7.5 7.5 8.5 9.5 10.5 11.5 12.5 12.5 13.5 14.5 14.5 21.5 21.5 22.5 22.5 25.5 27.5];

Deuxième emplacement

- tempsinfect2 = [0.5 0.5 0.5 0.5 0.5 0.5 2.5 2.5 3.5 6.5 15.5];
- obsercensur2 = [0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.5 1.5 1.5 1.5 2.5 2.5 2.5 2.5 3.5 3.5 3.5 3.5 3.5 4.5 4.5 4.5 5.5 5.5 5.5 5.5 5.5 5.5 6.5 7.5 7.5 7.5 8.5 8.5 8.5 9.5 9.5 10.5 10.5 10.5 11.5 11.5 12.5 12.5 12.5 12.5 14.5 14.5 16.5 16.5 18.5 19.5 19.5 19.5 20.5 22.5 24.5 25.5 26.5 26.5 28.5];

Ces données sont traitées par le programme MATALAB 7.7.0 (R2008b). La figure suivante montre les courbes des EKM calculés à partir des deux échantillons :

L'application de différents tests par Nehman et autres a conduit aux résultats suivants :

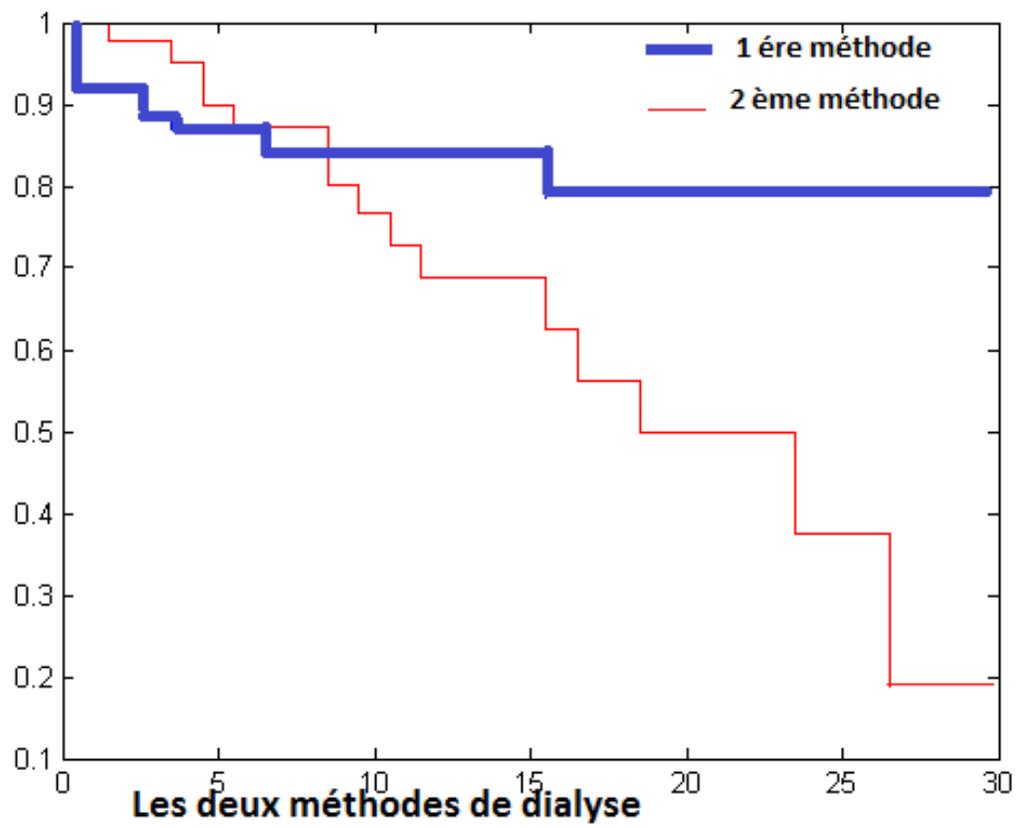


FIGURE 2.1 – Les deux courbes de survie

<i>Test</i>	$W(t_i)$	$Z_1(\tau)$	σ_{11}^2	χ^2	<i>p-value</i>
Log-Rank	1.0	3.96	6.21	2.53	0.112
Gehan	Y_i	-9	38862	0.002	0.964
Tarone-Ware	$Y_i^{1/2}$	13.20	432.83	0.40	0.526
Peto-Peto	$\tilde{S}(t_i)$	2.47	4.36	1.40	0.237
Modified Peto-Peto	$\tilde{S}(t_i) Y_i / (Y_i + 1)$	2.31	4.20	1.28	0.259
Fleming-Harrington $p = 0, q = 1$	$[1 - \hat{S}(t_{i-1})]$	1.41	0.21	9.67	0.002
Fleming-Harrington $p = 1, q = 0$	$\hat{S}(t_{i-1})$	2.55	4.69	1.39	0.239
Fleming-Harrington $p = 1, q = 1$	$\hat{S}(t_{i-1})[1 - \hat{S}(t_{i-1})]$	1.02	0.11	9.83	0.002
Fleming-Harrington $p = 0.5, q = 0.5$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^{0.5}$	2.47	0.66	9.28	0.002
Fleming-Harrington $p = 0.5, q = 2$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^2$	0.32	0.01	8.18	0.004

Les tests de Log-Rang et de Gehan donnent respectivement une p-valeur de 0.112 et 0.964, ce qui permet d'accepter l'hypothèse de l'équivalence des deux méthodes d'emplacement au niveau de signification $\alpha = 0.05$ tandis que les deux courbes dans la figure 2.1 s'éloignent remarquablement pour t assez grand.

En faisant la comparaison par le test de Fleming et Harrington pour $q > 0$, en donnant un grand poids aux derniers instants alors nous obtenons un rejet significatif au même seuil.

2.2 Autres tests de comparaison

2.2.1 Le test de Renyi

Ce test est analogue au test de Kolmogorov-Smirnov dans le cas des données complètes. Il est généralement utilisé pour les problèmes de comparaison dans le cas où les deux taux de hasard se croisent à partir d'un instant de l'étude car les tests étudiés précédemment ont une faible puissance dans ce cas.

Construction du test

- Soient 2 échantillons de tailles respectives n_1 et n_2 et $T_1 < T_2 < \dots < T_D$ les différents temps de vraies morts de l'échantillon global.

- d_{ij} et Y_{ij} sont respectivement le nombre de morts et le nombre d'individus à risque au temps t_i dans l'échantillon j ($j = 1, 2$).
- Soient $Y_i = Y_{i1} + Y_{i2}$ et $d_i = d_{i1} + d_{i2}$ respectivement le nombre total des individus à risque et le nombre total de morts au temps t_i .
- Soit $W(t)$ une fonction poids.

On pose :

$$U(t_i) = \sum_{t_k \leq t_i} W(t_k) \left[d_{k1} - Y_{k1} \left(\frac{d_k}{Y_k} \right) \right]; \quad i = 1, \dots, D. \quad (2.10)$$

La variance de $U(t_i)$ est

$$\hat{\sigma}^2 = \sum_{t_k \leq \tau} W(t_k)^2 \left(\frac{Y_{k1}}{Y_k} \right) \left(\frac{Y_{k2}}{Y_k} \right) \left(\frac{Y_k - d_k}{Y_k - 1} \right) d_k, \quad (2.11)$$

La statistique du test

La statistique de ce test est donnée par :

$$Q = \frac{\sup\{|U(t)|; t \leq \tau\}}{\sigma(\tau)}. \quad (2.12)$$

Sous l'hypothèse nulle H_0 , Q peut être approchée par :

$$\sup\{|\beta(x)|; 0 \leq x \leq 1\}$$

avec $(\beta(t) ; t \in [0, 1])$ le mouvement Brownien. Pour plus de détail voir Klein (1997) ([26]).

On a

$$P(\sup |\beta(t)| > y) = 1 - \frac{4}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k}{2k+1} \exp \left[- \frac{\pi^2(2k+1)}{2y^2} \right], \quad (2.13)$$

dont la preuve se trouve dans Billingsley [3]. Les valeurs données en (2.13) sont tabulées pour tout y .

2.2.2 Tests basés sur les tracés P-P et Q-Q

Dans ce paragraphe nous exposons les travaux de P. Deheuvels et J. H. J. Einmahl (1992) sur des tests de comparaisons de deux fonctions de survie.

Notation Soient $(X_i)_{i \geq 1}$, $(Y_i)_{i \geq 1}$, $(U_i)_{i \geq 1}$ et $(V_i)_{i \geq 1}$ des suites de variables aléatoires *i.i.d* positives ayant pour fonctions de répartition : $F(x) = P(X_i \leq x)$, $G(x) = P(Y_i \leq x)$, $H(x) = P(U_i \leq x)$ et $K(x) = P(V_i \leq x)$ pour tout $i \geq 1$ et $-\infty < x \leq +\infty$.

X_i (resp. Y_i) représente la durée de survie complète du $i^{\text{ème}}$ individu du 1^{er} (resp. 2^{ème}) échantillon.

U_i (resp. V_i) est le temps où le $i^{\text{ème}}$ individu est retiré de notre échantillon. c'est une variable de censure à droite.

Soit $Z'_i = \min(X_i, U_i)$, $Z''_i = \min(Y_i, V_i)$, $\delta'_i = 1_{\{X_i \leq U_i\}}$ et $\delta''_i = 1_{\{Y_i \leq V_i\}}$ pour tout $i \geq 1$.

Pour le modèle de censure à droite, on observe $\{(Z'_i, \delta'_i) | 1 \leq i \leq m\}$ et $\{(Z''_i, \delta''_i) | 1 \leq i \leq n\}$.

Les estimateurs EKM, de F et G , sont respectivement donnés par :

$$F_m(x) = 1 - \prod_{Z'_i \leq x} \left(\frac{1 - \delta'_i}{m - i + 1} \right) \quad -\infty < x \leq +\infty \quad (2.14)$$

et

$$G_n(x) = 1 - \prod_{Z''_i \leq x} \left(\frac{1 - \delta''_i}{n - i + 1} \right) \quad -\infty < x \leq +\infty \quad (2.15)$$

où $0 < Z'_{(1)} \leq \dots \leq Z'_{(m)}$ (resp. $0 < Z''_{(1)} \leq \dots \leq Z''_{(n)}$) sont les statistiques d'ordre de $(Z'_i)_{1 \leq i \leq m}$ (resp. $(Z''_i)_{1 \leq i \leq n}$). Nous supposons que que F (resp. G) est dérivable sur $]0, +\infty[$, de dérivée strictement positive et continue f (resp. g) avec $F(0) = 0$ (resp. $G(0) = 0$). Nous supposons aussi que H et K sont continues sur $] -\infty, +\infty[$ et que $H_-(+\infty) = \lim_{x \rightarrow +\infty} H(x)$ et $K_-(+\infty) = \lim_{x \rightarrow +\infty} K(x)$ sont chacune inférieure à 1. Dans le cas non censuré, U_i et V_i sont infinies avec probabilité un (1), (*i.e.* $H_-(+\infty) = K_-(+\infty) = 0$).

Soient

$$F^{inv}(s) = \inf\{x \geq 0 : F(x) \geq s\}$$

et

$$G^{inv}(s) = \inf\{x \geq 0 : G(x) \geq s\}$$

pour $0 \leq s \leq 1$, les fonctions quantiles respectives de F et G .

Posons $T_H = \sup\{x : H(x) < 1\}$ et $T_K = \sup\{x : K(x) < 1\}$ et introduisons :

$$h(s) = \begin{cases} \int_0^s \frac{1}{(1-u)^2 (1-H(F^{inv}(u)))} du & \text{si } 0 \leq s < F(T_H) \\ s & \text{si } s \leq 0. \end{cases} \quad (2.16)$$

et

$$k(s) = \begin{cases} \int_0^s \frac{1}{(1-u)^2} \frac{1}{1-K(G^{inv}(u))} du & \text{si } 0 \leq s < G(T_K) \\ s & \text{si } s \leq 0. \end{cases} \quad (2.17)$$

Considérons les processus empiriques $\alpha'_m(x) = m^{1/2}(F_m(x) - F(x))$ et $\alpha_n(x)'' = n^{1/2}(G_n(x) - G(x))$ pour $-\infty < x \leq +\infty$ et soient :

$$a'_m(s) = \alpha'_m(F^{inv}(s)) = m^{1/2}(F_m(F^{inv}(s)) - s). \quad \text{pour } 0 \leq s \leq 1. \quad (2.18)$$

et

$$a''_n(s) = \alpha''_n(G^{inv}(s)) = m^{1/2}(G_n(G^{inv}(s)) - s). \quad \text{pour } 0 \leq s \leq 1. \quad (2.19)$$

les processus empiriques réduits correspondants.

Burke, et autres (1988) ainsi que Major et Rejto (1988) ont donné des approximations fortes de a'_m et a''_n rappelées aux formules (2.20) et (2.21) ci dessous. Pour cela soient W' et W'' deux processus standard de Wiener, à deux paramètres, indépendants, tels que pour tout $\theta_F \in]0, F(T_H)[$ et $\theta_G \in]0, G(T_K)[$ on a presque sûrement lorsque $m \rightarrow +\infty$ et $n \rightarrow +\infty$:

$$\|a'_m - m^{-1/2}(1 - I)W'(h, m)\|_0^{\theta_F} =: \|a'_m - \Lambda'_m\|_0^{\theta_F} = O(m^{-1/2} \log^2 m) \quad (2.20)$$

et

$$\|a''_n - n^{-1/2}(1 - I)W''(k, n)\|_0^{\theta_G} =: \|a''_n - \Lambda''_n\|_0^{\theta_G} = O(n^{-1/2} \log^2 n). \quad (2.21)$$

où I est l'application identité et

$$\|\phi\|_c^d = \sup_{c \leq x \leq d} |\phi(x)|.$$

Dans le cas non censuré, on a $h = k = \frac{I}{1-I}$. En effet

$$h(s) = \int_0^s \frac{1}{(1-u)^2} \frac{1}{1-H(F^{inv}(u))} du = \int_0^s \frac{1}{(1-u)^2} du = \frac{s}{1-s} = k(s)$$

car $H(F^{inv}(s)) = 0K(G^{inv}(s)) = 0$.

Alors $m^{-1/2}(1 - I)W'(h, m)$ et $n^{-1/2}(1 - I)W''(k, n)$ dans (2.20) et (2.21) sont des ponts Browniens, dans ce cas (2.20) et (2.21) sont vérifiées pour $\theta_F = \theta_G = 1$ et coïncident avec l'approximation de Kiefer.

Pour tester l'hypothèse d'égalité de F et G on peut penser à utiliser la statistique de Kolmogorov-Smirnov suivante :

$$D_{FG:mn}^\nu = \left(\frac{mn}{m+n}\right)^{1/2} \|F_m - G_n\|_0^\nu \quad (2.22)$$

Malheureusement, non seulement la loi limite de $D_{FG:mn}^\nu$ dépend de F G mais en plus la courbe représentant $F_m - G_n$ ne fournit pas une bonne interprétation visuelle, nous choisissons de travailler avec les tracés quantile-quantile (Q-Q) et probabilité-probabilité (P-P), et définissons les fonctions quantiles empiriques F_m^{inv} et G_n^{inv} associés à F_m et G_n respectivement.

Le produit limite tracé Q-Q de F contre G est défini par :

$$\Delta_{FG:mn}(x) = F_m^{inv}(G_n(x)). \quad -\infty < x \leq +\infty. \quad (2.23)$$

Le produit limite tracé P-P de F contre G est défini par :

$$\tilde{\Delta}_{FG:mn}(s) = F_m(G_n^{inv}(s)). \quad 0 \leq s \leq 1. \quad (2.24)$$

Dans le paragraphe suivant nous présentons des approximations des processus, basés sur les tracés P-P et les tracés Q-Q, par des processus Gaussiens.

Approximations des tracés P-P et des tracés Q-Q considérons le cas $m = n$ et introduisons :

$$\Gamma'_n(s) = F_n(F_n^{inv}(s)) \quad ; \quad \Gamma''_n(s) = G_n(G_n^{inv}(s)), \text{ pour } 0 \leq s \leq 1. \quad (2.25)$$

et

$$\Gamma_n^{inv}(s) = F(F_n^{inv}(s)) \quad ; \quad \Gamma_n^{inv}(s) = G(G_n^{inv}(s)), \text{ pour } 0 \leq s \leq 1. \quad (2.26)$$

Nous définissons aussi le tracé produit limite P-P réduit de F contre G par :

$$\tilde{A}_n(s) = n^{1/2}(\Gamma'_n(\Gamma_n^{inv}(s)) - s). \quad 0 \leq s \leq 1. \quad (2.27)$$

et le processus du tracé Q-Q réduit par :

$$A_n(s) = n^{1/2}(\Gamma_n^{inv}(\Gamma''_n(s)) - s). \quad 0 \leq s \leq 1. \quad (2.28)$$

Dans le cas où $F = G$; $\tilde{A}_n(s)$ s'exprime d'une manière simple en fonction de $\tilde{\Delta}_{FF:nn}(x)$ comme suit :

$$\tilde{A}_n(s) = n^{1/2}(F_n(G_n^{inv}(s)) - s) = n^{1/2}(\tilde{\Delta}_{FF:nn}(s) - s) \text{ pour } 0 \leq s \leq 1$$

Par contre une relation simple n'existe pas entre le P-L Q-Q tracé réduit A_n et le tracé $\Delta_{FF:nn}$, cependant Deheuvels et Einmahl ont montré que si $F = G$, A_n peut-être "convenablement" approché par $\widehat{A}_n(s)$ donné par :

$$\begin{cases} \widehat{A}_n(s) = n^{1/2} f(F^{inv}(s))(F_n^{inv}(G_n(F^{inv}(s))) - F^{inv}(s)) \\ \quad = n^{1/2} f(F^{inv}(s))(\Delta_{FF:nn}(F^{inv}(s)) - F^{inv}(s)) & 0 < s < 1. \\ \widehat{A}_n(0) = 0 \end{cases} \quad (2.29)$$

Le processus suivant est donc une approximation de $\widetilde{A}_n, -A_n$ et $-\widehat{A}_n$.
Soit $\Theta = \min(F(T_H), G(T_K))$ et soit :

$$M_n(s) = n^{1/2}(1-s)(W'(h(s), n) - W''(k(s), n)) =: \Lambda'(s) - \Lambda''(s) \text{ pour } 0 \leq s < \Theta.$$

Le théorème suivant donne les taux des approximations de $\widetilde{A}_n, -A_n$ et $-\widehat{A}_n$ par M_n

Théorème 2 Si $F=G$. Alors pour tout $\theta \in]0, \Theta[$ on a pour $n \rightarrow +\infty$:

$$1. \quad \frac{n^{1/4}(\log n)^{-1/2} \|\widetilde{A}_n - M_n\|_0^\theta}{\left(\left\| \Lambda_n'' \left(\frac{1}{1-H(F^{inv})} + \frac{1}{1-K(F^{inv})} \right) \right\|_0^\theta \right)^{1/2}} \xrightarrow{P} 1. \quad (2.30)$$

$$2. \quad \frac{n^{1/4}(\log n)^{-1/2} \|A_n + M_n\|_0^\theta}{\left(\left\| \left(\Lambda_n' - \Lambda_n'' \right) \left(\frac{1}{1-H(F^{inv})} \right) \right\|_0^\theta \right)^{1/2}} \xrightarrow{P} 1. \quad (2.31)$$

On a aussi, si

$$\lim_{u \rightarrow 0} \left(u \log \left(\frac{1}{u} \right) \right)^{-1/2} \sup_{0 < s \leq F^{inv}(\theta); |s-t| \leq u} |f(t) - f(s)| = 0$$

et si f est continue à droite et positive en 0 alors on peut remplacer A_n par \widehat{A}_n dans (2.31)

Soit W un processus Standard de Wiener, pour $-\infty < s < \Theta$, posons :
 $l(s) = h(s) + k(s)$.

Corollaire 1 Pour $n \rightarrow +\infty$, nous avons

$$n^{1/4}(\log n)^{-1/2} \|\tilde{A}_n - M_n\|_0^\theta \xrightarrow{d} \left(\left\| \frac{(1-I)W(k)((1-H(F^{inv}))^{-1} + (1-K(F^{inv}))^{-1})}{1-H(F^{inv})} \right\|_0^\theta \right)^{1/2}$$

et :

$$n^{1/4}(\log n)^{-1/2} \|A_n - M_n\|_0^\theta \xrightarrow{d} \left(\left\| \frac{(1-I)W(l)}{1-H(F^{inv})} \right\|_0^\theta \right)^{1/2}$$

Remarque 1 (2.30) et (2.31) sont vérifiées en remplaçant Λ_n'' (resp. $\Lambda_n' - \Lambda_n''$) par a_n'' (resp. A_n).

Et maintenant nous démontrons le théorème (2) en utilisant la décomposition, avec $F = G$ on a :

$$\begin{aligned} \tilde{A}_n(s) &= n^{1/2}(\Gamma_n'(\Gamma_n''^{inv}(s)) - s) = \\ n^{1/2}(\Gamma_n'(\Gamma_n''^{inv}(s)) - \Gamma_n''^{inv}(s)) + n^{1/2}(\Gamma_n''^{inv}(s) - s) &=: a_n'(\Gamma_n''^{inv}(s)) + b_n''(s). \quad (\text{car } a_n'(s) = n^{1/2}(\Gamma_n'(s) - s) \text{ par définition}), \end{aligned}$$

$$=: a_n'(\Gamma_n''^{inv}(s)) - a_n''(\Gamma_n''^{inv}(s)) + R_{1,n}(s) \quad \text{avec} \quad R_{1,n}(s) = n^{1/2}(\Gamma_n''(\Gamma_n''^{inv}(s)) - s)$$

En effet :

$$\begin{aligned} -a_n''(\Gamma_n''^{inv}(s)) &= n^{1/2}(-\Gamma_n''(\Gamma_n''^{inv}(s)) + \Gamma_n''^{inv}(s)) \\ &= n^{1/2}(-\Gamma_n''(\Gamma_n''^{inv}(s)) + s + b_n''(s)) \\ \Leftrightarrow b_n''(s) &= -a_n''(\Gamma_n''^{inv}(s)) + n^{1/2}(\Gamma_n''(\Gamma_n''^{inv}(s)) - s) = -a_n''(\Gamma_n''^{inv}(s)) + R_{1,n}. \end{aligned}$$

$$=: \Lambda_n'(\Gamma_n''^{inv}(s)) - \Lambda_n''(\Gamma_n''^{inv}(s)) + \sum_{j=1}^2 R_{j,n}(s) =: M_n(s + n^{-1/2}\Lambda_n''(s)) + \sum_{j=1}^3 R_{j,n}(s)$$

$$=: M_n(s) + \sum_{j=1}^4 R_{j,n}(s).$$

Pour construire les $(R_{j,n}(s))_{j=\overline{1,4}}$ nous avons utilisé (2.19), (2.25) et (2.30), dans [9] on a montré que les $(R_{j,n}(s))_{j=\overline{1,4}}$ sont presque sûrement bornées lorsque $n \rightarrow +\infty$ ce qui prouve la 1^{ère} approximation, ensuite une pareille décomposition permet de conclure la 2^{ème} approximation du théorème 2.

Le test de type Bahadur Kiefer du tracé P-P Soit

$$\delta_n^\theta = \|n^{1/2}F_n(G_n^{inv}) + G_n(F_n^{inv}) - 2I\|_0^\theta \quad \text{pour } 0 < \theta < \Theta. \quad (2.32)$$

Dans le cas où $F = G$; on a d'après (2.25) (2.26) et (2.32) :

$$\delta_n^\theta = n^{1/2}\|\Gamma'_n(\Gamma_n^{inv}) + \Gamma_n''(\Gamma_n^{inv}) - 2I\|_0^\theta \quad \text{pour } 0 < \theta < \Theta. \quad (2.33)$$

on sait que : $\tilde{A}_n(s) = n^{1/2}(\Gamma'_n(\Gamma_n^{inv}(s)) - s)$ pour $0 \leq s \leq 1$ et on définit de la même manière :

$$\tilde{A}_n^*(s) = n^{1/2}(\Gamma_n''(\Gamma_n^{inv}(s)) - s) \quad \text{pour } 0 \leq s \leq 1$$

d'après le théorème (2) on a :

$$\|\tilde{A}_n - M\|_0^\theta = O_P(n^{-1/4}(\log n)^{1/2}) \quad \text{pour } n \rightarrow +\infty.$$

De la même manière on obtient : $\|\tilde{A}_n^* + M_n\|_0^\theta = O_P(n^{-1/4}(\log n)^{1/2})$ pour $n \rightarrow +\infty$.

Puisque : $\delta_n^\theta = \|\tilde{A}_n + \tilde{A}_n^*\|_0^\theta$ on a, sous l'hypothèse H_0 ; ($F = G$)

$$\delta_n^\theta = O_P(n^{-1/4}(\log n)^{1/2}) \xrightarrow{P} 0 \quad \text{si } n \rightarrow +\infty. \quad (2.34)$$

Remarque 2 D'après (2.34) il est clair que $(n^{1/4}(\log n)^{-1/2})\delta_n^\theta = O_P(1)$ pour $F = G$.

Théorème 3 Si $F = G$ on a pour tout $\theta \in]0, \Theta]$ et $n \rightarrow +\infty$

$$\frac{n^{1/4}(\log n)^{-1/2}\delta_n^\theta}{\left(\left\|\left(\Lambda'_n - \Lambda''_n\right)\left(\frac{1}{1-H(F^{inv})} + \frac{1}{1-K(F^{inv})}\right)\right\|_0^\theta\right)^{1/2}} \xrightarrow{P} 1. \quad (2.35)$$

Corollaire 2 Pour $n \rightarrow +\infty$ on a :

$$\frac{\delta_n^\theta}{\left(\log n \left\|\left(F_n(F^{inv}) - G_n(G^{inv})\right)\left(\frac{1}{1-H(F^{inv})} + \frac{1}{1-K(F^{inv})}\right)\right\|_0^\theta\right)^{1/2}} \xrightarrow{P} 1. \quad (2.36)$$

Proposition 2 Deheuvels et Einmahl dans ([9]) ont démontré que pour toute application $\{\rho(t) : 0 \leq t < \Theta\}$ continue et positive sur $[0, \theta]$ on a pour $F = G$ le résultat suivant :

$$n^{3/4}(\log n)^{-1/2}\|(F_n(G_n^{inv}) + G_n(F_n^{inv}) - 2I)\rho\|_0^\theta \xrightarrow{d} \left(\left\|\rho^2(1-I)W(l)\left(\frac{1}{1-H(F^{inv})}\right) + \left(\frac{1}{1-K(F^{inv})}\right)\right\|_0^\theta\right)^{1/2}$$

L'application en statistique Nous revenons au problème de tester l'hypothèse nulle que $F = G$ en ayant $\{(Z'_i, \delta'_i), 1 \leq i \leq m\}$ et $\{(Z''_i, \delta''_i), 1 \leq i \leq n\}$. Nous supposons que $F = G$; H et K sont inconnues. Partant des théorèmes 2 et 3 et de la proposition 2 nous proposons les statistiques :

$$S'_{1,n}(\rho, \theta) = \|(F_n(G_n^{inv}) - I)\rho\|_0^\theta \quad S''_{1,n}(\rho, \theta) = \|(G_n(F_n^{inv}) - I)\rho\|_0^\theta. \quad (2.37)$$

et

$$S_{2,n}(\rho, \theta) = \|(F_n(G_n^{inv}) + G_n(F_n^{inv}) - 2I)\rho\|_0^\theta \quad (2.38)$$

où $\{\rho(t) : 0 \leq t < \Theta\}$ est une fonction convenable.

Le problème maintenant est de trouver les valeurs critiques au seuil $\alpha = 0.05$.

L'hypothèse H_0 ($F = G$) est rejetée si $S'_{1,n}(\rho, \theta) \geq c'_\rho(\alpha, \theta)n^{-1/2}$ ou $S''_{1,n}(\rho, \theta) \geq c''_\rho(\alpha, \theta)n^{-1/2}$ ou $S_{2,n}(\rho, \theta) \geq c_\rho(\alpha, \theta)n^{-3/4}(\log n)^{1/2}$ avec $c'_\rho(\alpha, \theta) = c''_\rho(\alpha, \theta)$ et $c_\rho(\alpha, \theta)$ sont données par :

$$P(\|\rho W(l)(1 - I)\|_0^\theta \geq c'_\rho(\alpha, \theta)) = \alpha. \quad (2.39)$$

et

$$P\left(\left(\left\|\rho^2(1 - I)W(l)\left(\frac{1}{1 - H(F^{inv})} + \frac{1}{1 - K(F^{inv})}\right)\right\|_0^\theta\right)^{1/2} \geq c_\rho(\alpha, \theta)\right) = \alpha. \quad (2.40)$$

Il reste à estimer les valeurs inconnues l, H, K et F . Alors il est naturel d'introduire des estimateurs des termes inconnus dans (2.39) et (2.40). Pour cela posons, pour tout $s \in]-\infty, +\infty[$:

•

$$J'_n(s) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z'_i < s\}}(s).$$

•

$$J''_n(s) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z''_i < s\}}(s).$$

•

$$\tilde{J}'_n(s) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z'_i < s, \delta'_i = 1\}}(s).$$

•

$$\tilde{J}''_n(s) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z''_i < s, \delta''_i = 1\}}(s).$$

Pour :

$$0 \leq s < \mu_n := \min \left(\max\{F_n(Z'_i) : 1 \leq i \leq n, \delta'_i = 1\}, \max\{G_n(Z''_i) : 1 \leq i \leq n, \delta''_i = 1\} \right)$$

Posons :

•

$$h_n(s) = \int_0^{F_n^{inv}(s)} (1 - J'_n(t))^{-2} d\tilde{J}'_n(t)$$

•

$$k_n(s) = \int_0^{G_n^{inv}(s)} (1 - J''_n(t))^{-2} d\tilde{J}''_n(t)$$

Lemme 1 (cf Burke et autres (1981))

h_n et k_n sont des estimateurs uniformément et fortement consistants de h et k respectivement.

Moyennant ces résultats, nous choisissons un $\rho(= \rho_n)$, convenable à (2.39) :

$$\rho_n(s) = \frac{1}{(1-s)(l_n(\theta))^{1/2}} \quad \text{pour} \quad 0 \leq s < \mu_n \quad (l_n = k_n + h_n)$$

Maintenant nous pouvons dire que si $F = G$ on a :

$$\begin{aligned} \lim_{n \rightarrow \infty} P(n^{1/2} S'_{1,n}(\rho, \theta) \geq c) &= \lim_{n \rightarrow \infty} P(n^{1/2} S''_{1,n}(\rho, \theta) \geq c) = \lim_{n \rightarrow \infty} P(n^{1/2} \|\tilde{A}_n \rho\|_0^\theta \geq c) \\ &= \lim_{n \rightarrow \infty} P(n^{1/2} \|n^{1/2} n^{-1/2} (1 - I) W(l) \frac{1}{1 - I} \frac{1}{l_n(\theta)^{1/2}}\|_0^\theta \geq c) \end{aligned}$$

Ceci vient du fait que $\|\tilde{A}_n - M_n\|_0^\theta = O_P(n^{-1/4}(\log n)^{1/2})$ i.e pour n assez grand on peut approximativement considérer M_n au lieu de \tilde{A}_n (2) et donc on obtient que :

$$\begin{aligned} \lim_{n \rightarrow \infty} P(n^{1/2} S'_{1,n}(\rho, \theta) \geq c) &= P(\|W\|_0^1 \geq c) \\ &= 1 - \frac{4}{\pi} \sum_{j=0}^{+\infty} \frac{(-1)^j}{2k+1} \exp\left[-\frac{\pi^2(2j+1)^2}{8c^2}\right] \quad \text{pour } c > 0. \end{aligned}$$

Pour cette dernière égalité on peut se référer à Billingsley (1968) ([3]).

Ceci nous a donné une 1^{ère} solution pour tester l'égalité de deux fonctions de survie.

Pour la statistique $S_{2,n}$; un choix convenable de ρ est donné par :

$$\rho_n^*(s) = \left(\frac{1}{1-s}\right)^{1/2} (l_n(\theta))^{-1/4} \left(\frac{1}{1 - H_n(F_n^{inv}(s))} + \frac{1}{1 - K_n(G_n^{inv}(s))}\right)^{-1/2} \quad (2.41)$$

où H_n et K_n sont les estimateurs de Kaplan Meier de H et K respectivement.

D'autre part on peut remarquer que :

$$\begin{aligned} E(J'_n(s)) &= E\left(\frac{1}{n} \sum_{i=1}^n 1_{\{Z'_i < s\}}(s)\right) \\ &= P(Z'_i < s) = 1 - P(Z'_1 \geq s) = 1 - P(\max(X_1, U_1) \geq s) \\ &= 1 - P(X_1 \geq s)P(U_1 \geq s) = 1 - [1 - P(X_1 \leq s)][1 - P(U_1 \leq s)] \end{aligned}$$

(ceci provient de la continuité de X_1 et U_1)

$$= 1 - [1 - F(s)][1 - (H(s))]$$

Alors ce résultat nous montre que $J'_n(s)$ est un estimateur de $= 1 - [1 - F(s)][1 - (H(s))]$, et ce raisonnement peut être appliqué sur le 2^{ème} échantillon

par symétrie *i.e* $J_n''(s)$ est un estimateur de $= 1 - [1 - G(s)][1 - (K(s))]$ et cela permet d'écrire :

$$\begin{aligned} J_n'(s) = 1 - [1 - F_n(s)][1 - H_n(s)] &\Rightarrow [1 - F_n(s)][1 - H_n(s)] = 1 - J_n'(s) \\ &\Rightarrow 1 - H_n(s) = \frac{1 - J_n'(s)}{1 - F_n(s)}. \end{aligned}$$

et aussi :

$$\begin{aligned} J_n''(s) = 1 - [1 - G_n(s)][1 - K_n(s)] &\Rightarrow [1 - G_n(s)][1 - K_n(s)] = 1 - J_n''(s). \\ &\Rightarrow 1 - K_n(s) = \frac{1 - J_n''(s)}{1 - G_n(s)} \end{aligned}$$

On revient maintenant à la relation(2.41) en y injectant les relations précédentes on trouve que :

$$1 - H_n(F_n^{inv}) = \frac{1 - J_n'(F_n^{inv}(s))}{1 - F_n(F_n^{inv}(s))} = \frac{1 - J_n'(F_n^{inv}(s))}{1 - s}. \quad (2.42)$$

De même on a :

$$1 - K_n(G_n^{inv}) = \frac{1 - J_n''(G_n^{inv}(s))}{1 - s}. \quad (2.43)$$

On peut faire le remplacement de ρ_n^* par ρ_n où :

$$\rho_n(s) = \frac{1}{(1-s)^{1/2}} \frac{1}{(l_n(\theta))^{1/4}} \left(\frac{1}{1 - J_n'(F_n^{inv}(s))} + \frac{1}{1 - J_n''(G_n^{inv}(s))} \right)^{-1/2} \quad (2.44)$$

Alors sous l'hypothèse : $F = G$ nous avons ;

$$\lim_{n \rightarrow +\infty} P(n^{3/4}(\log(n))^{-1/2})S_{2,n}(\rho_n, \theta) \geq c$$

$$= \lim_{n \rightarrow +\infty} P(n^{3/4}(\log(n))^{-1/2} \|(F_n(G_n^{inv}) + G_n(F_n^{inv} - 2I)\rho_n)\|_0^\theta$$

. Deheuvels et Einmahl ont démontré dans [9] que pour tout $\{\rho(t) : 0 \leq t < \Theta\}$ continue et positive sur $[0, \theta]$:

$$n^{3/4}(\log(n))^{-1/2} \|(F_n(G_n^{inv}) + G_n(F_n^{inv} - 2I)\rho)\|_0^\theta \xrightarrow{d}$$

$$\|\rho^2(1-I)W(l)\left(\frac{1}{1-H(F^{inv})} + \frac{1}{1-K(F^{inv})}\right)\|_0^\theta)^{1/2}.$$

Moyennant (2.42) , (2.43) et la proposition nous pouvons écrire pour n assez grand :

$$\begin{aligned} & P\left(n^{3/4}(\log(n))^{-1/2}S_{2,n}(\rho_n, \theta) \geq c\right) \\ = & P\left(\left\|\frac{1}{(1-I)}\frac{1}{(l_n(\theta))^{1/2}}\left(\frac{1-I}{1-J'_n(F_n^{inv})} + \frac{1-I}{1-J''_n(F_n^{inv})}\right)^{-1} \geq c\right\| (1-I)W(l)\left(\frac{1}{1-H(F^{inv})} + \frac{1}{1-K(F^{inv})}\right)\right) \\ & = P\left(\left\|\frac{W(l)}{l_n(\theta)^{1/2}}\right\|_0^\theta \geq c\right) \\ & = P\left(\|W\|_0^\theta \geq c^2\right) \\ & = 1 - P\left(\|W\|_0^\theta \leq c^2\right). \\ & = 1 - \frac{4}{\pi} \sum_{j=0}^{+\infty} \frac{(-1)^j}{2j+1} \exp\left[-\frac{\pi^2(2j+1)^2}{8c^4}\right] \quad \text{pour } c > 0. \end{aligned}$$

La règle de décision de ce test est donnée comme suit :

On choisit une valeur critique $c_{\rho_n}(\alpha, \theta)$ qui rend la quantité a droite égale à α .

- Si $\|(F_n(G_n^{inv}) + G_n(F_n^{inv} - 2I)\rho_n)\|_0^\theta \geq c_{\rho_n}(\alpha, \theta)n^{-3/4}(\log(n))^{1/2}$ on rejette H_0 et donc on a une différence significative entre les deux distributions de survie.
- Si $\|(F_n(G_n^{inv}) + G_n(F_n^{inv} - 2I)\rho_n)\|_0^\theta < c_{\rho_n}(\alpha, \theta)n^{-3/4}(\log(n))^{1/2}$ on accepte H_0 et on a pas assez de raisons pour affirmer qu'il y a différence entre F et G . Les deux échantillons semblent être issus de la même population avec probabilité égale à $1 - \alpha$.

Ce test est basé sur de "beaux" résultats théoriques mais son application sur des données réelles demeure très difficile par rapport aux tests de Log-Rang qui sont simples à appliquer et très souvent intégrés dans les packages statistiques.

Chapitre 3

Tests basés sur des données censurées par intervalles

3.1 L'algorithme EM de Turnbull

Soit X la variable aléatoire d'intérêt, de fonction de répartition F . Nous allons supposer ici que X est discrète et prend les valeurs $0 < x_1 < \dots < x_k$. Sa loi de probabilité est donnée par :

$$p_j = P(X = x_j); \quad j = 1, \dots, k$$

et

$$F(x) = \sum_{x_j \leq x} p_j$$

le problème qui se pose est d'estimer le vecteur $p = (p_1, p_2, \dots, p_k)$. La valeur observée pour tout individu peut être censurée par intervalles, et s'écrit sous forme : $]X_L, X_R]$ avec $X_L < X \leq X_R$. Si X_L et X_R sont des valeurs consécutives dans $\{0, x_1, \dots, x_k\}$ alors on observe une donnée exacte ; le cas particulier $X_R = +\infty$ correspond à une observation censurée à droite.

Proposition 3 Soit $(]X_L^i, X_R^i])_{i=1, \dots, n}$ un échantillon de données censurées par intervalles de loi $p = (p_1, p_2, \dots, p_k)$ alors sa fonction de vraisemblance est :

$$L = \prod_{i=1}^n \left(\sum_{x_j > X_L^i}^{X_R^i} p_j \right)^{d_i} (1 - F(X_L^i))^{1-d_i}. \quad d_i = 1_{X_R^i < +\infty}. \quad (3.1)$$

Proposition 4 *Soit*

$$\alpha_{ij} = \begin{cases} 1 & \text{si } x_j \in]X_L^i, X_R^i] \\ 0 & \text{sinon} \end{cases}$$

alors une équation pour estimer $p_j, j = 1, \dots, k$ est donnée par :

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ij} p_j}{\sum_{l=1}^k \alpha_{il} p_l}, \quad j = 1, \dots, k. \quad (3.2)$$

Pour résoudre ce système, Turnbull dans [44] a proposé l'algorithme itératif suivant :

- Etape 1 : On prend un vecteur de valeur initial $p = (p_1^{(0)}, p_2^{(0)}, \dots, p_k^{(0)})$ (c'est un estimateur initial de $p = (p_1, p_2, \dots, p_k)$).
- Etape 2 : On obtient un estimateur $p_k^{(1)}$ en prenant :

$$\frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ij} p_j^{(0)}}{\sum_{l=1}^k \alpha_{il} p_l^{(0)}}, \quad j = 1, \dots, k.$$

- Etape 3 : On répète l'étape 1 en remplaçant $p_j^{(0)}$ par $p_j^{(1)}$.
- Etape 4 : On s'arrête si :

$$\max_{1 \leq j \leq k} |p_j^{(l)} - p_j^{(l-1)}| < 0.0001.$$

Cette procédure est facile et converge rapidement.

Théorème 4 *L'estimateur $\hat{p} = (\hat{p}_1, \dots, \hat{p}_k)$ issu de l'itération est l'unique estimateur par la méthode de Maximum de vraisemblance et est consistant.*

La preuve du théorème est détaillée dans [44].

3.2 Généralisation des tests de Wilcoxon et de Log-rang sur des données censurées par intervalles

Dans la suite nous allons généraliser les tests de log-rang et de Wilcoxon au cas des données censurées par intervalles.

3.2.1 Données censurées par intervalles

Dans ce paragraphe nous allons construire des tests permettant la comparaison de distribution de survie basées sur des données censurées par intervalles.

la statistique du test :

Supposons que

$$X_i \geq 0; \quad X_i \in E = \{x_1, \dots, x_k\}; i = 1, \dots, n$$

représentant le temps d'échec pour le k^{ème} individu où $0 < x_1 < \dots < x_k$; $k \leq 2$.

les observations sont alors $]X_L^i, X_R^i]$ $i = 1, \dots, n$; il est évident que si $X_L^i = 0$ (resp $X_R^i = \infty$) l'observation est censurée à gauche (resp à droite). On note $]X_L^i, X_R^i]$ $i = 1, \dots, n_1$; $]Y_L^i, Y_R^i]$ $i = 1, \dots, n_2$ les observations sur X et Y respectivement.

Soient $0 = s_1 < s_2 < \dots < s_m = \infty$ telles que $\{s_1, s_2, \dots, s_m\}$ forme la plus petite partie de $U \cup \{0\}$ contenant $X_L^i, X_R^i, Y_L^i, Y_R^i$ et : $S_j = \{x_r \in E : s_{j-1} < x_r \leq s_j\}_{j=1, \dots, m}$

On note $p_j = P(S_j/\theta = 0)$; $j = 1, \dots, m$. Soit $p = (p_1, \dots, p_m)$ alors l'estimateur de p par la méthode du maximum de vraisemblance obtenu par l'algorithme EM de Turnbull est $\hat{p} = (\hat{p}_1, \dots, \hat{p}_m)$. Pour construire un test de l'hypothèse H_0 ; réécrivant $]Y_L^i, Y_R^i] =]s_{r_1(j)}, s_{r_2(j)}]$ où $s_{r_1(j)}, s_{r_2(j)} \in S = \{0, s_1, \dots, s_m\}$ et on associe à l'observation $]Y_L^i, Y_R^i]$ le rang pondéré :

$$rang(]Y_L^i, Y_R^i]) = \sum_{l=1+r_1}^{r_2(j)} \frac{\hat{p}_l}{\hat{p}_{r_1(j)+1} + \hat{p}_{r_1(j)+2} + \dots + \hat{p}_{r_2(j)}} R_l$$

La statistique du test est définie par :

$$W = \sum_{j=1}^{n_2} \left(\sum_{l=1+r_1}^{r_2(j)} \frac{\hat{p}_l}{\hat{p}_{r_1(j)+1} + \hat{p}_{r_1(j)+2} + \dots + \hat{p}_{r_2(j)}} R_l \right). \quad (3.3)$$

On montre, grâce à une étude de simulation, que l'expression de $E(W)$ est donnée sous H_0 par :

$$E_{H_0}(W) \approx n_2 \sum_{k=1}^m p_k R_k,$$

contrairement à l'expression de la variance de W qui n'a pas une expression analytique mais pour chaque vecteur \hat{p} on donne une valeur associée à $Var(W)$, et finalement on a grâce au théorème central limite

$$\frac{W - E(W)}{\sqrt{Var(W)}} \sim N(0, 1) \quad (3.4)$$

3.3 Tests non paramétriques avec données doublement censurées par intervalles

3.3.1 Tests de Jianguo Sun (2001)

Plusieurs tests de comparaison de populations pour des données censurées par intervalles ont été proposés, voir Andersen et Ronn (1995), Sun et Kalbfleisch (1996), Finkelstein (1986), Fay (1999) et Sun (1996). Notre choix, parmi ces travaux, s'est porté sur le test de Sun (2001) parce que, d'une part le calcul de la statistique de test est assez facile et s'apprête, d'autre part, à une généralisation pour des données doublement censurées par intervalles, ce qui veut dire que la *v.a.* d'intérêt s'écrit comme la différence de deux variables censurées par intervalles.

Dans son travail (cf [40]) Sun a introduit des tests dans le cas de données doublement censurées pas intervalles (ce qui veut dire que la variable d'intérêt s'écrit comme la différence de deux variables censurées par intervalles). Ce qui correspond par exemple à la progression d'une maladie qui commence par un début précis puis un événement ultérieur comme la mort par exemple. Ici nous prenons le cas d'une population d'individus qui sont à risque d'être infectés par le virus d'immunodéficience humaine Type-1 (HIV-1), en ayant déjà ce virus les individus sont à risque d'avoir le syndrome du sida.

Pour une telle étude on considère n individus indépendants issus de p populations différentes. X_i et $T_i > 0$ représentent, respectivement, le temps initial (l'infection par le HIV-1) et le temps final (diagnostic du sida), ($i = 1, \dots, n$). La variable aléatoire $S_i = T_i - X_i$ correspond au temps d'intérêt .

Notons (pour simplifier) $\{1; 2; \dots; k + 1\}$ l'ensemble des valeurs possibles des X_i, T_i et S_i ; les observations sont alors $A_i = [a_{i1}, a_{i2}]$ qui contient la valeur de X_i et $I_i = [a_{i3}, a_{i4}]$ qui contient celle de T_i où $a_{ij} \in \{1; 2; \dots; k + 1\}$ et $a_{i1} \leq a_{i2} \leq a_{i3} \leq a_{i4}$. Nous considérons des variables discrètes, ce qui est assez plausible dans le cas des essais cliniques.

Soient $u_0 = 0 < u_1 < \dots < u_m < u_{m+1} = k + 1$ les valeurs ordonnées de l'ensemble $\{0, a_{i3} - a_{i2}, a_{i3} - a_{i2} + 1, \dots, a_{i4} - a_{i1}, k + 1; i = 1, \dots, n\}$ de tous les temps de survie possibles dans l'étude. Si $a_{i1} = a_{i2}, i = 1 \dots, n$ alors l'événement initial est observé exactement et c'est le cas usuel de données censurées par intervalles.

Dans la suite nous allons considérer le test de l'hypothèse H_0 que les p populations ont la même distribution de survie, nous supposons que pour tout i , X_i et S_i sont indépendantes, nous supposons aussi que le mécanisme de censure sur les A_i et les I_i est indépendant de la distribution de survie de X_i et S_i .

Soient d_j le nombre des échecs qui correspondent au temps u_j , n_j le nombre des individus à risque juste avant u_j , d_{jl} et n_{jl} sont les nombre des échecs et des individus à risque, respectivement, issus de la $l^{\text{ème}}$ population; $l = 1, \dots, p; j = 1, \dots, m$.

Alors comme nous avons vu précédemment la statistique de Log-rang est donnée par $T^{(1)} = (T_1^{(1)}, \dots, T_p^{(1)})^t$ où :

$$T_l^{(1)} = \sum_{j=1}^m \left(d_{jl} - n_{jl} \frac{d_j}{n_j} \right), \quad l = 1, \dots, p$$

Cette statistique est calculable pour des données censurées à droite mais ne l'est pas pour des données censurées par intervalles, Sun a donc introduit une statistique calculable dans ce dernier cas, qu'il généralisa ensuite au cas de données doublement censurées par intervalles

Test de Sun pour des données censurées par intervalles Dans ce paragraphe nous supposons que les X_i sont exactes.

Soit $P = (P_0, \dots, P_m)$ la distribution de survie des S_i sous H_0 et $\hat{P} = (\hat{P}_0, \dots, \hat{P}_m)$ l'estimateur du maximum de vraisemblance (MMV), obtenu par la méthode de Turnbull rappelée au début de ce chapitre.

Soient

$$\alpha_{ij} = 1_{\{u_j \in [a_{i3} - a_{i2}, a_{i4} - a_{i1}]\}} \quad j = 1, \dots, m + 1; i = 1, \dots, n$$

Pour construire notre test nous définissons pour tout $j = 1, \dots, m$.

$$d'_j = \sum_{i=1}^n [\alpha_{ij} (\hat{P}_{j-1} - \hat{P}_j)] / \sum_{u=1}^{m+1} \alpha_{iu} (\hat{P}_{u-1} - \hat{P}_u)$$

et

$$n'_j = \sum_{r=j}^{m+1} \sum_{i=1}^n [\alpha_{ir}(\hat{P}_{r-1} - \hat{P}_r) / \sum_{u=1}^{m+1} \alpha_{iu}(\hat{P}_{u-1} - \hat{P}_u)].$$

. De la même façon, pour chaque échantillon $l = 1, \dots, p$, posons

$$d'_{jl} = \sum_i^l [\alpha_{ij}(\hat{P}_{j-1} - \hat{P}_j) / \sum_{u=1}^{m+1} \alpha_{iu}(\hat{P}_{u-1} - \hat{P}_u)]$$

et

$$n'_{jl} = \sum_{r=j}^{m+1} \sum_i^l [\alpha_{ir}(\hat{P}_{r-1} - \hat{P}_r) / \sum_{u=1}^{m+1} \alpha_{iu}(\hat{P}_{u-1} - \hat{P}_u)],$$

où \sum_i^l désigne la somme sur les sujets i dans la $l^{\text{ème}}$ population, $j = 1, \dots, m$

La statistique du test est donnée par

$T^{(1)} = (T_1^{(2)}, \dots, T_p^{(2)})^t$ où :

$$T_l^{(2)} = \sum_{j=1}^m (d'_{jl} - n'_{jl} \frac{d'_j}{n'_j}), \quad l = 1, \dots, p \quad (3.5)$$

qui est la somme des différences entre les pseudo "morts" observées et les pseudo "morts" espérées conditionnellement aux observations. Le test de H_0 est basé sur la statistique : $\chi^{(2)} = T^{(2)t} V^{(2)-1} T^{(2)}$ qui suit approximativement une χ_{p-1}^2 où $V^{(2)}$ est l'estimateur de la matrice de covariance de $T^{(2)}$ et $V^{(2)-1}$ est son inverse généralisée, voir [38] pour le calcul de $V^{(2)}$. Cette méthode s'apprête à la généralisation suivante.

Test pour les données doublement censurée par intervalles Dans ce paragraphe nous donnons la généralisation du test précédent pour des données doublement censurées par intervalles. Nous gardons les même notations pour u_j, P et α_{ij} .

Soient $H(x) = P(X_i \leq x)$; \hat{H} et \hat{P} sont les estimateurs de H et P , obtenus par la MMV (cf sun (1997)) . Nous définissons aussi les quantités suivantes :

$$\alpha_{ij}^* = \begin{cases} \sum_{x+u_j \in]a_{i3}, a_{i4}[} \{\hat{H}(x) - \hat{H}(x-)\} & \text{si } u_j \in [a_{i3} - a_{i2}, a_{i4} - a_{i1}] \\ 1 & \text{sinon} \end{cases}$$

$$d_j'' = \sum_{i=1}^n [\alpha_{ij}\alpha_{ij}^*(\hat{P}_{j-1} - \hat{P}_j)] / \sum_{u=1}^{m+1} \alpha_{iu}\alpha_{ij}^*(\hat{P}_{u-1} - \hat{P}_u)$$

et

$$n_j'' = \sum_{r=j}^{m+1} \sum_{i=1}^n [\alpha_{ir}\alpha_{ir}^*(\hat{P}_{r-1} - \hat{P}_r)] / \sum_{u=1}^{m+1} \alpha_{iu}\alpha_{iu}^*(\hat{P}_{u-1} - \hat{P}_u).$$

De la même façon, pour chaque échantillon $l = 1, \dots, p$ on définit :

$$d_{jl}'' = \sum_i^l [\alpha_{ij}\alpha_{ij}^*(\hat{P}_{j-1} - \hat{P}_j)] / \sum_{u=1}^{m+1} \alpha_{iu}\alpha_{ij}^*(\hat{P}_{u-1} - \hat{P}_u).$$

et

$$n_{jl}'' = \sum_{r=j}^{m+1} \sum_i^l [\alpha_{ir}\alpha_{ir}^*(\hat{P}_{r-1} - \hat{P}_r)] / \sum_{u=1}^{m+1} \alpha_{iu}\alpha_{iu}^*(\hat{P}_{u-1} - \hat{P}_u).$$

où \sum_i^l désigne la somme sur les sujets i dans la $l^{\text{ème}}$ population, et d_j'', n_j'', d_{jl}'' et n_{jl}'' ont un sens analogue à d_j, n_j, d_{jl} et n_{jl} respectivement et généralisent d_j', n_j', d_{jl}' et n_{jl}' dans le cas où l'observation de l'événement initial est exacte, c'est à dire quand $a_{i1} = a_{i2}$ pour tout i .

Enfin nous obtenons la statistique $T^{(3)} = (T_1^{(3)}, \dots, T_p^{(3)})^t$ où :

$$T_l^{(3)} = \sum_{j=1}^m (d_{jl}'' - n_{jl}'' \frac{d_j''}{n_j''}), \quad l = 1, \dots, p \quad (3.6)$$

qui conserve la même forme de la statistique classique de Log-rang.

Pour le calcul de l'estimateur de la matrice de covariance de $T^{(3)}$ Sun à proposé un algorithme détaillé dans [40].

3.4 Généralisations des tests de log-rang par Jinheum Kim et autres (2006)

Nous Proposons d'étudier le test de type Log-Rang, introduit dans l'article de Jinheum Kim et autres (2006) (cf [24]). Il a pour but la comparaison de p populations sur la base des observations censurées par intervalles.

Le test proposé n'exige pas l'utilisation de l'algorithme EM, utilisé pour

l'estimation des survies proposée par Turnbull. il nécessite des poids qui dépendent uniquement de la taille de l'ensemble des individus à risque en tout point d'observation au lieu de dépendre des survies estimées ; cette technique réduit le temps de calcul et les résultats semble plus robustes à travers une étude de simulation proposée pour comparer entre ces tests.

3.5 Introduction

Dans une étude longitudinale avec des contrôles périodiques ; un sujet est examiné, par exemple, chaque semaine, pour voir s'il est arrivé à un événement défini par un changement d'état clinique précis.

Si un individu s'absente pendant quelques visites et revient après quelques semaines dans un état différent, alors le temps de l'événement considéré est censuré par intervalle ; on sait seulement que l'événement s'est déroulé pendant l'intervalle délimité par les deux visites cliniques concernées.

3.6 Tests de types de log-rang

Soient $T_{il} > 0, i = 1, \dots, n_l; l = 1, \dots, p; (p > 1)$; le véritable temps d'intérêt du $i^{\text{ème}}$ sujet dans la $l^{\text{ème}}$ population.

On suppose que les T_{il} sont censurées par intervalles ; ce qui veut dire que le $i^{\text{ème}}$ sujet dans la $l^{\text{ème}}$ population, on n'observe pas T_{il} mais on observe uniquement que $T_{il} \in A_{il}$ où $A_{il} =]L_{il}, R_{il}]$ est l'intervalle durant lequel l'événement d'intérêt s'est produit.

On note que pour les observations censurées à droite (resp. à gauche) $R_{il} = \infty$ (resp $L_{il} = 0$) et que $L_{il} = R_{il}$ correspond au cas d'une observation complète. Soit δ_{il} l'indicateur de censure défini par :

$$\delta_{il} = \begin{cases} 0 & \text{si l'observation du } i^{\text{ème}} \text{ sujet de la } l^{\text{ème}} \text{ population est censurée à droite.} \\ 1 & \text{sinon} \end{cases} \quad (3.7)$$

On suppose que le mécanisme de la censure est indépendant de la variable d'intérêt.

Les différents points L_{il} et R_{il} obtenus sont ordonnés et étiquetés :

$$0 = s_0 < s_1 < \dots < s_m = +\infty.$$

Pour les observations exactes ou censurées à droite, on redéfinit L_{il} comme

étant : $\max_{s_j \leq L_{il}}(s_j)$ et R_{il} comme étant L_{il} .

Pour

$$i = 1, \dots, n_l; l = 1, \dots, p \text{ et } j = 1, \dots, m$$

soient :

$$\alpha_{ijl} = \begin{cases} 1 & \text{si } L_{il} < s_j \leq R_{il} \\ 0 & \text{sinon} \end{cases} \quad (3.8)$$

L'hypothèse à tester est :

H_0 : les p populations se distribuent selon la même loi.

soit $I_j =]s_{j-1}, s_j]$; $j = 1, \dots, m$, alors chaque A_{il} peut s'écrire sous forme d'une réunion des intervalles disjoints suivants :

$$A_{il} = \bigcup_{k=f_{il}}^{f_{il}+m_{il}-1} I_k, \quad i = 1, \dots, n_l$$

où $f_{il} \in \{1; \dots; m\}$ et $m_{il} = \sum_{j=1}^m \alpha_{ijl}$ désigne le nombre de s_j appartenant à A_{il} .

Soit \mathfrak{R}_j le "pseudo ensemble à risque" , contenant les individus ayant une probabilité non-nulle d'être à risque dans l'intervalle I_j .

Soit \mathfrak{D}_j le "pseudo ensemble de mort", contenant les individus ayant une probabilité non nulle de mort (ou plus généralement de subir l'événement d'intérêt) pendant I_j .

On suppose que sous H_0 le vrai temps de mort du sujet i , sachant A_{il} , est uniformément distribué sur $\{s_k; k = f_{il}, \dots, f_{il} + m_{il} - 1\}$, c'est à dire :

$$P(T_{il} = s_k / T_{il} \in A_{il}) = \frac{1}{m_{il}}, \quad k = f_{il}, \dots, f_{il} + m_{il} - 1. \quad (3.9)$$

Sous cette l'hypothèse, on définit la probabilité conditionnelle que le $i^{\text{ème}}$ sujet soit à risque dans I_j et la probabilité conditionnelle que le $i^{\text{ème}}$ sujet subisse l'événement d'intérêt dans l'intervalle I_j respectivement comme :

$$w_{ijl}^{(r)} = P(T_{il} \geq s_j / A_{il}, i = 1, \dots, n_l) = \frac{\sum_{k=j}^m \alpha_{ikl}}{\sum_{h=1}^m \alpha_{ihl}}, \quad (3.10)$$

$$w_{ijl}^{(d)} = P(T_{il} \in I_j / A_{il}, i = 1, \dots, n_l) = \frac{\delta_{il} \alpha_{ijl}}{\sum_{h=1}^m \alpha_{ihl}}. \quad (3.11)$$

$w_{ijl}^{(r)}$ représente la quantité de contribution du $i^{\text{ème}}$ sujet dans la $l^{\text{ème}}$ population dans l'ensemble \mathfrak{R}_j et $w_{ijl}^{(d)}$ sa contribution dans \mathfrak{D}_j .

Pour $j = 1, \dots, m$ soit : $n_j = \sum_{l=1}^p \sum_{i=1}^{n_l} w_{ijl}^{(r)}$ et $d_j = \sum_{l=1}^p \sum_{i=1}^{n_l} w_{ijl}^{(d)}$; alors n_j et d_j représentent le pseudo-nombre des individus dans \mathfrak{R}_j et \mathfrak{D}_j respectivement.

De la même façon on définit le pseudo-nombre à risque et le pseudo-nombre de mort dans la $l^{\text{ème}}$ population, pendant l'intervalle I_j respectivement comme :

$$n_{jl} = \sum_{i=1}^{n_l} w_{ijl}^{(r)}, \text{ et } d_{jl} = \sum_{i=1}^{n_l} w_{ijl}^{(d)}.$$

Suivant les arguments de Sun (1996, 2001) et Zhaw et Sun (2004) ; on construit la statistique du test basée sur la loi conditionnelle de d_{j1}, \dots, d_{jp} sachant les morts et les censures jusqu'au temps s_j pour tout $j = 1, \dots, m$.

Pour cela soit $U = (U_1, \dots, U_{p-1})^T$, où $U_l = \sum_{j=1}^m \left[d_{jl} - n_{jl} \frac{(d_j)}{(n_j)} \right]$, $l = 1, \dots, p-1$. la statistique de log-rang U_l peut être interprétée comme la différence entre le pseudo-nombre de morts dans la $l^{\text{ème}}$ population et le nombres des mort "espérés" (ou attendus).

On note que U_l se réduit à la statistique de log-rang pour les données censurées à droite.

Pour tester H_0 ; on propose la statistique suivante :

$$P = U^T \hat{\Sigma}^{-1} U, \tag{3.12}$$

où $\hat{\Sigma}$ représente l'estimateur de la covariance de U ; on peut trouver dans Jinheum Kim et autres (2006) deux méthodes détaillées pour le calcul de $\hat{\Sigma}$. Ce test peut être mené grâce au résultat suivant :

Proposition 5 *Sous H_0 ; P suit une loi du χ^2 à $(p-1)$ degrés de liberté.*

Chapitre 4

Application

Nous avons fait dans ce chapitre des applications sur des tests déjà étudié dans ce mémoire, dans le cas des données censurées à droite.

4.1 Données réelles

Exemple 2 Nous reprenons, en premier lieu, l'exemple cité au chapitre 2 sur le problème d'équivalence de deux méthodes de dialyse. Nous avons écrit un programme, en langage MATLAB, qui permet de calculer les p – valeurs par application de trois tests : Test de Lograng, Test de Gehan et celui de Tarone et Ware et voilà le tableau des résultats :

Test	χ^2 observée	p – valeur
Lograng	2.53	0.112
Gehan	0.002	0.964
Tarone-Ware	0.4	0.526

Les résultats obtenus, par notre programme, sont identiques aux résultats publiés de cette étude et permettent de conclure, par les trois tests, à l'acceptation de l'équivalence des deux méthodes au seuil usuel de 0.05.

Exemple 3 Nous avons appliqué quelques tests, dans le cas de données censurées à droite, à un problème de comparaison de l'efficacité de l'insémination artificielle chez deux races de bovins, Pie Noire et Pie rouge, afin de détecter une éventuelle réaction différente entre les deux races. A cette fin, nous avons utilisé deux échantillons de vaches dans le territoire de la Daira de

Telerghma, wilaya de Mila. La variable d'intérêt est le nombre de fois que l'on pratique l'insémination artificielle pour que la vache soit en gestation. Les deux vecteurs 'echantillon1' et 'echantillon2' donnent les valeurs de notre variable d'intérêt chacun pour une race donnée. Par ailleurs, les éléments des vecteurs respectifs 'delta1' et 'delta2' prennent la valeur 0 si l'observation associée est censurée à droite, et la valeur 1 dans le cas d'une observation complète. Ici la censure correspond soit à la vente, soit à la mort d'une vache sur laquelle les tentatives d'insémination ont échoué. Signalons le fait que nous avons choisi des échantillons comparables à tout point de vu et en particulier à l'aptitude à la fertilité. Voici le tableau des données recueillies.

echantillon1=[1 2 2 2 1 1 1 1 3 2];

echantillon2=[5 2 1 2 1 1 1 1 4 2];

delta1=[0 0 0 0 1 1 1 1 1 1];

delta2=[1 1 0 0 1 1 1 1 0 1];

Les résultats des tests sont résumés dans le tableau suivant.

<i>Test</i>	χ^2 observée	<i>p – valeur</i>
<i>Lograng</i>	<i>1.3582</i>	<i>0.2438</i>
<i>Gehan</i>	<i>0.0699</i>	<i>0.7915</i>
<i>Tarone-Ware</i>	<i>0.0186</i>	<i>0.8915</i>

Les trois tests donnent des p – valeurs qui permettent d'accepter l'hypothèse de similitude des deux races, quant à l'effet de l'insémination pratiquée, au seuil 0.05. Les tracés des estimateurs des deux fonctions de survie sont donnés ci dessous.

Le graphique ci dessous montre bien que les deux estimateurs EKM se confondent dans une grande partie du plan et se trouvent dans les deux intervalles de confiance. Ceci concorde bien avec le fait que le test conclut à l'acceptation de l'hypothèse nulle.

L'estimateur de Kaplan-Meier de la fonction de survie

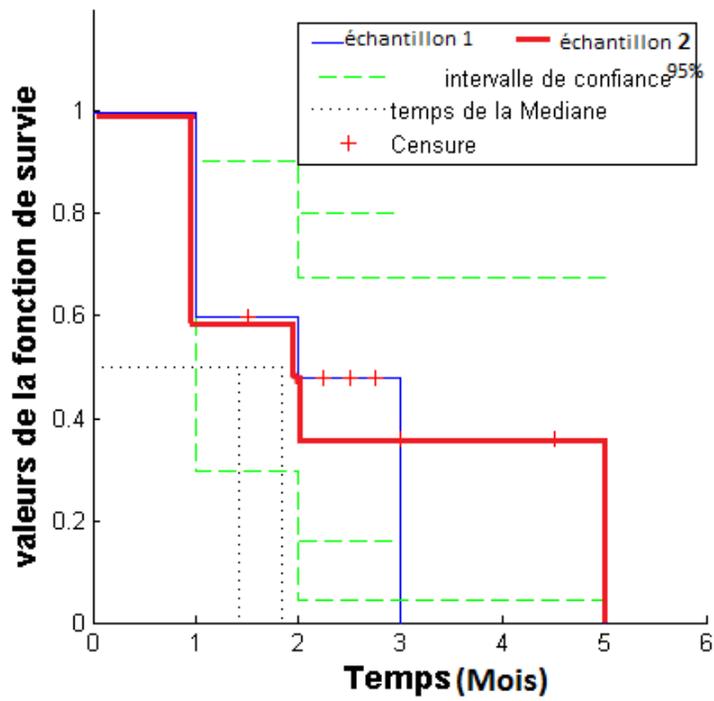


FIGURE 4.1 – les deux courbes de survie

4.2 Données simulées

Nous avons simulé deux échantillons extraits de la loi de Weibull de paramètres (λ, β) de fonction de survie $S(x) = \exp(-\lambda x^\beta)$.

Dans un premier temps, nous considérons le cas d'égalité des paramètres pour les deux échantillons. Puis nous faisons varier β en gardant le même λ . Finalement, nous faisons varier aussi bien λ que β d'un échantillon à l'autre. tous cela a été fait pour une taille de $n = 50$ et a été répété trois fois.

Les résultats obtenus, à l'issue du test sont rapportés dans les tableaux suivants.

Les échantillons de tailles $n = 50$

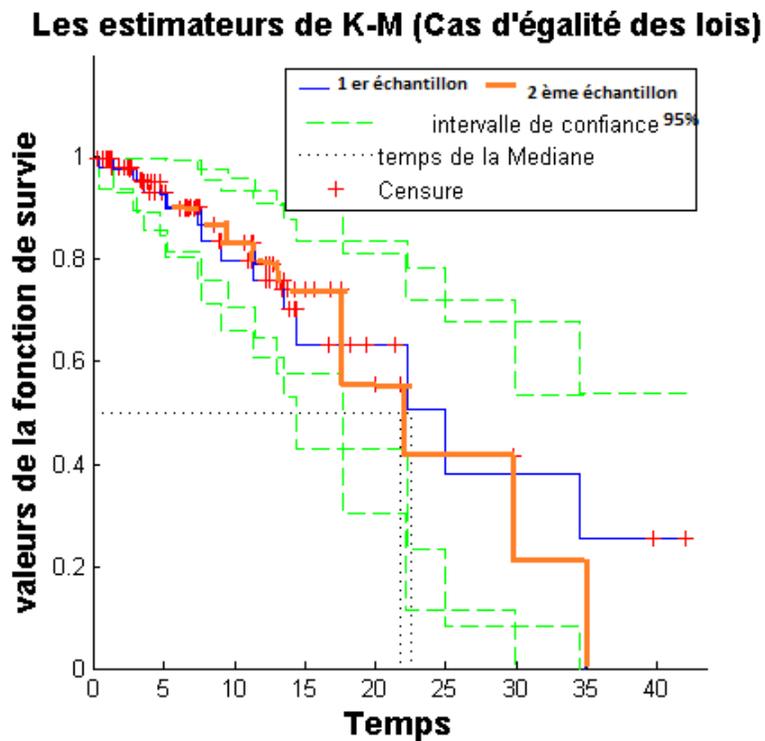


FIGURE 4.2 – Les deux courbes de survie

Même loi

Loi	Test de Lograng	Test de Gehan	Test de Tarone
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,1)	p-valeur=0.6667	p-valeur=0.8907	p-valeur=0.9735
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,1)	p-valeur=0.6826	p-valeur=0.5766	p-valeur=0.6030
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,1)	p-valeur=0.9338	p-valeur=0.7468	p-valeur=0.5379

Les trois tests donnent des $p - valeurs$ qui permettent d'accepter l'hypothèse H_0 au seuil 0.05 (heureusement !).

Même λ avec β différent

Loi	Test de Lograng	Test de Gehan	Test de Tarone
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,2)	p-valeur=0.0051	p-valeur=0.0329	p-valeur=0.0285
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,2)	p-valeur=0.0028	p-valeur=0.0273	p-valeur=0.0103
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (1/12,2)	p-valeur=0.0012	p-valeur=0.0009	p-valeur=0.0022

Nous pouvons remarquer que les trois tests donnent des $p - valeurs$ qui permettent de rejeter l'hypothèse H_0 au seuil 0.05.

λ et β différents

Loi	Test de Lograng	Test de Gehan	Test de Tarone
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (3,4)	p-valeur= $1.6 \cdot 10^{-8}$	p-valeur= $3 \cdot 10^{-7}$	p-valeur= $6.9 \cdot 10^{-8}$
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (3,4)	p-valeur= $5 \cdot 10^{-7}$	p-valeur= $7.1 \cdot 10^{-8}$	p-valeur= $2.3 \cdot 10^{-8}$
Echantillon 1 Weibull (1/12,1) Echantillon 2 Weibull (3,4)	p-valeur=0	p-valeur= $2 \cdot 10^{-9}$	p-valeur=0

les $p - valeurs$ obtenues permettent de rejeter l'hypothèse H_0 même au seuil 0.01 au profit de l'hypothèse de différence entre les lois, ce qui est attendu à cause de la différence significative des lois des échantillons comparés.

L'estimateur de Kaplan-Meier de la fonction de survie

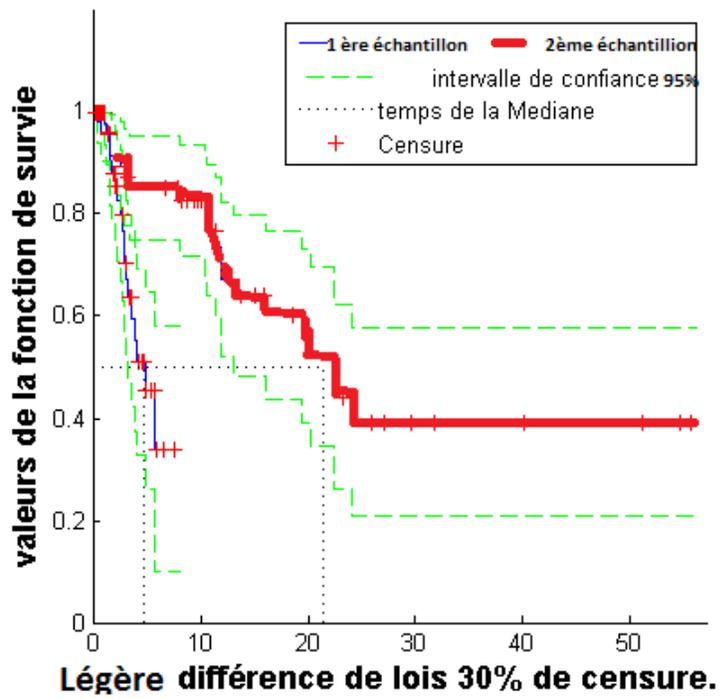


FIGURE 4.3 – Les deux courbes de survie

L'estimateur de Kaplan-Meier de la fonction de survie

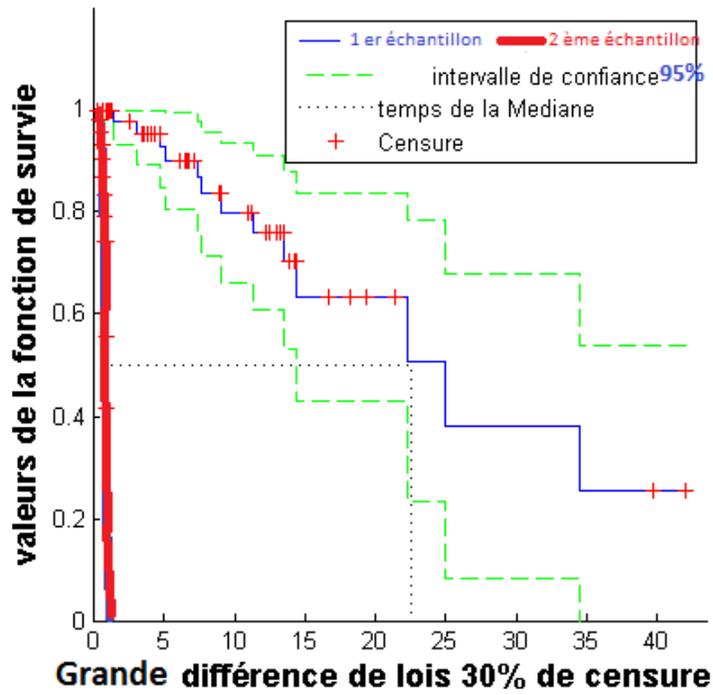


FIGURE 4.4 – Les deux courbes de survie

Bibliographie

- [1] Aalen O. O., *Nonparametric Inference for a Family of Counting Processes*. Annals of Statistics 6 : 701–726. (1978).
- [2] Andersen P. K. et Ronn B. B., *A nonparametric test for comparing two samples where all observations are either left- or right-censored* Biometrics 51 : 323–329. (1995).
- [3] Billingsley P., *Convergence of probability measures*. Wiley New York (1968).
- [4] Burke, M. D., Csörgö, S. et Horvath, L. *A correction to and improvement of Strong approximations of some biometric estimates under random censorship*. Probab. Theory Relat. Fields 19 : 51-57. (1988).
- [5] Breslow N. et Crowley J. *A large sample study of the life table and product limit estimates under random censorship*. The Annals of Statistics, 2(3) : 437–453, (1974).
- [6] Dacunha-Castelle D. et Duffo M., *Probabilités et statistique. Problèmes à temps fixe*. Masson, Paris (1982).
- [7] De Gruttola V. et Lagakos S. W., *Analysis of doubly-censored survival data, with application to AIDS*. Biometrics 45 : 1 – 12. (1989).
- [8] Deheuvels P. et Mason D. M., *A Bahadur-Kiefer-type two-sample statistic with applications to tests of goodness of fit*. In Colloq. Math. Sot. Jcinos Bolyai 57 : 157–172. North-Holland, Amsterdam. (1990).
- [9] Deheuvels P. et Einmahl H. J., *Approximations and Two-Sample Tests Based on P-P and Q-Q Plots of the Kaplan-Meier Estimators of Lifetime Distributions*. Journal of multivariate analysis 43, 200–217. (1992).
- [10] Efron B., *The Two Sample Problem with Censored Data*. In Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability. New York : Prentice-Hall, : 4, 831–853. (1967).

- [11] Efron B., *Generalized Wilcoxon Statistics for the two sample problem with censored data*. Technical report No. 25, Public health service grant 5 R01 GM21215-02. Division of biostatistics, Stanford University, Stanford, California, (1976).
- [12] Fay M. P., *Comparing several score tests for interval-censored data* *Statistics in Medicine* 18 : 273–285. (1999).
- [13] Finkelstein D. M. et Wolfe, R. A., *A semiparametric model for regression analysis of interval-censored failure time data*. *Biometrics* 41 : 933–945. (1985).
- [14] Fleming T. R., O’Fallon J. R. and O’Brien C. P., *Modified Kolmogorov-Smirnov Test Procedures with application to arbitrarily right censored data*. *Biometrics* 36 : 607–626. (1980)
- [15] Fleming T. R. et Harrington D. P., *Class of Hypothesis Tests for One and Two Samples of Censored Survival Data*. *Communications In Statistics* 10 763–794. (1981)
- [16] Fleming T. R. et Harrington D. P., *A Class of Rank Test Procedures for Censored Survival Data*. *Biometrika*, Vol.69, No.3 , pp.553–566. (1982).
- [17] Fleming T. R., Harrington D. P. et and O’Sullivan M. *Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics*. *Journal of the American Statistical Association* 82 : 312–320. (1987).
- [18] Fleming T. R. et Harrington D. P., *Counting Processes and Survival Analysis*. Wiley, New York, (1991).
- [19] Földes A. et Rejtó L. : *A LIL type result for the product limit estimator*. *Probability Theory and Related Fields* 56 (1) : 75–86, (1981).
- [20] Gehan E.A., *A Generalized Two-Sample Wilcoxon Test for Doubly-Censored Data*. *Biometrika* : 52 650–653. (1965).
- [21] Hájek J. et ŠíDák Z., *Theory of rank tests*. New York : Academic Press.
- [22] Kalbfleisch J.D. et Prentice R.L., *The Statistical Analysis of Failure Time Data*. Wiley New York, (2002).
- [23] Kaplan E.L. et Meier P., *Nonparametric estimation from incomplete observations*. *Jasa*, 457–481. (1958).
- [24] Kim J., Kang D. R. et Nam C. M., *Logranktype tests for comparing survival curves with interval-censored data*. *Computational Statistics & Data Analysis* 50 3165–3178. (2006).

- [25] Klein J. P., *Small-Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators*. Scandinavian Journal of Statistics 18 : 333–340. (1991).
- [26] Klein J.P., *Survival Analysis, Techniques for Censored and Truncated Data*. Springer, New York (1997).
- [27] Lagakos S. W., Barraj, L. M. et De gruttola V., *Nonparametric Analysis of Truncated Survival Data, with Application to AIDS*. Biometrika 75 : 515–523. (1988).
- [28] Lee E.T., *Statistical Methods for survival Data Analysis* .John Wiley & sons.Inc. (1992).
- [29] Lehmann E.L., *Elements of Large-Sample Theory*. Springer, New York (1999).
- [30] Major, P. et Rejtö L., *Strong embedding of the estimator of the distribution function under random censorship*. Ann. Statist. 16 : 1113–1132. (1988)
- [31] Mantel N., *Ranking procedures for arbitrarily restricted observations*. Biometrics 23 : 65–78. (1967).
- [32] Nehman N. S., Middendorf D. F., Bay W. H., McElligott R., Powell S. et Anderson J., *Modification of the Percutaneous Approach to Peritoneal Dialysis Catheter Placement Under Peritoneoscopic Visualization : Clinical Results in 78 Patients*. American Society of Nephrology 3 : 103–107. (1992)
- [33] Nelson W., *Theory and Applications of Hazard Plotting for Censored Failure Data*. Technometrics 14 : 945–965. (1972).
- [34] Patelia V. and Rolin J.M., *Product-Limit estimator of the survival function with twice censored data* The Annals of Statistics Vol. 34, No. 2 : 925–938. (2006).
- [35] Prentice L., *Linear rank tests with right-censored data*. Biometrika 65, 167–179. (1978).
- [36] Peto R. *Experimental survival curves for interval-censored data*. Applied Statistics 22 : 86–91. (1973).
- [37] Raison J. *Les principaux tests non paramétriques, Quelques généralités et références bibliographiques*. Revue de statistiques appliquée 7 : 83–106. (1959).

- [38] Sun J., *A nonparametric test for interval-censored failure time data with application to AIDS studies*. *Statistics in Medicine* 15 : 1387 – 1395. (1996).
- [39] Sun J., *Interval Censoring*. *Encyclopedia of Biostatistics* : 2090–2095. John Wiley, New York. (1998).
- [40] Sun J., *Nonparametric Test for Doubly Interval-Censored Failure Time Data*. *Lifetime Data Analysis*, 7 : 363 – 375. (2001).
- [41] Sun J., *The Statistical Analysis of Interval-censored Failure Time Data*. Springer Science-Business Media, Inc. (2006).
- [42] Shorack R. G. et Wellner W. J., *Empirical processes with applications to statistics*. John Wiley Sons, (1986).
- [43] Tarone R. E. et Ware J., *On distribution free tests for equality of survival distributios*. *Biometrika* 64 : 156–160. (1977).
- [44] Turnbull B. W., *Nonparametric Estimation of a Survivorship Function with Doubly Censored Data*. *Journal of the American Statistical Association*, Vol.69, No.345, pp.169–173. (1974).
- [45] Vo Khac Kh., *Théorie des probabilités*. Ellipses (1984).
- [46] Wilcoxon F. *A simplified method of evaluating dose-effect experiments*. *Pharmacol Exp. Ther.* 96 : 99–113. (1949).
- [47] Zhao Q. et Sun J., *Generalized log-rank test for mixed interval-censored failure time data*. *Statist. Med.* 23 : 1621–1629. (2004).