Dédicace

A ma mère, à ma mère, à ma mère Fatma et à mon père

A mes frères Mhamed et sa fille soulaf ,Abdelmoumen et son fils mostafa et

Abdelkader, mes sœurs et leurs familles

A Kwider et son file Bodour et abedalwadoude et spécialement hajarer

A mes collègues et mes amies

Abdelali, abdeallahe, , Kadiro, Moulay, Abdelrahman, Abdelkader,

Bakhaled, Lwibed Mohamed

Spécialement à Melle maref pour son encouragement, sa patience et son aide.

Ismail Arabi

Remerciements

Merci Dieu le tout puissant pour nous avoir donné la foi et le courage pour accomplir notre travail.

Nous tenons à remercier notre professeur et superviseur

Melle MAREF FOUZIA . pour la confiance qu'elle nous a accordé, les conseils qu'elle nous a donné, l'aide qu'elle nous a apporté et le suivi qu'elle a effectué.

Nous exprimons notre profonde gratitude envers les membres du jury, pour l'intérêt qu'ils ont bien voulu porter à ce projet en acceptant d'y participer.

Nos remercions tous les Enseignants de notre Département.

Ces remerciements ne seraient pas complets si nous ne les associons à nos familles qui nous ont soutenu et supporté pendant ces années d'études.

Enfin, nous souhaitons "bon courage "aux étudiants de notre promotion.

Ismail Arabi

Table des matières

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, 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 [10], le test de Mann-Whithney 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 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 étudions différents travaux dans ce contexte : Gehan (1965)[1], Tarone et Ware (1977)[8], Peto et Peto (1973) [5]. Au troisième chapitre, une généralisation des tests de rangs au cas de la censure par intervalles est présentée, une étude détaillée de l'article de Kim et autres (2006) [3] 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.

Chapitre 1

Rappels

Ce chapitre est divisé en deux sections. Dans la première section nous énonsons un certain nombre de généralités autour des tests d'hypothèses. Nous rappelons dans la deuxième section les différentes fonctions utilisées en analyse de survie et les différents schémas de censures ainsi que la constitution des observations ceci nous permettant d'introduire l'estimateur de Kaplan-Meier et l'estimateur de Nealson.

1.1 Tests d'hypothèses

Les tests statistiques constituent une approche décisionnelle de la statistique inférentielle. Un tel test a pour objet de décider sur la base d'un échantillon si une caractéristique de la loi mère (ou de la population) répond ou non à une certaine spécification que l'on appelle hypothèse, par exemple : on veut décider si un nouveau médicament est efficace, si une méthode pédagogique est meilleure qu'une autre, 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.

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

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.

1.1.1 Tests non paramétriques

On parle de tests non paramétriques lorsque l'on ne fait aucune hypothèse sur la distribution des variables. L'hypothèse alternative générique est leur différence.

La très grande majorité des tests non paramétriques reposent sur la notion de rangs. L'idée est de substituer aux valeurs leur numéro d'ordre dans l'ensemble des données, nous donnons tout d'abord quelques propriétés les concernant.

Les statistiques de rang

On considère un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de loi F. Pour des réalisations $(x_1, x_2, \ldots, x_n)$, le rang r_i d'une valeur x_i est la position qu'elle occupe quand les valeurs sont rangées dans l'ordre croissant. A tout vecteur de réalisations on peut donc associer le vecteur des rangs $(r_1, r_2, \ldots, r_n)$ qui consiste en une permutation des nombres $1, 2, \ldots, n$. Par exemple, avec $n = 5$, au vecteur $(8.2, 7.4, 9.2, 5.1, 6.7)$ on associe le vecteur des rangs $(4, 3, 5, 1, 2)$. Cette fonction appliquée à $(X_1, X_2, \ldots, X_n)$ procure les statistiques de rang $(R_1, R_2, \ldots, R_n)$. La v.a. R_i sera appelée le rang de X_i. Notons que si X_i est la statistique d'ordre k alors R_i est égal à k.

Proposition : Pour tout $i = 1, \ldots, n$, le rang R_i suit une loi discrète uniforme sur $\{1, 2, \ldots, n\}$. Ainsi :

$$E(R_i) = \frac{n+1}{2} \quad et \quad V(R_i) = \frac{n^2-1}{12}.$$

De plus on démontre que, pour tout i et tout j distincts,

$$Cov(R_i, R_j) = -\frac{n+1}{12}.$$

1.1.2 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.

Nous nous plaçons dans une optique non paramétrique au sens où les tests considérés devront s'appliquer quelle que soit la nature du modèle de loi mère envisagé. La fonction de répartition étant l'objet mathématique le plus approprié pour spécifier une loi, qu'elle soit discrète ou continue, il s'agit de tester

$$\begin{cases} H_0 : & F(x) = F_0(x) \\ H_1 : & F(x) \neq F_0(x) \end{cases}$$

où $F(x)$ est la fonction de répartition de la variable échantionnée et $F_0(x)$ est la fonction de répartition d'une variable aléatoire connue.

Nous présentons ici les deux tests les plus classiques, celui de Khi-deux et celui de Kolomogorov-Smirnov.

Test du khi-deux

Soit un échantillon aléatoire de taille n extrait d'une population et divisé en k classes d'effectifs respectifs $n_1, n_2, \ldots, n_k$ et de probabilités respectives $p_1, p_2, \ldots, p_k$ théoriques.

Considérons la statistique D^2 définie comme suit :

$$D^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Cette statistique est une mesure de l'écart aléatoire entre les effectifs réalisés et les effectifs espérés. D^2 dépend du nombre de termes de la somme k mais on remarque que tous ces termes ne sont pas indépendants puisque $\sum_{i=1}^{n} n_i = n$ il suffit d'en connaitre en fait $k-1$ degrés de liberté de D^2.
Le seuil de signification α étant fixé, on lit sur la table du χ^2 la valeur $\chi^2_{k-1}(\alpha)$.
Si $D^2 > \chi^2_{k-1}(\alpha)$, on rejette H_0.

Test de Kolmogorov-Smirnov

Ce test est fondé sur l'écart constaté entre la fonction de répartition empirique F_n et la fonction de répartition d'une loi continue connue F_0.
On détermine $D_n = \max |F_n(x) - F(x)|$ et on compare cet écart à des valeurs critiques particulières données par les tables $d(n)$.
On rejette H_0 si $D_n > d(n)$ à un seuil de signification α.

On présente maintenant le test de Wilcoxon et le test de Mann-Whitney au cas de deux échantillons afin de comparer leurs distributions. Soit $(X_1, \ldots, X_n)$ un n-échantillon de fonction de répartition F_X et $(Y_1, \ldots, Y_m)$ un m-échantillon de fonction de répartition F_Y. On suppose que les deux échantillons sont indépendants et que F_X et F_Y sont continues. On veut tester $H_0 : F_X = F_Y$ contre $F_X \neq F_Y$.

Tests de Wilcoxon et Mann-Whitney :

Ce test a été proposé initialement par Wilcoxon (1945). Par la suite Mann et Whitney (1947) ont proposé une forme équivalente qui permit de préciser ses propriétés. Soit deux échantillons indépendants $X_1, X_2, \ldots, X_n$ et $Y_1, Y_2, \ldots, Y_m$ issus respectivement de chaque loi. Considérons la fusion des $n + m$ valeurs en un seul échantillon et les rangs associés à celui-ci. La statistique de test de Wilcoxon est la somme des rangs de l'un des échantillons initiaux. Il est plus rapide de choisir celui de plus petite taille et nous supposerons qu'il s'agit du premier (soit $n \leq m$), notant alors la somme de ses rangs T_n. La valeur minimale pour T_n est atteinte lorsque toutes les réalisations $x_1, x_2, \ldots, x_n$ sont situées à gauche des réalisations $y_1, y_2, \ldots, y_m$ sur la droite réelle et elle vaut $1 + 2 + \ldots + n = n(n + 1)/2$. La valeur maximale est atteinte lorsque toutes les observations x_i sont situées à droite des observations y_j sur la droite réelle et elle vaut :

$$(m + 1) + (m + 2) + \ldots + (m + n) = nm + n(n + 1)/2.$$

Intuitivement on est enclin à rejeter H_0 lorsque la valeur de T_n s'approche de l'un ou l'autre de ces extrêmes (mais un seul d'entre eux pour un cas unilatéral). Pour déterminer les valeurs critiques il est nécessaire d'établir la loi de cette statistique sous H_0.
Si $n, m < 8$, on lit dans la table du test de Wilcoxon le nombre t tel que, sous (H_0),

$P(T > t) = \alpha$. On rejette (H_0) au risque d'erreur α si $T > t$. Autrement on accepte (H_0).

Pour les grandes tailles d'échantillons (en fait $n > 8$ et $m > 8$ suffisent) on peut utiliser une approximation gaussienne découlant du comportement asymptotique de T_n sous H_0. Pour cela il faut utiliser la moyenne et la variance de cette statistique sous H_0 :

$$E(T_n) = n(n + m + 1)/2,$$

$$V(T_n) = ((n + m)^2 - 1)/12.$$

En effet, on a, pour tout $i = 1, \ldots, n$ le rang R_i suit une loi discrète uniforme sur $1, 2, \ldots, n + m$, alors,

$$\begin{aligned} E(T_n) &= E(\sum_{i=1}^{n} R_i) \\ &= \sum_{i=1}^{n} E(R_i) \\ &= \sum_{i=1}^{n} (n + m + 1)/2 \\ &= n(n + m + 1)/2. \end{aligned}$$

Consernant la variance, on a, par extension de la formule sur la variance d'une somme de deux v.a. non indépendantes,

$$\begin{aligned} V(T_n) &= V(\sum_{i=1}^{n} R_i) \\ &= \sum_{i=1}^{n} V(R_i) + 2 \sum_{i<j} Cov(R_i, R_j) \\ &= \frac{n(n + m)^2 - 1}{12} + 2 \sum_{i<j} Cov(R_i, R_j) \end{aligned}$$

où la somme $\sum_{i<j}$ est à effectuer sur tous les $n(n - 1)/2$ couples (R_i, R_j) du premier échantillon tels que $i < j$. Comme $V(R_i) = ((n + m)^2 - 1)/12$ et

$Cov(R_i, R_j) = -(n+m+1)/12$ quels que soient i et j, on obtient finalement :

$$\begin{aligned} V(T_n) &= \frac{n(n+m)^2-1}{12} - 2\frac{n(n-1)}{2}\frac{(n+m+1)}{12} \\ &= \frac{nm(n+m+1)}{12}. \end{aligned}$$

La région critique est alors définie par :

$$|T_n - \frac{n(n+m+1)}{2}| > u_{\alpha/2}\sqrt{\frac{nm(n+m+1)}{12}}.$$

Exemple : On veut comparer les performances de deux groupes d'élèves à des tests d'habileté manuelle :
On choisit aléatoirement 8 individus du premier groupe et 10 du deuxième groupe. Les performances en minutes sont les suivantes :

Groupe1 : 22 31 14 19 24 28 27 28

Groupe2 : 25 13 20 11 23 16 21 18 17 26

On réordonne les 18 observations par ordre croissant. Les résultats du premier groupe sont soulignés :

Observations : 11 13 <u>14</u> 16 17 18 <u>19</u> 20 21 <u>22</u> 23 <u>24</u> 25 26 27 <u>28</u> <u>28</u> <u>31</u>

Rangs : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

La somme des rangs des individus du premier groupe est

$$T_n = 3+7+10+12+15+16+17+18 = 98$$

Si H_0 est vraie :

$$E(T_n) = \frac{8(8+10+1)}{2} = 76 \qquad \text{et} \qquad V(T_n) = \frac{8\times 10(8+10+1)}{12} = (11.25)^2$$

Comme $\dfrac{98-76}{11.25} = 1.96 \quad (> 1.65)$, on peux rejeter H_0 avec $\alpha = 0.1$ et conclure à une plus grande rapidité des élèves du groupe 2.

Test de 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'exequo).

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{\{X_i > Y_j\}}$$

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

$$U = T_n - n(n+1)/2$$

Les deux tests sont donc strictement équivalents. Pour l'exemple précédent, la statistique U prend la valeur $2 + 5 + 7 + 8 + 10 + 10 + 10 + 10 = 62 = 98 - 36$.

1.2 Analyse de survie

L'analyse de survie est un domaine des statistiques qui trouve sa place dans tous les champs d'application où l'on étudie la survenue d'un évènement. L'objectif de cette analyse réside dans l'analyse du délai de survenue d'un évènement dans un ou plusieurs groupes d'individus. Dans le domaine biomédical, par exemple, plusieurs évènements sont intéressants à étudier : le développement d'une maladie, la réponse à un traitement donné, la rechute d'une maladie ou le décés. Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, dans les enquêtes épidémiologiques, les données sont souvent recueillies de façons incomplètes. La censure et la troncature font partie de processus générant ce type de données.
Dans cette section, on va rappeler quelques notions de base utiles pour la suite de l'exposé.

1.2.1 Définitions et Notations

La durée de survie et la date d'origine

Soit T la variable aléatoire positive et continue qui représente la durée de survie ou délai, c'est-à-dire la durée écoulée jusqu'à la survenue de l'évènement d'intérêt. Pour définir cette durée, il faut définir une date d'origine qui est généralement propre aux sujets et

dont le choix va dépendre de l'évènement d'intérêt. Dans un contexte d'essais cliniques par exemple, si l'on souhaite comparer deux traitements, on choisit la date de mise sous traitement comme date d'origine. Lorsque l'évènement étudié est très dépendant de l'âge, on choisit souvent la date de naissance comme date d'origine et la variable T est alors un âge.

Censure et troncature :

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées (i.i.d.) de durées T, on observe la réalisation de la variable T soumise à diverses perturbations, indépendantes ou non du phénomène étudié. La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Pour l'individu i, considérons

- son temps de survie T_i,
- son temps de censure C_i,
- la durée réellement observée X_i.

Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées, pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

1. La censure de type I
   Soit C une valeur fixée, au lieu d'observer les variables $T_1, \ldots, T_n$ qui nous intéressent, on n'observe T_i uniquement lorsque $T_i \leq C$; sinon on sait uniquement que $T_i > C$. On utilise la notation suivante : $X_i = T_i \wedge C = \min(T_i, C)$. Par exemple,

Dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note T la durée d'apprentissage de cette langue. Pour certains étudiants nous allons observer leurs durées X_i d'apprentissage de la langue par contre pour d'autres leurs X_i ne seront pas observées car le stage est limité dans le temps.

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

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

3. La censure de type III (ou censure aléatoire de type I) :
   Soient $C_1, \ldots, C_n$ des variables aléatoires i.i.d. On observe les variables

$$X_i = T_i \wedge C_i.$$

   L'information disponible peut être résumée par :
   - la durée réellement observée X_i,
   - un indicateur $\delta_i = I_{T_i \leq C_i}$
   - $\delta_i = 1$ si l'événement est observé (d'où $X_i = T_i$). On observe les "vraies" durées ou les durées complètes.
   - $\delta_i = 0$ si l'individu est censuré (d'où $X_i = C_i$). On observe des durées incomplètes (censurées).

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

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

Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (X, δ) :

$$\begin{aligned} X &= T \vee C = max(T, C), \\ \delta &= I_{T \leq C}. \end{aligned}$$

Comme pour la censure à droite, on suppose que la censure C est indépendante de T. Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues.

La censure par intervalle se rencontre généralement dans les études de cohorte lorsque les sujets ne sont pas observés en temps continu mais par intermittence lors de visites. Par exemple, si l'on s'interesse a l'âge de survenue d'une maladie et que le sujet i est diagnostique malade au cours d'une visite, on sait seulement que $T_i \in [L_i, R_i]$ ou R_i est l'âge a la visite de diagnostic et L_i est l'âge a la visite précédente.

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 X conditionnellement à l'événement $\{X < C\}$ (respectivement $\{X > C\}$) 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 Fonctions de base en analyse de survie

Dans cette partie, nous allons définir des fonctions jouant un rôle très important en analyse de survie et nous allons voir comment elles sont interdépendantes.
Soit T une variable aléatoire non négative et continue qui représente la durée de survie d'un sujet dans une expérience. Plusieurs fonctions caractérisent la distribution de T :

Fonction de survie S

La fonction de survie est la probabilité que la durée de vie T soit supérieur à un temps t,

$$S(t) \quad = \quad \mathbb{P}(T > t), \qquad t \geqslant 0.$$

Fonction de répartition F

La fonction de répartition (ou c.d.f. pour "cumulative distribution function") est la probabilité de décéder entre 0 et t,

$$F(t) \quad = \quad \mathbb{P}(T \leqslant t) = 1 - S(t).$$

1.2.3 Densité de probabilité f

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

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

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

$$f(t) = \lim_{h \to 0} \frac{\mathbb{P}(t < T \leq t + h)}{h} = F^{'}(t) = -S^{'}(t).$$

Pour t fixé, la densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

1.2.4 Risque instantané h (ou taux de hasard)

La fonction de risque (parfois appelé aussi taux de hasard, taux de défaillance ou taux de survie) au point t s'interprète comme la probabilité instantanée de sortir de l'état que l'on observe (vie chômage, ...) à la date t, sachant que le sujet est encore dans cet état en t,

$$h(t) = \lim_{h \to 0} \frac{\mathbb{P}(t < T \leq t + h | T > t)}{h} = \frac{f(t)}{S(t)} = -(\ln(S(t)))^{'}.$$

Taux de hasard cumulé H

Le taux de hasard cumulé est l'intégrale du risque instantané h,

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

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

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

On en déduit que

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

1.2.5 Estimation des fonctions de base

La distribution du temps de survie que l'on cherche à modéliser est généralement inconnue. Deux approches sont possibles pour estimer cette distribution : l'inférence paramétrique et l'inférence non paramétrique. On s'intéresse, dans notre mémoire, aux méthodes non paramétriques. De nombreux estimateurs ont été développés afin de considérer les mécanismes de censure et troncature. Les plus connus sont l'estimateur de la fonction de survie de (Kaplan-Meier, 1958 [2]) et celui de la fonction de risque de Nelson-Aalen (Nelson, 1972, Aalen, 1975 [4]) pour traiter des données censurées à droite.

Estimateur de Kaplan-Meier de la survie

Cet estimateur est aussi appelé P-L (Produit-Limite) car il s'obtient comme la limite d'un produit. Il est fondé sur la remarque suivante : Si $t' < t$, la probabilité de survivre au-delà de l'instant t est égale au produit suivant :

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

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

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

et ainsi de suite. Donc, si on a $t_0 < t_1 < \ldots < t_n < t$, on obtient

$$S(t) = \mathbb{P}(T > t/T > t_n)\mathbb{P}(T > t_n/T > t_{n-1}) \ldots \mathbb{P}(T > t_1/T > t_0)\mathbb{P}(T > t_0).$$

Si l'on choisit pour les dates où l'on conditionne celles où il s'est produit un événement, qu'il s'agisse d'une mort ou d'une censure, on aura seulement à estimer des quantités de la forme :

$$\mathbb{P}(T > X_{(i)}/T > X_{(i-1)}) = p_i$$

Or p_i est la probabilité de survivre au-delà de l'intervalle de temps $I_i = [X_{(i-1)}, X_{(i)}[$ sachant qu'était vivant au début de l'intervalle.
Notons R_i le nombre des sujets qui sont vivants (donc "à risque" de mourir) juste avant l'instant $T_{(i)}$ et M_i le nombre des morts à l'instant $T_{(i)}$. On pose $q_i = 1 - p_i$. q_i est la probabilité de mourir durant l'intervalle I_i sachant que l'individu était vivant au début de cet intervalle. Alors l'estimateur naturel de q_i est

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

Alors l'estimateur de Kaplan-Meier est donné par

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

Supposons qu'il n'y ait pas d'ex-equo. Soit $\delta_{(i)}$ l'indicateur de censure associée à $X_{(i)}$, si $\delta_{(i)} = 1$, c'est qu'il y a eu un mort en $X_{(i)}$ et donc $M_i = 1$. Si $\delta_{(i)} = 0$, c'est qu'il y a eu une censure en $T = X_{(i)}$ et donc $M_i = 0$.
Par suite

$$\widehat{p_i} = \begin{cases} 1 - \dfrac{1}{R_i} & en \ \ cas \ \ de \ \ mort \ \ en \ \ X_{(i)} \\ 1 & en \ \ cas \ \ de \ \ censure \end{cases}$$

Comme $R_i = n - i + 1$, l'estimateur de Kaplan-Meier est dans ce cas donné par :

$$\widehat{S(t)} = \prod_{T = X_{(i)} \leq t} \left(1 - \frac{1}{n - i + 1}\right)^{\delta_{(i)}}.$$

Remarque

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

$$S_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{T_i > t\}}.$$

Estimateur de Nelson-Aalen du risque cumulé :

Si la variable T admet une densité, on a par définition du risque cumulé, du risque instantané et de la densité

$$H(t) = \int_0^t h(t)du = \int_0^t \frac{f(t)}{S(t)}dt.$$

Dans le cas où T n'admet pas de dérivée en tout point de R^+, on peut toujours définir le risque cumulé en utilisant la définition de la densité de T,

$$H(t) = -\int_0^t \frac{dS(t)}{S(t)}.$$

Considérons les quantités $Y(t) = \mathbb{P}(X > t)$ et $N(t) = \mathbb{P}(X > t, \delta = 1)$ et introduisons $G(t)$ la fonction de survie de la variable C. D'après l'hypothèse d'indépendance, on obtient les égalités suivantes

$$\begin{aligned} Y(t) &= \mathbb{P}(X > t) = \mathbb{P}(T > t, C > t) = S(t)G(t) \\ N(t) &= \mathbb{P}(X > t, \delta = 1) = \mathbb{P}(T > t, C \geq T) = E(1_{T>t}G(T-)) \\ &= \int_t^\infty G(u-)f(u)du = -\int_t^\infty G(u-)dS(u) \end{aligned}$$

Par conséquent, $dN(t) = G(t-)dS(t)$ et on obtient l'expression suivante pour le risque cumulé :

$$H(t) = \int_0^t \frac{dN(u)}{Y(u-)}$$

Un estimateur naturel s'obtient en remplaçant les fonctions Y et N par leurs équivalents empiriques (calculables car les variables X et δ sont observées). Soient

$$\widehat{Y}(t) = \frac{1}{n}\sum_{i=1}^n 1_{X_i>u} \quad et \quad \widehat{N}(t) = \frac{1}{n}\sum_{i=1}^n 1_{T=X_i>u,\delta_i=1}$$

l'estimateur de Nelson-Aalen est donné par les expressions suivantes

$$\widehat{H}(t) = \int_0^t \frac{\widehat{d}N(u)}{\widehat{Y}(u-)} = \sum_{T_i \leq x} \frac{M_{(i)}}{R_{(i)}}$$

où R_i représente le nombre d'individus à risque juste avant T_i et M_i représente le nombre de décès en T_i : L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut de taille M_i/R_i à chaque instant de décés.

Chapitre 2

Tests non paramétriques dans le cas des données censurées à droite

Le but de ce chapitre est de comparer les durées de vie de deux échantillons indépendants. Plus précisément, on dispose de deux échantillons indépendants, éventuellement censurés, et on souhaite tester l'hypothèse nulle d'égalité des fonctions de survie dans les deux échantillons (resp. plusieurs échantillons). En l'absence de censure, on dispose des tests classiques de rang (test de Wilcoxon, test de Khi-deux), que l'on va adapter à la présence de censure.

2.1 Statistique du Logrank

Pour simplifier, nous exposons le test de Logrank pour seulement deux groupes d'individus, avant d'aborder l'extension à plus de deux groupes.

2.1.1 Comparaison de deux fonctions de survies

On s'intéresse dans ce mémoire à une approche non-paramétrique. Le principe des tests consiste à comparer le nombre de décès observés dans chaque groupe au nombre de décès attendus (calculés sous l'hypothèse d'égalité des distributions de survie).

Notations :

Considérons les notations suivantes,

- $T_1 < \ldots < T_N$ les temps de décès ordonnés des deux échantillons réunis,

- d_{Ai} et d_{Bi} le nombre de décès observés au temps T_i dans chacun des groupes A et B,
- $d_i = d_{Ai} + d_{Bi}$, le nombre total de décès observés en T_i,
- Y_{Ai} et Y_{Bi} le nombre de sujets à risques en T_i dans les groupes A et B,
- $Y_i = Y_{Ai} + Y_{Bi}$, le nombre total de sujets à risques en T_i.

Pour chaque temps d'événement T_i, l'information peut être résumée sous forme de tableau :

	Décès en T_i	Vivant après T_i	
Groupe A	d_{Ai}	$Y_{Ai} - d_{Ai}$	Y_{Ai}
Groupe B	d_{Bi}	$Y_{Bi} - d_{Bi}$	Y_{Bi}

Statistiques de test

On cherche à tester l'hypothèse $H_0 : S_A(t) = S_B(t)$ qui est l'égalité des fonctions de survie dans les deux groupes. Ainsi, sous l'hypothèse H_0, la proportion attendue de décès (parmi les sujets à risque) est identique dans les deux groupes pour tous les temps de décès T_i : Pour chaque temps T_i, on peut comparer les pourcentages de décès parmi les sujets à risque dans chacun des groupes en utilisant le test du khi-deux.

Soit D_{Ai} (D_{Bi} et D_i) la variable dont la valeur est d_{Ai} (d_{Bi} et d_i), on peut montrer que D_{Ai} suit une loi hypergéométrique d'espérance :

$$\mathbb{E}(D_{Ai}) = \frac{Y_{Ai} \times d_i}{Y_i}$$

et de variance

$$\mathbb{V}(D_{Ai}) = \frac{Y_i - d_i}{Y_i - 1} \times \frac{d_i Y_{Ai}(Y_i - Y_{Ai})}{Y_i^2}$$

où $\mathbb{E}(D_{Ai})$ correspond au nombre de décès attendus dans le groupe A : Sous H_0, on montre que les variables $D_{Ai} - \mathbb{E}(D_{Ai})$ suivent asymptotiquement des lois $N(0, \mathbb{V}(D_{Ai}))$

($\frac{[D_{Ai} - \mathbb{E}(D_{Ai})]^2}{\mathbb{V}(D_{Ai})}$ suivent asymptotiquement des loi de χ^2_1).

Considérons des pondérations $w_i, i = 1, \ldots, N$, alors par indépendance entre les variables D_{Ai} et D_{Aj} (associées aux T_i et T_j), les variables

$$\sum_{i=1}^{N} w_i(D_{Ai} - \mathbb{E}(D_{Ai})) = \sum_{i=1}^{N} w_i(D_{Ai} - \frac{Y_{Ai} \times d_i}{Y_i})$$

suivent asymptotiquement des lois normales de moyennes nulles et de variances $\sum_{i=1}^{N} w_i^2 \mathbb{V}(D_{Ai})$. Par conséquent, sous H_0, les statistiques suivantes

$$\chi^2_0 = \frac{\left[\sum_{i=1}^{N} w_i(D_{Ai} - \frac{Y_{Ai} \times d_i}{Y_i})\right]^2}{\sum_{i=1}^{N} w_i^2 \frac{(Y_i - d_i)}{(Y_i - 1)} \frac{d_i Y_{Ai}(Y_i - Y_{Ai})}{Y_i^2}}$$

suivent asymptotiquement des lois de χ^2 à 1 degré de liberté.

Remarque 2.1.1 *Plusieurs statistiques de test ont été proposées*

- **Test du logrank :** $w_i = 1$,
  cette pondération attribue à chaque décès le même poids quel que soit l'instant où il survient. Le test compare le nombres de décès observés au nombre de décès attendus.

- **Test de Gehan :** $w_i = Y_i$,
  la pondération en T_i est égale au nombre d'individus à risque en T_i, donc les poids sont plus élévés pour les décès précoces que tardifs.

- **Test de Peto et Prentice :** $w_i = \prod_{k=1}^{i} \frac{Y_k}{Y_k + d_k}$,
  ces pondérations sont proches de l'estimateur de Kaplan-Meier de la survie. Elles attribuent des poids plus élévés aux décès précoces.

Remarque 2.1.2 *Il est important de noter que, par construction, ces tests sont valides uniquement si les fonctions de survie dans les deux groupes ne se croisent pas durant toute la période étudiée. Si les courbes se croisent, on observe une perte de puissance.*

Remarque 2.1.3 *Le test du Logrank est le test le plus souvent utilisé.*

Comparaison de trois groupes

Les tests de la section précédente se généralisent au cas de la comparaison des fonctions de survie de plusieurs échantillons. Dans cette section, seule l'extension du test du logrank sera envisagée (car c'est le test le plus utilisé).

Considérons le cas de trois groupes A, B et C, le tableau suivant résume les notations,

	Décès en T_i	Vivant après T_i	
Groupe A	d_{Ai}	$Y_{Ai} - d_{Ai}$	Y_{Ai}
Groupe B	d_{Bi}	$Y_{Bi} - d_{Bi}$	Y_{Bi}
Groupe C	d_{Ci}	$Y_{Ci} - d_{Ci}$	Y_{Ci}
	d_i	$Y_i - d_i$	Y_i

En suivant la même démarche que dans le cas de deux échantillons, on montre que le vecteur suivant :

$$V = \begin{pmatrix} \sum_{i=1}^{N} (D_{Ai} - \mathbb{E}(D_{Ai})) \\ \sum_{i=1}^{N} (D_{Bi} - \mathbb{E}(D_{Bi})) \end{pmatrix}$$

avec

$$E(D_{Ai}) = \frac{Y_{Ai} \times d_i}{Y_i}$$

$$E(D_{Bi}) = \frac{Y_{Bi} \times d_i}{Y_i}$$

suit asymptotiquement une loi normale dans $\mathbb{R}^2$ d'espérance nulle. On en déduit ensuite que la statistique suivante

$$\chi_0^2 = V' \begin{pmatrix} \sum_{i=1}^{N} \mathbb{V}(D_{Ai}) & \sum_{i=1}^{N} Cov(D_{Ai}, D_{Bi}) \\ \sum_{i=1}^{N} Cov(D_{Ai}, D_{Bi}) & \sum_{i=1}^{N} Cov(D_{Ai}, D_{Bi}) \end{pmatrix}^{-1} V$$

avec

$$\mathbb{V}(D_{Ai}) = \frac{Y_i - d_i}{Y_i - 1} \times \frac{d_i Y_{Ai}(Y_i - Y_{Ai})}{Y_i^2}$$

$$\mathbb{V}(D_{Bi}) = \frac{Y_i - d_i}{Y_i - 1} \times \frac{d_i Y_{Bi}(Y_i - Y_{Bi})}{Y_i^2}$$

$$Cov(D_{Ai}, D_{Bi}) = -\frac{Y_i - d_i}{Y_i - 1} \times \frac{d_i Y_{Ai} Y_{Bi}}{Y_i^2}$$

suit asymptotiquement une loi de de χ^2 à 2 degrés de liberté. Le raisonnement ci-dessus, avec le couple (A, B), est également possible avec les couples (A, C) ou (B, C) ce qui permet d'obtenir d'autres statistiques de test équivalentes.

Dans le cadre général de la comparaison de k groupes, la statistique de test s'obtient de la même façon et suit asymptotiquement une loi de χ^2 à $k - 1$ degrés de liberté.

2.1.2 Comparaison de plusieurs fonctions de survies

Nous allons traiter dans ce paragraphe le problème de comparaison k ($k \geq 2$) populations en se basant sur leurs fonction de survie $S_i(t)$ ($1 \leq i \leq k$), pour cela testons l'hypothèse suivante :

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t), \quad \textit{pour} \quad t \leq \tau,$$

contre H_1 : Au moins un des $S_j(t)$ 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 pour chacune des k populations.

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

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

Le test H_0 est basé sur des comparaisons pondérées de l'estimateur du taux de hasard de la $j^{ème}$ population et utilise l'estimateur de Nelson-Aalan.
Si l'hypothèse nulle est vraie alors le taux de hasard espéré dans la jème population est estimé par le taux de hasard dans l'échantillon global : $\frac{d_i}{Y_i}$. En utilisant les données du $j^{ème}$ 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 $Y_{ij} = 0$. Le test H_0 est basé sur les statistiques :

$$U_j(\tau) = \sum_{i=1}^{N} w_j(t_i)\{\frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_j}\}, \quad j = 1, \ldots, k$$

Si toutes les valeurs $U_j(\tau)$ sont proches de zéro alors il est peu probable que l'hypothèse nulle soit fausse, tandis que, si une des $U_j(\tau)$ est loin de zéro alors nous pouvons admettre que la jème 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 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 :

$$U_j(\tau) = \sum_{i=1}^{N} w(t_i)\{d_{ij} - Y_{ij}(\frac{d_i}{Y_i})\}, \quad j = 1, \ldots, k$$

On note qu'avec cette fonction de poids, la statistique est la somme des différences pondérées entre les nombres de morts observés et les nombres de morts espérés sous H_0 dans le jème échantillon. La variance de $U_j(\tau)$ est donnée par :

$$\hat{\sigma}_{jj} = \sum_{i=1}^{N} w(t_i)^2 \frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad j = 1, \ldots, k$$

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

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

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

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 $(d_i, p_j = \frac{Y_{ij}}{Y_i}), j = 1, \ldots, k$.

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

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

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 $k-1$: $\sum_{(k-1)\times(k-1)}$ contenant les $\hat{\sigma}_{jg}$. La statistique du test est donnée par la forme quadratique :

$$\chi^2 = (U_1(\tau), \ldots, U_{k-1}(\tau)) \sum^{-1} (U_1(\tau), \ldots, U_{k-1}(\tau))^t.$$

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 :

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

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. $H_1 : h_1(t) > h_2(t)$ pour un certain $t \leq \tau$, On rejette H_0 si $\chi^2 \geq \chi^2_\alpha$.
2. $H_1 : h_1(t) \neq h_2(t)$ pour un certain $t \leq \tau$. On rejette H_0 si $|\chi^2| \geq \chi^2_{\alpha/2}$.

2.2 Autres tests de comparaison

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

Test de Gehan(1965)

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

west 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.

Test de Peto et Peto

Ici le choix de la fonction de poids est $w(t_i) = \hat{S}(t_i)$, où :

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{Y_i + 1}).$$

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

$$w(t_i) = \frac{\hat{S}(t_i) Y_i}{Y_i + 1}$$

Chacun des poids définis 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.

Chapitre 3

Tests basés sur des données censurées par intervalles

Nous allons nous intéresser ici aux tests non paramétriques dans le cas où les données sont censurées par intervalle. Ce chapitre est divisé en deux sections, on présente dans la première section l'estimateur de maximum de vraisemblance de la fonction du survie et dans la deuxième section, on fait une comparaison de plusieurs fonctions du survies.

3.1 Procédures d'estimation

Rappelons que les temps de survie sont notés T_i, et qu'ils sont censurés dans les intervalles $A_i = (G_i, D_i], i = 1, \ldots, N$.

Supposons que $(T_i, 1, \ldots, N)$ est un échantillon aléatoire issu d'une certaine loi, dont la fonction de répartition est notée F. L'objectif de cette section réside dans l'estimation non paramétrique de la sonction de survie S, ou de façon équivalente de F.

Il existe diverses procédures d'estimation non paramétrique de la fonction de survie à partir de données censurées par intervalle. Nous verrons qu'elles se distinguent essentiellement par les algorithmes de calcul. Elles reposent cependant toutes sur une approche de maximum de vraisemblance.

3.1.1 Estimation du maximum de vraisemblance

Sous l'hypothèse que F est continue à droite, la contribution de la $i-$ème observation à la vraisemblance totale de l'échantillon $(T_i : i = 1, \ldots, N)$ est donnée par

$$L_i = F(D_i) - F(G_i).$$

La vraisemblance totale de l'échantillon s'écrit :

$$L(F) = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} (F(D_i) - F(G_i)).$$

Supposons maintenant que les intervalles de censure puissent s'écrire sous la forme d'une union finie de $k_i \geq 1$ intervalles disjoints :

$$A_i = \bigcup_{j=1}^{k_i} (G_{ij}, D_{ij}), i = 1, \ldots, N$$

La vraisemblance s'écrit dès lors

$$L(F) = \prod_{i=1}^{N} \sum_{j=1}^{k_i} (F(D_{ij}) - F(G_{ij})).$$

Construisons maintenant un ensemble d'intervalles disjoints dont les bornes gauche et droite appartiennent aux ensembles $\{G_{ij}, j = 1, \ldots, k_i\}$, $\{D_{ij}, j = 1, \ldots, k_i\}$, respectivement. Notons ces intervalles $(p_1, q_1], \ldots, (p_m, q_m]$ ils vérifient

$$0 \leq p_1 < q_1 < p_2 < q_2 < \ldots < p_m < q_m < \infty.$$

Définissons en outre les quantités

$$s_j = F(q_j) - F(p_j), j = 1, \ldots, m$$

On dispose alors du résultat suivant :

Théorème 3.1.1 *Le problème de maximisation de la vaisemblance $L(F)$ est équivalent au problème de maximisation de*

$$L(s_1, \ldots, s_m) = \prod_{i=1}^{N} \sum_{j=1}^{m} \alpha_{ij} s_j.$$

où

$$\alpha_{ij} = \begin{cases} 1 \quad si\ ]p_j, q_j] \subset A_i \\ \\ 0 \quad si\ sinon \end{cases}$$

Notons la log-vraisemblance

$$l(s) = \ln L(s_1, \ldots, s_m) = \sum_{i=1}^{N} \ln(\sum_{j=1}^{m} \alpha_{ij} s_j)$$

et notons l'ensemble

$$D_s = \{s \in R^m, s_j \geq 0, j = 1, \ldots, m, \sum_{j=1}^{m} s_j = 1\}.$$

L'estimateur du maximum de vraisemblance de la fonction de répartition F associée à la variable aléatoire T est donc obtenu par la résolution du problème d'optimisation sous contrainte suivant :

$$\max_{s \in D_s} l(s).$$

Une équation pour éstimer $s_j, j = 1, \ldots, m$ est donnée par :

$$\hat{s_j} = \frac{1}{N} \sum_{i=1}^{N} \frac{\alpha_{ij} s_j}{\sum_{l=1}^{m} \alpha_{il} s_l}, \quad j = 1, \ldots, m.$$

3.1.2 Algorithme de Turnbull

Turnbull (1976) fut l'un des premiers auteurs à s'intéresser au problème de l'estimation non paramétrique d'une fonction de survie à partir de données censurées par intervalle. L'algorithme qu'il a développé et que nous allons présenter maintenant reste à ce jour le plus populaire des algorithmes existants.

L'algorithme

- Etape 1 : On prend un vecteur de valeur initial $s = (s_1^{(0)}, s_2^{(0)}, \ldots, s_k^{(0)})$ (c'est un estimateur initial de $s = (s_1, s_2, \ldots, s_k)$).
- Etape 2 : On obtient un estimateur $s_k^{(1)}$ en prenant :

$$\frac{1}{n}\sum_{i=1}^{N}\frac{\alpha_{ij}s_j^{(0)}}{\sum_{l=1}^{k}\alpha_{il}s_l^{(0)}}, \quad j = 1, \ldots, k.$$

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

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

Cette procédure est facile et converge rapidement.

3.2 Comparaison de plusieurs fonctions du survies

La statistique de test qui va être introduite ici utilise toute l'estimé non paramétrique du maximum de vraisemblance de Turnbull de la fonction de survie. Afin d'alléger la présentation, nous parlerons simplement d'estimé de la fonction de survie pour désigner cet estimateur en particulier. On va présenter le test du log-rang pour données censurées par intervalle, décrit par Sun (1996).

3.2.1 Test de Logrank

Nous avons présenter ici la procédure de test développée par Sun (1996).

Un cas particulier

Supposons dans un premier temps que les intervalles de censure A_i sont disjoints.
Notons par ailleurs d_j le nombre de sujets qui connaissent l'événement dans l'intervalle I_j, et Y_j le nombre de ceux qui sont à risque dans ce même intervalle. Les sujets sont

dits "à risque" en un temps donné lorsqu'ils n'ont pas encore connu l'événement en ce temps. Supposons que l'échantillon des N sujets est divisé en G groupes que l'on veut comparer. Pour $g = 1, \ldots, G$, notons d_{jg}, et Y_{jg}, les quantités correspondant à d_j et Y_j pour le groupe g.

Soit l'hypothèse nulle H_0 suivant laquelle les distributions sous-jacentes du temps de survie dans chacun des G groupes sont identiques, i.e.

$$H_0 : S_1(.) = S_2(.) = \ldots = S_G(.),$$

la distribution des $d_{j1}, \ldots, d_{jG}$, conditionnellement aux échecs et à l'expérience de censure jusqu'en t_j, est alors la loi hypergéométrique

On déduit alors une statistique de type log-rang, qui s'écrit

$$R = (R_1, \ldots, R_g, \ldots, R_G)',$$

avec

$$R_g = \sum_{j=1}^{m} (d_{jg} - E(d_{jg}))$$

Sous H_0, et en notant V la matrice de variance-covariance du vecteur R, on aura :

$$R' V^{-1} R \rightsquigarrow \chi^2_{G-1}$$

On peut montrer en outre que, pour $g = 1, \ldots, G$, on aura

$$E(d_{jg}) = \frac{Y_{jg} \times d_j}{Y_j}$$

$$V(d_{jg}) = \frac{Y_{jg}(Y_j - Y_{jg}) d_j (Y_j - d_j)}{(Y_j^2)(Y_j - 1)}$$

$$cov(d_{jg1}, d_{jg2}) = -\frac{Y_{jg1} Y_{jg2} d_j (Y_j - d_j)}{Y_j^2 (Y_j - 1)}, g_1 \neq g_2;$$

Notons V la matrice de variance-covariance, de dimension G, associée au vecteur $(d_{jl}, \ldots, d_{jG})'$, dont les éléments se déduisent des équations ci-dessus. La matrice V se

déduit, sous H_0, c'est-à-dire en fait sous l'hypothèse d'indépendance des m tables de contingence pouvant être définies en chacun des m points du temps, de la façon suivante :

$$V = \sum_{j=1}^{m} V_j.$$

Nous avons supposé jusqu'ici une structure bien particulière des intervalles de censure, qu'il serait évidemment vain d'espérer retrouver en pratique.

Cas général

Soit $P = (P_1, \ldots, P_m)$ la fonction de survie commune aux G groupes sous H_0, i.e.

$$P_j = P(T > t_j), j = 0, 1, \ldots, m,$$

avec $P_0 = 1$ et $P_{m+1} = 0$.

Notons $\widehat{P} = (\widehat{P_1}, \ldots, \widehat{P_m})$ l'estimateur de P, obtenu par exemple par la méthode de Trunbull (1976). Définissons alors les quantités analogues aux d_j, n_j, d_{jg}, et Y_{jg}, introduites dans le cas particulier précédent, soit

$$d'_j = \sum_{i=1}^{N} \frac{\alpha_{ij}(\widehat{P}_{j-1} - \widehat{P}_j)}{\sum_{u=1}^{m+1} \alpha_{iu}(\widehat{P}_{u-1} - \widehat{P}_u)}$$

$$Y'_j = \sum_{r=j}^{m+1} \sum_{i=1}^{N} \frac{\alpha_{ir}(\widehat{P}_{r-1} - \widehat{P}_r)}{\sum_{u=1}^{m+1} \alpha_{iu}(\widehat{P}_{u-1} - \hat{P}_u)}$$

$$d'_{jg} = \sum_{i \in I_g} \frac{\alpha_{ij}(\widehat{P}_{j-1} - \widehat{P}_j)}{\sum_{u=1}^{m+1} \alpha_{iu}(\widehat{P}_{u-1} - \widehat{P}_u)}$$

$$Y'_{jg} = \sum_{r=j}^{m+1} \sum_{i \in I_g} \frac{\alpha_{ir}(\widehat{P}_{r-1} - \widehat{P}_r)}{\sum_{u=1}^{m+1} \alpha_{iu}(\widehat{P}_{u-1} - \widehat{P}_u)}$$

où l'on rappelle que I_g, est l'ensemble des indices des sujets appartenant au groupe g. Sous H_0, la statistique du log-rang s'écrit alors

$$R^* = (R_1^*, \ldots, R_g^*, \ldots, R_G^*)',$$

avec

$$R_g^* = \sum_{j=1}^{m} (d'_{jg} - \frac{n'_{jg} d'_j}{Y'_j})$$

et la distribution approximative sous H_0 est donnée par

$$R'^* V^{*-1} R^* \rightsquigarrow \chi^2_{G-1}$$

où V^* désigne la matrice de variance-covariance du vecteur T^*, et V^{*-1} son inverse généralisée.

Chapitre 4

Application

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

4.1 Application aux données de Freireich

Freireich, en 1963, a fait un essai thérapeutique ayant pour but de comparer les durées de rémission en semaines, de sujets atteint de leucémie selon qu'ils ont reçu ou non du 6-MP.

Durée de rémmission, en semaine, selon le traitement

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

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

L'estimateur empirique pour le groupe traité par un placebo (pas de censure) donne le

tableau suivant :

Semaine i	Nombre de rémissions à la semaine i	$\widehat{S}_{placebo}(t)$
0	21	1
1	19	19/21=0.90
2	17	17/21=0.81
3	16	0.76
4	14	0.66
5	12	0.57
8	8	0.38
11	6	0.26
12	4	0.19
15	3	0.14
17	2	0.09
22	1	0.05
23	0	0

L'estimateur de Kaplan-Meier de la fonction de survie S du groupe de 21 malades traité par le traitement 6-MP donne le tableau suivant :

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

Les calculs des statistiques de test peuvent être menés à partir du tableau suivant :

	6-MP		Placebo					
Durées	Y_{Ai}	d_{Ai}	Y_{Bi}	d_{Bi}	Y_i	d_i	$E(d_{Bi})$	$V(d_{Bi})$
1	21	0	21	2	42	2	1,00	0,49
2	21	0	19	2	40	2	0,95	0,49
3	21	0	17	1	38	1	0,45	0,25
4	21	0	16	2	37	2	0,86	0,48
5	21	0	14	2	35	2	0,80	0,47
6	21	3	12	0	33	3	1,09	0,65
7	17	1	12	0	29	1	0,41	0,24
8	16	0	12	4	28	4	1,71	0,87
10	15	1	8	0	23	1	0,35	0,23
11	13	0	8	2	21	2	0,76	0,45
12	12	0	6	2	18	2	0,67	0,42
13	12	1	4	0	16	1	0.25	0,19
15	11	0	4	1	15	1	0,27	0,20
16	11	1	3	0	14	1	0,21	0,17
17	10	0	3	1	13	1	0,23	0,18
22	7	1	2	1	9	2	0,44	0,30
23	6	1	1	1	7	2	0,29	0,20

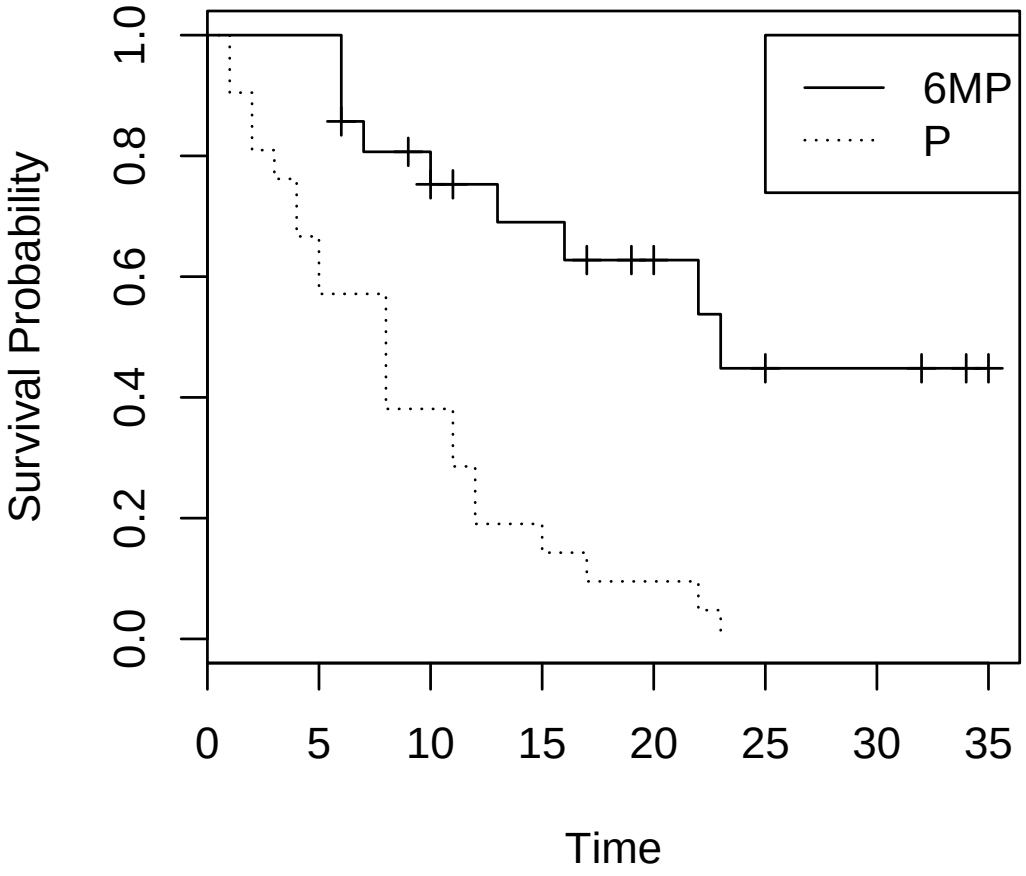

FIGURE 4.1 – Les deux courbes de survie

On obtient les résultats résumés ci-après :

	log-rank			Gehan		
Durées	Pondération	Coefficient	Variance	Pondération	Coefficient	Variance
1	1,00	1,00	0,49	42	42,00	860,49
2	1,00	1,05	0,49	40	42,00	777,54
3	1,00	0,55	0,25	38	21,00	357,54
4	1,00	1,14	0,48	37	42,00	653,33
5	1,00	1,20	0,47	35	42,00	570,71
6	1,00	-1,09	0,65	33	-36,00	708,75
7	1,00	-0,41	0,24	29	-12,00	204,00
8	1,00	2,29	0,87	28	64,00	682,67
10	1,00	-0,35	0,23	23	-8,00	120,00
11	1,00	1,24	0,45	21	26,00	197,60
12	1,00	1,33	0,42	18	24,00	135,53
13	1,00	-0,25	0,19	16	-4,00	48,00
15	1,00	0,73	0,20	15	11,00	44,00
16	1,00	-0,21	0,17	14	-3,00	33,00
17	1,00	0,77	0,18	13	10,00	30,00
22	1,00	0,56	0,30	9	5,00	24,50
23	1,00	0,71	0,20	7	5,00	10,00
		105,07	6,26		73441,00	5457,11

$$\chi_0^2 = 16,80 \qquad\qquad \chi_0^2 = 13,46$$

et si on compare ces chiffres aux valeurs critiques afférentes aux seuils de risque usuels de la distribution de khi-deux à un degré de liberté ($\chi^2 = 3.84$), on conclut que l'avantage du traitement est significatif.

Code R

Nous présentons ici le code des simulations informatiques relatives au chapitre 4.

Des outils ainsi que des données pour l'analyse des durées de vie sont disponibles dans les packages survival.

Les objets de survie sont crées au moyen de la fonction Surv(time, status) du package survival.

Pour créer des données censurées à droite, cette fonction a besoin de 2 arguments :

time : durée réellement observée.

status : indicatrice qui vaut 0 ou 1 (resp FALSE ou TRUE, resp 1 ou 2) selon que l'observation correspond à une censure ou non.

Estimateur de Kaplan-Meier

```
X=c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35,1,1,2,2,3,4,4,5,5,8,8,8,
8,11,11,12,12,15,17,22,23)
C=c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
t=c(rep("6MP",21),rep("P",21))
f=data.frame(X,C,t)
library(survival)
s=survfit(Surv(X,C) t, data=f)
plot(s, lty=c(1,3), xlab="Time", ylab="Survival Probability")
legend(10, 1.0, c("6MP", "P") , lty=c(1,3) )
```

Test de Logrank

ff(formula = Surv(X, C) t)

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
t=6MP	21	9	19.3	5.46	16.8
t=p	21	21	10.7	9.77	16.8

Bibliographie

[1] Gehan E.A. A generalized two-sample wilcoxon test for doubly censored data. Biometrica 1965, 52 : 650-653.

[2] kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958, 53 : 457-481.

[3] Kim J., Kang D. R. et Nam C. M. Logranktype tests for comparing survival curves with interval-censored data. Computational Statistics. Data Analysis (2006), 50 : 3165-3178.

[4] Nelson W. Theory and applications of hazard plotting for censored failure data. Technometrics 1972, 14 : 265-275.

[5] Peto R., Peto J . Asymptotically efficient rank invariant test procedures. J. R. Stat. Soc. A 1972, 135, 185-207.

[6] Philippe Saint Pierre. Introduction à l'analyse des durées de survie. Université Pièrre et Marie Curie, (2015).

[7] Sun J. A nonparametric test for interval-censored failure time data with application to AIDS studies. Statistics in Medicine 15 : 1387- 1395. (1996).

[8] Tarone R. E. etWare J. On distribution free tests for equality of survival distributios. Biometrika 64 : 156-160. (1977).

[9] 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).

[10] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin, Vol. 1, No. 6, pp. 80-83. 1945.

[11] Wilcoxon F. A simplified method of evaluating dose-effect experiments. Pharmacol Exp. Ther. 96 : 99-113. (1949).