République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Mathématiques

Mémoire présenté en vue de l'obtention du

DIPLÔME DE MASTER EN MATHÉMATIQES

Option : **Probabilités et Statistique**

Par

Younes Aicha Hibat Errahmane

Thème

Analyse des durées du chômage

Membres du Comité d'Examen

Pr.	Brahim Brahimi	UMKB	Président
Dr.	Louiza Soltane	UMKB	Encadreur
Dr.	Imen Benelmir	UMKB	Examinateur

2 Juin 2025

À mes chers Parents,

À mes soeurs, et mon seul frère,

À tous ceux que j'aime, et qui m'aiment.

REMERCIEMENTS

J e tiens à remercier avant tout ALLAH de m'avoir donné le courage d'achever ce travail, la volonté et la patience pour compléter mes études de licence et master.

Je remercie également Docteure Soltane Louiza pour l'aide et les conseils concernant les missions évoquées dans ce travail, qu'elle m'a apportés lors des différents suivis.

Mes vifs remerciements vont également aux membres du jury : Pr. Brahim Brahimi et Dr. Imen Benelmir pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Je remercie tous les enseignants qui ont contribué à notre formation, ainsi que tous les employés du département de mathématiques.

Je tiens aussi à remercier mes parents, qui trouvent en moi la source de leur fierté et qui n'ont cessé de me donner avec amour le nécessaire pour que je puisse arriver à ce que je suis aujourd'hui. Je remercie également mes amies pour leur soutien et leurs encouragements.

Enfin, je tiens à remercier toute personne qui m'a aidé de près ou de loin afin de réaliser ce travail.

Hibat Errahmane

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

L'analyse de survie est un domaine des statistiques qui s'intéresse à mesurer le temps jusqu'à évènement particulier, souvent appelé temps d'échec, ou temps de survie. L'analyse de survie est basée sur l'estimation de la fonction de survie

La méthode actuarielle, première méthode d'analyse de survie, est apparue en 1912. Sa première utilisation a eu lieu dans le domaine médical. Ensuite, l'analyse de survie est devenue une branche indispensable dans divers domaines, tels que l'actuariat ou l'étude des séismes. Elle est également appliquée dans d'autres champs, comme l'économie, l'assurance et la sociologie. L'étude de la durée du chômage, par exemple, constitue une application typique de l'analyse de survie. Elle permet notamment de déterminer combien de temps un individu reste au chômage.

La base de toute analyse statistique est l'échantillon auquel il arrive parfois d'être censuré. Il existe plusieurs mécanismes de censure dont la forme la plus couramment rencontrée est la censure aléatoire à droite.

En 1958, Kaplan et Meier présentent d'importants résultats concernant l'estimation non paramétrique de la fonction de survie.

Le mémoire est composé de deux chapitres :

- Le premier chapitre constitue le cadre théorique du mémoire. Il est composé de trois sections qui abordent : les bases de l'analyse de survie , la définition de la censure, notamment basée sur la censure à droite de type 3 aléatoire , les distributions de la durée de survie, ainsi que l'estimation non paramétrique (telle que les méthodes de Kaplan-Meier et Nelson-Aalen).
- Le second chapitre est le chapitre d'application qui est la présentation d'un cas pratique basé sur des données réelles et portant sur des individus au chômage.

CHAPITRE 1

Analyse de Survie et Données Censurées

Ddans ce chapitre, on va aborder toutes les méthodes statistiques qui vont être utilisées dans la partie pratique. Les fonctions empiriques de survie, de risque, ... etc, dans le cas où les données sont censurées, constituent la partie principale de ce travail. Ainsi, ce chapitre débute par quelques rappels et définitions couramment utilisées dans ces études. Ces définitions peuvent être trouvées en détail dans les ouvrages suivants Mohamed Elsherif [9], **Göran Broström** [3] et **Måns Thulin** [7].

1.1 Bases de l'Analyse de Survie

1.1.1 Analyse de survie

Définition 1.1.1 (Analyse de survie)

L'analyse de survie est une méthode statistique utilisée pour étudier la durée de vie d'un individu, d'un produit, d'un patient, ... etc. Elle permet de modéliser et d'analyser les données

de survie qu'elles sont sous forme de temps allant de l'origine de temps à la survenance d'un point final spécifique.

$$\textit{Début} \overset{temps}{\Longrightarrow} \textit{événement.}$$

1.1.2 Temps de survie

Définition 1.1.2 (Temps de survie)

Un temps de survie est défini comme le temps nécessaire à la survenue d'un événement, mesuré à partir d'un événement de départ bien défini.

Il y a trois éléments de base doivent être bien définis pour préciser le temps de survie : date d'origine, date de point et date des dernières nouvelles, elles sont présentées dans la **Figure 1.1**, qu'on abordera en détail comme suit :

1. Date d'origine : C'est l'origine du début de l'analyse de survie, elle correspond du temps égal à zéro, elle peut être date de diagnostic, date de début de traitement, date de naissance d'un individu, date d'entrée en service, date d'une opération chirurgical. Chaque individu a une date d'origine différente, ce qui n'est pas important car c'est la durée qui nous interesse.
2. Date de point : C'est une date au-dela de laquelle nous arrêtons l'observation et nous tenons plus compte de l'état du sujet après cet instant. La date de point peut être : date de fin de traitement, date de fin de survie, date de décès, date de dernière observation.
3. Date des dernières nouvelles : Pour faire l'analyse des résultats, chaque individu dispose d'une date des dernières nouvelles, qui est la date la plus récente où les informations ont été recueillies, cela peut être la date de survenue de l'évènement étudié.
4. Recul : C'est le délai écoulé entre la date d'origine et la date de point c'est à dire le délai maximal d'observation du sujet.Il est dit aussi le délai de censure.
5. Temps de participation : C'est le temps écoulé entre :

- la date d'origine et la date de dernières nouvelles, si cette dernière est antérieure à la date du point (décés d'avant la date de point).
- la date d'origine et la date de point, si celle-ci est antérieure à la date de dernières nouvelles (vivant aux dernières nouvelles). (source : M'ziou Imane [10]).

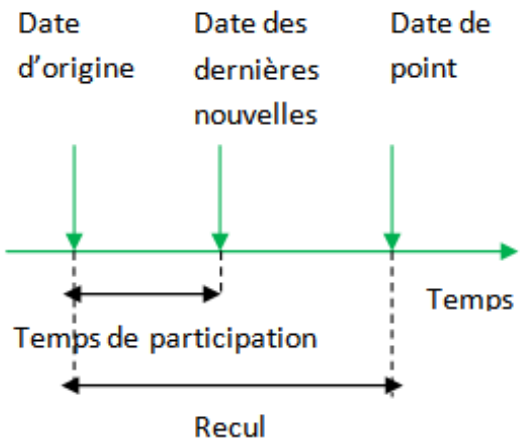


FIG. 1.1 – Schéma représentant les principales définitions relatives à l'analyse de la durée de survie.

1.1.3 Données de survie

Définition 1.1.3 (Données de survie)

Les données de survie (Survival Data) sont des données qui contiennent des informations sur la durée de vie d'un individu, d'un produit ou d'un système.Ces données mesurent le temps jusqu'à un évènement d'intêret, ce temps est connu par le temps de survie qui est toujours une variable réelle des variables positives.

Ici quelques exemples sur ces données :

Exemple 1.1.1 (En médicale)

Dans la recherche médicale l'origine du temps est le point de référence à partir duquel on mesure dans une étude, comme les essais médicaux pour comparer des défférents de groupe de patient ou médicament (tester l'éficacité d'un médicament), si le point final est ***la mort*** *donc les données résultantes sont les* ***temps de survie*** *(Survival Time), mais si le point final n'est pas la mort, il peut s'agir de tout autre événement, alors dans ce cas les données sont appelées données de temps sur évènement (Time to Event Data).*

- *Survie d'un individu après l'apparition des maladies grave.*
- *Temps de rémission après opération chirurgicale.*
- *Âge de décès chez des patients atteints de diabète...*

Exemple 1.1.2 (D'autres domaines)

- *Industrie : étudier la durée de vie des produits et des équipements.*
- *Sciences sociales : étude du temps jusqu'à des évènements comme le mariage, ...*
- *Finance : étudier la survie des entreprises et des investisments.*
- *Economie : durées des périodes de travail ou de chômage, temps avant faillite,...*
- *Assurance : durée de cotisation avant le premier remboursement.*

1.2 Censure

dans l'analyse de survie, la censure est un phénomène courant où les données ne sont pas toujours complètement observées, parce que pour certains individus de l'événement du début et /ou de fin n'est pas observé (absence d'événement), c'est-à-dire privées d'une partie de l'information (données incomplètes). Elle se produit lorsque l'observation d'un sujet est interrompue, que ce soit parce qu'il est perdu de vue, se retire de l'étude, ou que l'étude se termine avant la survenue de l'événement d'intérêt (décès, récidive, etc.). Dans ce cas on parle alors des données censurées (censored data). Ces données, bien qu'incomplètes, fournissent

une information précieuse : les sujets censurés sont considérés comme n'ayant pas connu l'événement pendant la période où ils ont été suivis. L'explication de ce dernier paragraphe, est une synthèse inspirée des travaux de (Mohamed Elsherif, 2021) et (Heddar, 2022). L'exemple suivant est cité du livre de Mohamed Elsherif [9].

Exemple 1.2.1

Soit une étude clinique réalisée sur une durée d'un an, comme illustrée dans la figure ci-dessous.

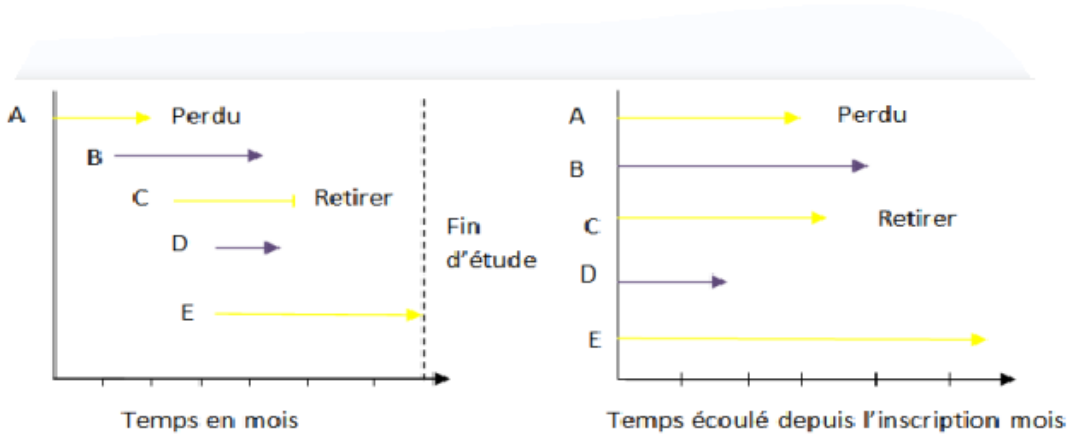


FIG. 1.2 – Schéma représentant les cas de la censure

- *Patient A : a été perdu de vue après* 3 *mois, sans évènement.*
- *Patient B : a eu un événement* $4,5$ *mois après son inscription.*
- *Patient C : a retiré de l'étude* $3,5$ *mois après son inscription.*
- *Patient D : a vécu un événement* 2 *mois après son inscription.*
- *Patient E : n'a eu aucun événement avant la fin de l'étude (il a été suivi pendant* 6 *mois).*

La durée exacte d'un évènement ne pouvait être enregistré que pour les patients B et D (leurs données ne sont pas censurées). Pour les autres patients, on ignore s'ils ont ou n'ont pas

présenté un événement après la fin de l'étude. La seule information valable disponible pour les patients A, C et E est qu'ils n'ont présenté aucun événement jusqu'à leur dernier suivi. Par conséquent, leurs données sont censurées.

Plusieurs types de données censurées sont discutés dans la littérature, y compris dans les références suivantes [2] et [6]**.** Cependant, dans ce travail, On va concentrer uniquement sur les données censurées à droite du type 3 aléatoire.

Définition 1.2.1 (Variable censure)

la variable de censure C est définie par la non-observation de l'évenement étudie si au lieu d'observer X, on observe C, et quel' on sait que $X > C$ (réspectivement $X < C, C_1 < X < C_2$), on dit qu'il ya censure à droite (réspectivement censure à gauche, censure par intervalle)

Mathématiques, soit $X_1, ..., X_n$ un échantillon d'une va positive X, on dit qu'il y a censure aléatoire de cet échantillon s'il existe une autre va positive elle aussi C d'échantillon $C_1, ..., C_n$. Pour un individu donné i, on va considérer

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

Dans ce cas au lieu d'observer les $X_i's$, on observe un couple de va's (T_i, δ_i) avec

$$T_i := \min(X_i, C_i) \quad \text{et} \quad \delta_i := \mathbb{1}\{X_i \leq C_i\} \quad \text{pour } i = 1, ...n, \tag{1.1}$$

où δ_i l'indicateur de censure, qui détermine si X a été censuré ou non :

- si $\delta_i = 1$, la durée d'intérêt est observée $(T_i = X_i)$.
- si $\delta_i = 0$, elle est censurée $(T_i = C_i)$. On observe des durées incomplètes.

Supposons que la durée de survie X soit une variable positive ou nulle et absolument continue, alors la distribution de X peut être définie par l'une des fonctions suivantes :

1.2.1 Fonction de répartition

Définition 1.2.2 (Fonction de répartition)

La fonction de répartition représente pour t fixé, la probabilité de mourir avant l'instant t, c'est à dire :

$$F(t) := P(X \leq t). \tag{1.2}$$

Propriété 1.2.1

La fonction de répartition F est une fonction croissante, monotone et continue à droite telle que :

$$\lim_{t \longrightarrow 0} F(t) = 0 \quad et \quad \lim_{t \longrightarrow \infty} F(t) = 1.$$

1.2.2 Fonction de survie

Définition 1.2.3 (Fonction de survie)

La fonction de survie, qui est notée par ($S(t)$ ou $\overline{F}(t) := 1 - F(t)$), est pour t fixée c'est la probabilité de survivre jusqu'à l'instant t, c'est à dire pour $t \geq 0$ on a :

$$S(t) := P(X > t), \tag{1.3}$$

où X la durée de vie (aléatoire) étudiée.

Propriété 1.2.2

$S(t)$ est une fonction monotone, décroissante et continue avec :

$$\lim_{t \longrightarrow \infty} S(t) = 0 \quad et \quad S(0) = 1.$$

la fonction $S(t)$ est définie sur $\mathbb{R}_+$.

1.2.3 Fonction de densité

Définition 1.2.4 (Fonction de densité)

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

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

Remarque 1.2.1

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

$$\begin{aligned} f(t) &= \lim_{dt \longrightarrow 0} \frac{P(t \leq X < t + dt)}{dt} \\ &= \lim_{dt \longrightarrow 0} \frac{F(t + dt) - F(t)}{dt}. \end{aligned}$$

D'autre façon :

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t). \tag{1.4}$$

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 Fonctions de risque et de risque cumulé

Définition 1.2.5 (Fonction de risque)

La fonction de risque notée par $h(t)$ est définie par :

$$\begin{aligned} h(t) &= \lim_{dx \longrightarrow 0} \frac{P(t \leq X < t + dx / X \geq t)}{dx} \\ &= \lim \frac{S(t) - S(t + dx)}{dx S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1.5}$$

Définition 1.2.6 (Fonction de risque cumulé)

La fonction de risque cumulé notée par $H(t)$, c'est l'intégrale de fonction de risque.

$$H(t) = \int_0^t h(x)dx = \int_0^t \frac{dF(x)}{S(x)}dx = -\ln(S(t)). \tag{1.6}$$

Remarque 1.2.2

On peut déduire de cette équation une expression de la fonction de survie en fonction de taux de hazard cumulé

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u)du\right). \tag{1.7}$$

Remarque 1.2.3

La fonction de risque cumulé est importante car elle est relativement facile à estimer de manière non paramétrique.

1.2.5 Fonctions empiriques de répartition et de survie

Soit $X_1, ..., X_n$ un échantillon de variable iid définies sur un espace de probabilité (Ω, A, P), à valeurs dans $\mathbb{R}$, avec pour fonction de répartition F et fonction de survie S. Les fonctions empiriques de répartition et de survie F_n et S_n sont respectivement définies par :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}, \forall t \geq 0, \tag{1.8}$$

et

$$S_n(t) = 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > t\}}, \quad \forall t \geq 0, \tag{1.9}$$

où $\mathbb{1}\{A\}$ est la fonction indicatrice de l'évènement A.

Les fonctions précédentes s'écrivent en termes des valeurs de statistiques d'ordre[1] comme

[1] Les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$, sont obtenues en classant ces va's par ordre croissant $X_{1:n} \leq ... \leq X_{n:n}$

suit :

$$F_n(t) = \begin{cases} 0 & \text{si} \quad t < X_{1:n}, \\ \dfrac{i}{n} & \text{si} \quad X_{i:n} \leq t < X_{i+1:n}, \\ 1 & \text{si} \quad t \geq X_{n:n}, \end{cases} \quad \text{et} \quad S_n(t) = \begin{cases} 1 & \text{si} \quad t < X_{1:n}, \\ 1 - \dfrac{i}{n} & \text{si} \quad X_{i:n} \leq t < X_{i+1:n}, \\ 0 & \text{si} \quad t \geq X_{n:n}, \end{cases} .$$

1.3 Méthodes non Paramétriques

Les méthodes non paramétriques sont des méthodes très faciles et simples à comprendre par rapport aux méthodes paramétriques.

Les principales estimations sont :

- Estimation de Kaplan-Meier pour la fonction de survie (1.3).
- Estimation de Nelson-Aalen pour le risque cumulé (1.6).

1.3.1 Méthode de Kaplan-Meier

L'estimation de Kaplan-Meier est une méthode utilisée pour estimer la fonction de survie à partir des données censurées, cette estimation découle l'idée suivante : survivre après un temps t, c'est être en vie juste avant t et ne pas mourir au temps t, soient $0 < t^{''} < t^{'} < t$, on a :

$$\begin{aligned} S(X) &= P(X > t) \\ &= P(X > t^{'}, X > t) \\ &= P(X > t/X > t^{'}) \times P(X > t^{'}) \\ &= P(X > t/X > t^{'}) \times P(X > t^{'}/X > t^{''}) \times P(X > t^{''}). \end{aligned}$$

En concidérant les temps d'évènement (décès et censure) distincts T_i tel que T_i, $i = 1, ..., n$

rangés par ordre croissant, on obtient :

$$P(X > T_i) = \prod_{k=1}^{i} P(X > T_k / X > T_{k-1}) \text{ avec } T_0 = 0.$$

Concidérant les notations suivantes :

- n_i le nombre d'individus à risque de subir l'évènement juste avant le temps T_i.
- d_i le nombre de décès en T_i.

Alors la probabilité p_i de mourir dans l'intervalle $\left]T_{i-1}, T_i\right]$, sachant que l'on était vivant en T_{i-1} c'est à dire, $p_i = P(X \leq T_i / X > T_{i-1})$ estimé par : $\widehat{p_i} = \dfrac{d_i}{n_i}$. Comme les temps d'évènement sont supposés distincts, on a $d_i = 0$ en cas de censure en T_i quand $\delta_i = 0$, $d_i = 1$ en cas de décès en T_i quand $\delta_i = 1$. On obtient alors l'estimation de Kaplan-Meier :

$$S_n^{KM}(t) = \overline{F}_n^{KM}(t) = \prod_{\substack{i=1,...,n \\ T_i \leq t}} \left(1 - \frac{d_i}{n_i}\right) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{n_i}\right) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{n - i + 1}\right), \quad (1.10)$$

où $T_1 \leq ... \leq T_n$ sont les statistiques d'ordre associées à $T_1, ..., T_n$.

Remarque 1.3.1

Il existe d'autre forme pour l'estimateur de Kaplan-Meier :

$$S_n^{KM}(t) = \prod_{T_i \leq t} \left(\frac{n - i}{n - i + 1}\right)^{\delta_i} = \prod_{i=1}^{n} \left(1 - \frac{\delta_i}{n - i + 1}\right)^{1_{\{T_i \leq t\}}}, \ i = 1, ..., n.$$

$S^{KM}(t)$ est une fonction décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car le temps de décès ne sont pas connus).

Remarque 1.3.2

Dans le cas où il y a des ex-aequo :

1. *Si ce sont des évènements de nature différente,on considère que les observations non censurées ont lieu avant les censurées.*
2. *S' il y a plusieurs décès au même temps T, alors $d_i > 1$ et on a :*

$$S_n^{KM}(t) = \prod_{i=1}^{n} \left(1 - \frac{d_i}{n_i}\right) = \prod_{i=1}^{n} \widehat{p}_i.$$

1.3.2 Méthode de Nelson-Aalen

L'estimateur de Nelson-Aalen est une méthode statistique non paramétrique utilisée pour estimer la fonction de risque cumulé à partir des données censurées, il est introduit par Nelson en (1972) et généralisé par Aalen en (1978). Soit $N(t) = P(T > t, \delta = 0)$ et $N_1(t) = P(T > t, \delta = 1)$ et introduisant $G(t)$ la fonction de survie de la variable C.

D'après l'hypothèse générale d'ndipendance, on obtient :

$$\begin{aligned} N(t) &= P(T > t, C > t) = S(t)G(t) \\ N_1(t) &= P(T > t, \delta = 1) = P(X > t, C \geq X) \\ &= \int_t^{\infty} \overline{G}(u) f(u) du = -\int_t^{\infty} \overline{G}(u) S(du). \end{aligned}$$

Par conséquent, $N_1(dt) = \overline{G}(t)S(dt)$ et on obtient l'expression suivante pour le risque cumulé :

$$H(t) = -\int_0^t \frac{N_1(du)}{\overline{N}(u)}.$$

Définition 1.3.1 (Estimateur de Nelson-Aalen)

L'estimateur de Nelson-Aalen $\widehat{H}(t)$ de H basé sur l'échantillon $\{(T_i, \delta_i)\}, 1 \leq i \leq n\}$ donné

par :

$$\widehat{H}(t) = -\int_0^t \frac{\widehat{N}_1(du)}{\overline{N}(u)} = \begin{cases} \sum_{T_{i.n}\leq t}^{n} \frac{\delta_{[i.n]}}{n-i+1} & si \quad t < T_{i.n} \\ 1 & si \quad t \geq T_{i.n}, \end{cases}$$

$$= \sum_{i:T_i\leq t} \frac{d_i}{n_i},$$

où $\widehat{N}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i>u\}}$ *et* $\widehat{N}_1(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i>u,\delta_i=1\}}$ *et* n_i *représente le nombre d'individus à risque juste avant* T_i *et* d_i *représente le nombre de décès en* T_i*.*

Se référer à l'article de Meraghni et al. **[8].**

Remarque 1.3.3

L'estimateur de risque cumulé peut être obtenu à partir de l'estimateur de Kaplan-Meier en utilisant la relation (1.6) *:*

$$\widehat{H}(t) = \ln(\widehat{S}(t))$$

$$= -\sum_{i:T_i\leq t} \ln\left(1 - \frac{d_i}{n_i}\right),$$

cette écriture est l'écriture de Breslow de risque cumulé.

Remarque 1.3.4

A partir de la relation (1.7) *et de l'estimateur de Nelson-Meier, on peut en déduire un autre*

estimateur de la fonction de survie :

$$\begin{aligned}\widehat{S}(t) &= \exp(-\widehat{H}(t)) \\ &= \prod_{i:T_i \leq t}^{n} \exp\left(-\frac{d_i}{n_i}\right) \\ &\approx \prod_{i:T_i \leq t}^{n} \left(1 - \frac{d_i}{n_i}\right), \ si \ \frac{d_i}{n_i} \to 0,\end{aligned}$$

cette écriture est l'écriture de Harrington et Fleming.

CHAPITRE 2

Exemple d'Application

DDans ce chapitre, on applique quelques unes des méthodes vues aux Chapter 1 sur un exemple de données réelles. Les résultats numériques et les représentations graphiques sont obtenus à l'aide du package "summarytools", "explore", "survival", "survminer", et "tidyverse" du logiciel d'analyse statistique RStudio.

2.1 Aperçu des Données

Pour illustrer notre procédure d'estimation, on analyse un jeu de données réelles sur la durée du chômage aux États-Unis, disponible dans l'article d'Ani Katchova (2013) intitulé "Survival Analysis Example" [5]. Un sous ensemble de données est présenté dans le Table 2.1. Ces données proviennent des "Displaced Workers Supplements-DWS" de l'enquête sur la population actuelle (Current Population Survey-CPS) de janvier pour les années 1986, 1988, 1990 et 1992. Elles permettent de suivre la durée nécessaire aux individus pour trouver un emploi à plein temps après une période de chômage. Ces données concernent 3343 individus et 43

variables numériques (total des observations 143749). L'ensemble de données se concentre principalement sur deux variables dépendantes :

- "*spell*" durée : nombre de périodes pendant lesquelles un individu reste au chômage en semaine.
- "*event*" l'événement correspondant : indiquant si un individu a trouvé un emploi = 1 ou toujours au chômage = 0 (autrement dit, les données ont été censurées).
- *censor2, censor3, censor4* : Types de censure. Ce sont des variables binaires (0 ou 1) qui indiquent probablement différentes raisons pour lesquelles la censure a pris fin sans que l'événement (trouver un emploi) ne se produise. On distingue selon l'arbre de censure (voir la Figure 2.5)
  - Cas1 : $\delta_i = 0 \Longrightarrow censor2 = 0, censor3 = 0, censor4 = 0$. Il y a une censure à droite, lorsque la période du chômage était toujours en cours à la fin d'étude.
  - Cas2 : $\delta_i = 0 \Longrightarrow censor2 = 1, censor3 = 1, censor4 = 1$.

Les autres variables sont des variables indépendantes résumées dans l'Annex C 2.3.

spell t	*event* δ_i	*cens.*2	*cens.*3	*cens.*4	*iu*	*rep.*	*logw.*	*ten.*	*mar.*	*fem.*	*chi.*	*age*	...
05	1	0	0	0	0	0.18	6.89	03	1	0	1	43	...
13	1	0	0	0	1	0.53	5.29	06	1	0	1	24	...
21	1	0	0	0	1	0.21	6.77	01	1	0	1	32	...
03	1	0	0	0	1	0.45	5.98	03	1	0	1	35	...
09	0	0	1	0	1	0.32	6.32	0	1	1	0	31	...
11	0	0	0	1	1	0.19	6.85	09	1	0	1	26	...
01	0	0	0	0	0	0.52	5.61	01	1	0	0	49	...
03	1	0	0	0	0	0.37	6.16	0	0	1	0	39	...
07	1	0	0	0	1	0.52	5.29	02	0	1	0	40	...
05	0	0	0	1	1	0.52	5.29	01	1	0	1	20	...
07	0	0	0	1	1	0.52	5.76	02	1	0	0	39	...
15	0	0	0	1	1	0.22	6.67	21	0	0	0	38	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

TAB. 2.1 – Sous ensemble des données du chômage

2.2 Analyse Univariée

Dans cette section, on va faire une analyse d' ensembles de données "pour résumer les principales caractéristiques des variables, souvent à l'aide de tableaux de statistiques descriptives et de graphiques pour la visualisation des variables individuelles.

Ici, on va utiliser l'approche descriptive pour décrire les données d'étude. Cette approche représente une étape nécessaire dans l'analyse et l'interprétation des données. Pour cela, à l'aide du Table 2.2 et de la **Figure 2.1**, on donne l'interprétation des résultats en général des variables d'analyse de survie (*"spell"* et *"event"*) :

Valid	*Missing*	*Min*	*Q1*	*Mean (sd)*	*Median*	*Q3*	*Max*	*Outliers*	*Skewness*
3343	0	1	2	6.28 (5.61)	5	3	28	130	1.52

TAB. 2.2 – Statistiques descriptives pour les variables d'analyse de survie

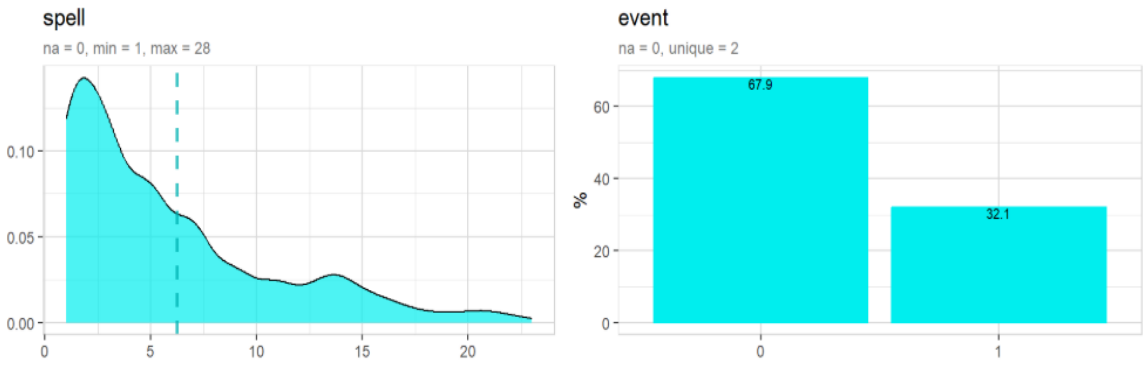


FIG. 2.1 – Densité et diagrammes en barres pour les variables d'analyse de survie

- La durée moyenne de chômage est 6.28 semaines avec un écart-type 5.6. Cet écart-type, relativement élevé par rapport à la moyenne, indique une grande variation de la durée de chômage entre les individus.
- On constate que la durée minimale de chômage enregistrée est d'une semaine, alors que la durée maximale de chômage enregistrée est de 28 semaines. Cela signifie qu'il y a 28 valeurs distinctes.

- La majorité des individus trouvent un emploi relativement rapidement ($Median = 5$ semaine), mais il existe un groupe d'individus qui mettent beaucoup plus de temps à trouver un emploi, ce qui tire la moyenne (6.28 semaine) plus élevée que la médiane et contribue à la forte dispersion des données.
- Le panneau gauche de la Figure 2.1 montre la distribution de densité de la durée du chômage. La plupart des individus ont de courtes périodes de chômage (les valeurs se regroupent à l'extrémité gauche du graphique), et quelques individus ont de très longues périodes de chômage (avec une longue queue s'étendant vers la droite).
- Selon le panneau droit de la Figure 2.1 montre que 32% (1073) des individus de l'échantillon ont retrouvé un emploi a tandis que 68% (2270) restants au chômage, au cours de la période d'étude.

À partir des résultats précédents et du graphe Q-Qplot, on peux conclure que la normalité de variable numérique *"spell"* dont est issu l'échantillon ne suit pas loi normale. Les statistiques descriptives pour les autres variables sont présentées en détail dans l'Annex C 2.3.

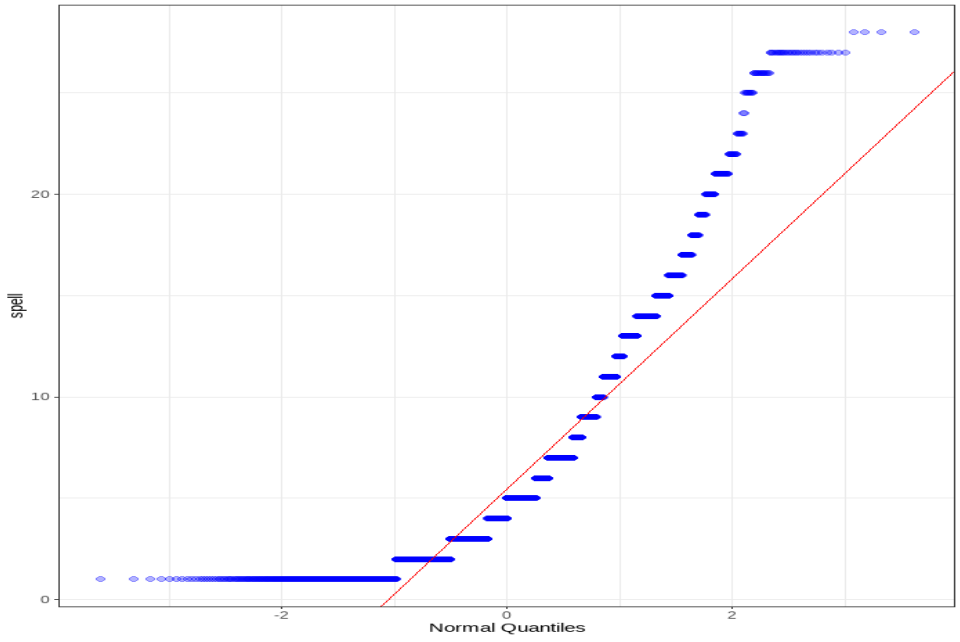


FIG. 2.2 – Q-Qplot pour la variable numérique *"durée"*

2.3 Méthodes non Paramétriques

Dans cette Section, on va utiliser une méthode non paramétrique pour l'analyse de survie des données de chômage, à l'aide du tableau de survie (Life Table) Table 2.3. Ce dernier fournit une estimation détaillée de l'évolution de la probabilité d'un individu de rester au chômage au fil du temps.

La relation de l'estimateur (1.10) permet de calculer l'estimation de la probabilité de survie par la méthode de Kaplan Meier au Table 2.1. Les résultats obtenus sont donnés dans le Table 2.3, où la $1^{ère}$ colonne contient les 27 cas réellement observées, parmi les 3343 individus qui sont analysés. La $2^{ème}$ colonne indique le nombre des individus n_i à risque sur l'intervalle de temps écoulé. Le nombre d'évènements observé d_i est indiqué dans la $3^{ème}$ colonne. Dans la $4^{ème}$ et $5^{ème}$ colonne on donne l'estimation de KM pour la fonction de survie $\mathbf{S}_n^{KM}(t)$ et l'erreur standard (err.std**)**, où la dernière colonne contient, pour chaque instant t_i (Temps), l'intervalle d'estimation à 95%. La représentation graphique associée au Table 2.3 est présentée dans la Figure 2.3, où la courbe de survie estimée pour les 3343 données de individus en escaliers décroissants (avec les bornes de confiance à 95%), sur lequel on constate que la plus grande des valeurs non censurées pour estimation $\mathbf{S}_n^{KM}(27) = 0.31 \neq 0$. Ceci témoigne de l'existence de données censurées au delà de 27. On peut conclure du tableau :

- Qu'une grande proportion trouve un emploi dans les premières semaines, et cela est dû au fait que l'estimation de la probabilité de survie diminue relativement rapidement au début.
- Il montre également la croissante de l'erreur standard dans les estimations de $\mathbf{S}_n^{KM}(t)$ pour les durées plus longues en raison de la diminution de la taille effective de l'échantillon.

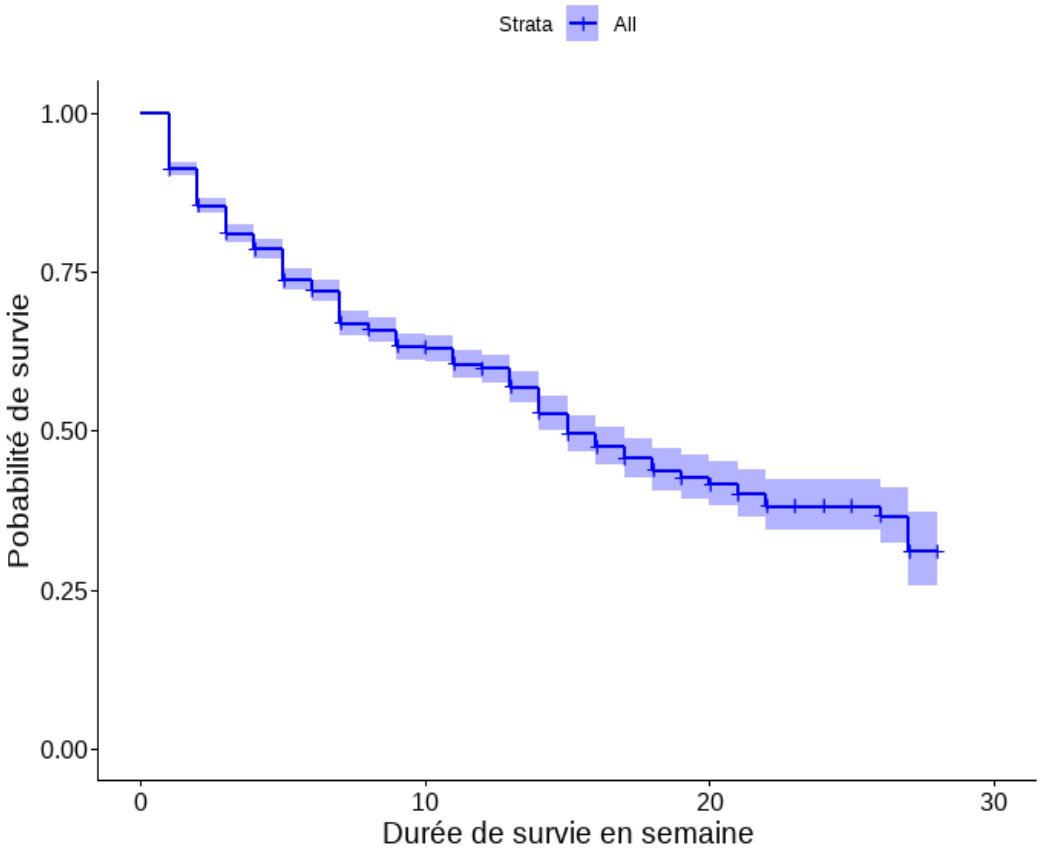


FIG. 2.3 – Courbe de survie estimée pour les 3343 données du chômage.

Durée	*À risque*	*n.évén.*	*n.censuré*	*pro.survie*	*err.std.*	95% *CI*	
semaine (t_i)	n_i	d_i		$\mathbf{S}_n^{KM}(t)$		*Lower*	*Upper*
1	3343	294	246	0.912	0.005	0.903	0.922
2	2803	178	304	0.850	0.007	0.842	0.866
3	2321	119	305	0.810	0.009	0.797	0.824
4	1897	56	165	0.786	0.010	0.772	0.801
5	1676	104	233	0.738	0.010	0.721	0.754
6	1339	32	111	0.720	0.010	0.703	0.737
7	1196	85	178	0.669	0.010	0.650	0.688
8	933	15	70	0.658	0.020	0.639	0.678
9	848	33	98	0.632	0.020	0.612	0.654
10	717	3	55	0.630	0.020	0.609	0.651
11	659	26	77	0.605	0.020	0.583	0.627
12	556	7	40	0.597	0.020	0.575	0.620
13	509	25	69	0.568	0.020	0.544	0.593
14	415	30	74	0.527	0.030	0.501	0.554
15	311	19	40	0.495	0.030	0.467	0.524
16	252	10	41	0.475	0.030	0.446	0.506
17	201	8	24	0.456	0.040	0.426	0.489
18	169	7	13	0.437	0.040	0.405	0.472
19	149	4	15	0.426	0.040	0.393	0.461
20	130	3	18	0.416	0.040	0.382	0.452
21	109	4	23	0.400	0.050	0.365	0.439
22	82	4	9	0.381	0.050	0.343	0.423
23	69	0	9	0.381	0.050	0.343	0.423
24	60	0	2	0.381	0.050	0.343	0.423
25	58	0	10	0.381	0.050	0.343	0.423
26	48	2	13	0.365	0.060	0.324	0.412
27	33	5	24	0.310	0.1	0.257	0.374

TAB. 2.3 – Tableau de survie pour les données du chômage

Conclusion

L'objectif de mon travail est d'étudier la durée de chômage à travers l'estimation de la fonction de survie.

- Notre analyse indique que la majorité des individus trouvent un emploi avec succès dans un délai relativement court après avoir commencé leur recherche.
- Ce résultat confirme notre observation selon laquelle la probabilité de survie au chômage diminue fortement au cours des premières semaines.

Cette étude s'est concentrée sur les variables d'analyse de survie et a suivi sa propre méthodologie. Cependant, des limitations liées au fait de ne pas analyser l'effet d'autres variables, en plus de noter la présence des valeurs extrêmes dans les données. Il serait intéressant de faire le même travail, car ce cas mérite d'être considéré attentivement.

BIBLIOGRAPHIE

[1] Ben Kouider, L. (2013). Mémoire de master. Introduction à l'analyse de survie. Centre Universitaire, Mila.

[2] Elisa, T.and John Wenyu Wang (2003). Statistical methods for survival analysis using R. New York, NY : Springer.

[3] Göran Broström, (2021), Event History Analysis with R. Second Edition. .

[4] Heddar, C. (2022). Mémoire de master. Sur l'estimation de la fonction de survie. Université Mohamed Khider, Biskra.

[5] Katchova, A. (2013). Survival Analysis Example. Retrieved from Econometrics Academy website.

[6] Klein, M. (2003). Survival analysis techniques for censored

[7] Måns Thulin, (2025), Modern Statistics with R, From wrangling and exploring data to inference and predictive modelling. Second edition. Chapman & Hall/CRC Press. ISBN 9781032512440.

[8] Meraghni, D., Necir, A., and Soltane, L. (2025). Nelson-Aalen tail product-limit process and extreme value index estimation under random censorship. Sankhya A, $1 - 49$.

[9] Mohamed Elsherif, (2021), Applied Medical Statistics for Beginners. Second edition.

[10] M'ziou, I. (2014). Mémoire de master. Estimation non paramétrique, Université Mohamed Khider, Biskra.

Annexe A : Logiciel *R*

R est un système, communément appelé langage de programmation statistique, qui offre plusieurs fonctionnalités, notamment :

- Il permet de réaliser des analyses statistiques.
- Vaste écosystème de packages "CRAN". Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, et les différentes méthodes d'analyse des données.
- Il comporte des moyens qui rendent possible la manipulation des données, ainsi que les calculs et les représentations graphiques ("dplyr", "tidyverse", "ggplot2", ...)
- R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers texte, des pages web, ...
- Il dispose d'une interface utilisateur, RStudio, qui facilite le travail pour tout utilisateur.
- C'est un langage gratuit, open-source et interactif avec une grande communauté active de développeurs.

Tout utilisateur peut installer R et RStudio depuis le site web suivant posit.

Annexe B : Abréviations et Notations

Les différentes abréVariations et notation utulisées tout au long de cette thèse sont expliquées ci-dessous.

S ou $\overline{F}$	:	Fonction de survie.
S_n	:	Fonction de survie empirique.
F	:	Fonction de répartition.
F_n	:	Fonction de répartition empirique.
H	:	Fonction de risque cumulé.
S_n^{KM}	:	Estimateur de Kaplan-Meier.
$\widehat{H}$	:	Estimateur de Nelson-Aalen.
iid	:	Indépendantes et identiquement distribué.
$\mathbb{1}_A$	:	Fonction indicatrice de l'ensemble A.
$X_{1:n}, ..., X_{n:n}$	:	Statistique d'ordre associées à $X_1, ..., X_n$.
$\mathbb{R}$	:	Ensemble des nombres réels.
KM	:	Kaplan-Meier.
err-std	:	Erreur standard.
IC	:	Intervalle de confiance.

Annexe C : Résumant des Variables

Une image et un tableau qui montrent les noms et l'organisation des variables avec une explication simple pour chaque variable.

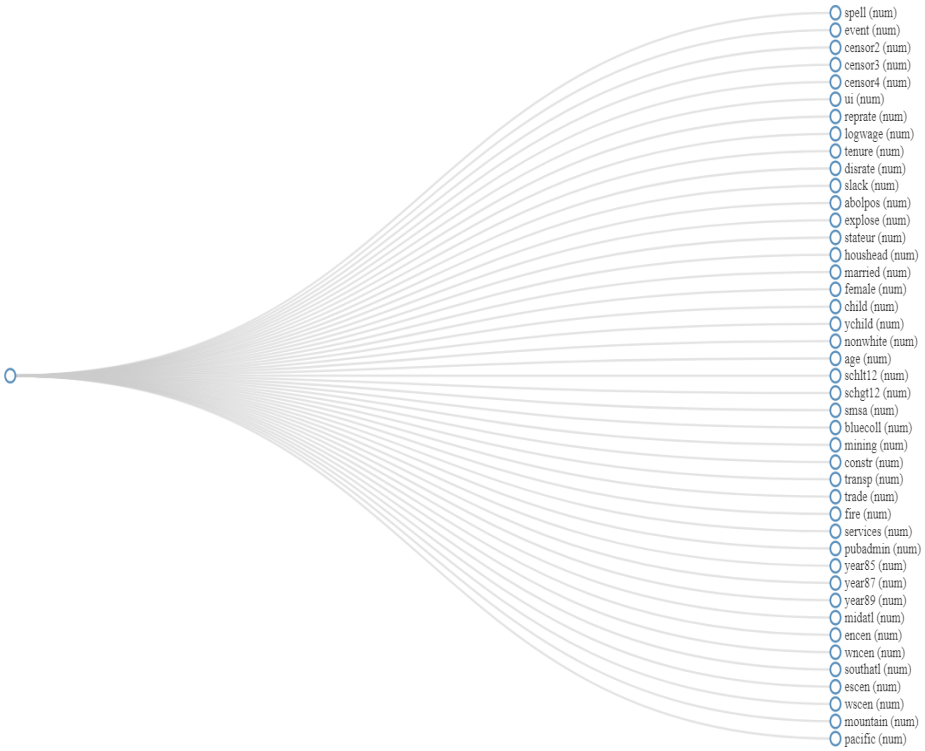


FIG. 2.4 – Noms des variables

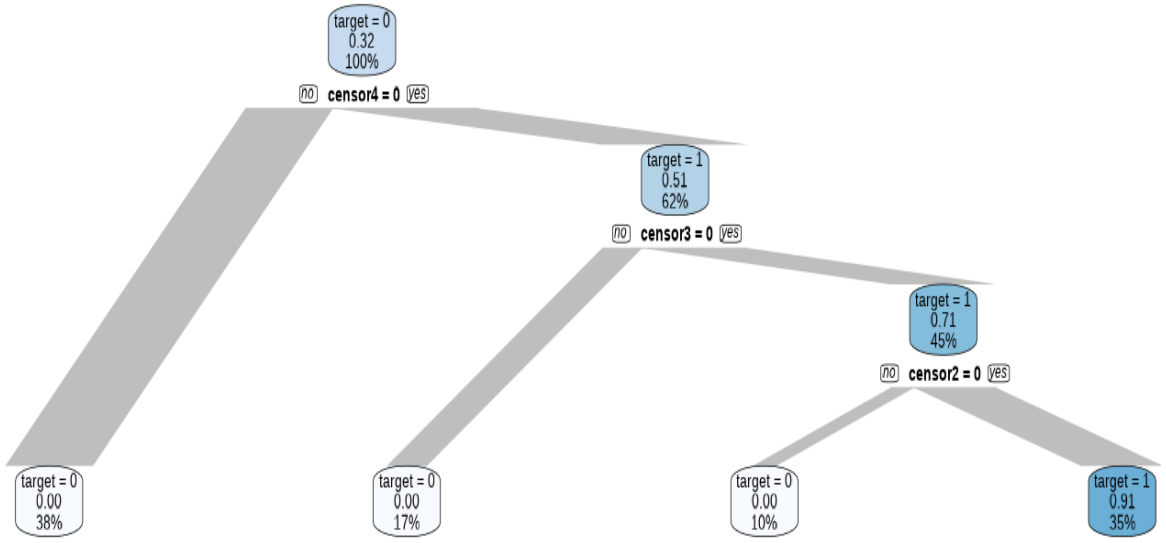
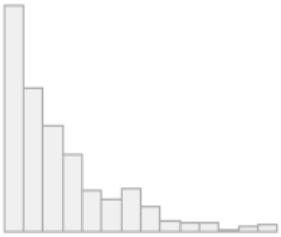
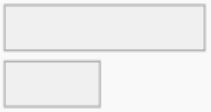
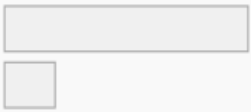
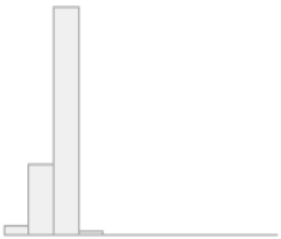


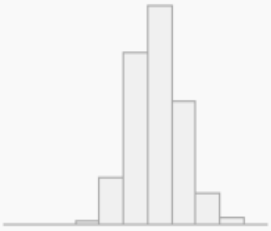
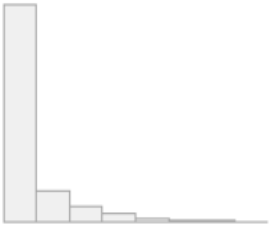
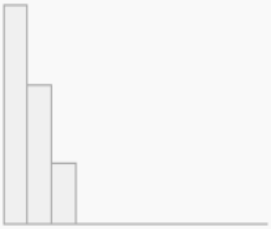
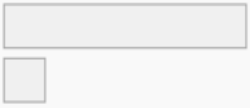
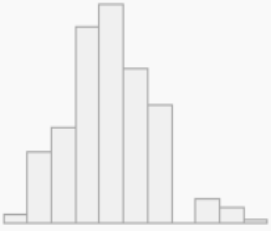
FIG. 2.5 – Arbre de censure

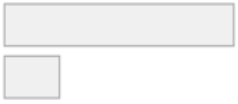
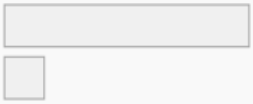
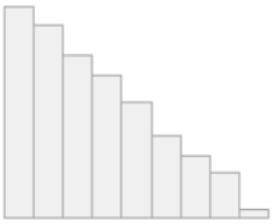
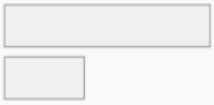
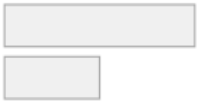
Catégorie	Nom de la variable	Explication
Variables d'Analyse de Survie	spell	Durée
	event	Événement
	censor2	Censure Type 2
	censor3	Censure Type 3
	censor4	Censure Type 4
Caractéristiques de l'Assurance Chômage	ui	Assurance Chômage (UI)
	reprate	Taux de Remplacement
	logwage	Logarithme du Salaire
	tenure	Ancienneté
	disrate	Taux de Chômage/Déplacement
	slack	Marché du Travail Détendu ("Slack")
	abolpos	Poste Supprimé
	explose	Licenciement Collectif
	stateur	Taux de Chômage État
Caractéristiques	houshead	Chef de Famille
Démographiques/Sociales	married	Marié(e)
	female	Femme
	child	Enfant(s)
	ychild	Jeune(s) Enfant(s)
	nonwhite	Non-Blanc
	age	Âge
Éducation	schlt12	Scolarité < 12ans
	schgt12	Scolarité > 12ans
Localisation et Emploi Précédent	smsa	Zone Urbaine (SMSA)
	bluecoll	Col Bleu
Industrie (Emploi Précédent)	mining	Industrie : Mines
	constr	Industrie : Construction
	transp	Industrie : Transports
	trade	Industrie : Commerce
	fire	Industrie : FIRE
	services	Industrie : Services
	pubadmin	Industrie : Administration Publique
Variables Temporelles	year85	Année 1985
et Géographiques (Contrôle)	year87	Année 1987
	year89	Année 1989
Région Géographique	midatl	Région : Mid-Atlantic
	encen	Région : East North Central
	wncen	Région : West North Central
	southatl	Région : South Atlantic
	escen	Région : East South Central
	wscen	Région : West South Central
	mountain	Région : Mountain
	pacific	Région : Pacific

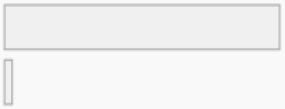
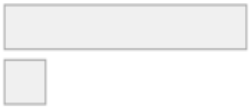
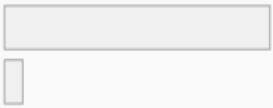
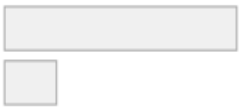
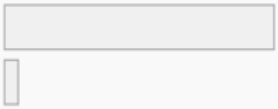
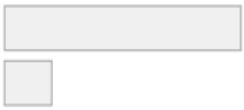
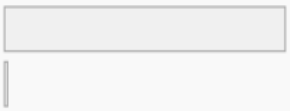
TAB. 2.4 – Variables des données du chômage

Le tableau ci-dessous présente un résumé des statistiques descriptives élémentaires relatives au chômage, ainsi que des histogrammes et des diagrammes en barres pour chaque variable.

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
spell [numeric]	Mean (sd) : 6.2 (5.6) min ≤ med ≤ max: 1 ≤ 5 ≤ 28 IQR (CV) : 7 (0.9)	28 distinct values		3343 (100.0%)	0 (0.0%)
event [numeric]	Min : 0 Mean : 0.3 Max : 1	0 : 2270 (67.9%) 1 : 1073 (32.1%)		3343 (100.0%)	0 (0.0%)
censor2 [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 3004 (89.9%) 1 : 339 (10.1%)		3343 (100.0%)	0 (0.0%)
censor3 [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2769 (82.8%) 1 : 574 (17.2%)		3343 (100.0%)	0 (0.0%)
censor4 [numeric]	Min : 0 Mean : 0.4 Max : 1	0 : 2088 (62.5%) 1 : 1255 (37.5%)		3343 (100.0%)	0 (0.0%)
ui [numeric]	Min : 0 Mean : 0.6 Max : 1	0 : 1495 (44.7%) 1 : 1848 (55.3%)		3343 (100.0%)	0 (0.0%)
reprate [numeric]	Mean (sd) : 0.5 (0.1) min ≤ med ≤ max: 0.1 ≤ 0.5 ≤ 2.1 IQR (CV) : 0.1 (0.3)	419 distinct values		3343 (100.0%)	0 (0.0%)

logwage [numeric]	Mean (sd) : 5.7 (0.5) min ≤ med ≤ max: 2.7 ≤ 5.7 ≤ 7.6 IQR (CV) : 0.8 (0.1)	499 distinct values		3343 (100.0%)	0 (0.0%)
tenure [numeric]	Mean (sd) : 4.1 (5.9) min ≤ med ≤ max: 0 ≤ 2 ≤ 40 IQR (CV) : 5 (1.4)	40 distinct values		3343 (100.0%)	0 (0.0%)
disrate [numeric]	Mean (sd) : 0.1 (0.1) min ≤ med ≤ max: 0 ≤ 0.1 ≤ 1 IQR (CV) : 0.1 (0.7)	246 distinct values		3343 (100.0%)	0 (0.0%)
slack [numeric]	Min : 0 Mean : 0.5 Max : 1	0 : 1710 (51.2%) 1 : 1633 (48.8%)		3343 (100.0%)	0 (0.0%)
abolpos [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2856 (85.4%) 1 : 487 (14.6%)		3343 (100.0%)	0 (0.0%)
explose [numeric]	Min : 0 Mean : 0.5 Max : 1	0 : 1663 (49.7%) 1 : 1680 (50.3%)		3343 (100.0%)	0 (0.0%)
stateur [numeric]	Mean (sd) : 6.6 (1.8) min ≤ med ≤ max: 2.5 ≤ 6.6 ≤ 13 IQR (CV) : 2 (0.3)	64 distinct values		3343 (100.0%)	0 (0.0%)

houshead [numeric]	Min : 0 Mean : 0.6 Max : 1	0 : 1297 (38.8%) 1 : 2046 (61.2%)		3343 (100.0%)	0 (0.0%)
married [numeric]	Min : 0 Mean : 0.6 Max : 1	0 : 1384 (41.4%) 1 : 1959 (58.6%)		3343 (100.0%)	0 (0.0%)
female [numeric]	Min : 0 Mean : 0.3 Max : 1	0 : 2180 (65.2%) 1 : 1163 (34.8%)		3343 (100.0%)	0 (0.0%)
child [numeric]	Min : 0 Mean : 0.5 Max : 1	0 : 1838 (55.0%) 1 : 1505 (45.0%)		3343 (100.0%)	0 (0.0%)
ychild [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2689 (80.4%) 1 : 654 (19.6%)		3343 (100.0%)	0 (0.0%)
nonwhite [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2878 (86.1%) 1 : 465 (13.9%)		3343 (100.0%)	0 (0.0%)
age [numeric]	Mean (sd) : 35.4 (10.6) min ≤ med ≤ max: 20 ≤ 34 ≤ 61 IQR (CV) : 16 (0.3)	42 distinct values		3343 (100.0%)	0 (0.0%)
schlt12 [numeric]	Min : 0 Mean : 0.3 Max : 1	0 : 2403 (71.9%) 1 : 940 (28.1%)		3343 (100.0%)	0 (0.0%)
schgt12 [numeric]	Min : 0 Mean : 0.3 Max : 1	0 : 2221 (66.4%) 1 : 1122 (33.6%)		3343 (100.0%)	0 (0.0%)

smsa [numeric]	Min : 0 Mean : 0.7 Max : 1	0 : 922 (27.6%) 1 : 2421 (72.4%)		3343 (100.0%)	0 (0.0%)
bluecoll [numeric]	Min : 0 Mean : 0.6 Max : 1	0 : 1325 (39.6%) 1 : 2018 (60.4%)		3343 (100.0%)	0 (0.0%)
mining [numeric]	Min : 0 Mean : 0 Max : 1	0 : 3245 (97.1%) 1 : 98 (2.9%)		3343 (100.0%)	0 (0.0%)
constr [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2848 (85.2%) 1 : 495 (14.8%)		3343 (100.0%)	0 (0.0%)
transp [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 3127 (93.5%) 1 : 216 (6.5%)		3343 (100.0%)	0 (0.0%)
trade [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2725 (81.5%) 1 : 618 (18.5%)		3343 (100.0%)	0 (0.0%)
fire [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 3171 (94.9%) 1 : 172 (5.1%)		3343 (100.0%)	0 (0.0%)
services [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2775 (83.0%) 1 : 568 (17.0%)		3343 (100.0%)	0 (0.0%)
pubadmin [numeric]	Min : 0 Mean : 0 Max : 1	0 : 3311 (99.0%) 1 : 32 (1.0%)		3343 (100.0%)	0 (0.0%)

year85 [numeric]	Min : 0 Mean : 0.3 Max : 1	0 : 2448 (73.2%) 1 : 895 (26.8%)		3343 (100.0%)	0 (0.0%)
year87 [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2616 (78.3%) 1 : 727 (21.7%)		3343 (100.0%)	0 (0.0%)
year89 [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2675 (80.0%) 1 : 668 (20.0%)		3343 (100.0%)	0 (0.0%)
midatl [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2979 (89.1%) 1 : 364 (10.9%)		3343 (100.0%)	0 (0.0%)
encen [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2865 (85.7%) 1 : 478 (14.3%)		3343 (100.0%)	0 (0.0%)
wncen [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 3128 (93.6%) 1 : 215 (6.4%)		3343 (100.0%)	0 (0.0%)
southatl [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 2549 (76.2%) 1 : 794 (23.8%)		3343 (100.0%)	0 (0.0%)
escen [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 3165 (94.7%) 1 : 178 (5.3%)		3343 (100.0%)	0 (0.0%)
wscen [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2861 (85.6%) 1 : 482 (14.4%)		3343 (100.0%)	0 (0.0%)
mountain [numeric]	Min : 0 Mean : 0.1 Max : 1	0 : 2982 (89.2%) 1 : 361 (10.8%)		3343 (100.0%)	0 (0.0%)
pacific [numeric]	Min : 0 Mean : 0 Max : 1	0 : 3256 (97.4%) 1 : 87 (2.6%)		3343 (100.0%)	0 (0.0%)

FIG. 2.6 – Statistiques descriptives du chômage

ملخص

الهـدف من هذه المذكرة هو تقـدير دالة البقـاء لتحليـل مـدة البطـالة قبـل العثـور على وظيفـة في ظـل البيـانات غير المكتملـة من اليميـن بشكـل عشـوائي. في الفـصل الاول قـدمنا تعـاريف ومفـاهيم أساسيـة لتحليـل البقـاء، الرقـابة والتقـدير اللامعلمـي، والذي يتـم تقديمـه عمـومًا بتقـدير كابلان-مايـر وتقـدير نيلسـون-آليـن. في الفصل الثاني طبـقنـا دراسـة عن حـالة مبنيـة على بيـانات حقيقيـة حـول الأفـراد العاطليـن عن العمـل.

الكلمات المفتاحية: الرقـابة، دالـة البقـاء، التقـدير اللامعلمـي، مقدر كابلان-مايـر، مدة البطالة.

Résumé

L'objectif de ce mémoire est d'estimer la fonction de survie pour analyser la durée du chômage avant de trouver un emploi sous des données censurées aléatoirement à droite. Dans le premier chapitre nous avons présenté les définitions et concepts de base de l'analyse de survie, la censure et l'estimation non paramétrique, qui est généralement présenté par l'estimation de Kaplan-Meier et l'estimation de Nelson-Aalen. Dans le deuxième chapitre, nous avons mené une étude de cas basée sur des données réelles sur des individus au chômage.

Les mots clés : La censure, La fonction de survie, l'estimation non paramétrique, estimateur de Kaplan-Meier, La durée de chômage.

Abstract

This thesis aims to estimate the survival function to analyse the duration of unemployment before finding a job under right-censored data. In the first chapter, we presented the basic definitions and concepts of survival analysis, censoring, and non-parametric estimation, which the Kaplan-Meier and Nelson-Aalen estimators generally give. In the second chapter, we conducted a case study based on real data concerning unemployed individuals.

Keywords: Censoring, the survival function, non-parametric estimation, Kaplan-Meier estimator, the duration of unemployment.