RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEURE ET DE LA RECHERCHE SCIENTIFIQUE

Université Batna 2, Batna
Faculté de Mathématiques et de l'informatique
Département de Mathématiques

Modèles de durées et analyse de survie

Support de cours

Par
FATEH MERAHI

1ère année Master (SAD)
2020-2021

Table des matières

Introduction

Coefficients : 3. **Crédits : 6**

Objectifs de l'enseignement : A l'issue de ce cours, l'étudiant sera familiarisé avec les concepts et modèles de base en analyse de survie. En outre, l'étudiant sera capable d'analyser des données réelles à l'aide de logiciels.

Connaissances préalables recommandées : Des connaissances de base de statistique mathématique et de probabilités qui sont nécessaires.

0- Introduction.

1- Concepts et spécificités de l'analyse de survie.

2- Phénomènes de censure et de troncature.

3- Estimateur non paramétrique de Kaplan-Meier : Construction.

4- Estimateur non paramétrique de Kaplan-Meier : Propriétés.

5- Table de mortalité et lissage.

6- Lissage et estimation paramétrique.

Références

1- Cox, D.R. et Oakes, D. (1984). Analysis of survival data, Chapman and Hall, New York.

2- Hougaard, P. (2000). Analysis of multivariate survival data. Springer, New-York.

3- Klein, J.P. et Moeschberger, M.L. (1997). Survival analysis, techniques for censored and truncated data, Springer, New York.

Mode d'évaluation : Contrôle continu (40%), Examen (60%).

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant la mort pour des organismes biologiques (l'espérance de vie) ou le temps restant avant l'échec ou la panne dans les systèmes artificiels, ce que l'on représente graphiquement sous la forme d'une courbe de survie. On parle aussi d'analyse de la fiabilité en ingénierie, d'analyse de la durée en économie ou d'analyse de l'histoire d'événements en sociologie. La représentation des données de survie se fait souvent sous la forme graphique d'une courbe de survie. Dans les cas où les événements d'intérêt ne se sont pas produits avant la fin de la période d'observation (e.g., la maladie n'est pas apparue chez un malade) on parle de censure de la série de données.

L'analyse des modèes de survie a pris son développement depuis la deuxième guerre mondiale. La première méthode d'analyse de survie, la méthode actuarielle, est apparue en 1912. Elle est utilisée dans le domaine médical pour la première fois en 1950. La seconde méthode, dite de Kaplan-Meier, est apparue en 1958.

Elle est largement appliquée dans de divers domaines, ils constituent un outil utilisé dans l'assurance : durée de la vie humaine, durée de l'arrêt de travail, durée de chômage, mais aussi durée d'attente entre deux sinistres, durée avant la ruine. Dans la fiabilité, la durée de survie est, par exemple, définie comme le temps qui sépare la mise en marche d'une machine de la panne de celle-ci. Elle est aussi utilisée dans d'autres domaines comme l'économie, les assurance etc ...

Chapitre 1

Concepts et spécificités de l'analyse de survie

1.1 Distributions de la durée de survie

Supposons que la durée de survie T soit une variable positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions) :

1.1.1 Fonction de survie S

La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est- à-dire

$$S(t) = P(T > t), \ t \geq 0.$$

S est donc une fonction décroissante telle que $S(0) = 1$ (si $P(T = 0) = 0$, ce que nous supposerons) et $\lim_{t\to\infty} S(t) = 0$

1.1.2 Fonction de répartition F

La fonction de répartition (ou $c.d.f.$ pour "*cumulative distribution function*") représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à-dire

$$F(t) = P(T \leq t) = 1 - S(t).$$

Notations : Dans les cas où F a des sauts (la variable aléatoire T, n'est pas absolument continue), on a les notations suivantes :

La limite à gauche de F est notée : $F^{-}(t) = P(T < t)$.

La limite à droite de F est est notée : $F^{+}(t) = P(T \leq t)$.

La limite à gauche de S est notée : $S^{-}(t) = P(T > t)$.

La limite à droite de S est est notée : $S^{+}(t) = P(T \geq t)$.

Remarquons que

$$F^{-} \leq F^{+} \quad \text{et} \quad S^{-} \geq S^{+}.$$

1.1.3 Survie conditionnelle

On pose tout d'abord

$$S_u(t) = P(T > u + t \mid_{T>u})$$

la fonction de survie conditionnelle, on s'intéresse donc à la survie d'un élément après un instant $u + t$, sachant qu'il a déjà fonctionné correctement jusqu'en u. En revenant à la définition de la probabilité conditionnelle on peut écrire :

$$S_u(t) = P(T > u + t \mid_{T>u}) = \frac{P(T > u + t)}{P(T > u)} = \frac{S(u + t)}{S(u)}.$$

La fonction de survie conditionnelle s'exprime donc simplement à l'aide de la fonction de survie.

1.1.4 Densité de probabilité f

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

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

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

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

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

1.1.5 La fonction de hasard h

La fonction de hasard (ou taux de hasard, taux de panne, taux de défaillance, taux de décès, risque instantané, etc.) est par définition, pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

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

1.1.6 La fonction de hasard cumulée H

La fonction de hasard cumulée (ou le taux de hasard cumulé) est l'intégrale du risque instantané λ :

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

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

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

1.1.7 Quantités associées à la distribution de survie

Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbb{E}(T)$ et la variance de la durée de survie $\mathbb{V}(T)$ sont définis par les quantités suivantes

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

$$\mathbb{V}(T) = 2\int_0^\infty tS(t)\, dt - (\mathbb{E}(T))^2.$$

Ainsi on peut déduire l'espérance et la variance à partir de n'importe laquelle des fonctions F, S, f, ; (mais pas l'inverse).

Preuve On suppose que l'espérance existe. On écrit que $\mathbb{E}(T) = \int_0^\infty t dF(t) = \lim_{u\to\infty} \int_0^u t dF(t)$, en intégrant par parties on peut écrire

$$\begin{aligned} \int_0^u t dF(t) &= -\int_0^u t dS(t) \\ &= -uS(u) + \int_0^u S(t)\, dt; \end{aligned}$$

l'inégalité de Markov $\left((P(T > t) \leq \frac{\mathbb{E}(T)}{t}\right)$ assure alors que $tS(t) \leq \mathbb{E}(T)$ et donc le terme $uS(u)$ est borné. On en déduit que l'intégrale $\int_0^\infty S(t)\, dt$ converge, ce qui implique que $\lim_{t\to\infty} tS(t) = 0$ et en passant à la limite on obtient le résultat attendu.

On montre de la même manière que :

$$\mathbb{V}(T) = 2\int_0^\infty tS(t)\, dt - (\mathbb{E}(T))^2.$$

Quantiles de la durée de survie

La médiane de la durée de survie est le temps t pour lequel la probabilité de survie $S(t)$ est égale à 0.5, c'est-à-dire, la valeur t_m qui satisfait $S(t_m) = 0.5$.

La fonction quantile de la durée de survie est définie par

$$\begin{aligned} q(p) &= \inf\{t : F(t) \geq p\},\ 0 < p < 1, \\ &= \inf\{t : S(t) \leq 1 - p\}. \end{aligned}$$

Lorsque la fonction de répartition F est strictement croissante et continue alors

$$\begin{aligned} q(p) &= F^{-1}(p),\ 0 < p < 1, \\ &= S^{-1}(1 - p). \end{aligned}$$

Le quantile $q(p)$ est le temps où une proportion p de la population a disparu.

1.1.8 Cas des variables discrètes

Si la variable aléatoire T prend des valeurs entières $t_1, t_2, ..., t_j, ...$, sa distribution est décrite par les $p_j = P(T = t_j)$, pour $j = 1, 2, ...$ avec $t_j > 0$. La fonction de survie s'écrit simplement

$$S(t_k) = \sum_{t_j > t_k} P(T = t_j) = \sum_{j>k} p_j = \sum_{j \geq k+1} p_j.$$

L'interprétation de la fonction de hasard donnée dans le cas continu conduit naturellement à poser dans le cas discret :

$$h(t_k) = P\left(T = t_k \mid_{T > t_{k-1}}\right) = \frac{P(T = t_k)}{P(T > t_{k-1})} = \frac{p_k}{S(t_{k-1})}.$$

La fonction de hasard au point k s'interprète donc comme le taux de décès à l'âge k. De l'expression ci-dessus on tire que

$$1 - h(t_k) = \frac{S(t_k)}{S(t_{k-1})},$$

puis, par récurrence :

$$S(t_k) = \prod_{j=1}^{k} (1 - h(t_j)).$$

1.1.9 Les lois paramétriques usuelles

Le modèle exponentiel

La spécification la plus simple consiste à poser $h(t) = \lambda$, avec $\lambda > 0$. On en déduit immédiatement que

$$S(t) = e^{-\lambda t}.$$

Il est également caractérisé par le fait que la fonction de survie est multiplicative, au sens où

$$S(u + t) = S(u) S(t)$$

On vérifie aisément par un calcul direct que

$$\mathbb{E}(T) = \frac{1}{\lambda} \quad \text{et} \quad \mathbb{V}(T) = \frac{1}{\lambda^2}.$$

L'estimation du paramètre λ est donné par

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i} = \frac{1}{\overline{T}}.$$

Le modèle de Weibull

On suppose ici que la fonction de hasard est de la forme

$$h(t) = \lambda\alpha t^{\alpha-1}, \quad \text{avec } \alpha, \lambda > 0.$$

λ est un paramètre d'échelle et α.

Il est utilisé en physique pour modéliser la durée de vie de certaines particules ou le bruit en sortie de certains récepteurs de transmissions.

La distribution de T est alors la distribution de Weibull $W(\alpha, \lambda)$, dont la fonction de survie s'écrit

$$S(t) = e^{-\lambda t^{\alpha}}, \ t > 0.$$

La moyenne et la variance sont données par

$$\mathbb{E}(T) = \lambda^{-\frac{1}{\alpha}}\Gamma\left(\frac{\alpha+1}{\alpha}\right) \quad \text{et} \quad \mathbb{V}(T) = \lambda^{-\frac{2}{\alpha}}\left[\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha+1}{\alpha}\right)\right],$$

où $\Gamma(x) = \int_0^{+\infty} u^{x-1}e^{-u}du$ est la fonction gamma.

Le modèle Gamma

La densité f de la loi gamma $G(\alpha, \lambda)$ est donnée par

$$f(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}, \quad t \geq 0 \ \text{ et } \ \alpha, \lambda > 0.$$

L'espérance et la variance d'une loi Gamma sont données par :

$$\mathbb{E}(T) = \frac{\alpha}{\lambda} \quad \text{et} \quad \mathbb{V}(T) = \frac{\alpha}{\lambda^2}.$$

1.2 Phénomènes de censure et de 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.

1.2.1 Censure

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 $T_1, T_2, ..., T_n$ 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

$$X_i = T_i \wedge C = \min(T_i, C).$$

Exemple 1.2.1 *On peut tester la durée de vie de n objet identiques (ampoules) sur un intervalle d'observation fixé [0,u].*

2. La censure de type II (Attente jusqu'au k-ième décés)

On observe les durées de vie de n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment. Si on ordonne les durées $T_1, T_2, ..., T_n$, on obtient les statistiques d'ordre des $T_{(1)}, T_{(2)}, ..., T_{(n)}$. La date de censure est donc $X_{(k)}$ et on observe $X_{(1)} = T_{(1)}$, $X_{(2)} = T_{(2)}, ..., X_{(k)} = T_{(k)}$, $X_{(k+1)} = T_{(k)}, ...,$ $X_{(n)} = T_{(k)}$.

2. La censure de type III (censure aléatoire de type I)

Soient $C_1, ..., C_n$ 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 $\delta_i = \mathbb{I}_{\{T_i \leq C_i\}}$.

  · $\delta_i = 1$ 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, à chaque patient i on lui associe non seulement son temps de survie T_i mais aussi son temps de censure C_i. Ce temps de censure sera propre à chaque patient et peut-être dû à plusieurs raisons dont :

- **Perte de vue** : le patient quitte l'étude en cours et on ne le revoit plus. Ce sont des patients "*perdus de vue*".
- **Arrêt ou le changement du traitement** : Il peut y avoir des effets secondaires si désastreux pour le patient qu'on est obligé d'arrêter le traitement. Ces patients sont "*exclus*" de l'étude.
- **Fin de l'étude** : l'étude se termine alors que certains patients sont toujours vivants. Ce sont des patients "exclus-vivants".

Ce qu'on observe alors c'est le couple (X_i, δ_i) où $X_i = \min(T_i, C_i)$ et $\delta_i = \mathbb{I}_{\{T_i \leq C_i\}}$ (l'indicatrice de non censure).

Remarque 1.2.1 *Dans la suite de ce cours, nous ferons l'hypothèse que la censure est indépendante de l'événement, c'est-à-dire, que T_i est indépendant de C_i. Cette hypothèse est très utile d'un point de vue mathématique et indispensable aux modèles classiques d'analyse de survie.*

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 (X, δ)

$$\begin{aligned} X &= T \vee C = \max(T, C), \\ \delta &= \mathbb{I}_{\{T \geq C\}} \end{aligned}$$

Exemple 1.2.2 *On peut ignorer la date exacte à la quelle s'est d'éclanchéé la maladie.*

Remarque 1.2.2 *Les modèles présentés dans ce cours traitent le cas de la censure à droite. Très peu de travaux s'intéressent à la censure à gauche car beaucoup moins fréquente.*

Censure mixte

C'est la censure à gauche et à droite on même temps.

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. On sait que $T \in I$ (I intervalle).

· Pour la censure à droite si je ne connai pas T je sais que $T \in [c, +\infty[$.

· Pour la censure à gauche si je ne connai pas T je sais que $T \in [0, c]$.

1.2.2 Troncature

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui même. Ainsi, une variable T est tronquée par un sous ensemble éventuellement aléatoire $A \subset \mathbb{R}_+$ si au lieu de T, on observe T uniquement si $T \in A$. Les points de l'échantillon "tronqué" appartiennent tous à A, S'il y a troncature, une partie des individus (donc des T_i) ne sont pas observables.

La troncature à gauche

Soit Z une variable aléatoire indépendante de T, on dit qu'il y a troncature à gauche lorsque T n'est observable que si $T > Z$. On observe le couple (T, Z) , avec $T > Z$.

La troncature à droite

De même, il y a troncature à droite lorsque T n'est observable que si $T < Z$. On observe le couple (T, Z) , avec $T < Z$.

La troncature par intervalle

Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.

1.2.3 Fonction de vraisemblance

Cas d'une censure à droite

Considérons le cas d'une censure aléatoire droite C indépendante de la durée d'intérêt T. Supposons que les variables T et C ont pour densités respectives f et g et pour survies S et G. La distribution de T est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (X_i, δ_i), où $X_i = \min(T_i, C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = \mathbb{I}_{\{T_i \leq C_i\}}$. Ainsi, la contribution à la vraisemblance pour l'individu i est

$$\begin{aligned} L_i &= P(X_i \in [t_i, t_i + dt], \delta_i = 1|_\theta)^{\delta_i} \times P(X_i \in [t_i, t_i + dt], \delta_i = 0|_\theta)^{1-\delta_i} \\ &= P(T_i \in [t_i, t_i + dt], T_i \leq C_i|_\theta)^{\delta_i} \times P(C_i \in [t_i, t_i + dt], T_i > C_i|_\theta)^{1-\delta_i} \\ &= \left[f(t_i|_\theta) G\left(t_i^-\right)\right]^{\delta_i} \times [g(t_i|) S(t_i|_\theta)]^{1-\delta_i}. \end{aligned}$$

le paramètre d'intérêt θ n'apparaît pas dans la loi de la censure, alors La partie utile de la vraisemblance se réduit alors à

$$L = \prod_{i=1}^{n} [f(t_i|_\theta)]^{\delta_i} \times [S(t_i|_\theta)]^{1-\delta_i}.$$

Remarque 1.2.3 *Notons que la présence de données censurées doit être prise en compte dans l'écriture de la vraisemblance. En effet, en raisonnant sur le sous échantillon des données non censurées, la vraisemblance est*

$$\overline{L} = \prod_{i=1}^{n} [f(t_i|_\theta)].$$

L'estimateur obtenu en maximisant $\overline{L}$ est asymptotiquement biaisé.

Cas de troncature

Considérons le cas de données tronquées à gauche de manière aléatoire. Les variables T_i sont soumises à troncature par les variables aléatoires Z_i supposées indépendantes des T_i. On dispose d'un échantillon $(T_i, Z_i)_{i=1,..,N}$ où N est aléatoire puisqu'on sélectionne (on observe) les individus pour lesquels $T_i \geq Z_i$ dans une population de taille inconnue n. La vraisemblance conditionnelle par rapport à N et aux valeurs de la variable de troncature observées est

$$L_T = \prod_{i=1}^{N} P(T_i \in [t_i, t_i + dt]|_{T \geq z_i}) = \prod_{i=1}^{N} \frac{P(T_i \in [t_i, t_i + dt], T \geq z_i)}{P(T \geq z_i)} = \prod_{i=1}^{N} \frac{f(t_i)}{S(z_i)}.$$

Par construction, on a $t_i \geq z_i$ pour tout i.

TRAVAUX PRATIQUES

1.**Burr Distribution** :

La densité de probabilité de *la loi de Burr* est donnée par

$$f(x) = \frac{ck}{\lambda} \frac{\left(\frac{x}{\lambda}\right)^{c-1}}{\left(1+\left(\frac{x}{\lambda}\right)^{c}\right)^{k+1}}, \quad x \in]0, +\infty[.$$

et sa fonction de répartition est :

$$F(x) = 1 - \left(1+\left(\frac{x}{\lambda}\right)^{c}\right)^{-k}, \quad x \in]0, +\infty[.$$

Calculer et tracer **la fonction de survie** d'une distribution de **Burr** avec des paramètres 50, 3, et 1 :

Algorithme :

```
x = 0 :0.1 :200 ;
figure()
plot(x,1-cdf('Burr',x,50,3,1))
xlabel('Failure time') ;
ylabel('Survival probability') ;
```

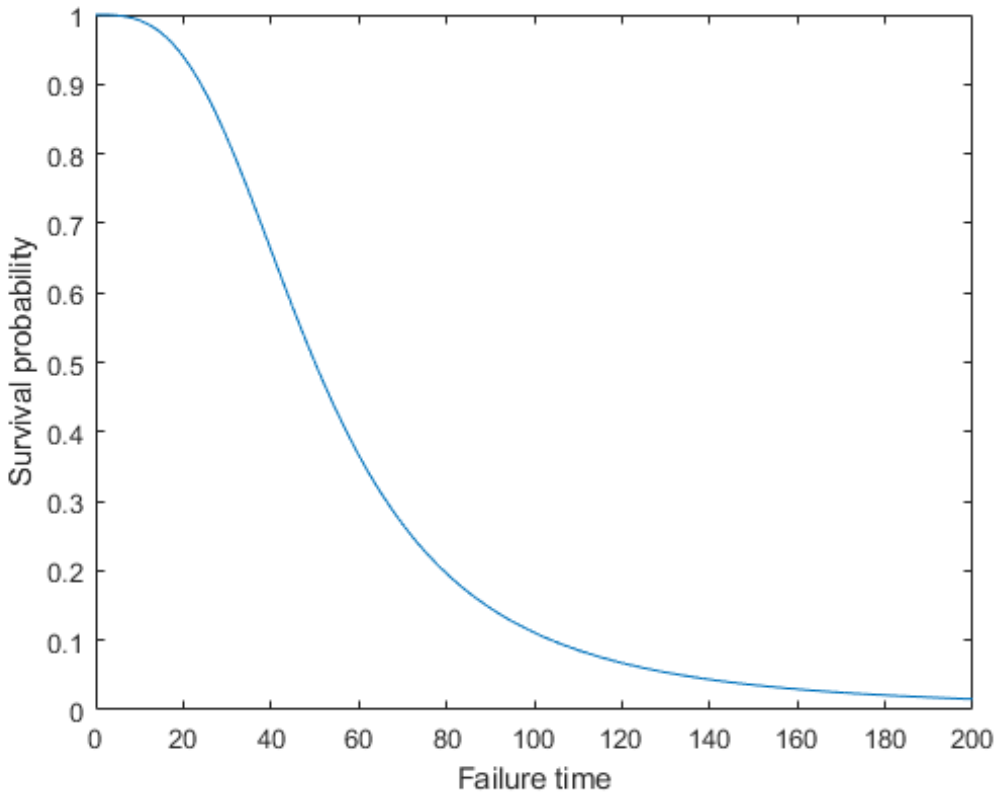
Calculer et tracer **la fonction de hasard** d'une distribution de **Burr** avec des paramètres 50, 3, et 1.

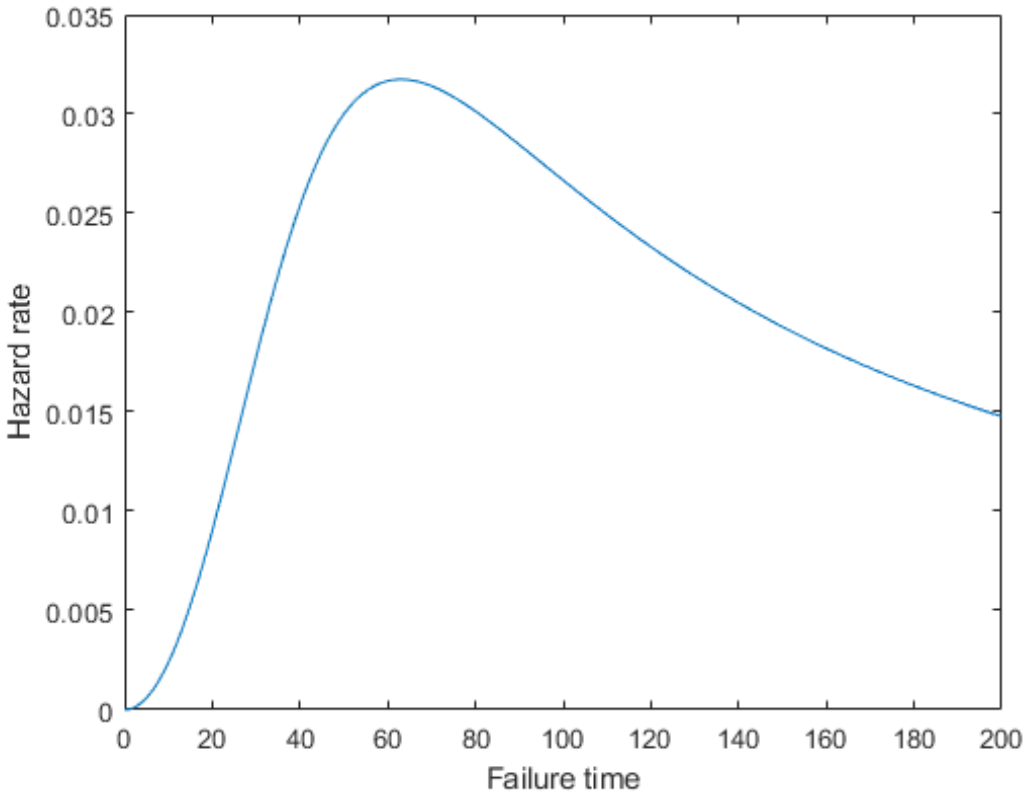
Algorithme :

```
x = 0 :1 :200 ;
Burrhazard = pdf('Burr',x,50,3,1)./(1-cdf('Burr',x,50,3,1)) ;
figure()
plot(x,Burrhazard)
xlabel('Failure time') ;
ylabel('Hazard rate') ;
```

2.**Weibull Distribution**

Calculer et tracer **la fonction de hasard** d'une distribution de **Weibull** avec des paramètres 3, 0.6, et 9, 4 et 2.5, 1.


Survival probability
Failure time
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160
180
200


0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Hazard rate
0
20
40
60
80
100
120
140
160
180
200
Failure time

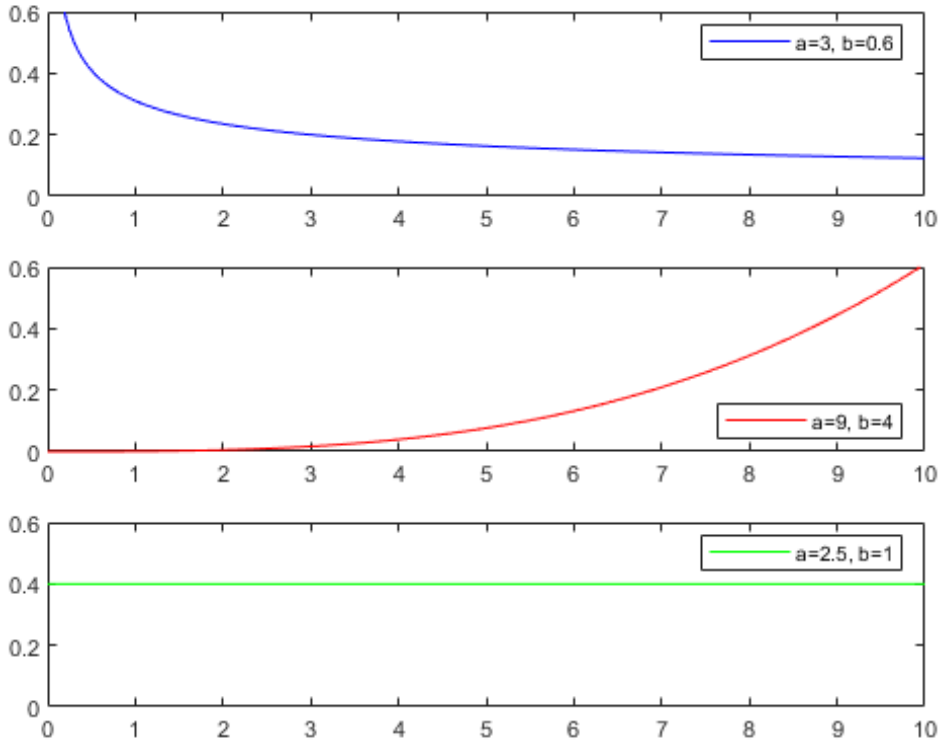
Algorithme :

```
figure
ax1 = subplot(3,1,1);
x1 = 0 :0.05 :10;
hazard1 = pdf('wbl',x1,3,0.6)./(1-cdf('wbl',x1,3,0.6));
plot(x1,hazard1,'color','b')
set(ax1,'Ylim',[0 0.6]);
legend(ax1,'a=3, b=0.6');
ax2 = subplot(3,1,2);
x2 = 0 :0.05 :10;
hazard2 = pdf('wbl',x2,9,4)./(1-cdf('wbl',x2,9,4));
plot(x2,hazard2,'color','r')
set(ax2,'Ylim',[0 0.6]);
legend(ax2,'a=9, b=4','location','southeast');
ax3 = subplot(3,1,3);
x3 = 0 :0.05 :10;
hazard3 = pdf('wbl',x3,2.5,1)./(1-cdf('wbl',x3,2.5,1));
plot(x3,hazard3,'color','g')
set(ax3,'Ylim',[0 0.6]);
legend(ax3,'a=2.5, b=1');
```

Nom de distribution de probabilité :

1. Les distrubutions discrètes :

Nom	*Distribution*
Binomial	Binomial distribution
Geometric	Geomeric distribution
Hypergeometric	Hypergeometric distribution
Discrete Uniform	uniform distribution (discrete)
Poisson	Poisson distribution
Multinomial	Multinomial distribution
Negative Bionomial	Negative Bionomial distribution


a=3, b=0.6
a=9, b=4
a=2.5, b=1
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9 10

1. Les distrubutions continues :

Nom	*Distribution*
Beta	Beta distribution
Burr	Burr distribution
Chisquare	Chisquare distribution
Gamma	Gamma distribution
Lognormal	Lognormal distribution
Lognormal	Lognormal distribution
Normal	Normal distribution
T	Student distribution
Uniform	Uniform distribution (continue)
Weibull	Weibull distribution

Fonctions Matlab :

Fonctions	
makedist	Créer un objet de distribution de probabilité
fitdist	Ajuster l'objet de distribution de probabilité aux données
cdf	Fonction de distribution cumulative
pdf	Fonction de densité de probabilité

Exemple 1.2.3 ***1.*** $pd = makedist('Poisson', 'lambda', lambda);$

$x = -3 : .1 : 3;$

$pdf = pdf(pd, x);$

2.$y = cdf('Normal', x, mu, sigma)$

$x = -3 : .1 : 3;$

$p = cdf(y, x);$

$plot(x, p)$

Cumulative Distribution Function and Survival Probability :

Une table de mortalité (life table) se compose généralement de :

- t : Temps d'échec (**Failure times**)
- d : Nombre d'éléments ayant échoué à un moment (**Number Failed**)

- Nombre d'éléments censurés à un moment / une période
- r : Nombre d'éléments à risque au début d'une période (**Number at Risk**)

Le nombre à risque est le nombre total de survivants au début de chaque période. Le nombre à risque au début de la première période correspond à tous les individus participant à l'étude à vie. Au début de chaque période restante, le nombre à risque est réduit du nombre d'échecs plus les individus censurés à la fin de la période précédente.

Exemple 1.2.4 *(cas non censuré)*

Cette table de mortalité (life table) présente des données de survie fictives. Au début du premier temps de panne, ***7*** *éléments sont à risque. Au temps* ***4****,* ***3*** *échouent. Donc, au début du temps* ***7****, il y a* ***4*** *éléments à risque. Un seul* ***1*** *échec au temps* ***7****, donc le nombre à risque au début du temps* ***11*** *est de* ***3****.* ***2*** *échouent au temps* ***11****, donc au début du temps* ***12****, le nombre à risque est* ***1****. L'élément restant au moment de l'échec* ***12****.*

t	d	r	$h = d/r$	*Survival Probability : S*	*cdf : F = 1 − S*
4	*3*	*7*	$\frac{3}{7}$	$1 - \frac{3}{7} = 0.5714$	*0.4286*
7	*1*	*4*	$\frac{1}{4}$	$\left(1 - \frac{1}{4}\right) 0.5714 = 0.4286$	*0.5714*
11	*2*	*3*	$\frac{2}{3}$	$\left(1 - \frac{2}{3}\right) 0.4286 = 0.1429$	*0.8571*
12	*1*	*1*	$\frac{1}{1}$	$\left(1 - \frac{1}{1}\right) 0.1429 = 0$	*1*

t : ***Failure times***

d : ***Number Failed***

r : ***Number at Risk***

h : ***Hazard rate***

t=(4,7,11,12)

Number Failed (d)=(3,1,2,1)

Number at Risk (r)=(7,4,3,1)

Algorithme :

```
t = [4 7 11 12];
freq = [3 1 2 1];
% F=cdf,
[F,x] = ecdf(t,'frequency',freq)
```

S=1-F ;

Résultats :

x =

4

4

7

11

12

F =

0

0.4286

0.5714

0.8571

1.0000

S =

1.0000

0.5714

0.4286

0.1429

0

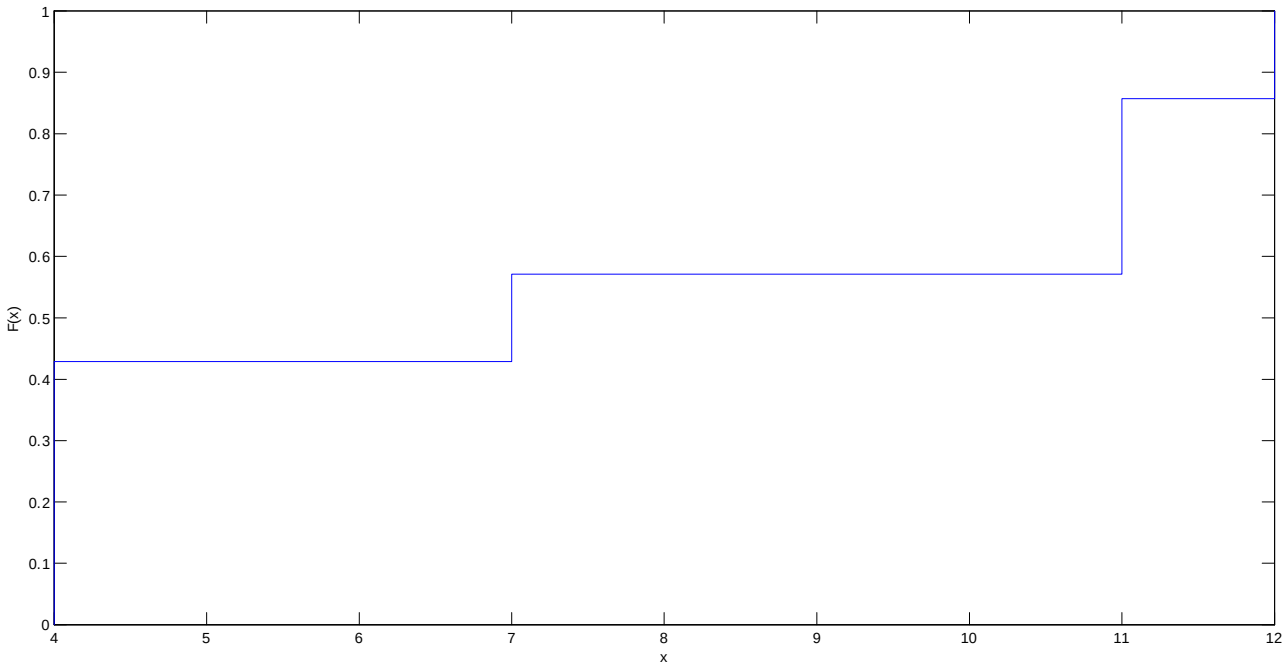


FIG. 1.1 – La courbe de la fonction de répartition F

Exercices

Exercice 1. Dans le cas absolument continu, supposons que la densité f est continue sur $\mathbb{R}_+^*$. Montrer que pour tout $t > 0$ tel que $S(t) \neq 0$, on a

$$h(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid_{T \geq t})}{h} = \frac{f(t)}{S(t)}.$$

Exercice 2. Dans le cas absolument continu, montrer que

1) S est continue.

2) S est décroissante.

3) $\lim\limits_{t \to 0} S(t) = 1$ et $\lim\limits_{t \to \infty} S(t) = 0$.

Exercice 3. Dans le cas absolument continu, supposons que la densité f est continue sur $\mathbb{R}_+^*$ et que $A = \{t : \ S(t) \neq 0\}$. Montrer que les conditions suivantes sont équivalentes :

1) $h(t) = \frac{f(t)}{S(t)}$, $\forall t \in A$.

2) $h(t) = (-\ln S(t))'$, $\forall t \in A$.

3) $S(t) = \exp(-H(t)), \forall t \in A$.

4) $f(t) = h(t) \exp(-H(t)), \forall t \in A$.

Exercice 4. Soit $T \rightsquigarrow \mathcal{E}xp(\alpha)$, la densité f de la loi exponentielle est donnée par

$$f(t) = \alpha \exp(-\alpha t), \text{ pour } t \in]0, +\infty[$$

Montrer que la fonction de hasard h est constante.

Exercice 5. Dans le cas discret, supposons que la variable aléatoire T prend les valeurs $1, 2, 3$ avec $P(T = j) = \frac{1}{3}$, pour $j = 1, 2, 3$. Calculer $S(x)$ et $h(x)$ et tracer la courbe de $S(x)$.

Solutions

Exercice 1. on a

$$\begin{aligned} h(t) &= \lim_{h\to 0}\frac{P(t\leq T<t+h\,|_{T\geq t})}{h}=\lim_{h\to 0}\frac{1}{h}\frac{P(\{t\leq T<t+h\}\cap\{T\geq t\})}{P(T\geq t)} \\ &= \lim_{h\to 0}\frac{1}{h}\frac{P(t\leq T<t+h)}{P(T\geq t)}=\frac{1}{P(T\geq t)}\lim_{h\to 0}\frac{P(t\leq T<t+h)}{h}=\frac{f(t)}{S(t)} \end{aligned}$$

Exercice 2. On a $S(t)=1-F(t)$, alors

1) La fonction de répartition $F(t)$ est continue, donc $S(t)$ est continue.

2) $S(t)$ est décroissante car $F(t)$ est croissante.

3) $\lim\limits_{t\to 0}S(t)=\lim\limits_{t\to 0}[1-F(t)]=1-\lim\limits_{t\to 0}F(t)=1-0=1$ et $\lim\limits_{t\to\infty}S(t)=\lim\limits_{t\to\infty}[1-F(t)]=1-\lim\limits_{t\to\infty}F(t)=1-1=0$.

Exercice 3.

1) $\Longrightarrow$ 2) $h(t)=\frac{f(t)}{S(t)}=\frac{-S'(t)}{S(t)}=-(\ln S(t))'$.

2) $\Longrightarrow$ 3) $-(\ln S(t))'=h(t)\Longrightarrow -\int\limits_0^t(\ln S(u))'du=\int\limits_0^t h(u)du\Longrightarrow -[\ln S(t)-\ln S(0)]=H(t)\Longrightarrow -\ln S(t)+\ln 1=H(t)\Longrightarrow S(t)=\exp(-H(t))$.

3) $\Longrightarrow$ 4) d'après 1) $S(t)=\frac{f(t)}{h(t)}$ et d'après 3) $S(t)=\exp(-H(t))\Longrightarrow f(t)=h(t)\exp(-H(t))$.

4) $\Longrightarrow$ 1) $\frac{f(t)}{h(t)}=\exp(-H(t))$ et et d'après 3) $\exp(-H(t))=S(t)\Longrightarrow\frac{f(t)}{h(t)}=S(t)$ alors $h(t)=\frac{f(t)}{S(t)}$.

Exercice 4.

· $x\in]-\infty,1[$: $S(x)=P(T>x)=\sum\limits_{x_j>x}P(T=x_j)=P(T=1)+P(T=2)+P(T=3)=\frac{1}{3}+\frac{1}{3}+\frac{1}{3}=1$.

· $x\in[1,2[$: $S(x)=P(T>x)=\sum\limits_{x_j>x}P(T=x_j)=P(T=2)+P(T=3)=\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$.

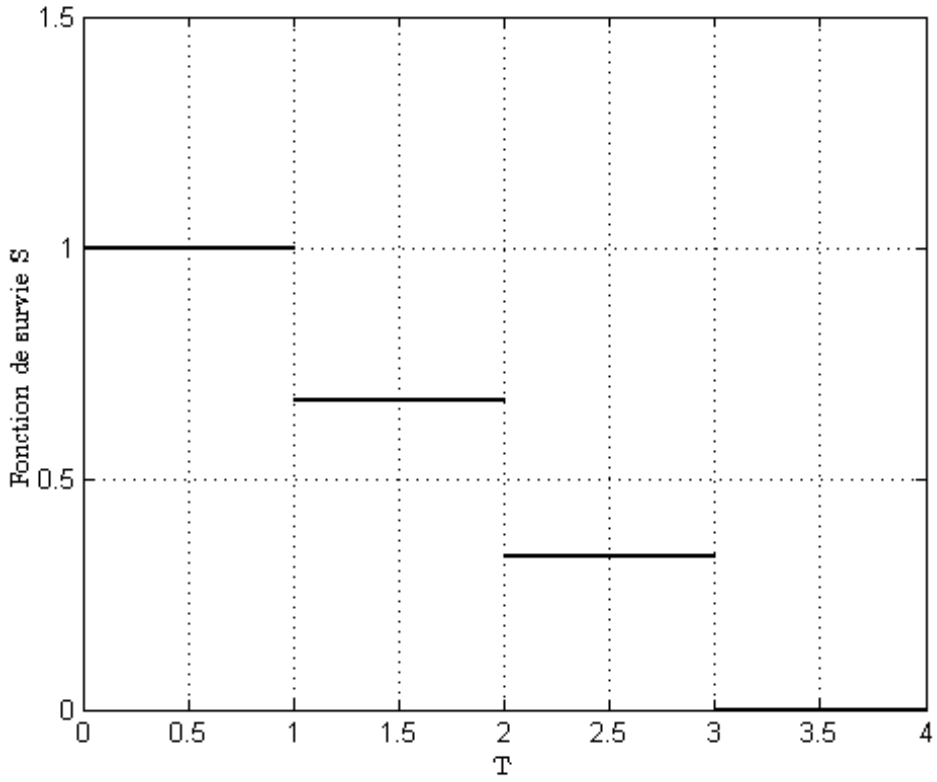
· $x\in[2,3[$: $S(x)=P(T>x)=\sum\limits_{x_j>x}P(T=x_j)=P(T=3)=\frac{1}{3}$.

· $x \in [3, +\infty[: S(x) = 0$.

Alors

$$h(x_j) = \frac{p_j}{S(x_{j-1})}$$

$h(1) = \frac{\frac{1}{3}}{1} = \frac{1}{3}$, $h(2) = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$, $h(3) = \frac{\frac{1}{3}}{\frac{1}{3}} = 1$.


1.5
1
0.5
0
Fonction de survie S
0
0.5
1
1.5
2
2.5
3
3.5
4
T

Chapitre 2

Estimateur non paramétrique de Kaplan-Meier

2.1 Estimation de la survie

1. **Modèle complet** :

Soit les variables d'intérêt $(T_1, T_2, ..., T_n)$ de fonction de répartition F, et de survie S. L'estimateur ponctuel est donné par

$$\forall t \geq 0,\ \widehat{S}_n(t) = 1 - \widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i > t\}},$$

avec $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i \leq t\}}$ la fonction de répartition empirique.

Ecriture des moments empiriques :

$$\overline{\overline{T}}_n = \frac{1}{n}\sum_{i=1}^{n} T_i = \int_0^{\infty} t d\widehat{F}_n(t).$$

Propriétés

1. L'estimateur $\widehat{S}_n(t)$ est sans biais (Absence de biais)

$$E\left\{\widehat{S}_n(t)\right\} = S(t).$$

2. Convergence presque sûre (p.s) uniforme

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\widehat{S}_n(t) - S(t)\right| = 0, \quad (p.s)\,.$$

3. Normalité asymptotique

$$\sqrt{n}\left(\widehat{S}_n(.) - S(.)\right) \xrightarrow{\mathcal{L}} W(.)\,.$$

Où W est un processus gaussien centré (pont Brownien) de fonction de var-cov :

$$\rho(s,t) = F(s) \wedge F(t) - F(s)\,F(t)\,.$$

2. Modèle censuré à droite :

Considérons le cas d'une censure aléatoire droite C indépendante de la durée d'intérêt T. Supposons que les variables T et C ont pour densités respectives f et g et pour survies S et G, F la fonction de répartition de T. La distribution de T est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (X_i, δ_i), où $X_i = \min(T_i, C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = \mathbb{I}_{\{T_i \leq C_i\}}$. La fonction de répartition de X est $F_X(.) = F_T(.).F_C(.)$, alors $S_X(.) = S(.)\,G(.)$.

2.2 Estimateur de Kaplan-Meier de la survie

L'estimateur de Kaplan-Meier (EKM) découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t ; c'est-à-dire, si $t'' < t' < t$

$$\begin{aligned} P(T > t) &= P\left(T > t, T > t'\right) \\ &= P\left(T > t|_{T>t'}\right) P\left(T > t'\right) \\ &= P\left(T > t|_{T>t'}\right) P\left(T > t'|_{T>t''}\right) P\left(T > t''\right) \end{aligned}$$

En considérant les temps d'événements (décès et censure) distincts $X_{(i)}$ $(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$P\left(T > X_{(j)}\right) = \prod_{k=1}^{j} P\left(T > X_{(k)}|_{T>X_{(k-1)}}\right),$$

avec $X_{(0)} = 0$. Considérons les notations suivantes :

· r_i : le nombre d'individus à risque de subir l'événement juste avant le temps $X_{(i)}$.

· d_i : le nombre de décès en $X_{(i)}$.

· $p_i = P\left(T \leq X_{(i)}|_{T>X_{(i-1)}}\right)$: la probabilité de mourir dans l'intervalle $]X_{(i-1)}, X_{(i)}]$ sachant que l'on était vivant en $X_{(i-1)}$.

Alors la probabilité p_i peut être estimée par

$$\widehat{p}_i = \frac{d_i}{r_i}$$

Comme les temps d'événements sont supposés distincts, on a

$$\begin{aligned} d_i &= 0 \text{ en cas de censure en } X_{(i)}, \text{ i.e. quand } \delta_i = 0. \\ d_i &= 1 \text{ en cas de décès en } X_{(i)}, \text{ i.e. quand } \delta_i = 1. \end{aligned}$$

On obtient alors l'estimateur de Kaplan-Meier :

$$\widehat{S}_{KM}(t) = \prod_{\substack{i=1,\ldots,n \\ X_{(i)} \leq t}} \left(1 - \frac{\delta_i}{r_i}\right) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\delta_i}{n-(i-1)}\right) = \prod_{i:X_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}, \ \forall t < X_{(1)}, \ \widehat{S}_{KM}(t) = 1.$$

2.2.1 Propriétés

· $\widehat{S}_{KM}(t)$ est une fonction en escalier décroissante, continue à droite.

· $\widehat{S}_{KM}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit.

· On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance.

· Si il y a plusieurs décès au même temps $X_{(i)}$, alors $d_i > 1$ et on a

$$\widehat{S}_{KM}(t) = \prod_{\substack{i=1,\ldots,n \\ X_{(i)} \leq t}} \left(1 - \frac{d_i}{r_i}\right)$$

2.2.2 Estimation de la variance de $\widehat{S}_{KM}(t)$

L'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier est

$$\widehat{Var}\left\{\widehat{S}_{KM}(t)\right\} = \left(\widehat{S}(t)\right)^2 \sum_{i:X_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)}.$$

2.2.3 Absence de biais asymptotique

L'estimateur de Kaplan-Meier est asymptotiquement sans biais :

$$\lim_{n\to\infty} E\left\{\widehat{S}_{KM}(t)\right\} = S(t).$$

2.2.4 Convergence presque sûre (p.s) uniforme

L'estimateur de Kaplan-Meier converge uniformément presque sûrement :

$$\lim_{n\to\infty} \sup_{t<\tau} \left|\widehat{S}_{KM}(t) - S(t)\right| = 0, \quad (p.s).$$

avec $\tau = \inf\{x \geq 0,\ F_X(x) = 1\}$.

2.2.5 La propriété de normalité asymptotique de l'estimateur de Kaplan-Meier.

Théorème 2.2.1 *En tout point de continuité de S, $t_0 \in [0,\tau]$ et $S(\tau^-) > 0$,*

$$\sqrt{n}\left(\widehat{S}_{MK}(t_0) - S(t_0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, V^2(t_0)\right),$$

avec

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{S(du)}{S^2(u)\,G(u)},$$

où $G(t)$ la fonction de survie de la variable C.

Exemple 2.2.1 *Freireich, en 1963, a fait un essai thérapeutique pour comparer les durées de rémission, en semaines, de patients atteints de leucémie selon qu'ils ont reçu ou non un médicament appelé 6 M-P ; le groupe témoin a reçu un placebo.*

6 M-P : $6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$.

placebo : $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$.

Les nombres suivis du signe + correspondent à des données censurées. Les observation $\{(X_i, \delta_i)\}_{i=1,\ldots,21}$ *sont*

$Obs = \{(6,1),(6,1),(6,1),(6,0),(7,1),(9,0),(10,1),(10,0),(11,0),(13,1),$
$(16,1),(17,0),(19,0),(20,0),(22,1),(23,1),(25,0),(32,0),,(32,0),(34,0),(35,0)\}.$

L'estimateur de Kaplan − Meier $\widehat{S}_{KM}(t)$ *est donné par :*

$\widehat{S}_{KM}(t) = 1$ *si* $0 \leq t < 6$

$\widehat{S}_{KM}(t) = \left(1 - \frac{3}{21}\right)\widehat{S}(6^-) = \left(1 - \frac{3}{21}\right)\widehat{S}(0) = \left(1 - \frac{3}{21}\right)1 = 0.857$ *si* $6 \leq t < 7$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{17}\right)\widehat{S}(7^-) = \left(1 - \frac{1}{17}\right)\widehat{S}(6) = \left(1 - \frac{1}{17}\right)0.857 = 0.807$ *si* $7 \leq t < 10$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{15}\right)\widehat{S}(10^-) = \left(1 - \frac{1}{15}\right)\widehat{S}(7) = \left(1 - \frac{1}{15}\right)0.807 = 0.753$ *si* $10 \leq t < 13$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{12}\right)\widehat{S}(13^-) = \left(1 - \frac{1}{12}\right)\widehat{S}(10) = \left(1 - \frac{1}{12}\right)0.753 = 0.690$ *si* $13 \leq t < 16$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{11}\right)\widehat{S}(16^-) = \left(1 - \frac{1}{11}\right)\widehat{S}(13) = \left(1 - \frac{1}{11}\right)0.690 = 0.627 \;\; si\; 16 \leq t < 22$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{7}\right)\widehat{S}(22^-) \;\; = \left(1 - \frac{1}{7}\right)\widehat{S}(16) = \left(1 - \frac{1}{7}\right)0.627 \;\; = 0.538 \; si\; 22 \leq t < 23$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{6}\right)\widehat{S}(23^-) \; = \left(1 - \frac{1}{6}\right)\widehat{S}(22) \; = \left(1 - \frac{1}{6}\right)0.538 \;\; = 0.448 \; si\; 23 \leq t$

t	d	*Censoring*	r	$h = d/r$	*Survival Probability : S*	F
6	*3*	*1*	***21***	$\frac{3}{21}$	$\left(1 - \frac{3}{21}\right)\widehat{S}(0) = \left(1 - \frac{3}{21}\right)1 = 0.857$	0.1430
7	*1*	*0*	***17***	$\frac{1}{17}$	$\left(1 - \frac{1}{17}\right)\widehat{S}(6) \; = \left(1 - \frac{1}{17}\right)0.857 = 0.807$	0.1930
9	*0*	*1*	*16*	**0**	0.807	0.1930
10	*1*	*1*	***15***	$\frac{1}{15}$	$\left(1 - \frac{1}{15}\right)\widehat{S}(7) \;\; = \left(1 - \frac{1}{15}\right)0.807 \;\; = 0.753$	0.2470
11	*0*	*1*	*13*	**0**	0.753	0.2470
13	*1*	*0*	***12***	$\frac{1}{12}$	$\left(1 - \frac{1}{12}\right)\widehat{S}(10) = \left(1 - \frac{1}{12}\right)0.753 = 0.690$	0.3100
16	*1*	*0*	***11***	$\frac{1}{11}$	$\left(1 - \frac{1}{11}\right)\widehat{S}(13) = \left(1 - \frac{1}{11}\right)0.690 = 0.627$	0.3730
17	*0*	*1*	*10*	**0**	0.627	0.3730
19	*0*	*1*	*9*	**0**	0.627	0.3730
20	*0*	*1*	*8*	**0**	0.627	0.3730
22	*1*	*0*	***7***	$\frac{1}{7}$	$\left(1 - \frac{1}{7}\right)\widehat{S}(16) = \left(1 - \frac{1}{7}\right)0.627 = 0.538$	0.4620
23	*1*	*0*	***6***	$\frac{1}{6}$	$\left(1 - \frac{1}{6}\right)\widehat{S}(22) \; = \left(1 - \frac{1}{6}\right)0.538 \; = 0.448$	0.5520
25	*0*	*1*	*5*	**0**	0.448	0.5520
32	*0*	*2*	*4*	**0**	0.448	0.5520
34	*0*	*1*	*2*	**0**	0.448	0.5520
35	*0*	*1*	*1*	**0**	0.448	0.5520
Total	***n=21***		/	/	/	/

pour le groupe traité avec placebo on obtient :

t	d	r	$h = d/r$	*Survival Probability : S*	*cdf : F = 1 − S*
1	*2*	*21*	$\frac{2}{21}$	$\left(1-\frac{2}{21}\right)\widehat{S}(0) = \left(1-\frac{2}{21}\right)1 = 0.9048$	$1-0.9048 = 0.0952$
2	*2*	*19*	$\frac{2}{19}$	$\left(1-\frac{2}{19}\right)\widehat{S}(1) = \left(1-\frac{2}{19}\right)0.9048 = 0.8095$	$1-0.8095 = 0.1905$
3	*1*	*17*	$\frac{1}{17}$	$\left(1-\frac{1}{17}\right)\widehat{S}(2) = \left(1-\frac{1}{17}\right)0.8095 = 0.7619$	$1-0.7619 = 0.2381$
4	*2*	*16*	$\frac{2}{16}$	$\left(1-\frac{2}{16}\right)\widehat{S}(3) = \left(1-\frac{2}{16}\right)0.7619 = 0.6667$	$1-0.6667 = 0.3333$
5	*2*	*14*	$\frac{2}{14}$	$\left(1-\frac{2}{14}\right)\widehat{S}(4) = \left(1-\frac{2}{14}\right)0.6667 = 0.5714$	$1-0.5714 = 0.4286$
8	*4*	*12*	$\frac{4}{12}$	$\left(1-\frac{4}{12}\right)\widehat{S}(5) = \left(1-\frac{4}{12}\right)0.5714 = 0.3810$	$1-0.3810 = 0.6190$
11	*2*	*8*	$\frac{2}{8}$	$\left(1-\frac{2}{8}\right)\widehat{S}(8) = \left(1-\frac{2}{8}\right)0.3810 = 0.2857$	$1-0.2857 = 0.7143$
12	*2*	*6*	$\frac{2}{6}$	$\left(1-\frac{2}{6}\right)\widehat{S}(11) = \left(1-\frac{2}{6}\right)0.2857 = 0.1905$	$1-0.1905 = 0.8095$
15	*1*	*4*	$\frac{1}{4}$	$\left(1-\frac{1}{4}\right)\widehat{S}(12) = \left(1-\frac{1}{4}\right)0.1905 = 0.1429$	$1-0.1429 = 0.8571$
17	*1*	*3*	$\frac{1}{3}$	$\left(1-\frac{1}{3}\right)\widehat{S}(15) = \left(1-\frac{1}{3}\right)0.1429 = 0.0952$	$1-0.0952 = 0.9048$
22	*1*	*2*	$\frac{1}{2}$	$\left(1-\frac{1}{2}\right)\widehat{S}(17) = \left(1-\frac{1}{2}\right)0.0952 = 0.0476$	$1-0.0476 = 0.9524$
23	*1*	*1*	$\frac{1}{1}$	$\left(1-\frac{1}{1}\right)\widehat{S}(22) = 0.0000$	$1-0.0000 = 1$
Total	***n=21***	/	/	/	/

t : ***Failure times***

d : ***Number Failed***

r : ***Number at Risk***

h : ***Hazard rate***

TRAVAUX PRATIQUES

Exemple 2.2.2 *Freireich, en 1963, a fait un essai thérapeutique pour comparer les durées de rémission, en semaines, de patients atteints de leucémie selon qu'ils ont reçu ou non un médicament appelé 6 M-P ; le groupe témoin a reçu un placebo.*

6 M-P : $6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$.

placebo : $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$.

Les nombres suivis du signe + correspondent à des données censurées. Les observation $\{(X_i, \delta_i)\}_{i=1,...,21}$ *sont*

$Obs = \{(6,1), (6,1), (6,1), (6,0), (7,1), (9,0), (10,1), (10,0), (11,0), (13,1),$
$(16,1), (17,0), (19,0), (20,0), (22,1), (23,1), (25,0), (32,0), , (32,0), (34,0), (35,0)\}$.

Algorithme :

```
time(t) =[6  7  9  10  11  13  16  17  19  20  22  23  25  32  34  35];
Number failed (d) =[3  1  0  1  0  1  1  0  0  0  1  1  0  0  0  0];
% Enter 1 for censored data, and enter 0 for exact failure time.
Cens=[0  0  0  1  0  1  0  1  1  0  0  1  1  1  0  0  1  1  1  1  1];
Number at Risk (r) =[21 17  16  15  13  12  11  10  9  8  7  6  5  3  2  1]
y=[6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35]
[F,x] = ecdf(y,'censoring',cens);
S=1-F;
figure()
ecdf(y,'censoring',cens,'function','survivor');
```

Résultats

```
x =
6
6
7
10
13
```

16

22

23

$S =$

1.0000

0.8571

0.8067

0.7529

0.6902

0.6275

0.5378

0.4482

Exemple 2.2.3 *(censured case)*

Lorsque on a des données censurées, la table de mortalité est donnée par :

t	d	*Censoring*	r	$h = d/r$	*Survival Probability : S*	*cdf : F*
4	*2*	*1*	*7*	$\frac{2}{7}$	$1 - \frac{2}{7} = 0.7143$	*0.2857*
7	*1*	*0*	*4*	$\frac{1}{4}$	$\left(1 - \frac{1}{4}\right) 0.7143 = 0.5357$	*0.4643*
11	*1*	*1*	*3*	$\frac{1}{3}$	$\left(1 - \frac{1}{3}\right) 0.5357 = 0.3571$	*0.6429*
12	*1*	*0*	*1*	$\frac{1}{1}$	$\left(1 - \frac{1}{1}\right) 0.3571 = 0$	*1*

t : ***Failure times***

d : ***Number Failed***

r : ***Number at Risk***

h : ***Hazard rate***

t=(4,7,11,12)

Number Failed (d)=(2,1,1,1)

Number at Risk (r)=(7,4,3,1)

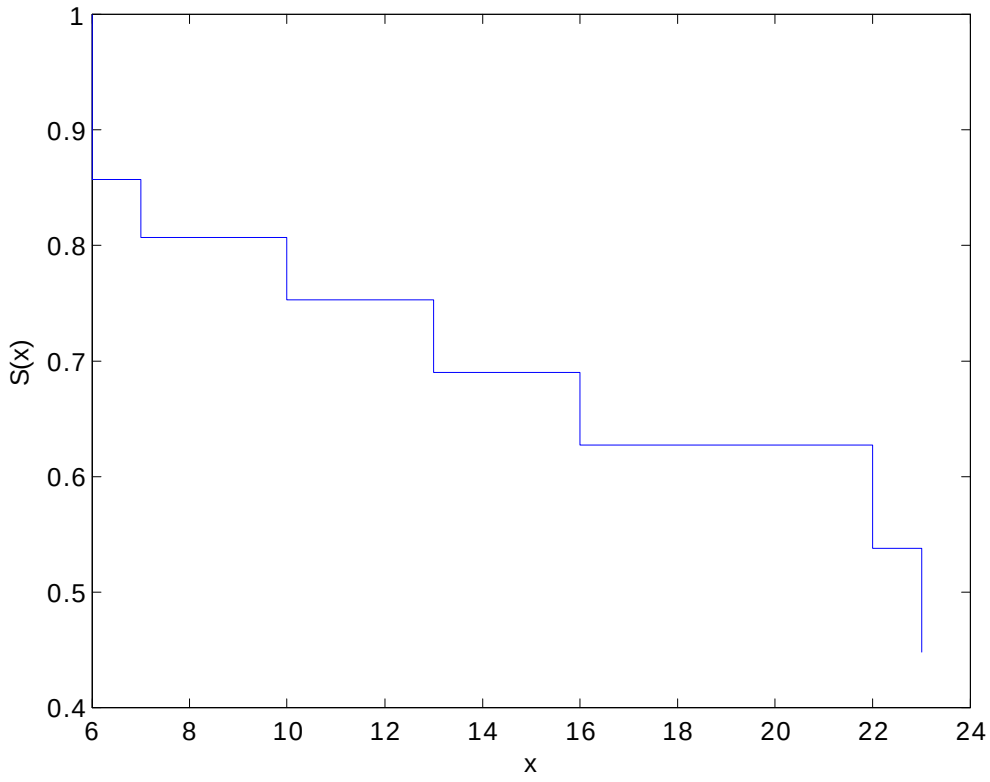


FIG. 2.1 – La fonction de survie dans l'exemple "essai de Freireich 1963".

Algorithme :

```
t= [4  4  4  7  11  11  12];
cens = [0  1  0  0  1  0  0];
[F,x] = ecdf(t,'censoring',cens)
S=1-F;
disp[(x,F,S)]
figure()
ecdf(y,'censoring',cens,'function','survivor');
%figure()
%ecdf(y,'censoring',cens,'function','cumulative hazard');
```

Résultats

x	F	S
4	0.0000	1.0000
4	0.2857	0.7143
7	0.4643	0.5357
11	0.6429	0.3571
12	1.0000	0.0000

-

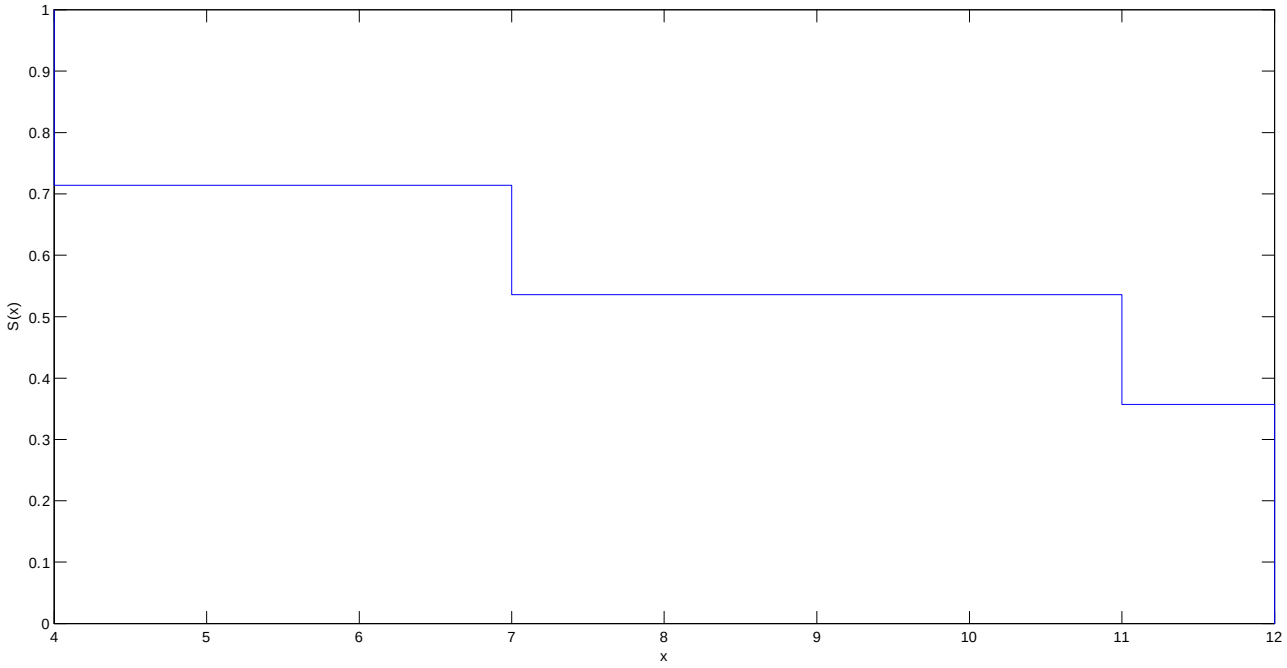


FIG. 2.2 – La fonction de survie S

Exercices

Exercice 1. Soit $Y_1, Y_2, ..., Y_n$ des variables aléatoires $i.i.d.$

Déterminer la loi de $U = \min(Y_1, Y_2, ..., Y_n)$ et la loi de $V = \max(Y_1, Y_2, ..., Y_n)$.

Exercice 2. On observe les durées de séjour de 10 épisodes de chômage exprimés en mois

$$1, 2, 4^+, 5, 7^+, 8, 9, 10^+, 11, 13^+$$

Les nombres suivis du signe + correspondent à des données censurées. Calculer l'estimateur de $Kaplan - Meier$ $\widehat{S}_{KM}(t)$.

Exercice 3. Soit les observations (les durées) suivantes $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$.

Calculer la fonction de survie $S(t)$.

Exercice 4. Soit les observation suivantes $1, 1, 2, 2^+, 3, 4^+, 4, 5, 5, 8^+, 8, 8, 8, 11, 11, 12^+, 12, 15, 17, 22^+, 23$.

Les nombres suivis du signe + correspondent à des données censurées. Calculer l'estimateur de $Kaplan - Meier$ $\widehat{S}_{KM}(t)$.

Solutions

Exercice 1. On a

$$\begin{aligned}
F_U(t) &= P(U \leq t) \\
&= P(\min(Y_1, Y_2, ..., Y_n) \leq t) \\
&= 1 - P(\min(Y_1, Y_2, ..., Y_n) > t) \\
&= 1 - P(Y_1 > t, Y_2 > t, ..., Y_n > t,) \\
&= 1 - P(\{Y_1 > t\} \cap \{Y_2 > t\} \cap ... \cap \{Y_n > t\}) \\
&= 1 - P(Y_1 > t) P(Y_2 > t) ... P(Y_n > t) \qquad (\text{car } Y_1, Y_2, ..., Y_n \text{ son indépendantes}) \\
&\quad 1 - [1 - P(Y_1 \leq t)][1 - P(Y_2 \leq t)] ... [1 - P(Y_n \leq t)] \\
&= 1 - [1 - F_Y(t)][1 - F_Y(t)] ... [1 - F_Y(t)] \quad (\text{car } Y_1, Y_2, ..., Y_n \text{ ont la même distribution } F_Y) \\
&= 1 - [1 - F_Y(t)]^n .
\end{aligned}$$

et

$$\begin{aligned}
F_V(t) &= P(V \leq t) \\
&= P(\max(Y_1, Y_2, ..., Y_n) \leq t) \\
&= P(Y_1 \leq t, Y_2 \leq t, ..., Y_n \leq t,) \\
&= P(\{Y_1 \leq t\} \cap \{Y_2 \leq t\} \cap ... \cap \{Y_n \leq t\}) \\
&= P(Y_1 \leq t) P(Y_2 \leq t) ... P(Y_n \leq t) \qquad (\text{car } Y_1, Y_2, ..., Y_n \text{ son indépendantes}) \\
&= F_Y(t) F_Y(t) .... F_Y(t) \\
&= [F_Y(t)]^n
\end{aligned}$$

Exercice 2. Les observation $\{(X_i, \delta_i)\}_{i=1,\dots,21}$ sont

$$Obs = \{(1,1),(2,1),(4,0),(5,1),(7,0),(8,1),(9,1),(10,0),(11,1),(13,0)\}.$$

L'estimateur de $Kaplan - Meier$ $\widehat{S}_{KM}(t)$

$\widehat{S}_{KM}(t) = 1 \;\; si \;\; 0 \leq t < 1$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{10}\right)\widehat{S}(1^-) \;=\; \left(1 - \frac{1}{10}\right)\widehat{S}(0) \;= 0.9 \;\; si \;\; 1 \leq t < 2$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{9}\right)\widehat{S}(2^-) \;=\; \left(1 - \frac{1}{9}\right)\widehat{S}(1) \;= 0.8 \;\; si \;\; 2 \leq t < 5$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{7}\right)\widehat{S}(5^-) \;=\; \left(1 - \frac{1}{7}\right)\widehat{S}(2) \;= 0.6857 \;\; si \;\; 5 \leq t < 8$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{5}\right)\widehat{S}(8^-) \;=\; \left(1 - \frac{1}{5}\right)\widehat{S}(5) \;= 0.5485 \;\; si \; 8 \leq t < 9$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{4}\right)\widehat{S}(9^-) \;=\; \left(1 - \frac{1}{4}\right)\widehat{S}(8) \;= 0.411 \;\; si \;\; 9 \leq t < 11$

$\widehat{S}_{KM}(t) = \left(1 - \frac{1}{2}\right)\widehat{S}(11^-) \;=\; \left(1 - \frac{1}{2}\right)\widehat{S}(9) \;= 0.205 \;\; si \;\; 11 \leq t$

t	d	Censoring	r	$h = d/r$	Survival Probability : S	cdf : $F = 1 - S$
1	1	0	10	$\frac{1}{10}$	$\left(1 - \frac{1}{10}\right)\widehat{S}(1^-) = \left(1 - \frac{1}{10}\right)\widehat{S}(0) = 0.9$	0.1000
2	1	0	9	$\frac{1}{9}$	$\left(1 - \frac{1}{9}\right)\widehat{S}(2^-) \;=\; \left(1 - \frac{1}{9}\right)\widehat{S}(1) \;= 0.8$	0.2000
4^+	0	1	8	$\frac{0}{8}$	0.8	0.2000
5	1	0	7	$\frac{1}{7}$	$\left(1 - \frac{1}{7}\right)\widehat{S}(5^-) \;=\; \left(1 - \frac{1}{7}\right)\widehat{S}(2) \;= 0.6857$	0.3143
7^+	0	1	6	$\frac{0}{6}$	$0,6857$	0.3143
8	1	0	5	$\frac{1}{5}$	$\left(1 - \frac{1}{5}\right)\widehat{S}(8^-) \;=\; \left(1 - \frac{1}{5}\right)\widehat{S}(5) = 0.5485$	0.4515
9	1	0	4	$\frac{1}{4}$	$\left(1 - \frac{1}{4}\right)\widehat{S}(9^-) \;=\; \left(1 - \frac{1}{4}\right)\widehat{S}(8) \;= 0.411$	0.5890
10^+	0	1	3	$\frac{0}{3}$	$0,411$	0.5890
11	1	0	2	$\frac{1}{2}$	$\left(1 - \frac{1}{2}\right)\widehat{S}(11^-) \;=\; \left(1 - \frac{1}{2}\right)\widehat{S}(9) \;= 0.205$	0.7950
13^+	0	1	1	$\frac{0}{1}$	$0,205$	0.7950
Total	**n=10**		/	/	/	/

$S(0) = 1$

t : **Failure times**

d : **Number Failed**

r : **Number at Risk**

h : **Hazard rate**

Exercice 3. $S(0) = 1$.

t	d	r	$h = d/r$	Survival Probability : S	cdf : $F = 1 - S$
1	2	21	$\frac{2}{21}$	$\left(1 - \frac{2}{21}\right) \widehat{S}(0) = \left(1 - \frac{2}{21}\right) \times 1 = 0.9048$	0.0952
2	2	19	$\frac{2}{19}$	$\left(1 - \frac{2}{19}\right) \widehat{S}(1) = \left(1 - \frac{2}{19}\right) \times 0.9048 = 0.8096$	0.1904
3	1	17	$\frac{1}{17}$	$\left(1 - \frac{1}{17}\right) \widehat{S}(2) = \left(1 - \frac{1}{17}\right) \times 0.8096 = 0.7620$	0.2380
4	2	16	$\frac{2}{16}$	$\left(1 - \frac{2}{16}\right) \widehat{S}(3) = \left(1 - \frac{2}{16}\right) \times 0.7620 = 0.6667$	0.3333
5	2	14	$\frac{2}{14}$	$\left(1 - \frac{2}{14}\right) \widehat{S}(4) = \left(1 - \frac{2}{14}\right) \times 0.6667 = 0.5715$	0.4285
8	4	12	$\frac{4}{12}$	$\left(1 - \frac{4}{12}\right) \widehat{S}(5) = \left(1 - \frac{4}{12}\right) \times 0.5715 = 0.3810$	0.6190
11	2	8	$\frac{2}{8}$	$\left(1 - \frac{2}{8}\right) \widehat{S}(8) = \left(1 - \frac{2}{8}\right) \times 0.3810 = 0.2858$	0.7142
12	2	6	$\frac{2}{6}$	$\left(1 - \frac{2}{6}\right) \widehat{S}(11) = \left(1 - \frac{2}{6}\right) \times 0.2858 = 0.1905$	0.8095
15	1	4	$\frac{1}{4}$	$\left(1 - \frac{1}{4}\right) \widehat{S}(12) = \left(1 - \frac{1}{4}\right) \times 0.1905 = 0.1429$	0.8571
17	1	3	$\frac{1}{3}$	$\left(1 - \frac{1}{3}\right) \widehat{S}(15) = \left(1 - \frac{1}{3}\right) \times 0.1429 = 0.0953$	0.9047
22	1	2	$\frac{1}{2}$	$\left(1 - \frac{1}{2}\right) \widehat{S}(27) = \left(1 - \frac{1}{2}\right) \times 0.0953 = 0.0476$	0.9524
23	1	1	$\frac{1}{1}$	$\left(1 - \frac{1}{1}\right) \widehat{S}(22) = \left(1 - \frac{1}{1}\right) \times 0.0476 = 0$	1
Total	**n=21**		/	/	/

t : **Failure times**

d : **Number Failed**

r : **Number at Risk**

h : **Hazard rate**

Exercice 4.

t	d	Censoring	r	$h = d/r$	Survival Probability : S	cdf : $F = 1 - S$
1	2	0	21	$\frac{2}{21}$	$\left(1-\frac{2}{21}\right)\widehat{S}(0) = \left(1-\frac{2}{21}\right) \times 1 = 0.9048$	0.0952
2	1	1	19	$\frac{1}{19}$	$\left(1-\frac{1}{19}\right)\widehat{S}(1) = \left(1-\frac{1}{19}\right) \times 0.9048 = 0.8572$	0.1428
3	1	0	17	$\frac{1}{17}$	$\left(1-\frac{1}{17}\right)\widehat{S}(2) = \left(1-\frac{1}{17}\right) \times 0.8572 = 0.8068$	0.1932
4	1	1	16	$\frac{1}{16}$	$\left(1-\frac{1}{16}\right)\widehat{S}(3) = \left(1-\frac{1}{16}\right) \times 0.8068 = 0.7564$	0.2436
5	2	0	14	$\frac{2}{14}$	$\left(1-\frac{2}{14}\right)\widehat{S}(4) = \left(1-\frac{2}{14}\right) \times 0.7564 = 0.6483$	0.3517
8	3	1	12	$\frac{3}{12}$	$\left(1-\frac{3}{12}\right)\widehat{S}(5) = \left(1-\frac{3}{12}\right) \times 0.6483 = 0.4862$	0.5138
11	2	0	8	$\frac{2}{8}$	$\left(1-\frac{2}{8}\right)\widehat{S}(8) = \left(1-\frac{2}{8}\right) \times 0.4862 = 0.3647$	0.6353
12	1	1	6	$\frac{1}{6}$	$\left(1-\frac{1}{6}\right)\widehat{S}(11) = \left(1-\frac{1}{6}\right) \times 0.3647 = 0.3039$	0.6961
15	1	0	4	$\frac{1}{4}$	$\left(1-\frac{1}{4}\right)\widehat{S}(12) = \left(1-\frac{1}{4}\right) \times 0.3039 = 0.2279$	0.7721
17	1	0	3	$\frac{1}{3}$	$\left(1-\frac{1}{3}\right)\widehat{S}(15) = \left(1-\frac{1}{3}\right) \times 0.2279 = 0.1519$	0.8481
22	0	1	2	$\frac{0}{2}$	$\left(1-\frac{0}{2}\right)\widehat{S}(27) = \left(1-\frac{0}{2}\right) \times 0.1519 = 0.1519$	0.8481
23	1	0	1	$\frac{1}{1}$	$\left(1-\frac{1}{1}\right)\widehat{S}(22) = \left(1-\frac{1}{1}\right) \times 0.1519 = 0$	1
Total	**n=21**			/	/	/

t : **Failure times**

d : **Number Failed**

r : **Number at Risk**

h : **Hazard rate**

Chapitre 3

Table de mortalité et lissage

Une table de mortalité (aussi appelée table de survie) est une construction qui permet de suivre minutieusement le destin d'une population. Cet outil est surtout utilisé en démographie et en actuariat afin d'étudier le nombre de décès, les probabilités de décès ou de survie et l'espérance de vie selon l'âge et le sexe. Il existe deux types de tables de mortalité : **la table de mortalité du moment** (ou **transversales)** et **la table de mortalité par génération.**

Les tables de mortalité utilisées par les assureurs pour leurs tarifs et leurs provisions. Du point de vue de l'assureur, on peut distinguer les tables **réglementaires**, qui jouent un rôle particulier dans la détermination du tarif et des provisions, et les tables **d'expérience** ; d'un point de vue technique, on distingue les tables **transversales**, ou "tables du **moment**" et les tables **prospectives**, intégrant l'aspect dynamique de la mortalité.

3.1 L'analyse de la mortalité

On s'intéresse à la variable aléatoire T représentant la durée de vie d'un individu, on suppose les individus de la population dans un premier temps identiques, de sorte qu'on pourra disposer d'échantillons issus de la loi de T.

3.1.1 Notations et définitions

- F : **La fonction de répartition** de T :

$$F(t) = P(T \leq t).$$

- S : **La fonctionde de survie de** e T :

$$S(t) = 1 - F(t) = P(T > t).$$

- T_x : représentant **la durée de vie résiduelle** d'un individu conditionnellement au fait qu'il soit vivant à l'âge $x \geq 0$, i.e.

$$T_x \overset{déf}{=} [T - x | T > x].$$

- F_x : **La fonction de répartition** de T_x :

$$F_x(t) = P(T_x \leq t).$$

- S_x : **La fonctionde de survie de** e T_x :

$$S_x(t) = 1 - F_x(t) = P(T_x > t).$$

- On peut obtenir :

$$P(T_x \leq t) = \frac{P(x < T \leq x + t)}{P(T > x)},$$

alors

$$F_x(t) = \frac{F(x + t) - F(x)}{S(x)}.$$

Ainsi on utilise $S_x(t) = 1 - F_x(t)$,

$$S_x(t) = \frac{S(x + t)}{S(x)},$$

et on peut écrir

$$S(x + t) = S(x) S_x(t).$$

- $p_{(x,t)}$: **la probabilité de survie** entre x et $x + t$:

$$p_{(x,t)} = P(T_x > t) = P(T > x + t | T > x),$$

- $q_{(x,t)}$: **le quotient de mortalité** entre x et $x + t$:

$$q_{(x,t)} = 1 - p_{(x,t)} = P(T_x \leq t) = P(T \leq x + t | T > x)$$

• Lorsque $t = 1$, on écrit plus simplement :

$$p_x = p_{(x,t)} \quad \text{et} \quad q_x = q_{(x,t)}.$$

• Les quotients $p_{(x,t)}$ et $q_{(x,t)}$ s'expriment simplement à l'aide de la fonction de survie de T :

$$p_{(x,t)} = \frac{S(x+t)}{S(x)} \quad \text{et} \quad q_{(x,t)} = 1 - \frac{S(x+t)}{S(x)}$$

• $d_{(x,t)}$: **le nombre de décès entre** x et $x + t$:

$$d_{(x,t)} = I_x - I_{x+t},$$

avec $I_x = S(x)$.

• $L_{(x,t)}$: **l'effectif moyen sur la période,** dans le cadre de l'analyse statistique de la mortalité d'une **cohorte** (groupe) on mesure **le temps vécu par les individus de la cohorte** entre x et $x + t$, défini par

$$L_{(x,t)} = \int_0^t I_{x+u} du.$$

• E_x : **la durée de vie résiduelle**, qui est un indicateur caractéristique de la table de mortalité :

$$E_x = \int_0^\infty I_{x+u} du = \sum_{t=1}^{\infty} L_{(x,t)}$$

• Le quotient de mortalité $q_{(x,t)}$ est calculé en rapportant **un nombre de décès sur la période** $d_{(x,t)}$ à **l'effectif en début de période** I_x, i.e. :

$$q_{(x,t)} = P(T \leq x + t | T > x) = \frac{S(x) - S(x+t)}{S(x)} = \frac{I_x - I_{x+t}}{I_x} = \frac{d_{(x,t)}}{I_x}$$

• $m_{(x,t)}$: **le taux de mortalité**, obtenu en rapportant **le nombre de décès** à **l'effectif moyen sur la période**, soit :

$$m_{(x,t)} = \frac{d_{(x,t)}}{L_{(x,t)}}.$$

• h_{x+t} : **le taux instantané de mortalité** :

$$h_{x+t} = \lim_{\Delta t \to 0} \frac{P(t < T_x \leq t + \Delta t | T_x > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{q_{(x,t+\Delta t)} - q_{(x,t)}}{p_{(x,t)}\ \Delta t} = \frac{1}{p_{(x,t)}} \frac{dq_{(x,t)}}{dt} = -\frac{1}{p_{(x,t)}} \frac{dp_{(x,t)}}{dt},$$

ce qui implique

$$p_{(x,t)}\ h_{x+t} = -\frac{dp_{(x,t)}}{dt} = f_x(t),$$

donc lorsque Δt est petit,

$$q_{(x,\Delta t)} \simeq \Delta t \; h_x \quad \text{et} \quad p_{(x,\Delta t)} \simeq 1 - \Delta t \; h_x \; . \textbf{? ? ?}$$

Le lien entre **le taux instantané de mortalité** et **le taux de mortalité** est direct :

$$h_x = \lim_{\Delta t \to 0} m_{(x,\Delta t)}.$$

• La relation entre **fonction de survie conditionnelle** et **fonction de hasard** s'écrit avec les notations utilisées ici :

$$p_{(x,t)} = \exp\left(-\int_0^t h_{x+s} ds \right).$$

• F_x la distribution de T_x et f_x sa densité de probabilité, donc

$$F_x(t) = \int_0^t f_x(s) ds,$$

et on peut écrir

$$q_{(x,t)} = \int_0^t p_{(x,s)} \; h_{x+s} ds.$$

• Dans le cas $t = 1$, on obtient

$$q_x = \int_0^1 p_{(x,s)} \; h_{x+s} ds.$$

Quand q_x est petit, $p_x \approx 1$ donc $p_{(x,s)} \approx 1$ pour $0 \leq s \leq 1$, alors

$$q_x \approx \int_0^1 h_{x+s} ds \approx h_{x+\frac{1}{2}},$$

la relation est obtenue par l'application de la méthode "**the mid-point rule**" pour l'intégration numérique.

• **La moyenne et l'écart-type :**

· $e_x^0 = E\{T_x\}$: la moyenne de T, on a

$$p_{(x,t)} \; h_{x+t} = -\frac{dp_{(x,t)}}{dt} = f_x(t),$$

et d'après la définition de l'espérance mathématiques, on a

$$e_x^0 = \int_0^\infty t f_x(t)\, dt = \int_0^\infty t\ p_{(x,t)}\ h_{x+t} dt.$$

On utilise la relation $p_{(x,t)}\ h_{x+t} = -\frac{dp_{(x,t)}}{dt} = f_x(t)$, et par intégration par parties nous obtenons

$$\begin{aligned} e_x^0 &= -\int_0^\infty t \left(\frac{dp_{(x,t)}}{dt}\right) dt \\ &= -\left(\left[t\ p_{(x,t)}\right]_0^\infty - \int_0^\infty p_{(x,t)} dt\right). \end{aligned}$$

Sous l'hypothèse $\lim_{t\to\infty} t p_{(x,t)} = 0$, on a

$$e_x^0 = \int_0^\infty p_{(x,t)} dt.$$

• De même pour $E\left\{T_x^2\right\}$, on a

$$E\left\{T_x^2\right\} = 2\int_0^\infty t\ p_{(x,t)} dt.$$

La variance de T_x est évaluée par

$$Var\left\{T_x\right\} = E\left\{T_x^2\right\} - \left(E\left\{T_x\right\}\right)^2 = E\left\{T_x^2\right\} - \left(e_x^0\right)^2,$$

et l'écart-type σ_x de T_x est donné par

$$\sigma_x = \sqrt{Var\left\{T_x\right\}}.$$

• K_x : "The curtate future lifetime random variable" est définie par

$$K_x = [T_x],$$

avec $t \longmapsto [t]$ est la fonction de la partie entière de t.

Alors pour les entiers $k = 0, 1, 2, 3, ...$, on a

$$\begin{aligned} P(K_x = k) &= P(k \le T_x < k+1) \\ &= q_{(x,k)} \\ &= p_{(x,k)} - p_{(x,k+1)} \\ &= p_{(x,k)} - p_{(x,k)} p_{x+k} \\ &= p_{(x,k)} q_{x+k} \end{aligned}$$

e_x : L'espérance mathématiques de K_x est donnée par

$$\begin{aligned}
e_x &= E\{K_x\} \\
&= \sum_{k=0}^{\infty} k\ P(K_x = k) = \sum_{k=0}^{\infty} k\ \left(p_{(x,k)} - p_{(x,k+1)}\right) \\
&= \sum_{k=1}^{\infty} k\ p_{(x,k)} - \sum_{k=1}^{\infty} k\ p_{(x,k+1)} = \sum_{k=1}^{\infty} k\ p_{(x,k)} - \sum_{k=2}^{\infty} (k-1)\ p_{(x,k)} \\
&= p_{(x,1)} + \sum_{k=2}^{\infty} k\ p_{(x,k)} - \sum_{k=2}^{\infty} (k-1)\ p_{(x,k)} = p_{(x,1)} + \sum_{k=2}^{\infty} (k\ -(k-1))\ p_{(x,k)} \\
&= p_{(x,1)} + \sum_{k=2}^{\infty} p_{(x,k)} \\
&= \sum_{k=1}^{\infty} p_{(x,k)}.
\end{aligned}$$

De même pour $E\left\{K_x^2\right\}$, on peut obtenir

$$\begin{aligned}
E\left\{K_x^2\right\} &= \sum_{k=0}^{\infty} k\ P(K_x = k) \\
&= \sum_{k=0}^{\infty} k^2\ \left(p_{(x,k)} - p_{(x,k+1)}\right) \\
&= 2\sum_{k=1}^{\infty} k\ p_{(x,k)} - e_x.
\end{aligned}$$

• **La relation entre e_x^0 et e_x**: On a

$$e_x^0 = \int_0^{\infty} p_{(x,t)}dt = \sum_{j=0}^{\infty} \int_j^{j+1} p_{(x,t)}dt,$$

par l'application de la méthode de trapez " **the trapezium rule**" pour l'intégration numérique, on obtient

$$\int_j^{j+1} p_{(x,t)}dt \approx \frac{1}{2}\left(p_{(x,j)} + p_{(x,j+1)}\right),$$

alors

$$\begin{aligned}
e_x^0 &\approx \sum_{j=0}^{\infty} \frac{1}{2}\left(p_{(x,j)} + p_{(x,j+1)}\right) = \frac{1}{2}p_{(x,0)} + \frac{1}{2}\left(\sum_{j=1}^{\infty} p_{(x,j)} + \sum_{j=0}^{\infty} p_{(x,j+1)}\right) \\
&\approx \frac{1}{2} + \frac{1}{2}\left(\sum_{j=1}^{\infty} p_{(x,j)} + \sum_{j=1}^{\infty} p_{(x,j)}\right) = \frac{1}{2} + \sum_{j=1}^{\infty} p_{(x,j)},
\end{aligned}$$

et on obtient une approximation utilisée en pratique

$$e_x^0 \approx \frac{1}{2} + e_x.$$

3.1.2 Table de mortalité

• Avec un modèle de survie donné et de probabilté de survie $p_{(x,t)}$, on peut construire une table de mortalité (ou de survie) du modèle pour certain âge initial x_0 ($x_0 = 0$ en général).

• l_{x_0} : **"Radix"** de la table de survie. On commence par l_{x_0} nombre initial de la cohorte (e.g. 100 000).

• l_x : le nombre d'individus qu'ils soient vivant à l'âge x, on a la définition:

$$l_{x_0+t} = l_{x_0}\ S_{x_0}(t) = l_{x_0}\ p_{(x_0,t)},$$

et pour $x_0 \leq x \leq x+t$, on on peut déduire

$$\begin{aligned}
l_{x+t} &= l_{x_0}\ p_{(x_0,x+t-x_0)} \\
&= l_{x_0}\ S_{x_0}(x+t-x_0) \\
&= l_{x_0}\ P(T_{x_0} > x+t-x_0) \\
&= l_{x_0}\ P(T_{x_0} > x-x_0, T_{x_0} > x+t-x_0) \\
&= l_{x_0}\ P(T_{x_0} > x-x_0, T_x > t) \\
&= l_{x_0}\ P(T_{x_0} > x-x_0)\ P(T_x > t) \\
&= l_{x_0}\ S_{x_0}(x-x_0)\ S_x(t) \\
&= \underbrace{l_{x_0}\ p_{(x_0,x-x_0)}}\ p_{(x,t)} \\
&= l_x\ p_{(x,t)},
\end{aligned}$$

ce qui implique

$$p_{(x,t)} = \frac{l_{x+t}}{l_x}.$$

• Le nombre de survivant à l'âge $x+t$ est une variable aléatoire Binômiale de paramètres l_x et $p_{(x,t)}$ notée $\mathcal{L}_t$, i.e.

$$\mathcal{L}_t \rightsquigarrow B\left(l_x, p_{(x,t)}\right).$$

L'espérance du nombre de survivant à l'âge $x+t$ est

$$E\{\mathcal{L}_t\} = l_x \; p_{(x,t)} = l_{x+t}.$$

• Soit d_x : le nombre de décès entre x et $x+1$, on a alors

$$\begin{aligned} d_x &= l_x - l_{x+1} \\ &= l_x \left(1 - \frac{l_{x+1}}{l_x}\right) \\ &= l_x \left(1 - p_x\right) \\ &= l_x \; q_x. \end{aligned}$$

• **Hypothèse d'âge fractionnaire :**

Trois hypothèses pour l'âge fractionnaire sont :

· $\mathcal{H}_1$: **La distribution uniforme de décès** (**UUD**).

· $\mathcal{H}_2$: **Le taux instantané de mortalité constant** (**CF**).

· $\mathcal{H}_3$: **Balducc** (**Bal**).

∘$\mathcal{H}_1$ (**UUD**) : **La distribution uniforme de décès (L'hypothèse Linéaire)**:

· **UUD1 :** For integer x, et pour $0 \leq s \leq 1$, assume that

$$q_{(x,s)} = sq_x.$$

· **UUD2** : on a $K_x = [T_x]$, on définit la variable aléatoire R_x tel que:

$$T_x = K_x + R_x.$$

L'hypothèse **UUD2** est donnée par :

$$\text{pour l'entier } x, \; R_x \rightsquigarrow \mathcal{U}(0,1) \quad \text{et} \quad R_x \; \text{ est indépendantes de } \; K_x.$$

Alors, sous l'hypothèse **UUD1,** pour l'entier x, et pour $0 \leq s \leq 1$,

$$\begin{aligned} P(R_x \leq s) &= \sum_{k=0}^{\infty} P(R_x \leq s \text{ and } K_x = k) \\ &= \sum_{k=0}^{\infty} P(k \leq T_x \leq k+s) \\ &= \sum_{k=0}^{\infty} p_{(x,k)}\ q_{(x+k,s)} \\ &= \sum_{k=0}^{\infty} p_{(x,k)}\ sq_{x+k} \quad (\text{ on utilise } \mathbf{UUD1}) \\ &= s\sum_{k=0}^{\infty} p_{(x,k)}\ q_{x+k} \\ &= s\sum_{k=0}^{\infty} P(T_x = k) \\ &= s, \end{aligned}$$

ce qui prouve $R_x \rightsquigarrow \mathcal{U}(0,1)$, et pour prouve l'indépendnce de R_x et K_x on a

$$\begin{aligned} P(R_x \leq s \text{ and } K_x = k) &= P(k \leq T_x \leq k+s) \\ &= p_{(x,k)}\ q_{(x+k,s)} \\ &= s\ p_{(x,k)}\ q_{x+k} \\ &= P(R_x \leq s)\ P(K_x = k), \end{aligned}$$

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

Maintenant, pour prouver l'inverse: **UUD2$\Longrightarrow$UUD1,** on a pour l'entier x, et pour $0 \leq s \leq 1$,

$$\begin{aligned} q_{(x,s)} &= P(T_x \leq s) \\ &= P(R_x \leq s \text{ and } K_x = 0\ ) \\ &= P(R_x \leq s)\ P(K_x = 0) \\ &= s\ q_x. \end{aligned}$$

· Les résultas de **UUD1** et **UUD2** :

◦ On a on a pour l'entier x, et pour $0 \leq s \leq 1$,

$$\begin{aligned} q_{(x,s)} &= 1 - \frac{l_{x+s}}{l_x} \\ &= \frac{l_x - l_{x+s}}{l_x}, \end{aligned}$$

et par **UUD1,**

$$q_{(x,s)} = sq_x = s\frac{d_x}{I_x},$$

ce qui implique

$$\begin{aligned} l_{x+s} &= l_x - sd_x \\ &= l_x - s\left(l_x - l_{x+1}\right) \\ &= \left(1-s\right)l_x + sl_{x+1}. \end{aligned}$$

◦ Sous l'hypothèse **UUD1,** on a

$$\frac{d}{ds}q_{(x,s)} = q_x,\ 0 \leq s \leq 1,$$

et d'autre part

$$\begin{aligned} \frac{d}{ds}q_{(x,s)} &= \frac{d}{ds}P\left(T_x \leq s\right) \\ &= \frac{d}{ds}F_x\left(s\right) \\ &= f_x\left(s\right) = p_{(x,s)}\ h_{x+s}, \end{aligned}$$

ce qui implique que

$$f_x\left(s\right) \quad \text{est constante pour } s \in [0,1]\,,$$

et

$$q_x = p_{(x,s)}\ h_{x+s},$$

pour $0 \leq s \leq 1$.

Comme q_x est constant pour s et $p_{(x,s)}$ est décroissante pour s, on peut déduire que h_{x+s} est croissante pour s.

L'hypothèse la distribution uniforme de R_x implique aussi $f_x\left(s\right)$ est constante pour $s \in [0,1]\,.$

$\circ\mathcal{H}_2$ (**CF**) : **Le taux instantané de mortalité constant (L'hypothèse Exponentielle) :**

· Pour l'entier x, et pour $0 \leq s \leq 1$, on suppose que h_{x+s} ne dépond pas de s et on note $h_{x+s} = h_x^*$.

· Pour obtenir la valeur de h_x^*, on utilise

$$p_x = \exp\left(-\int_0^1 h_{x+s}ds\right),$$

si on remplace h_{x+s} par h_x^*, on obtient

$$p_x = e^{-h_x^*}.$$

· Sous l'yhpothèse "Le taux instantané de mortalité constant", on obtient pour l'entier x, et pour $0 \leq s \leq 1$,

$$p_{(x,s)} = \exp\left(-\int_0^s h_x^* ds\right) = e^{-sh_x^*} = (p_x)^s,$$

et pour $t, s > 0$ avec $t + s < 1$, on a

$$p_{(x+t,s)} = \exp\left(-\int_0^s h_x^* ds\right) = e^{-sh_x^*} = (p_x)^s.$$

$\circ\mathcal{H}_3$ (**Bal**) : **Balducci (L'hypothèse hyperbolique) :**

· Cette hypothèse (**Bal**) suppose que l'inverse de l_{x+s} est linéaire en fonction de s :

$$\frac{1}{l_{x+s}} = \frac{1}{l_x} - s\left(\frac{1}{l_x} - \frac{1}{l_{x+1}}\right), \quad 0 \leq s < 1.$$

· Sous cette hypothèse, on peut obtenir les relations suivantes :

$$q_{(x,s)} = \frac{sq_x}{1-(1-s)\,q_x} \quad \text{et} \quad q_{(x+s,1-s)} = (1-s)\,q_x,$$

car on a d'après l'hypothèse $\mathcal{H}_3$,

$$\frac{l_{x+1}}{l_{x+s}} = 1 + (s-1)\,q_x \;\text{ et }\; q_{(x,s)} = \frac{sq_x}{\frac{l_{x+1}}{l_{x+s}}} = \frac{sq_x}{1-(1-s)\,q_x},$$

et

$$\frac{l_{x+1}}{l_{x+s+1}} = 1 + sq_x \text{ et } q_{x+s} = \frac{q_x}{\frac{l_{x+1}}{l_{x+s+1}}} = \frac{q_x}{1+sq_x} \text{et } \; q_{(x+s,1-s)} = \frac{(1-s)\,q_{x+s}}{1-sq_{x+s}} = (1-s)\,q_x.$$

3.2 Exemples

3.3 Diagramme de Lexis

• Le diagramme de Lexis est un outil, utilisé essentiellement en démographie et en actuariat, qui permet de localiser, sur une figure à deux axes, des événements (naissances, décès...) et des effectifs de population en fonction du temps (âge, période, génération).

• Sur le diagramme de Lexis, l'axe horizontal est dévolu au temps et l'axe vertical, à l'âge. Pour repérer un événement sur ce diagramme, il suffit donc d'en connaitre les coordonnées en termes de date et d'âge.

• Dans les études de mortalité, les informations sur l'âge aux décès et les dates de décès ne sont pas exactes en général, ces informations et ces données sont le plus souvent disponible sous forme arrondie, en âge entier et année entière, alors on utilise le diagramme de Lexis pour déterminer correctement les taux bruts de mortalité.

• L'analyse de la mortalité d'un groupe donée fait intervenir trois mesures de temps, ces mesures sont des dimension importants dans la détermination du niveau de la mortalité et influent sur le risque de décès :

- L'âge des individus : cette variable influe le risque de décès.
- La date d'observation : le risque de décès peut varier en fonction d'une épidémie, ...
- La génération (la date de naissance) : amélioration des conditions sanitaires, les progrès de la médecine,.... influent le risque de mortalité à un âge donné.

3.3.1 Figures et interprétation

• Figure 3.1 : Les points mortuaires qui se situent dans le carré associés aux décès à l'âge x au cours de l'année t :

· $[x, x+1[$: Les individus à l'âge x.

· Les individus concernés appartient aux générations $t-x$ et $t-x-1$.

• Figure 3.2 : Les décès se sont produits au cours des années $g+x$ et $g+x+1$.

• Figure 3.3 : La représentation du nombre de décès au cours de l'année t parmi les individus de la génération g.

• Figure 3.4 : Le nombre de décès à l'âge x parmi la génération g au cours de l'anné t.

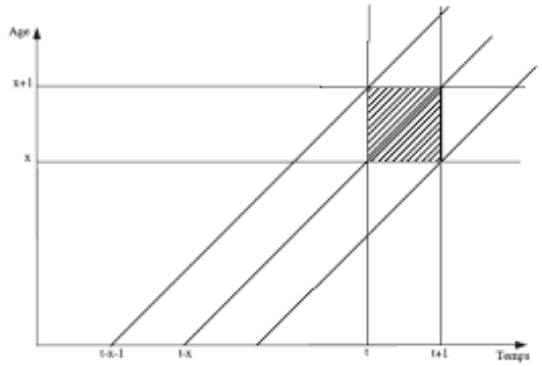


FIG. 3.1 – Identification des décès à l'âge x l'année t.

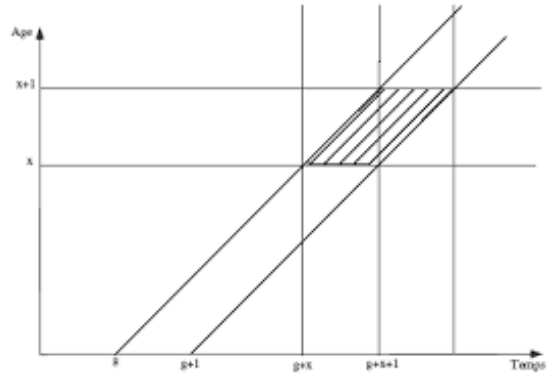


FIG. 3.2 – Identification des décès à l'âge x dans la génération g.

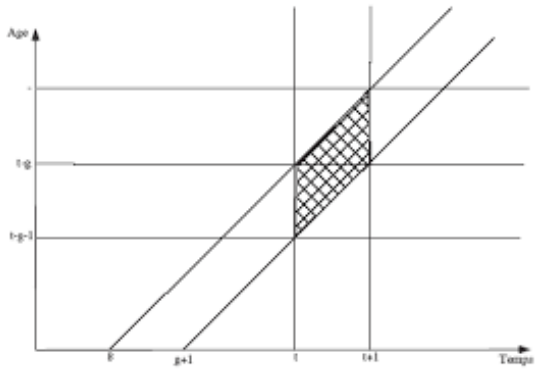


FIG. 3.3 – Identification des décès dans la génération g l'année t.

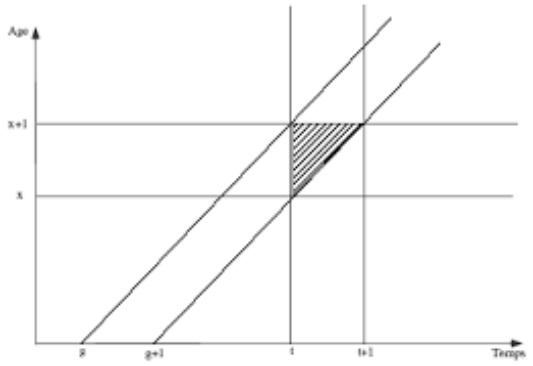


FIG. 3.4 – Identification des décès à l'âge x l'année t dans la génération g.

Bibliographie

[1] Cox, D.R. et Oakes, D. (1984). Analysis of survival data, Chapman and Hall, New York.

[2] Hougaard, P. (2000). Analysis of multivariate survival data. Springer, New York.

[3] Klein, J.P. et Moeschberger, M.L. (1997). Survival analysis, techniques for censored and truncated data, Springer, New York.

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

[5] Frédéric Planchet. (2021). Modèles de durée-Tables de mortalité (Support de cours 2020-2021). Institut de science financière et d'assurance, Lyon, France.

[6] Frédéric Planchet. (2021). Modèles de durée-Introduction (Support de cours 2020-2021). Institut de science financière et d'assurance, Lyon, France.

[7] Frédéric Planchet. (2021). Modèles de durée-Statistique des modèles paramétriques et semi-paramétriques (Support de cours 2020-2021). Institut de science financière et d'assurance, Lyon, France.

[8] Frédéric Planchet. (2021). Modèles de durée-Statistique des modèles non paramétriques (Support de cours 2020-2021). Institut de science financière et d'assurance, Lyon, France.

[9] Frédéric Planchet. (2021). Modèles de durée-Méthodes de lissage et d'ajustement (Support de cours 2020-2021). Institut de science financière et d'assurance, Lyon, France.

[10] Loïc Desquilbet, (2020). Introduction à l'analyse de survie. Ecole nationale vétérinaire d'Alfort. Maisons-Alfort, France.

[11] Frédéric PLANCHET, Laurent FAUCILLON et Marc JUILLARD. (2006). QUANTIFICATION DU RISQUE SYSTEMATIQUE DE MORTALITE POUR UN REGIME DE RENTES EN COURS DE SERVICE. ISFA-Laboratoire SAF, Université Claude Bernard- Lyon 1. France.

[12] Laureline HERBAUT. (2017). Étude de la mortalité des personnes victimes d'accidents médicaux. Mémoire présenté devant l'Institut des Actuaires.

[13] Lyasmine HARROUCHE. (2018). Analyse statistique des modèles de survie (Mémoire de Master). UNIVERSITE MOULOUD MAMMERI de TIZI-OUZOU, Algérie.

[14] LEILA BOURMOUCHE. (2016). MODÈLES MULTI-ÉTATS MARKOVIENS EN ANALYSE DE SURVIE (Mémoire). UNIVERSITÉ DU QUÉBEC À MONTRÉAL, Canada.