Table des matières

Remerciement

Je remercie en premier lieu mon dieu qui a bien voulu me donner la force pour effectuer le présent travail.

Mes remerciements vont second lieu à M^{elle} F. Maref, pour avoir accept l'encadreur de mon travail, pour la confiance qu'elle nous accordones a réaliser ce projet ainsi que pour sa grande attention et sa patience tout au long de ce travail.

Je tiens à exprimer ma reconnaissance aux membres du jury de m.avoir fait l'honneur de participer à ma soutenance.

Je remercie vivement tous les enseignats du département mathématiques qui ont contribué dans mon formation.

Je me permets également de remercier mes parents, mon frère pour leur soutient moral et leur encouragement tout au long de mes études.

Introduction

L'analyse de données de survie constitue un domaine de la statistique qui s'intéresse à mesurer le temps jusqu'à un événement particulier, souvent appelé temps d'échec, ou temps de survie. Les applications de telles analyses sont multiples, nous pouvons citer comme exemples de temps d'échec la durée de fonctionnement de pièces avant une défaillance en fiabilité in dustrielle, la durée de grèves ou de périodes de non-emploi en économie, la durée de vie de patients lors d'essais cliniques. Ce type d'analyse est particulièrement utile dans la recherche biomédicale. Des modèles peuvent être construits pour essayer de mieux comprendre le développement de certaines maladies. L'analyse de données de survie permet également d'évaluer l'efficacité de divers traitements ou la résistance du patient face à une maladie, en observant par exemple le temps entre le début d'un traitement et la guérison, ou le temps avant une rechute.

Une des difficultés principales dans l'analyse de survie réside dans le fait que l'on ne connaît pas nécessairement le temps d'échec de tous les individus. Lors d'essais médicamenteux par exemple, l'évolution de la maladie est observée pendant une certaine période, cependant il est probable que la plupart des patients soient toujours en vie à la fin de l'étude. Pour ceux-ci, on ne connaîtra donc pas le temps de mort. De telles données sont dites censurées.

L'analyse de la survie est née au vingtième siècle, et a connu un développement important dans la seconde moitié du siècle. En 1951, Weibull conçoit un modèle paramétrique dans le domaine de la fiabilité, à cet effet, il fournit une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de la survie : la « loi de Weibull ». En 1958, Kaplan et Meier présentent d'importants résultats concernant l'estimation nonparamétrique de la fonction de survie, ils étudient l'espérance, la variance et les propriétés asymptotiques. Mantel (1966), a étudié la statistique du log-rank pour comparer deux distributions de survie.

Dans ce mémoire nous présentons une étude sur les modèles de survie en abordant principalement l'estimation de la fonction de survie. Les méthodes d'estimation de ces fonctions sont de type non paramétriques et sont basées sur les estimateurs à noyau et l'estimateur de Kaplan-Meier.

Ce mémoire est présenté en trois chapitre. Dans le premier chapitre, nous rappelons des préliminaires sur les modèles de survie. Nous introduisons les principales fonctions en analyse de survie : fonction de survie, fonction de répatition par et les différentes formes du taux de risque ect.... Nous donnons aussi les différents types de censure (censure à droite, censure à gauche, censure par intervalle, troncature ect....). Dans le deuxième chapitre, en première partie, nous rappelons les résultats de l'estimation de la fonction de survie S et l'estimation de la densité f des durées T dans le cas de données non censurées par la méthode du noyau en se basant essentiellement sur le livre de Bosq [1]. Dans la deuxième partie, nous étudions l'estimation de la fonction de survie S des données T dans le cas de la censure à droite. Cet estimateur est construit sur l'estimateur de Kaplan-Meier. Il est également connu sous le nom de l'estimateur produit-limite. En recherche médicale, il est souvent utilisé pour mesurer la fraction de patients en vie pour une certaine durée après leur traitement. Il est également utilisé en économie et en écologie. Dans le troisième chapitre nous nous intéressons essentiellement aux tests non-paramétriques visés, dans le cas de données censurées à droite, Des exemples d'illustration des méthodes étudiées complètent l'étude théorique aussi bien pour des données réelles.

Chapitre 1

Généralités

L'analyse des données de survie s'intéresse à l'étude statistique des données qui proviennent des expériences sur des durées de vie ou durée de fonctionnement. La durée de survie désigne le temps qui s'ecoule depuis un instant initial (début du traitement, diagnostic,...) jusqu'à la survenue d'un évènement d'intérêt final (décès du patient, rechute, rémission, guérison, ...). Dans ce chapitre, nous allons ouvrir une parenthèse assez rapide pour rappeler quelques définitions utilisés le long de ce mémoire.

1.1 Définitions

Quelques définitions sont couramment utilisées dans les études de survie.

- Date d'origine : elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine diférente (pas important car c'est la durée qui nous intérèsse).

- Date de point : c'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

- Date des dernières nouvelles : c'est la date la plus récente où des informations sur un sujet ont été recueillies.

1.2 Distributions de la durée de survie

Supposons que la durée de survie X 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) :

1.2.1 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

$$S(t) = \mathbb{P}(X > t), t \geq 0.$$

1.2.2 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

$$F(t) = \mathbb{P}(X \leq t) = 1 - S(t).$$

Remarque 1.2.1. *Il est arbitraire de décider que $S(t) = \mathbb{P}(X \geq t)$ ou $S(t) = \mathbb{P}(X > t)$. Cela n'a aucune importance quand la loi de X est continue car $\mathbb{P}(X > t) = \mathbb{P}(X \geq t)$.*
Dans les cas où F a des sauts (quand le temps est discret, par exemple, compté en mois ou semaine), on utilise les notations suivantes :

$$F^-(t) = \mathbb{P}(X < t) \quad et \quad F^+(t) = \mathbb{P}(X \leq t)$$

où F^- est la limite gauche et F^+ la limite à droite de F (définitions et notations sont identiques pour la fonction S) : Remarquons que $F^- \leq F^+$ et $S^- \geq S^+$.

1.2.3 Densité de probabilité f

C'est la fonction $f(t) > 0$ telle que pour tout $t > 0$

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

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

$$f(t) = \lim_{h \longrightarrow 0} \frac{\mathbb{P}(t \leq X < t + h)}{h} = F'(t) = (1 - S(t))' = -S'(t).$$

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

1.2.4 Risque instantané h (ou taux de hasard)

Le risque instantané (ou taux d'incidence), 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) :

$$h(t) = \lim_{h \longrightarrow 0} \frac{\mathbb{P}(t \leq X < t+h | X > t)}{h} = \lim_{h \longrightarrow 0} \frac{1}{h} \frac{\mathbb{P}(t \leq X < t+h)}{\mathbb{P}(X > t)} = \frac{f(t)}{S(t)} = -(\ln S(t))'.$$

1.2.5 Taux de hasard cumulé H

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

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

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

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

On en déduit que

$$\begin{aligned} f(t) &= -S'(t). \\ &= -(\exp(-H(t)))'. \\ &= -\left(\exp(-\int_0^t h(u)du)\right)'. \\ &= h(t)\exp(-\int_0^t h(u)du). \\ &= h(t)S(t). \end{aligned}$$

1.2.6 Quantités associées à la distribution de survie

Moyenne et variance de la durée de survie

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

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

et

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

En effet :

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^\infty tf(t)dt. \\
&= -\int_0^\infty tS'(t)dt. \\
&= -\int_0^\infty tdS(t). \\
&= -tS(t)|_0^{+\infty} + \int_0^{+\infty} S(t)dt. \\
&= \int_0^{+\infty} S(t)dt.
\end{aligned}$$

En ce conserne la variance, on a,

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2. \\
&= \int_0^\infty t^2 f(t)dt - (\mathbb{E}(X))^2. \\
&= -\int_0^\infty t^2 S'(t)dt - (\mathbb{E}(X))^2. \\
&= -\int_0^\infty t^2 dS(t) - (\mathbb{E}(X))^2. \\
&= -t^2 S(t)|_0^{+\infty} + \int_0^{+\infty} 2tS(t)dt - (\mathbb{E}(X))^2. \\
&= 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2.
\end{aligned}$$

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

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 $S(t_m) = 0.5$: Dans le cas où l'estimateur est une fonction en escalier (ex : Kaplan-Meier), il se peut qu'il y ait un intervalle de temps vérifiant

$S(t_m) = 0.5$: Il faut alors être prudent dans l'interprétation, notamment si les deux événements encadrant le temps médian sont éloignés. Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[B_i, B_s]$ un intervalle de confiance de niveau α de $S(t_m)$; alors un intervalle de confiance de niveau α du temps médian t_m est

$$[S^{-1}(B_s), S^{-1}(B_i)].$$

- La fonction quantile de la durée de survie est définie par :

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

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

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

1.3 Censure et troncature

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées (i.i.d.) de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

1.3.1 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 X_i,
- son temps de censure C_i,
- la durée réellement observée T_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 (Censure fixe)
   Soit C une valeur fixée, au lieu d'observer les variables $X_1, \ldots, X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq C$, sinon on sait uniquement que $X_i > C$: On utilise la notation suivante :

   $$T_i = X_i \wedge C = min(X_i, C); i = 1, \ldots, n.$$

   Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles, par exemple dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note X la durée d'apprentissage de cette langue. Pour certains étudiants, nous allons observer leurs durées X_i d'apprentissage de la langue par contre pour d'autres, leurs X_i ne seront pas observées car le stage est limité dans le temps.

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

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

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

   $$T_i = X_i \wedge C_i; i = 1, \ldots, n.$$

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

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

(a) la perte de vue : le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients perdus de vue.

(b) l'arrêt ou le changement du traitement : les effets secondaires ou l'inifécacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.

(c) la fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients exclus-vivants. Les perdus de vue (et les exclusions) et les exclus-vivants correspondent à des observations censurées mais les deux mécanismes sont de nature différente (la censure peut être informative chez les perdus de vue).

Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (T, δ) :

$$T = X \vee C = max(X, C),$$

$$\delta = \mathbb{I}_{\{X \geq C\}}.$$

Comme pour la censure à droite, on suppose que la censure C est indé pendante de X. Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum

entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

Remarque 1.3.1. *Les modèles présentés dans ce travail traitent le cas de la censure à droite. Très peu de travaux s'intéressent à la censure à gauche car beaucoup moins fréquente. Certains auteurs ont proposé de "renverser" l'échelle de temps, c'est-à-dire de considérer la variable* $\tau - T = \tau - (X \vee C) = (\tau - X) \wedge (\tau - C)$ *au lieu de la variable* T *; où* τ *est un réel positif choisi de sorte que les observations* $\tau - T$ *restent dans* $\mathbb{R}_+$.

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. Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'événement s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à de la censure à droite.

1.3.2 Troncature

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

Exemple

Durée de vie après la retraite : on étudie la durée de vie après la retraite de sujets qui entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite et l'instant de l'enquête, la durée de vie après la retraite est donc tranquée à gauche par ce délai. Elle peut être censurée à droite si la fin de l'enquête à lieu alors que le sujet est toujours vivant.

Chapitre 2

Estimation non paramétrique

Si l'on peux pas supposer a priori que la loi de la durée de survie obéit à un modèle paramétrique, on peut estimer la fonction de survie S grâce à pleusieurs méthodes non paramétriques dont la plus intéréssante est celle de Kaplan-Meier. Nous allons cependant donne d'abord l'estimateur empirique de la fonction de survie dans le cas où les données sont complètes.

2.1 Données non censurées

2.1.1 Estimation de la fonction de survie

Soit $X \backsim F$, avec $F(t) = P(X \leq t)$ la fonction de répartition de X.
Soit $X_1, X_2, \ldots, X_n$ un échantillon indépendantes et identiquement distribuées (i.i.d.) de F et $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ les observations ordonnées. Supposons que F soit complètement inconnue.

Un bon estimateur pour F est la fonction de répartition empirique, notée F_n et définie par

$$\begin{aligned} F_n(t) &= \frac{\text{nombre d'observations} \leq t}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_{(i)} \leq t\}} \\ &= \begin{cases} 0 & si \quad t < X_{(1)} \\ \dfrac{k}{n} & si \quad X_{(k)} \leq t < X_{(k+1)}, \quad k = 1, \ldots, n-1 \\ 1 & si \quad t \geq X_{(n)} \end{cases} \end{aligned}$$

Pour estimer la fonction de survie, on utilise la formule suivante :

$$S(t) = 1 - F(t).$$

Alors, l'estimateur empirique $S_n(t)$ de $S(t)$ est définie par

$$S_n(t) = \begin{cases} 1 & \text{si} \quad t \leq X_{(1)}, \\ \dfrac{n-k}{n} & \text{si} \qquad t \in [X_{(k)}, X_{(k+1)}[, \quad k = 1, \ldots, n-1 \\ 0 & \text{si} \quad t > X_{(n)} \end{cases}$$

On peut aussi écrire

$$(1) \qquad S_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(X_i > t)}.$$

Exemple :

Freireich, en 1663, a fait un essai thérapeutique ayant pour but de comparer les durées de rémission en semaines, de sujets atteint de leucémie selon qu'ils ont reçu ou non du 6-MP.

Durée de rémmission, en semaine, selon le traitement

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

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

L'estimateur (1) pour le groupe traité par un placebo (pas de censure) donne le tableau suivant :

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

Propriétés :

- **Biais de l'estimateur** $S_n(t)$

  Pour tout point t, $S_n(t)$ est un estimateur sans biais de $S(t)$, car

  $$\mathbb{E}(S_n(t)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbb{I}_{(X_i>t)} = \mathbf{P}(X > t) = S(t).$$

- **Variance de l'estimateur** $S_n(t)$

  La variance de l'estimateur $S_n(t)$ est donnée par :

  $$\mathbb{V}(S_n(t)) = \frac{S(t)(1-S(t))}{n}.$$

  En effet

  $$nS_n(t) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i>t\}} \sim \text{BIN}(n, S(t)).$$

  Alors

  $$\mathbb{V}(nS_n(t)) = nS(t)(1-S(t))$$

  et

  $$\mathbb{V}(S_n(t)) = \frac{S(t)(1-S(t))}{n}.$$

- **La loi des grands nombres, nous donne**

$$\forall t \in \mathbb{R}, S_n(t) \xrightarrow{P} S(t), \quad si \quad n \to \infty.$$

- **Le théorème central-limite donne**

$$\frac{nS_n(t) - nS(t)}{\sqrt{nS(t)(1-S(t))}} \xrightarrow{L} N(0,1).$$

2.1.2 Estimation de la densité de survie

a) Histogramme de densité

On choisi un point d'origine t_0 et une longueur de classe h ($h > 0$). Les classes sont définies par :

$$B_k = [t_k, t_{k+1}[, \quad k \in \mathbb{Z} \quad (la \quad k^{\underline{\text{ème}}} \quad classe).$$

avec

$$t_{k+1} = t_k + h, \quad k \in \mathbb{Z}.$$

Un estimateur de f est donné par

$$\widehat{f}_H(x) = \frac{1}{nh} \#\{i : X_i \quad \text{est dans la classe qui contient} \quad x\}.$$

Si nous notons le nombre d'observations dans une classe B_k par ν_k, l'estimateur du type histogramme de densité s'écrit

$$\widehat{f}_H(x) = \frac{\nu_k}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}_{[t_k, t_{k+1}[}(X_i) \quad \text{pour} \quad x \in B_k$$

- L'histogramme de densité est un estimateur trés élémentaire, mais peut quand même déjà donner une premiére idée assez bonne de la forme de la densité f. Par contre, si on voulait utiliser cet estimateur dans d'autres analyses statistiques (comme par exemple l'estimation d'un taux de hasard, etc), il vaudrait mieux démarrer avec un estimateur plus précis.

- L'histogramme de densité est une fonction étagée, et donc discontinue.

b) Estimateur simple

Rappelons que la densité de probabilité f est égale à la dérivée de la fonction de répartition F (si cette dérivée existe). On peut donc écrire

$$\begin{aligned} f(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h}. \\ &= \lim_{h\to 0} \frac{P(x-h<X\leq x+h)}{2h}. \end{aligned}$$

Un estimateur de $f(x)$ est alors

$$\begin{aligned} \widehat{f}(x) &= \frac{1}{2h}\frac{\#\{i : x-h<X_i\leq x+h\}}{n}. \\ &= \frac{1}{2hn}\sum_{i=1}^{n} \mathbb{I}_{\{x-h<X_i\leq x+h\}}. \\ &= \frac{1}{2hn}\sum_{i=1}^{n} \mathbb{I}_{\{-1\leq \frac{x-X_i}{h}<1\}}. \end{aligned}$$

Notons que cet estimateur peut encore s'écrire comme

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\omega\left(\frac{x-X_i}{h}\right)$$

où

$$\omega(y) = \begin{cases} 1/2 & si \qquad y\in[-1,1[\\ 0 & sinon \end{cases}$$

Propriétés de l'estimateur simple :

Remarquons que

$$\widehat{f}(x) = \frac{F_n(x+h)-F_n(x-h)}{2h}$$

avec F_n la fonction de répartition empirique. Le paramétre de lissage h dépend de la taille de l'échantillon n, c'est-à-dire $h = h_n$.

Nous savons que

$$nF_n(x) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i\leq x\}} \sim \text{BIN}(n, F(x))$$

et

$$2nh_n\widehat{f}(x) = nF_n(x+h_n) - nF_n(x-h_n) \sim \text{BIN}(n, F(x+h_n) - F(x-h_n)).$$

Alors,

$$\mathbb{E}\{2nh_n\widehat{f}(x)\} = n[F(x+h_n) - F(x-h_n)]$$

et

$$\mathbb{E}\{\widehat{f}(x)\} = \frac{1}{2h_n}[F(x+h_n) - F(x-h_n)].$$

Pour la variance nous trouvons

$$Var\{2nh_n\widehat{f}(x)\} = n[F(x+h_n) - F(x-h_n)][1 - F(x+h_n) - F(x-h_n)]$$

Alors,

$$Var\{\widehat{f}(x)\} = \frac{1}{4nh_n^2}[F(x+h_n) - F(x-h_n)][1 - F(x+h_n) - F(x-h_n)].$$

Remarquons que, si $n \to \infty$ et $h_n \to \infty$, alors

$$\mathbb{E}\{\widehat{f}(x)\} \to f(x)$$

et

$$nh_n Var\{\widehat{f}(x)\} \to \frac{1}{2}f(x)$$

Le risque quadratique moyen de l'estimateur $\widehat{f}(x)$ de $f(x)$ est donné par

$$\begin{aligned}
\mathbb{E}\{\widehat{f}(x) - f(x)\}^2 &= \mathbb{E}\left(\widehat{f}(x) - \mathbb{E}\{\widehat{f}(x)\} + \mathbb{E}\{\widehat{f}(x)\} - f(x)\right)^2. \\
&= Var\{\widehat{f}(x)\} + \left[\mathbb{E}\{\widehat{f}(x)\} - f(x)\right]^2. \\
&= Var\{\widehat{f}(x)\} + \left[\text{Biais}\{\widehat{f}(x)\}\right]^2.
\end{aligned}$$

Si $h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, on obtient

$$\mathbb{E}\{\widehat{f}(x) - f(x)\}^2 \to \infty$$

pour tout point x. L'estimateur simple $\widehat{f}(x)$ est alors un estimateur consistent de $f(x)$.

Remarque 2.1.1. • *On n'a plus le probléme du choix d'un point d'origine (un point t_0) comme dans le cas d'un histogramme de densité.*

- *L'estimateur*

$$\widehat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} \mathbb{I}_{\{x-h<X_i\leq x+h\}} = \frac{1}{2hn} \sum_{i=1}^{n} \mathbb{I}_{\{X_i-h\leq x<X_i+h\}}$$

est une fonction discontinue, avec des discontinuités aux points $X_i \pm h$*, et constante entre ces points.*

c) Estimateur à noyau

Rappelons l'estimateur simple :

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \omega \left(\frac{x - X_i}{h} \right).$$

avec

$$\omega(y) = \begin{cases} 1/2 & si \\ 0 & sinon \end{cases} \qquad y \in [-1, 1[$$

la densité de probabilité uniforme sur l'intervalle $[-1, 1[$. Cet estimateur peut être généralisé en remplaçant la fonction de poids $w(\Delta)$ (la densité de probabilité uniforme) par une fonction de poids plus générale K (par exemple une densité de probabilité quelconque). Ceci résulte en l'estimateur

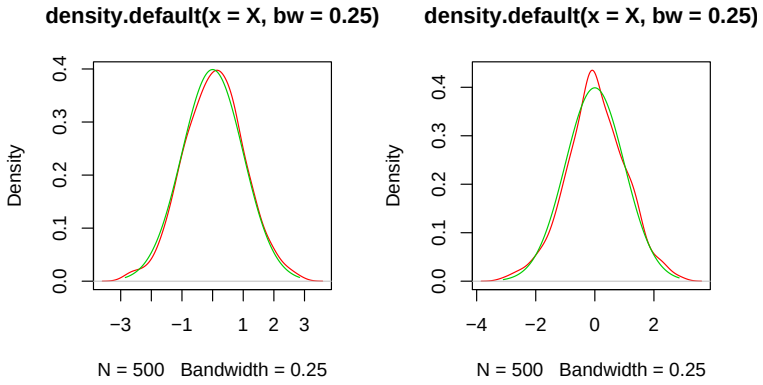
$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left(\frac{x - X_i}{h} \right)$$

où K est un noyau, h le paramétre de lissage. Souvent on prend pour K une densité de probabilité symétrique.

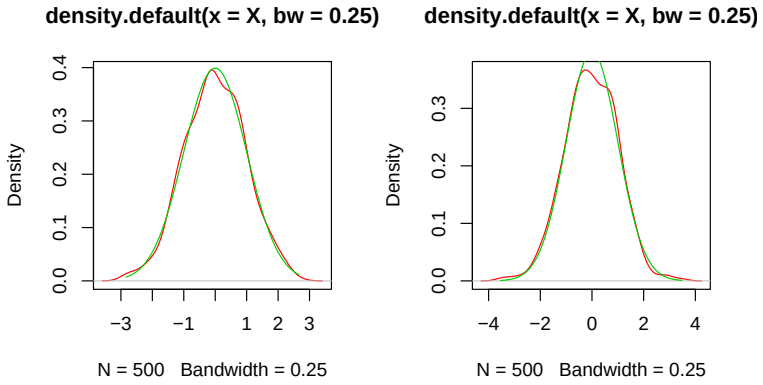
Remarque :

– Le noyau K détermine la forme des 'bosses', et la fenêtre h détermine la largeur des 'bosses'.

– Le choix du noyau a beaucoup moins d'importance que celui de h.
Une faible largeur de fenêtre implique une faible degré de lissage et résulte en une fonction de densité irrégulière. Une large valeur de h conduit à une estimation lisse.



Noyau gaussian Noyau Epanqunikov



Noyau triangulaire Noyau rectangulaire

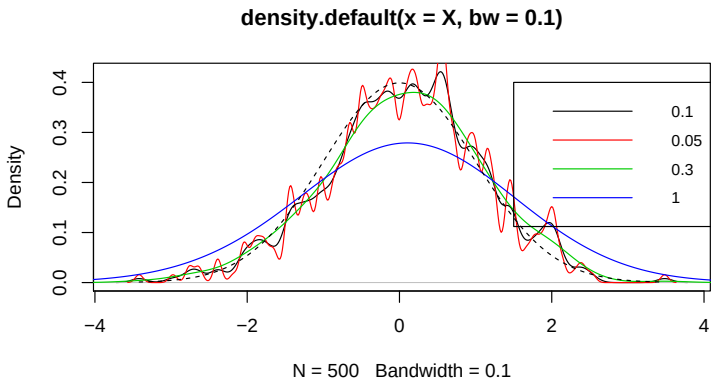


FIGURE 2.1 –

Exemples de noyaux

(1) Noyau gaussien

$$\forall u \in \mathbb{R}, K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{u^2}{2}}$$

(2) Noyau uniforme

$$\forall u \in \mathbb{R}, K(u) = \frac{1}{2} \mathbb{I}_{|u| \leq 1}$$

(3) Noyau d'Epanechnikov

$$\forall u \in \mathbb{R}, K(u) = \frac{3}{4}(1 - u^2) \mathbb{I}_{|u| \leq 1}$$

(4) Noyau triangulaire

$$\forall u \in \mathbb{R}, K(u) = (1 - |u|) \mathbb{I}_{|u| \leq 1}$$

(5) Noyau quadratique

$$\forall u \in \mathbb{R}, K(u) = \frac{15}{16}(1 - u^2)^2 \mathbb{I}_{|u| \leq 1}$$

Propriétés :

Il est facile de voir que l'estimateur à noyau

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

posséde les propriétés suivantes :

- Si K est une densité de probabilité, alors $\widehat{f}$ est aussi une densité de probabilité.
- $\widehat{f}$ a les mêmes propriétés de continuité et de différentiabilité que K :
  - Si K est continue, $\widehat{f}$ sera une fonction continue.
  - Si K est différentiable, $\widehat{f}$ sera une fonction différentiable.
  - Si K peut prendre des valeurs négatives, alors $\widehat{f}$ pourra aussi prendre des valeurs négatives.

Expressions du biais et de la variance :

Considérons l'estimateur à noyau

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i),$$

où nous avons introduit la notation

$$K_h(.) = \frac{1}{h}K\left(\frac{.}{h}\right),$$

pour une version transformée de K.

Pour calculer le biais de l'estimateur à noyau, remarquons d'abord que

$$\begin{aligned}\mathbb{E}\{\widehat{f}(x)\} &= \mathbb{E}\{K_h(x-X)\} \text{ car les } X_i \text{ sont identiquement distribuées.}\\ &= \int K_h(x-X)f(y)dy.\end{aligned}$$

La convolution entre deux fonctions f et g est définie par

$$(f*g)(x) = \int f(x-y)g(y)dy.$$

Dés lors, nous avons

$$\mathbb{E}\{\widehat{f}(x)\} - f(x) = (K_h * f)(x) - f(x)$$

$$\begin{aligned}Var\{\widehat{f}(x)\} &= \mathbb{E}\{\widehat{f}^2(x)\} - [\mathbb{E}\{\widehat{f}(x)\}]^2.\\ &= \mathbb{E}[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(x-X_i)K_h(x-X_j)] - \{\mathbb{E}\{K_h(x-X)\}\}^2.\\ &= \frac{1}{n}\mathbb{E}\{K_h^2(x-X)\} + \frac{1}{n^2}n(n-1)\{\mathbb{E}\{K_h(x-X)\}\}^2 - \{\mathbb{E}\{K_h(x-X)\}\}^2.\\ &= \frac{1}{n}\mathbb{E}\{K_h^2(x-X)\} - \frac{1}{n}[\mathbb{E}\{K_h(x-X)\}]^2.\\ &= \frac{1}{n}\{\mathbb{E}\{K_h^2(x-X)\} - [\mathbb{E}\{K_h(x-X)\}]^2\}.\\ &= \frac{1}{n}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\}.\end{aligned}$$

L'erreur quadratique moyenne (en anglais : "Mean squared error", MSE) de l'estimateur à noyau est donnée par :

$$\begin{aligned}
\text{MSE}\{\widehat{f}(x)\} &= \mathbb{E}\{\widehat{f}(x) - f(x)\}^2. \\
&= Var\{\widehat{f}(x)\} + [Biais(\widehat{f}(x))]^2. \\
&= \frac{1}{n}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\} + \{(K_h * f)(x) - f(x)\}^2. \\
&= \frac{1}{n}(K_h^2 * f)(x) + \left(1 - \frac{1}{n}\right)(K_h * f)^2(x) - 2(K_h * f)(x) + f^2(x).
\end{aligned}$$

Expression exacte de l'erreur quadratique moyenne integrée

L'expression exacte de l'erreur quadratique moyenne integrée (en anglais : "Mean Integrated Squared Error", MISE) peut être obtenue à partir de

$$\text{MISE}\{\widehat{f}\} = \int MSE\{\widehat{f}(x)\}dx$$

et est égale à

$$\text{MISE}\{\widehat{f}(.)\} = \frac{1}{n}\int (K_h^2 * f)(x)dx + (1-\frac{1}{n})\int (K_h * f)^2(x)dx - 2\int (K_h * f)(x)dx + \int f^2(x)dx.$$

Comme

$$\begin{aligned}
\int (K_h^2 * f)(x)dx &= \int \frac{1}{h^2}\left\{\int K^2\left(\frac{x-y}{h}\right)f(y)dy\right\}dx. \\
&= \frac{1}{h}\int\int K^2(u)f(x-uh)dudx, \quad \text{avec} \quad u = \frac{x-y}{h}. \\
&= \frac{1}{h}\int K^2(u)\left\{\int f(x-uh)dx\right\}du. \\
&= \frac{1}{h}\int K^2(u)du.
\end{aligned}$$

nous trouvons

$$\text{MISE}\{\widehat{f}(.)\} = \frac{1}{h}\int K^2(u)du + (1-\frac{1}{n})\int (K_h * f)^2(x)dx - 2\int (K_h * f)(x)dx + \int f^2(x)dx.$$

Malgré le fait qu'on ait des expressions exactes pour $\text{MSE}\{\widehat{f}(x)\}$ et $\text{MISE}\{\widehat{f}(.)\}$, ces expressions ne sont pas très attrayantes, car elles dépendent de manière très complexe du paramètre de lissage h. Pour cette raison on cherche des expressions asymptotiques qui pourraient dépendre de h de manière plus simple.

Expressions asymptotiques du biais et de la variance :

Une approximation asymptotique de l'espérance de l'estimateur $\widehat{f}(x)$ est donnée (sous certaines conditions sur f et K) par

$$\begin{aligned}
\mathbb{E}\{\widehat{f}(x)\} &= \int K_h(x-X)f(y)dy \\
&= \int K(u)f(x-uh)du, \quad \text{avec} \quad u=\frac{x-y}{h}, \quad du=-\frac{1}{h}dy \\
&= \int K(u)[f(x)-f'(x)uh+\frac{1}{2}f''(x)u^2h^2+\ldots]du \quad \text{par Taylor} \\
&= f(x)\int K(u)du - f'(x)h\int K(u)udu + \frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2).
\end{aligned}$$

Théorème 2.1. *Supposons maintenant que le noyau K satisfait*

$$K \geq 0, \quad \int K(u)du = 1, \quad \int K(u)udu = 0, \quad 0 < \int K(u)u^2du < \infty.$$

Alors

$$\mathbb{E}\{\widehat{f}(x)\} - f(x) = \frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2).$$

Preuve :

Comme

$$Var\{\widehat{f}(x)\} = \frac{1}{n}\{\mathbb{E}K_h^2(x-X) - [\mathbb{E}K_h(x-X)]^2\}$$

et

$$\begin{aligned}
\mathbb{E}K_h^2(x-X) &= \frac{1}{h^2}\int K^2\left(\frac{x-y}{h}\right)f(y)dy \\
&= \frac{1}{h}\int K^2(u)f(x-uh)du, \quad \text{avec} \quad u=\frac{x-y}{h} \\
&= \frac{1}{h}\int K^2(u)[f(x)-f'(x)uh+\frac{1}{2}f''(x)u^2h^2+\ldots]du \quad \text{par Taylor} \\
&= \frac{1}{h}f(x)\int K^2(u)du - f'(x)\int K^2(u)udu + o(1).
\end{aligned}$$

Nous trouvons que

$$Var\{\widehat{f}(x)\} = \frac{1}{nh}f(x)\int K^2(u)du + o\left(\frac{1}{nh}\right).$$

Nous avons donc établi que

$$\text{Bais}\{\widehat{f}(x)\} = \frac{1}{2}f''(x)\mu_2 h^2 + o(h^2), \quad \mu_2 = \int K(u)u^2du$$

$$Var\{\widehat{f}(x)\} = \frac{1}{nh}f(x)R(K) + o\left(\frac{1}{nh}\right), \quad R(K) = \int K^2(u)du.$$

Si $h_n \to 0$ quand $n \to \infty$, alors

$$\text{Bais}\{\widehat{f}(x)\} \to 0.$$

Si $h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, alors

$$Var\{\widehat{f}(x)\} \to 0.$$

Remarquons que

Si h décroît alors le $(bias)^2 \searrow$ et la variance $\nearrow$.

Si h augmente alors $(bias)^2 \nearrow$ et la variance $\searrow$.

Il faut donc essayer de choisir un h qui fasse un compromis entre le $(bias)^2$ et la variance.

Les expressions asymptotiques du biais et de la variance de $\widehat{f} = \widehat{f}_n$ nous permettent de trouver des expressions asymptotiques pour la MSE et la MISE. Rappelons ces expressions asymptotiques du biais et de la variance :

(2) $$\text{Bais}\{\widehat{f}(x)\} = \frac{1}{2}f''(x)\mu_2 h^2 + o(h^2),$$

(3) $$Var\{\widehat{f}(x)\} = \frac{1}{nh}f(x)R(K) + o\left(\frac{1}{nh}\right)$$

où $\mu_2 = \int K(u)u^2du$, $R(K) = \int K^2(u)du$ et $R(g) = \int g^2udu$, pour une fonction g de carré intégrable.

Ces expressions ont été obtenues sous certaines conditions sur K :

$$K(t) \geq 0, \quad \int K(u)du = 1, \quad \int K(u)udu = 0, \quad 0 < \int u^2K(u)du < \infty.$$

et en supposant que la densité de probabilité f avait toutes les dérivées (continues) nécessaires.
A partir de (2) et (3) on peut obtenir facilement les approximations asymptotiques suivantes pour la MSE et la MISE

$$\text{MSE}\{\widehat{f}(x)\} = \frac{1}{4}h^4\mu_2^2\{f''(x)\}^2 + \frac{1}{nh}f(x)R(K) + o\left(h^4 + \frac{1}{nh}\right)$$

$$\text{MISE}\{\widehat{f}(.)\} = \frac{1}{4}h^4\mu_2^2\int\{f''(x)\}^2dx + \frac{1}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right),$$

sous des conditions appropriées d'intégrabilité de f et ses dérivées.
On note l'approximation asymptotique de la MSE par

$$\text{AMSE}\{\widehat{f}(x)\} = \frac{1}{4}h^4\mu_2^2\{f''(x)\}^2 + \frac{1}{nh}f(x)R(K),$$

et l'approximation asymptotique de la MISE par

$$\text{AMISE}\{\widehat{f}(.)\} = \frac{1}{4}h^4\mu_2^2\int\{f''(x)\}^2dx + \frac{1}{nh}R(K),$$

Le paramètre de lissage optimal est la valeur de h qui minimise la MISE. Il est facile de vérifier à partir de (2.3) que

$$^h\text{AMISE} = \left\{\frac{f(x)R(K)}{\mu_2^2R(f'')}\right\}^{1/5}n^{-1/5}$$

et

$$^h\text{MISE} \sim \left\{\frac{R(K)}{\mu_2^2R(f'')}\right\}^{1/5}n^{-1/5}.$$

Les choix des valeurs $hAMISE$ et $hMISE$ sont des choix théoriques, qui ne sont pas utilisables en pratique car ils dépendent des quantitées inconnues f et f''. En pratique, on utilise la méthode de validation croisée (voir [1]).

2.2 Données censurées

2.2.1 Estimateur de Kaplan- Meier

Kaplan et Meier ont proposé un estimateur de S nommé aussi estimateur produit-limite. Il repose sur l'idée suivante : un "individu" est en vie après

l'instant t, c'est être en vie juste avant l'instant t et ne pas "mourir" en t. Cette idée se traduit par les relations suivantes :

$$\begin{aligned} S(t) &= \mathbb{P}(T > t) \\ &= \mathbb{P}(T > t/T > t-1)\mathbb{P}(T > t-1) \\ &= \dots \\ &= \mathbb{P}(T > t/T > t-1)\dots\mathbb{P}(T > 1/T > 0)\mathbb{P}(T > 0) \end{aligned}$$

Si l'on choisit les instants de conditionnement où il se produit un évènement t_i (mort, panne ou censure ...) on aura à estimer des quantités de la forme :

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

où les $t_{(i-1)} < t_{(i)}$ et p_i est la probabilité de survivre pendant l'intervalle de temps $I_i = [t_{(i-1)}, t_{(i)}[$ sachant qu'on était "vivant" au début de cet intervalle. Notons, d_i est le nombre de morts observées à l'instant $t_{(i)}$ et n_i est le nombre des individus ni morts ni censurés juste avant $t_{(i)}$, dits à risque (de mourrir). Or $q_i = 1 - p_i$ c'est la la probabilité de mourir durant l'intervalle $[t_{i-1}, t_i[$, sachant que l'individu était vivant en t_{i-1}. Un estimateur naturel de q_i est la fréquence

$$\widehat{q_i} = \frac{d_i}{n_i}.$$

On obtient alors l'estimateur de Kaplan-Meier de la fonction de survie S

$$\widehat{S}_{KM}(t) = \prod_{t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right). \quad (4)$$

Exemple

L'estimateur de Kaplan-Meier 4 de la fonction de survie S du groupe de 21 malades traité par le traitement 6-MP donne le tableau suivant :

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

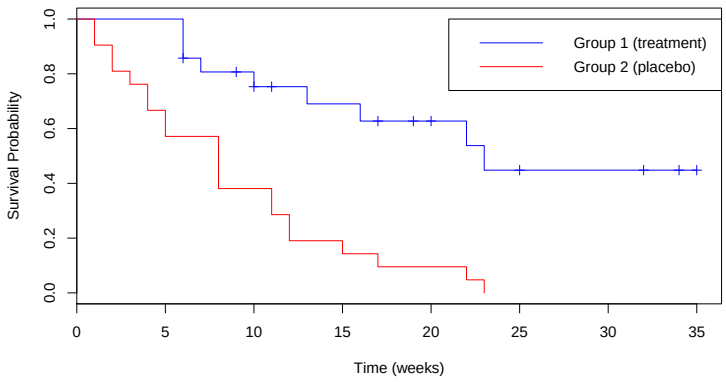


FIGURE 2.2 – Courbe de survie de Kaplan pour les données de Freireich

Représentations graphiques

Le graphe suivant présente les estimateurs de la fonction de survie par Kaplan-Meier (1) et par (4) (pour les deux traitements).

Paramètres denposition et de dispersion

Une durée de survie est une variable quantitative continue, il est donc possible de calculer les paramètres de position et de dispersion habituels : moyenne, médiane, écart-type, étendue, etc. Cependant, les distributions de ces variables sont généralement asymétriques et on préfèrera utiliser des paramètres robustes tels que médiane et quartiles plutôt que moyenne et écart-type.
Pour tenir compte de la présence de données censurées, le calcul des paramètres usuels (moyenne, médiane, quartiles) est basé sur le calcul de l'estimateur de Kaplan-Meier de la fonction de survie S(t).

- Le quantile d'ordre α est obtenu par :

$$q_\alpha = \inf\{t : 1 - \widehat{S}(t) \geq \alpha\}$$

  $\widehat{S}(t)$ est l'estimateur de Kaplan-Meier de $S(t)$.
- Notons que l'on peut utiliser l'estimateur de la fonction de survie pour estimer une durée moyenne : puisque l'espérance de la durée peut généralement s'écrire :

$$\mathbb{E}(X) = \int_0^\infty uf(u)du = \int_0^\infty S(u)du,$$

  on peut utiliser l'estimateur suivant :

$$\widehat{\mu} = \sum_{i=1}^{n} \widehat{S}(t_{i-1})(t_i - t_{i-1}).$$

- Il est intéressant de pouvoir calculer la variance ou l'écart-type de l'estimateur pour apprécier sa qualité (comme toujours dés que l'on fait de l'estimation ponctuelle).
  L'erreur standard de l'estimateur de Kaplan-Meier se calcule selon la formule de Greenwood, au temps de décés $t_i, i = 1, \ldots, n$ par :

$$\widehat{\sigma}(\widehat{S}(t_i)) = \widehat{S}(t_i)\sqrt{\sum_{j=1}^{n} \frac{d_i}{n_j(n_j - d_j)}}$$

  où d_i : nombre de décés observés au temps t_i
  et n_i : nombre de sujets exposés au risque de décés au temps t_i. En

effet :

$$\begin{aligned}
Var(\widehat{S}(t)) &= Var\left[\prod_i (1-d_i/n_i)\right] \\
&= \mathbb{E}\left[\prod_i (1-d_i/n_i)\right]^2 - \mathbb{E}^2\left[\prod_i (1-d_i/n_i)\right] \\
&= \prod_i \mathbb{E}\left[(1-d_i/n_i)\right]^2 - \prod_i \mathbb{E}^2\left[(1-d_i/n_i)\right] \\
&= \prod_i Var\left[(1-d_i/n_i)\right] + \mathbb{E}^2\left[\prod_i (1-d_i/n_i)\right] - \prod_i \mathbb{E}^2\left[(1-d_i/n_i)\right] \\
&= \prod_i (\frac{p_i q_i}{n_i} + p_i^2) - \prod_i p_i^2 \\
&= S^2(t) \prod_i (\frac{q_i}{r_i p_i} + 1) - S^2(t) \\
&= S^2(t) \prod_i (\frac{q_i}{r_i p_i}) \\
&= \widehat{S}(t) \sum_{i=1}^{n} \frac{d_i}{n_i(n_i - d_i)}.
\end{aligned}$$

2.2.2 Prporiétés de l'estimateur de Kaplan

- Biaisé : $\mathbb{E}(\widehat{S}(t)) \to_{n\to\infty} S(t)$,
- Converge en presque sure : $\lim_{n\to\infty} \widehat{S}(t) = S(t)$ p.s.
- On a,
$$\widehat{S}(t) \approx \mathcal{N}(S(t), Var(\widehat{S}(t))).$$

Estimation de la densité de survie

Nous reprenons les mêmes notations. On suppose que F admet une densité f par rapport à la mesure de Lebesgue qu'on se propose d'estimer en utilisant les observations (X_i, D_i), $i = 1, \ldots, n$. En se basant sur l'estimateur

de Kaplan-Meier, Blum et Susarla (1980) ont proposé un estimateur de la densité f par la méthode du noyau donné par

$$f_n(t) = \frac{1}{b_n} \int_0^\infty K(\frac{t-s}{b_n}) d\widehat{F}_n(s)$$

où $\widehat{F}_n$ est une fonction empirique, $(b_n)_{n\geq 1}$ est la fenêtre avec $b_n \to_{n\to\infty} 0$ et K un noyau de support $[-1, 1]$.

Chapitre 3

Tests de comparaison

La théorie des tests est l'une des deux branches de la statistique mathématique. Elle se subdivise en deux volets principaux, les tests paramétriques et les tests non-paramétriques. Ces derniers n'imposent aucune forme à la loi de probabilité des phénomènes étudiés contrairement au cas paramétrique qui requiert un modèle à fortes contraintes (comme la normalité des distributions, l'égalité des moyennes,...). Le but de ce chapitre est de comparer les durées de vie respectives de deux échantillons indépendants. Plus précisément, on dispose de deux échantillons indépendants, éventuellement censurés et on souhaite tester l'hypothèse nulle d'égalité des fonctions de survie dans les deux échantillons.

3.1 Comparaison de deux groupes

Notons S_A et S_B les fonctions de survie dans deux groupes A et B. On souhaite tester

$$(H_0) : S_A = S_B \quad \text{contre} \quad (H_1) : S_A \neq S_B.$$

Si il n'y avait pas de données censurées, on pourrait utiliser :

- Test de Kolomogorov Smirnov de comparaison de lois.
- Test de la somme des rangs.

En présence de données censurées, on généralise des tests non-paramétriques usuels. Deux tests (les plus courants) sont utilisés : le test de test de Gehan et le test du log-rank.

Considérons les notations suivantes,

1. $T_1 < \ldots < T_N$ les temps de décès ordonnés des deux échantillons réunis.
2. d_{Ai} et d_{Bi} le nombre de décès observés au temps T_i dans chacun des groupes A et B.
3. $d_i = d_{Ai} + d_{Bi}$ le nombre total de décès observés en T_i.
4. n_{Ai} et n_{Bi} le nombre de sujets à risques en T_i dans les groupes A et B.
5. $n_i = n_{Ai} + n_{Bi}$ le nombre total de sujets à risques en T_i.

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

	Décès en T_i	Vivant après T_i	Total
Groupe A	d_{Ai}	$n_{Ai} - d_{Ai}$	n_{Ai}
Groupe B	d_{Bi}	$n_{Bi} - d_{Bi}$	n_{Bi}
Total	d_i	$n_i - d_i$	n_i

Statistiques de test

On cherche à tester l'hypothèse H_0 : $S_A(t) = S_B(t)$ qui est l'égalité des fonctions de survie dans les deux groupes. Ainsi, sous l'hypothèse H_0, la proportion attendue de décès (parmi les sujets à risque) est identique dans les deux groupes pour tous les temps de décès T_i : Pour chaque temps T_i, on peut comparer les pourcentages de décès parmi les sujets à risque dans chacun des groupes en utilisant le test du Chi-2.

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

$$\mathbb{E}(D_{Ai}) = \frac{n_{Ai} d_i}{n_i}$$

et la variance

$$Var(D_{Ai)} = \frac{n_i - d_i}{n_i - 1} \frac{d_i n_{Ai} n_{Bi}}{n_i^2}.$$

où $\mathbb{E}(D_{Ai})$ correspond au nombre de décès attendus dans le groupe A : Sous H_0, on montre que les variables $D_{Ai} - \mathbb{E}(D_{Ai})$ suivent asymptotiquement des lois $N(0, V(D_{Ai}))$, $\left(\frac{D_{Ai} - \mathbb{E}(D_{Ai})}{V(D_{Ai})}\right)$ suivent asymptotiquement des loi de χ_1). Considérons des pondérations $w_i, i = 1, \ldots, N$, alors par indépendance entre les variables D_{Ai} et D_{Aj} (associées aux T_i et T_j), les variables

$$\sum_{i=1}^{N} \omega_i (D_{Ai} - \mathbb{E}(D_{Ai})) = \sum_{i=1}^{N} \omega_i \left(D_{Ai} - \frac{n_{Ai} \times d_i}{n_i} \right)$$

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

$$\chi_C^2 = \frac{\left[\sum_{i=1}^{N} \omega_i \left(d_{Ai} - \frac{n_{Ai} d_i}{n_i}\right)\right]^2}{\sum_{i=1}^{N} \omega_i^2 d_i \frac{n_i - d_i}{n_i - 1} \frac{n_{Ai} n_{Bi}}{n_i^2}}$$

qui suit asymptotiquement un $\chi^2(1)$.

On a

$$P(\chi_C^2 < \chi_1^2(\alpha)) = 1 - \alpha.$$

La règle de décision est : On accepte l'hypothèse H_0 si :

$$\chi_C^2 < \chi_1^2(\alpha)$$

Sinon on la rejette.

Remarque 3.1.1. *Les tests sont établis conditionnellement aux marges des tableaux et en supposant que les temps d'événements sont fixés. Les tableaux peuvent alors être traités comme des tableaux indépendants. Plusieurs statistiques de test ont été proposées*

1. *Test du logrank : $w_i = 1$, Cette pondération attribue à chaque décès le même poids quel que soit l'instant où il survient. Le test compare le nombres de décès observés au nombre de décès attendus.*
2. *Test de Gehan : $w_i = n_i$, La pondération en T_i est égale au nombre d'individus à risque en T_i donc les poids sont plus élévés pour les décès précoces que tardifs.*
3. *Test de Peto et Prentice : $w_i = \prod_i^k \frac{n_k}{n_k + d_k}$, Ces pondérations sont proches de l'estimateur de Kaplan-Meier de la survie. Elles attribuent des poids plus élévés aux décès précoces.*

3.2 Exemple : Application aux données de Freireich

Nous allons tester une différence de survie sans rechute entre les groupes placebo et traité pour les données sur la durée de rémission de sujets atteints de leucémie (Freireich, 1963). Le graphique des estimateurs des deux fonctions

de survie, présenté au chapitre précédent, montre que la survie sans rechute du groupe traité est toujours supérieure à la survie sans rechute du groupe sous placebo. On va donc effectuer un test du log-rank pour pouvoir affirmer que la différence entre les fonctions de survie est significative.

	6-MP	Placebo	Total	6-MP	Plac	Tatal		
Durées	n_{i1}	n_{i2}	n_i	d_{i1}	d_{i2}	d_i	$E(d_{i2})$	$V(d_{i2})$
1	21	21	42	0	2	2	1.00	0.49
2	21	19	40	0	2	2	0.95	0.49
3	21	17	38	0	1	1	0.45	0.25
4	21	16	37	0	2	2	0.86	0.48
5	21	14	35	0	2	2	0.80	0.47
6	21	12	33	3	0	3	1.09	0.65
7	17	12	29	1	0	1	0.41	0.24
8	16	12	38	0	4	4	1.71	0.87
10	15	8	23	1	0	1	0.35	0.23
11	13	8	21	0	2	2	0.76	0.45
12	12	6	20	0	2	2	0.67	0.42
13	12	4	16	1	0	1	0.25	0.19
15	11	4	15	0	1	1	0.27	0.20
16	11	3	14	1	0	1	0.21	0.17
17	10	3	13	0	1	1	0.23	0.18
22	7	2	9	1	1	2	0.44	0.30
23	6	1	7	1	1	2	0.29	0.20

Test du logrank $= (21 - 10,75)2/6,26 = 16,79$

La statistique de test vaut $16,79$, sous H_0 elle suit une loi du χ^2 à 1 degré de liberté (région critique pour les valeurs supérieures à $3,84$). On rejette donc l'hypothèse nulle. On conclut donc que le $6 - MP$ à un effet (positif) sur la fonction de survie.

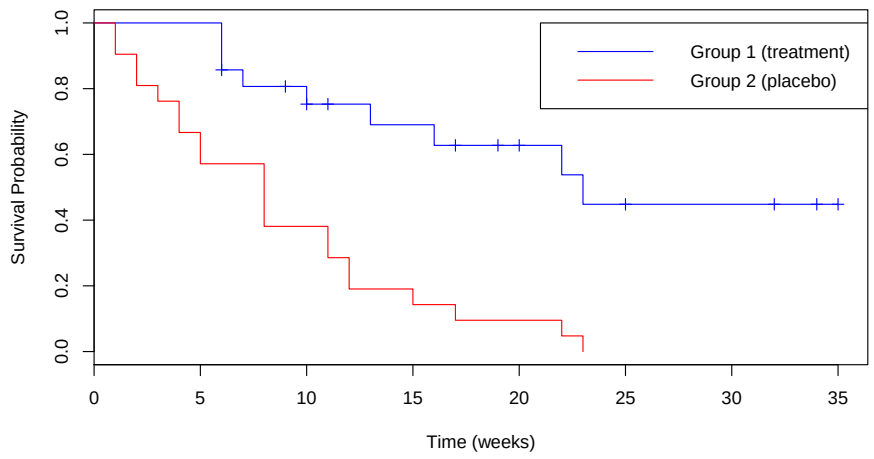


FIGURE 3.1 – Courbe de survie de Kaplan pour les données de Freireich

Rejet de H_0 : les 2 courbes de survie sont différentes $\rightarrow$ effet traitement. On conclut que le 6-MP à un effet positif sur la fonction de survie.

On utilise maintenant le test de Gehan, on obtient le tableau suivant :

	long-rank	Gehan		
Durées	Variance	Pondération	Coefficient	Variance
1	0.49	42	42.00	860.49
2	0.49	40	42.00	777.54
3	0.25	38	21.00	357.00
4	0.48	37	42.00	653.33
5	0.47	35	42.00	570.71
6	0.65	33	-36.00	708.75
7	0.24	29	-12.00	204.00
8	0.87	28	64.00	682.67
10	0.23	23	-8.00	120.00
11	0.45	21	26.00	197.60
12	0.42	18	24.00	135.53
13	0.19	16	-4.00	48.00
15	0.20	15	11.00	44.00
16	0.17	14	-3.00	33.00
17	0.18	13	10.00	30.00
22	0.30	9	5.00	24.50
23	0.20	7	5.00	10.00

$\chi^2 = 13.46$, on rejet alors H_0.

Conclusion

Le test du logrank est le test le plus populaire pour comparer 2 ou plusieurs courbes de survie. Il permet de prendre en compte toute l'information sur l'ensemble du suivi sans nécessité de faire des hypothèses sur la distribution des temps de survie. Il consiste à comparer le nombre d'événements observés au nombre d'événements attendus sous l'hypothèse nulle d'égalité de fonctions de survie des groupes. La statistique de test suit sous cette hypothèse approximativement une distribution du Chi2. Les autres tests sont plus aptes à déceler une différence entre les groupes en présence de nombreux décès précoces. Par ailleurs, la comparaison de pondérations utilisées montre que la statistique de Gehan dépend davantage de la distribution des censures que la statistique de Peto-Prentice. En conclusion, le test de Log-rank est le test le plus employé. Cependant, l'interpétation des résultats des tests doit prendre en considération la taille de l'effectif étudié ainsi que le profil et la distribution des censures. Quand le sujet est faible, les résultats des tests doivent être interprétés avec prudence.

Bibliographie

[1] Bosq, D., Lecoutre, J.P., Théorie de l'estimation fonctionnelle, Edition Economica, Paris, 1987.

[2] Catherine Huber, Analyse des durées de survie : http :www.biomedicale.univ-paris5.fr/survie/enseign/survie-sansi.pdf

[3] Catherine Huber, Cours de Modélisation Biostatistique en Splus : http :www.biomedicale.univ-paris5.fr/survie/enseign/cours-stat-avec-splus.pdf

[4] C.Hill, C. Com-Nougué, A. Kramar, T. Moreau, J. O'Quigle, R. Senousi,Cl. Chastang, Flammarion Sciences, 1996,3ème édition, 2000. "Analyse Statistique des Données de Survie"

[5] Codes de statistiques de l'université Penn State. http ://www.astro.psu.edu/statcodes/

[6] Jean-David Fermanian, Modéles de durées, téléchargeable sur la page : http : //www.crest.fr/ses.php/user=2975

[7] Michel Fioc (Michel.Fioc@iap.fr, www2.iap.fr/users/fioc/enseignement/analyse-de-survie/)

[8] Sahi, N. *"Modèles de survie Estimation et Applications"* Mémoire de magister en probabilités statistiques, (octobre 2011).