

# Examen de « Survie »

Master Mathématiques et Applications

Parcours Statistique

Année 2023 - 2024

*Durée : 2h. Les calculatrices et téléphones portables sont interdits*

---

**Exercice 1** – Donner la définition de l'estimateur de Kaplan-Meier. Vous prendrez soin de définir clairement et précisément toutes les notations requises.

**Exercice 2** – Pour  $n = 9$  individus on dispose des données suivantes.

0.6	1.7 <sup>+</sup>	1.6 <sup>+</sup>	1.6	1.2	1.1	1.2	0.9 <sup>+</sup>	1.4
-----	------------------	------------------	-----	-----	-----	-----	------------------	-----

Ces valeurs correspondent aux  $n = 9$  observations  $\{t_1^*, \dots, t_n^*\}$  de la variable aléatoire  $T^* = \min(T, C)$ . L'exposant + signifie que la valeur a été censurée à droite (i.e., l'événement n'est pas observé).

- 1) Donner la valeur  $m$  ainsi que l'ensemble  $\{t_{(1)}^* < \dots < t_{(m)}^*\}$  des observations distinctes et ordonnées.

On commence par ordonner les observations

0.6	0.9 <sup>+</sup>	1.1	1.2	1.2	1.4	1.6 <sup>+</sup>	1.6	1.7 <sup>+</sup>
-----	------------------	-----	-----	-----	-----	------------------	-----	------------------

On remarque qu'il y a  $m = 7$  observations distinctes :

$t_{(1)}^*$	$t_{(2)}^*$	$t_{(3)}^*$	$t_{(4)}^*$	$t_{(5)}^*$	$t_{(6)}^*$	$t_{(7)}^*$
0.6	0.9	1.1	1.2	1.4	1.6	1.7

Pour tout  $i \in \{1, \dots, m\}$ , on note  $O_i$  le nombre d'événements observés à l'instant  $t_{(i)}^*$  et  $N_i$  le nombre d'individus à risque à l'instant  $t_{(i)}^*$ .

- 2) Donner les valeurs  $O_1, \dots, O_m$  et  $N_1, \dots, N_m$ .

Les valeurs des  $O_i$  sont :

$O_1$	$O_2$	$O_3$	$O_4$	$O_5$	$O_6$	$O_7$
1	0	1	2	1	1	0

Les valeurs des  $N_i$  sont :

$N_1$	$N_2$	$N_3$	$N_4$	$N_5$	$N_6$	$N_7$
9	8	7	6	4	3	1

- 3) Calculer la valeur observée de l'estimateur de Kaplan-Meier au point  $t_{(3)}^*$ .  
(Vous donnerez le résultat sous forme d'une fraction irréductible)

La valeur observée est

$$\left(1 - \frac{1}{9}\right) \times \left(1 - \frac{0}{8}\right) \times \left(1 - \frac{1}{7}\right) = \frac{8}{9} \times \frac{6}{7} = \frac{16}{21}.$$

**Exercice 3** – Soient  $T$  et  $C$  deux variables aléatoires indépendantes et positives. On pose  $T^* := \min(T, C)$  et  $\Delta = \mathbb{I}_{\{T \leq C\}}$ . On suppose que  $T$  suit une loi exponentielle de paramètre  $\lambda > 0$  (i.e., d'espérance  $1/\lambda$ ) et que  $C$  suit une loi exponentielle de paramètre  $\theta\lambda$  avec  $\theta > 0$ . On note  $f_T(\cdot)$  (resp.  $f_C(\cdot)$ ) la densité de la loi de  $T$  (resp.  $C$ ) et  $S_T(\cdot)$  (resp.  $S_C(\cdot)$ ) la fonction de survie de la loi de  $T$  (resp.  $C$ ).

- 1) Montrer que

$$\mathbb{P}(T \leq C) = \int_0^\infty S_C(x) f_T(x) dx = \int_0^\infty (1 - S_T(x)) f_C(x) dx.$$

(Aide : utiliser les probabilités conditionnelles  $\mathbb{P}(T \leq C \mid T = x)$  et  $\mathbb{P}(T \leq C \mid C = x)$ ).

On a

$$\mathbb{P}(T \leq C) = \int_0^\infty \mathbb{P}(T \leq C \mid T = x) f_T(x) dx = \int_0^\infty \mathbb{P}(C \geq x \mid T = x) f_T(x) dx.$$

En utilisant l'indépendance entre  $T$  et  $C$ , il vient

$$\mathbb{P}(T \leq C) = \int_0^\infty \mathbb{P}(C \geq x) f_T(x) dx = \int_0^\infty S_C(x) f_T(x) dx.$$

On montre l'autre égalité de la même façon, en conditionnant par rapport à la variable aléatoire  $C$ .

- 2) En déduire que  $\mathbb{P}(T \leq C) = 1/(1 + \theta)$ .

On a pour tout  $x > 0$ ,  $S_C(x) = \exp(-\lambda\theta x)$  et  $f_T(x) = \lambda \exp(-\lambda x)$ . Ainsi,

$$\mathbb{P}(T \leq C) = \int_0^\infty S_C(x) f_T(x) dx = \lambda \int_0^\infty \exp(-\lambda(1+\theta)x) dx = \frac{1}{1+\theta}.$$

3) Quelle est la loi de  $T^*$  ?

Soit  $t > 0$ . En utilisant l'indépendance entre  $T$  et  $C$ , on a :

$$\mathbb{P}(T^* > t) = \mathbb{P}(T > t)\mathbb{P}(C > t) = \exp(-\lambda(1+\theta)t).$$

Ainsi,  $T^*$  suit une loi exponentielle de paramètre  $\lambda(1+\theta)$ .

Soient  $(\mathbf{T}^*, \mathbf{\Delta}) = \{(T_1^*, \Delta_1), \dots, (T_n^*, \Delta_n)\}$  des couples indépendants de même loi que le couple  $(T^*, \Delta)$ . On note  $(\mathbf{t}^*, \mathbf{d}) = \{(t_1^*, d_1), \dots, (t_n^*, d_n)\}$  les observations associées. On pose

$$\bar{T}_n^* := \frac{1}{n} \sum_{i=1}^n T_i^* \quad \text{et} \quad \bar{\Delta}_n := \frac{1}{n} \sum_{i=1}^n \Delta_i.$$

Les observations des variables aléatoires  $\bar{T}_n^*$  et  $\bar{\Delta}_n$  sont notées  $\bar{t}_n^*$  et  $\bar{d}_n$ .

4) Quel est l'effet du paramètre  $\theta$  sur  $n\mathbb{E}(\bar{\Delta}_n)$  qui est le nombre attendu d'événements dans un échantillon de taille  $n$  ?

Remarquons tout d'abord que  $n\mathbb{E}(\bar{\Delta}_n) = n\mathbb{P}(T \leq C) = n/(1+\theta)$ . Ainsi, plus  $\theta$  est proche de 0, plus on s'attend à observer d'événements.

5) En utilisant la question 2), proposez un estimateur (i.e., une fonction de  $(\mathbf{T}^*, \mathbf{\Delta})$ ) du paramètre  $\theta$ .

On a montré à la question 2) que  $\mathbb{P}(T \leq C) = \mathbb{E}(\Delta) = 1/(1+\theta)$ . L'estimateur naturel de  $\mathbb{E}(\Delta)$  est  $\bar{\Delta}_n$  ce qui nous conduit à l'estimateur de  $\theta$  donné par  $\hat{\theta}_n = (1 - \bar{\Delta}_n)/\bar{\Delta}_n$ .

6) En utilisant l'expression de l'espérance de  $T^*$ , proposez un estimateur de  $\lambda$  dont l'expression dépendra de  $\bar{T}_n^*$  et de l'estimateur de  $\theta$  obtenu à la question précédente.

Il suffit de remarquer que

$$\mathbb{E}(T^*) = \frac{1}{\lambda(1+\theta)}.$$

En estimant  $\mathbb{E}(T^*)$  par  $\bar{T}_n^*$ , un estimateur de  $\lambda$  est donné par

$$\hat{\lambda}_n = \frac{1}{\bar{T}_n^*(1+\hat{\theta}_n)}.$$

Pour tout  $h > 0$ , on introduit la fonction de vraisemblance observée

$$L_h(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) := \frac{1}{h^n} \prod_{i=1}^n \mathbb{P}(\{t_i^* - h < T^* \leq t_i^*\} \cap \{\Delta = d_i\}).$$

7) Montrer, en justifiant correctement chaque étape, que

$$\begin{aligned} L_h(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) &:= \frac{1}{h^n} \prod_{i=1}^n [\mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{T \leq C\})]^{d_i} \\ &\quad \times [\mathbb{P}(\{t_i^* - h < C \leq t_i^*\} \cap \{T > C\})]^{1-d_i} \end{aligned}$$

Tout d'abord, on remarque que

$$\begin{aligned} L_h(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) &:= \frac{1}{h^n} \prod_{i=1}^n [\mathbb{P}(\{t_i^* - h < T^* \leq t_i^*\} \cap \{\Delta = d_i\})]^{d_i} \\ &\quad \times [\mathbb{P}(\{t_i^* - h < T^* \leq t_i^*\} \cap \{\Delta = d_i\})]^{1-d_i} \end{aligned}$$

Or, le premier facteur n'intervient que lorsque  $d_i = 1$  i.e., lorsque  $\Delta = 1$  et donc  $T^* = T$ . De même, le second facteur n'intervient que pour  $d_i = 0$  i.e.,  $\Delta = 0$  et donc  $T^* = C$ . Ainsi,

$$\begin{aligned} L_h(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) &:= \frac{1}{h^n} \prod_{i=1}^n [\mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{\Delta = 1\})]^{d_i} \\ &\quad \times [\mathbb{P}(\{t_i^* - h < C \leq t_i^*\} \cap \{\Delta = 0\})]^{1-d_i}. \end{aligned}$$

Il reste à remarquer que  $\{\Delta = 1\} = \{T \leq C\}$  et  $\{\Delta = 0\} = \{T > C\}$ .

8) Donner l'expression de la limite

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{T \leq C\}).$$

Vous justifierez correctement votre réponse.

On remarque que  $\{C \geq t_i^*\} \cap \{T \leq t_i^*\} \subset \{T \leq C\} \cap \{T \leq t_i^*\}$ . De plus,  $\{T \leq C\} \cap \{T > t_i^* - h\} \subset \{C \geq t_i^* - h\} \cap \{T > t_i^* - h\}$ . Ainsi,

$$\begin{aligned} \mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{C \geq t_i^*\}) &\leq \mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{T \leq C\}) \\ &\leq \mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{C \geq t_i^* - h\}). \end{aligned}$$

En utilisant l'indépendance entre  $T$  et  $C$ , il vient

$$\begin{aligned}\mathbb{P}(\{t_i^* - h < T \leq t_i^*\}) S_C(t_i^*) &\leq \mathbb{P}(\{t_i^* - h < T \leq t_i^*\} \cap \{T \leq C\}) \\ &\leq \mathbb{P}(\{t_i^* - h < T \leq t_i^*\}) S_C(t_i^* - h).\end{aligned}$$

Il reste à remarquer que lorsque  $h \rightarrow 0$ ,  $\mathbb{P}(\{t_i^* - h < T \leq t_i^*\})/h \rightarrow f_T(t_i^*)$  et  $S_C(t_i^* - h) \rightarrow S_C(t_i^*)$ .

9) En déduire que la vraisemblance observée du modèle est

$$\begin{aligned}L_0(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) &:= \lim_{h \rightarrow 0} L_h(\lambda, \theta \mid (\mathbf{t}^*, \mathbf{d})) \\ &= \prod_{i=1}^n [f_T(t_i^*) S_C(t_i^*)]^{d_i} [f_C(t_i^*) S_T(t_i^*)]^{1-d_i}.\end{aligned}$$

C'est une conséquence directe des questions 7) et 8).

10) Donner l'expression (en fonction de  $\lambda$ ,  $\theta$ ,  $\bar{T}_n^*$  et  $\bar{\Delta}_n$ ) de la log-vraisemblance

$$\mathcal{L}_0(\lambda, \theta \mid (\mathbf{T}^*, \mathbf{\Delta})) := \log(L_0(\lambda, \theta \mid (\mathbf{T}^*, \mathbf{\Delta}))).$$

En remplaçant  $f_T(\cdot)$ ,  $f_C(\cdot)$ ,  $S_T(\cdot)$  et  $S_C(\cdot)$  par leur expression, on obtient

$$\mathcal{L}_0(\lambda, \theta \mid (\mathbf{T}^*, \mathbf{\Delta})) = n \log(\lambda) + n(1 - \bar{\Delta}_n) \log(\theta) - n\lambda(1 + \theta)\bar{T}_n^*.$$

11) Donner l'expression des estimateurs  $\hat{\lambda}_n$  et  $\hat{\theta}_n$  et vérifier qu'ils coïncident avec ceux trouvés aux questions 5) et 6).

En dérivant la log-vraisemblance, on montre que les estimateurs du maximum de vraisemblance  $\hat{\lambda}_n$  et  $\hat{\theta}_n$  sont solutions du système d'équations d'inconnus  $\lambda$  et  $\theta$  suivant

$$\begin{cases} \lambda^{-1} = \bar{T}_n^*(1 + \theta) \\ \theta = (1 - \bar{\Delta}_n)/(\lambda \bar{T}_n^*).\end{cases}$$

En résolvant ce système, on trouve  $\hat{\lambda}_n = \bar{\Delta}_n/\bar{T}_n^*$  et  $\hat{\theta}_n = (1 - \bar{\Delta}_n)/\bar{\Delta}_n$ .

Ils sont égaux aux estimateurs trouvés aux questions 5) et 6).

En supposant que les hypothèses de régularité requises sont satisfaites, on rappelle que

$$[\mathcal{I}_n(\theta, \lambda)]^{1/2} \left[ \begin{pmatrix} \hat{\theta}_n \\ \hat{\lambda}_n \end{pmatrix} - \begin{pmatrix} \theta \\ \lambda \end{pmatrix} \right] \xrightarrow{d} \mathcal{N}(0, I_2)$$

où  $I_2$  est la matrice identité de dimension 2 et  $\mathcal{I}_n(\theta, \lambda)$  est la matrice d'infor-

mation de Fisher donnée par

$$\mathcal{I}_n(\theta, \lambda) = -\mathbb{E} \left( \begin{array}{cc} \frac{\partial^2}{\partial \theta^2} \mathcal{L}_0(\lambda, \theta | (\mathbf{T}^*, \mathbf{\Delta})) & \frac{\partial^2}{\partial \theta \partial \lambda} \mathcal{L}_0(\lambda, \theta | (\mathbf{T}^*, \mathbf{\Delta})) \\ \frac{\partial^2}{\partial \theta \partial \lambda} \mathcal{L}_0(\lambda, \theta | (\mathbf{T}^*, \mathbf{\Delta})) & \frac{\partial^2}{\partial \lambda^2} \mathcal{L}_0(\lambda, \theta | (\mathbf{T}^*, \mathbf{\Delta})) \end{array} \right)$$

- 12) Donner l'expression de la matrice d'information de Fisher en fonction de  $n$ ,  $\theta$ ,  $\lambda$ ,  $\mathbb{E}(T^*)$  et  $\mathbb{P}(T > C)$ .

Un simple calcul donne

$$\mathcal{I}_n(\theta, \lambda) = n \begin{pmatrix} \mathbb{P}(T > C)/\theta^2 & \mathbb{E}(T^*) \\ \mathbb{E}(T^*) & 1/\lambda^2 \end{pmatrix}$$

- 13) Montrer que

$$\frac{\overline{\Delta}_n^2}{1 - \overline{\Delta}_n} \xrightarrow{\mathbb{P}} \frac{\mathbb{P}(T > C)}{\theta^2}.$$

On sait que  $\mathbb{P}(T > C) = 1 - \mathbb{P}(T \leq C) = \theta/(1 + \theta)$ . L'estimateur du maximum de vraisemblance de  $\theta$  étant faiblement consistant, on en déduit que

$$\frac{1}{\widehat{\theta}_n^2} \frac{\widehat{\theta}_n}{1 + \widehat{\theta}_n} = \frac{1}{\widehat{\theta}_n(1 + \widehat{\theta}_n)} \xrightarrow{\mathbb{P}} 1.$$

Il reste à remarquer que

$$\frac{1}{\widehat{\theta}_n(1 + \widehat{\theta}_n)} = \frac{\overline{\Delta}_n^2}{1 - \overline{\Delta}_n}.$$

On introduit la matrice

$$\widehat{\mathcal{I}}_n = n \begin{pmatrix} \overline{\Delta}_n^2/(1 - \overline{\Delta}_n) & \overline{T}_n^* \\ \overline{T}_n^* & (\overline{T}_n^*/\overline{\Delta}_n)^2 \end{pmatrix}$$

On admettra que  $\widehat{\mathcal{I}}_n^{-1} \mathcal{I}_n(\theta, \lambda) \xrightarrow{\mathbb{P}} I_2$ .

- 14) Montrer que

$$\sqrt{n} \frac{\overline{T}_n^*}{\overline{\Delta}_n} (\widehat{\lambda}_n - \lambda) \xrightarrow{d} \mathcal{N}(0, 1).$$

En utilisant le lemme de Slutsky, on a

$$[\widehat{\mathcal{I}}_n]^{1/2} \left[ \begin{pmatrix} \widehat{\theta}_n \\ \widehat{\lambda}_n \end{pmatrix} - \begin{pmatrix} \theta \\ \lambda \end{pmatrix} \right] \xrightarrow{d} \mathcal{N}(0, I_2).$$

En ne considérant que la seconde composante, on obtient le résultat attendu.

- 15) Proposer un test de l'hypothèse nulle  $H_0 : \mathbb{E}(T) = 1$  contre l'hypothèse alternative  $H_1 : \mathbb{E}(T) \neq 1$  avec un risque de première espèce égal à  $\alpha$ . Il faut tout d'abord remarquer que  $\mathbb{E}(T) = 1/\lambda$ . Ainsi, on a  $H_0 : \lambda = 1$ . Evidemment, le test sera basé sur la variable aléatoire  $\hat{\lambda}_n$ . Une valeur observée de  $\hat{\lambda}_n$  très différente de 1 nous conduira à ne pas accepter l'hypothèse nulle. Or, sous  $H_0$ , on sait d'après la question précédente que

$$\widehat{W}_n := \sqrt{n} \frac{\overline{T}_n^*}{\Delta_n} (\hat{\lambda}_n - 1) \xrightarrow{d} \mathcal{N}(0, 1).$$

On va donc rejeter l'hypothèse nulle si la valeur observée de  $\widehat{W}_n$  est significativement différente de 0 i.e., si la valeur observée de  $|\widehat{W}_n|$  est strictement supérieure à  $z_{1-\alpha/2}$  où  $z_{1-\alpha/2}$  est le quantile d'ordre  $1 - \alpha/2$  d'une loi normale centrée et réduite.