

Introduction à la biostatistique : Analyse des données de Survie

Adeline Samson

Département STID, Université Paris Descartes

1

¹Transparents élaborées à partir des cours de Elodie Brunel et Marie-Luce Taupin



Données de survie

En médecine ou en biologie, on s'intéresse souvent à des durées :

- 1 durée sans symptôme de patients infectés par le VIH
- 2 durée de fièvre chez un patient atteint de pneumonie

On distingue l'évènement d'intérêt :

- 1 Début des symptômes chez un patient séropositif
- 2 Fin de la fièvre chez un patient atteint de pneumonie

de la variable à expliquer : "durée avant l'apparition de l'évènement"

- 1 Temps écoulé sans symptôme
- 2 Temps écoulé avant la fin de la fièvre

Problématique

Supposons que l'étude soit un essai clinique portant sur deux groupes de patients, recevant 2 types de traitements.

↔ **question 1** : L'un des deux traitements est-il plus efficace que l'autre en terme d'amélioration de la survie des patients ?

↔ **question 2** : Peut-on mettre en évidence des **facteurs pronostiques** c'est-à-dire qui améliorent/ détériorent la survie ?

Exemples : âge, sexe, tabagisme, taux de cholestérol, antécédents familiaux, etc

Problématique statistique

↔ **Pour répondre à la question 1** : méthodes statistiques de comparaison des deux groupes de patients

↔ **Pour répondre à la question 2** : modèle reliant la durée de survie des patients à des variables explicatives et mise en évidence des facteurs pronostiques.

Difficultés

Cohorte/Essai clinique

- Durées d'observation différentes selon les patients (instants différents d'entrée dans l'étude)
- Temps de survie peut- être censuré

Si on enlève les données censurées \implies perte d'information.

\implies Techniques statistiques habituelles ne s'appliquent pas

Problème de la censure

Définition (*Censure*)

La durée T est dite censurée si

- le patient est toujours vivant à la fin de l'étude (exclus-vivants)
- le patient est perdu de vue : il a quitté l'étude avant qu'on ait pu observer l'évènement d'intérêt

Notations

Pour l'individu i avec $i = 1, \dots, n$, on appelle

- T_i : le temps de survie de l'individu i
- C_i : le temps de censure de l'individu i .
- δ_i : l'indicateur de censure de l'individu i .

Pour un individu i , on observe

$$X_i = \min(T_i, C_i), \quad \delta_i = \mathbf{1}_{T_i \leq C_i}$$

Les observations sont donc $(X_1, \delta_1), \dots, (X_n, \delta_n)$

$$\text{avec } (X_i, \delta_i) = \begin{cases} (T_i, 1) & \text{si } C_i \geq T_i \quad \text{non censuré} \\ (C_i, 0) & \text{si } C_i < T_i \quad \text{censuré.} \end{cases}$$

Deux fonctions essentielles

1) **Fonction de survie** : probabilité de survivre au moins jusqu'à la date t

$$S(t) = \mathbb{P}(T > t)$$

Propriétés

- $S(t) = 1 - F_T(t)$ où F_T est la fonction de répartition de T
- $S(t)$ est décroissante
- $S(0) = 1$
- $S(\infty) = 0$

Deux fonctions essentielles

2) **Risque instantané de décès $h(t)$** : probabilité que l'évènement survienne dans un petit intervalle de temps après t , sachant qu'il n'a pas eu lieu jusqu'à l'instant t .

$$h(t) = \lim_{dt \rightarrow 0} \frac{1}{dt} \mathbb{P}(t < T \leq t + dt | T > t) \quad (1)$$

ex : Si t est mesuré en jours, $h(t)$ approxime la probabilité d'un individu vivant le jour t de mourir le jour suivant

Quelques propriétés du risque instantané

Propriétés

Soit $H(t) = \int_0^t h(u)du$, le risque cumulé de décès. On a

- $S(t) = \exp(-H(t)) = \exp[-\int_0^t h(s)ds]$.
- $H(t) = -\ln(S(t))$.
- $h(t) = (H(t))' = \frac{(-S(t))'}{S(t)} = \frac{f_T(t)}{S(t)}$, où f_T est la densité de T .

ex : $h(t) = \lambda \Leftrightarrow S(t) = \exp(-\lambda t) \Leftrightarrow T \sim \mathcal{E}(\lambda)$.

- Ces deux fonctions $S(t)$ et $h(t)$ caractérisent la loi de T .
- Elles sont **inconnues**.
- On va les estimer en utilisant les observations

$$(X_i, \delta_i) = (\min(T_i, C_i), \mathbb{1}_{T_i \leq C_i}), \quad i = 1, \dots, n.$$

Estimateur de la fonction de Survie

- Si les données ne sont pas censurées, $S(t)$ peut être estimée par la proportion d'individus ayant survécu à l'instant t .

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{T_i > t} = \text{proportion d'individus tels que } T_i > t.$$

- Si les données sont censurées, la fonction de survie est estimée par la méthode de Kaplan-Meier.

Idée :

$$\begin{aligned} \mathbb{P}(\text{avoir de la fièvre à la } i\text{-ème semaine}) = \\ \mathbb{P}(\text{avoir de la fièvre à la } (i-1)\text{ème semaine}) \times \\ \mathbb{P}(\text{pas d'arrêt de la fièvre lors de la } i\text{-ème semaine}) \end{aligned}$$

Estimateur de Kaplan Meier

On définit l'estimateur de Kaplan Meier $\hat{S}(t)$. Pour $X(i) \leq t < X(i+1)$,

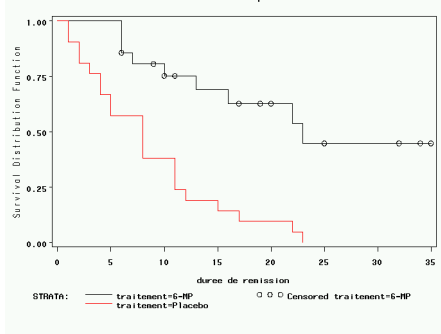
$$\hat{S}(t) = \prod_{j=1}^i \frac{R_j - d_j}{R_j}.$$

où

- d_j est le nombre de décès observé au temps $X(j)$
- R_j est le nombre d'individus à risque (exposés au risque de décès) juste avant $X(j)$.

Courbes de survie

Durees de remission avant rechute de patients atteints de leucemie



Outils de comparaison ?

Tests de rangs

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

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

Sans données censurées :

- Test de Kolomogorov Smirnov de comparaison
- Test de la somme des rangs
- Test de Mann-Whitney

Avec données censurées, généralisation des tests non-paramétriques usuels :

- le **test de Wilcoxon généralisé** (ou test de Gehan)
- et le **test du log-rank**.

Construction du test

Soit t_1, \dots, t_k temps de décès ordonnés des 2 groupes A et B réunis,

$$U = \sum_{i=1}^k w_i (d_{B,i} - e_{B,i})$$

- w_i : poids.
- $d_{B,i}$: nombre de décès observés au temps t_i dans le groupe B
- $e_{B,i}$: nombre de décès attendus au temps t_i dans le groupe B sous H_0

$$e_{B,i} = \frac{R_{B,i}}{R_{A,i} + R_{B,i}} (d_{A,i} + d_{B,i})$$

- $R_{A,i}, R_{B,i}$: nbre de sujets exposés au risque de décès juste avant t_i , dans les groupes A et B avec $R_{A,i} + R_{B,i} = R_i$.

Statistique de test et zone de rejet

Sous H_0 , $E(U) = 0$ et :

$$\text{sous } H_0 \frac{U - E(U)}{\sqrt{V(U)}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\underset{H_0}{\rightarrow}}} \mathcal{N}(0, 1)$$

avec $V(U) = \sum_{i=1}^k w_i^2 v_i$ et $v_i = d_i \frac{R_i - d_i}{R_i - 1} \frac{R_{A,i} R_{B,i}}{R_i^2}$.

La statistique de test est donc

$$T_n = U^2 / V(U)$$

avec sous H_0 , $T_n = U^2 / V(U) \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\underset{H_0}{\rightarrow}}} \chi^2(1)$.

Tests de rangs

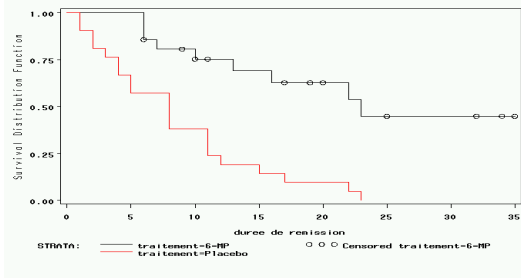
Choix des poids w_i :

- **test du log-rank** : $w_i = 1, \forall i = 1, \dots, k$.
- **test de Wilcoxon généralisé** (ou test de Breslow) :
 $w_i = n_i, \forall i = 1, \dots, k$ où n_i est le nombre de sujets exposés à la date t_i .
poids plus important pour les décès précoces
- **test de Peto** : $w_i = \hat{S}(t_i)$

Influence du choix des poids w_i sur la puissance des tests.

Exemple : comparaison suivant traitement

Durees de remission avant rechute de patients atteints de leucemie



test	Khi2	DF	Pr > Khi2
Log-Rank	16.8246	1	<0.001
Wilcoxon	13.5383	1	0.002

Rejet de H_0 : les 2 courbes de survie sont différentes \Rightarrow effet trait^t
On conclut que le traitement B à un effet positif sur la fonction de survie.

Modèle de Cox

Prise en compte de variables supplémentaires

- Introduction d'un modèle de régression adapté aux données censurées
- Modèle de Cox

Hypothèses

- Modèle à risque proportionnel : ratio des risques pour 2 individus est indépendant du temps
- Forme exponentielle du rapport des risques

Le modèle de Cox

Définition (*Modèle de Cox*)

Dans le modèle de Cox le risque instantané pour un individu i s'écrit :

$$h(t|Z_i) = \alpha_0(t) \exp(\beta_1 Z_{i,1} + \dots, \beta_p Z_{i,p}) = \alpha_0(t) e^{\beta^T Z_i}$$

où $\alpha_0(t)$ est une fonction quelconque qui ne dépend que du temps (risque de base) et $\beta_1, \beta_2, \dots, \beta_p$ sont des constantes.

- $\alpha_0(t)$, le risque de base **inconnu**, est indépendant de Z_i .
- $\beta = (\beta_1, \dots, \beta_p)^T$ est le paramètre de régression **inconnu**.

Le modèle de Cox

Hypothèses (*de validité du modèle de Cox*)

- 1 Rapport des **risques instantané de décès** (hazard rate) de deux patients **indépendant du temps** : hypothèse des **risques proportionnels**

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{\alpha_0(t)e^{\beta^T Z_i}}{\alpha_0(t)e^{\beta^T Z_j}} = \frac{e^{\beta^T Z_i}}{e^{\beta^T Z_j}}$$

- 2 $\log(h(t|Z_i)) = \log(\alpha_0(t)) + \beta^T Z_i$: hypothèse de **log-linéarité**.

Modèle de Cox avec une covariable binaire

$$Z = \begin{cases} 0 & \text{si groupe A} \\ 1 & \text{si groupe B} \end{cases} \quad \text{et } h(t|Z) = \alpha_0(t)e^{\beta Z}.$$

Rapport des risques (HR, Hazard Ratio)

$$HR = \frac{h(t|Z=1)}{h(t|Z=0)} = \frac{\alpha_0(t)e^{\beta}}{\alpha_0(t)} = e^{\beta}.$$

Interprétation :

$HR = 1 \iff \beta = 0 \iff$ même risque pour les 2 groupes.

$HR > 1 \iff \beta > 0 \iff$ risque plus élevé pour le groupe B.

$HR < 1 \iff \beta < 0 \iff$ risque plus faible pour le groupe B.

Estimation

Objectifs :

- Estimation de β
- Estimation de HR , $e^{\hat{\beta}}$.
- Calcul d'intervalles de confiance pour β et pour HR .
- Test de

$$(H_0) : \beta = 0 \text{ contre } (H_1) : \beta \neq 0.$$

Vraisemblance partielle de Cox

Le paramètre β est estimé par

$$\hat{\beta} = \arg \max L_n^{(p)}(\beta) \text{ avec } L_n^{(p)}(\beta) = \prod_{i=1}^k \frac{\exp(\beta Z_i)}{\sum_{j \in R(t_i)} \exp(\beta Z_j)},$$

où $R(t_i)$ est l'ensemble des individus à risque juste avant t_i
et t_1, \dots, t_k sont les k instants de décès.

Propriétés asymptotiques

- $\hat{\beta}$ est un estimateur consistant de β ,

$$\hat{\beta} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \beta.$$

- $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, V(\beta)).$
- On en déduit un intervalle de confiance

$$\left[\hat{\beta} - u_{\alpha} \frac{\hat{V}(\hat{\beta})}{\sqrt{n}}; \hat{\beta} + u_{\alpha} \frac{\hat{V}(\hat{\beta})}{\sqrt{n}} \right].$$

Test

On souhaite tester $(H_0) : \beta = 0$ contre $(H_1) : \beta \neq 0$. Trois tests classiques

- 1 Le test du **rapport de vraisemblance**
- 2 Le test du **score** (ou test de **Rao**)
- 3 Le test de **Wald**

Ces trois tests sont basés sur 3 statistiques qui suivent asymptotiquement, **sous H_0** , des lois du $\chi^2(p)$ à p degrés de liberté si le vecteur des paramètres β est de **dimension p** . Ici $p = 1$

Test de Wald

On teste

$$(H_0) : \beta = 0 \text{ contre } (H_1) : \beta \neq 0.$$

Statistique de test :

$$T_n = \left(\frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}} \right)^2 \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \text{sous } H_0 \mathcal{X}^2(1).$$

Zone de rejet de H_0 au niveau α :

$$ZR_\alpha = \{T_n \geq x_\alpha\} \text{ avec } \mathbb{P}(\mathcal{X}^2(1) \geq x_\alpha) = \alpha.$$

Règle de décision : on rejette H_0 si $T_n \geq x_\alpha$ ou si p -value $< \alpha$.

Résultats

Test de l'hypothèse nulle globale : $BETA=0$

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	15.1935	1	<.0001
Score	15.9069	1	<.0001
Wald	13.5589	1	0.0002

Analyse des estimations de la vraisemblance maximum

Variable	DF	Résultat estimé des paramètres	Erreur std	Khi 2	Pr > Khi 2	Rapport de risque
groupe	1	-1.50844	0.40965	13.5589	0.0002	0.221
variable groupe		Rapport de risque	95% Limites de confiance du rapport de risque			
		0.221	0.099 - 0.494			

Modèle de Cox avec une variable catégorielle Z

Codage de la variable

- première possibilité :

$$Z = \text{disease}, Z \in \{0, 1, 2, 3\} \implies h(t|Z) = \alpha_0(t)e^{\beta Z}.$$

Rapport des risques des individus des catégories 1 et 0

$$\frac{h(t|Z=1)}{h(t|Z=0)} = \frac{\exp(1\beta)}{\exp(0\beta)} = \exp(\beta)$$

Rapport des risques des individus des catégories 2 et 0

$$\frac{h(t|Z=2)}{h(t|Z=0)} = \frac{\exp(2\beta)}{\exp(0\beta)} = \exp(2\beta) = \left(\frac{h(t|Z=1)}{h(t|Z=0)} \right)^2$$

Modèle de Cox avec une variable catégorielle Z

- deuxième possibilité

$$dis1 = 1 \text{ si } disease = 1, 0 \text{ sinon}$$

$$dis2 = 1 \text{ si } disease = 2, 0 \text{ sinon}$$

$$dis3 = 1 \text{ si } disease = 3, 0 \text{ sinon}$$

$$h(t|Z) = \alpha_0(t) \exp(\beta_1 dis1 + \beta_2 dis2 + \beta_3 dis3)$$

Rapport des risques des individus des catégories 1 et 0

$$\frac{h(t|Z=1)}{h(t|Z=0)} = \frac{\exp(\beta_1)}{\exp(0)} = \exp(\beta_1)$$

Rapport des risques des individus des catégories 2 et 0

$$\frac{h(t|Z=2)}{h(t|Z=0)} = \frac{\exp(\beta_2)}{\exp(0)} = \exp(\beta_2) \neq \left(\frac{h(t|Z=1)}{h(t|Z=0)} \right)^2$$

Extensions

Modèle avec plusieurs covariables

- Introduction d'un terme d'interaction entre 2 variables catégorielles
- Interprétation du rapport des risques pour une variable quantitative
- Risque relatif de décès lié à l'exposition du facteur Z_i

$$HR = \frac{h(t|Z_i = z + 1, Z_{j \neq i})}{h(t|Z_i = z, Z_{j \neq i})} = e^{\beta_i}$$

Sélection de variables

- Modèle de Cox avec p covariables

$$h(t|Z) = \alpha_0(t) \exp(\beta_1 Z_1 + \dots, \beta_p Z_p).$$

Question : quelles sont les variables les plus pertinentes ?

- Critère AIC

Validation a posteriori du Modèle de Cox

Vérification de l'hypothèse de risques proportionnels

- Définition des résidus (Cox-Snell, Martingale, Deviance, etc)
- Représentations graphiques

Vérification de l'hypothèse de log-linéarité

- Graphe des résidus en fonction de chaque variable explicative Z_j
(aspect totalement aléatoire autour d'une droite horizontale)
- Recherche de sujets marginaux : estimation $\hat{\beta}_{(i)}$ sans le sujet i puis
graphe de $\hat{\beta} - \hat{\beta}_{(i)}$ en fonction du temps