Estimation de la survie - Tests d'hypothèse de comparaison de distribution de survie

Notion de distribution - Estimateur de Kaplan-Meier

Juste Goungounga [1] Cédric Bationo [2]

[1] Méthodologiste Biostatisticien, MD PhD, Aix Marseille Univ, Université de Bourgogne

[2] Méthodologiste Biostatisticien, Msc PhDc, Aix Marseille Univ

Dernière mise à jour : 2021-05-01

Objectifs à atteindre à la fin de cet UE

Comprendre l'analyse de survie (ou de durées de vie)

- Estimer la probabilité de survenue d'un évènement censuré donné : décès, rechute, sans récidive
- Comparer deux ou plusieurs distributions de survie
- Tester l'effet de covariables sur la probabilité de survenue d'un évènement
- Tester l'effet de traitements ou stratégies thérapeutiques sur le risque de survenue d'un évènement

Section 1

Introduction

Exemple d'application

On souhaite évaluer l'effet du tabagisme sur le délai de survenue du décès.

- Age en années
- sexe : male, female
- tabac : oui, non
- delai : nombre d'années
- status : 0 vivant, 1 décédé

age	sexe	tabac	delai	status
57.18481	female	non	0.6872005	0
52.93672	male	oui	0.3363586	0
53.92778	female	oui	0.1012370	1
51.39696	female	non	0.3252191	1
53.13683	male	oui	1.1287078	0

Exemple d'application

- Es-ce-que le tabagisme réduit le délai d'apparition du décès ?
- Quelle méthode semble appropriée pour répondre à cette question ?
- Un test de comparaison de moyennes ?
- Une regression logistique ?
- Une regression linéaire ?

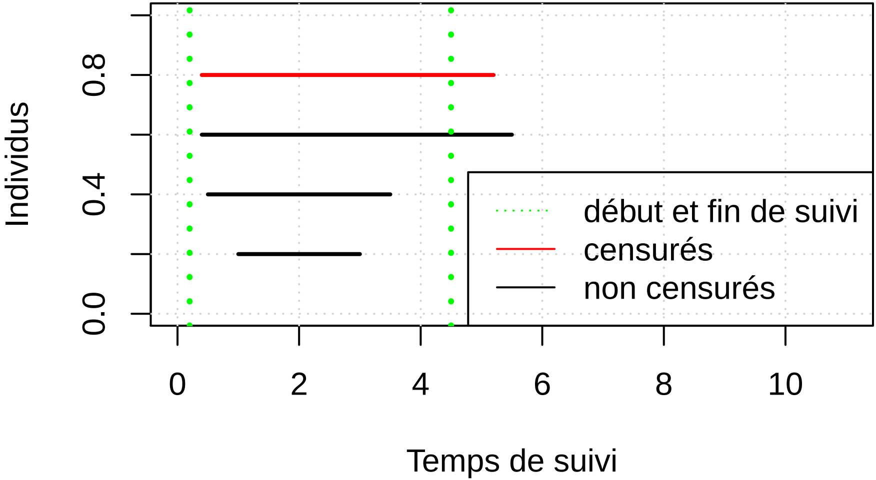
La régression linéaire

Le modèle s'écrit sous la forme suivante :

$$E[T|X_1, ..., X_k] = \beta_0 + B_1 X_1 + ... + B_k X_k$$

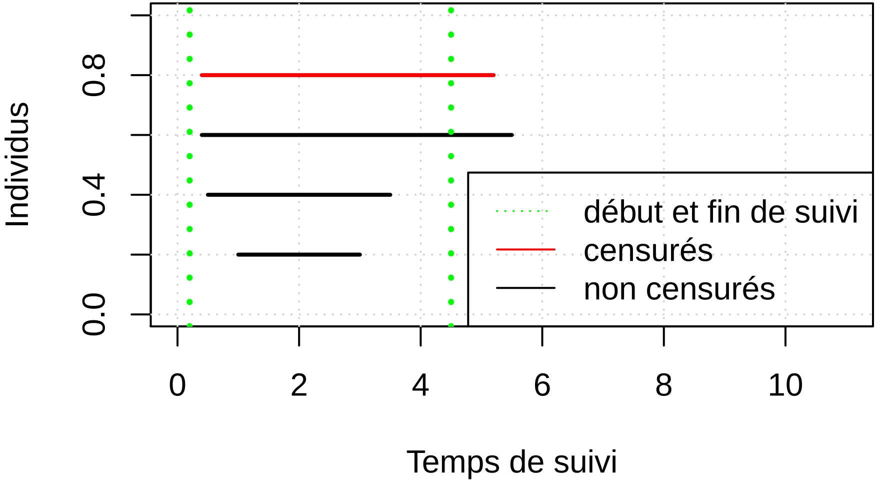
- T : Variable aléatoire - temps à l'évènement-délai écoulé jusqu'à la survenue de l'évènement (en jours, mois ou **en années**)
- $X_1, ..., X_k$ les variables explicatives d'intérêts
- $\beta_1, ..., \beta_k$ les paramétres de régressions associés aux variables d'intérêts $X_1, ..., X_k$

La régression linéaire : limites



- On ne connait pas les temps d'apparition des évènements pour certains individus

La régression linéaire : limites



- On a des données dites censurées : le suivi est arrêté avant l'apparition de l'évènement

Section 2

Quelques définitions

Dates

- date de dernières nouvelles : date la plus récente où l'on a récueilli des informations sur un individu
- date de point : date au-delà de laquelle on ne tiendra pas compt des informations sur le sujet et pour laquelle on cherchera à connaitre l'état de chaque individu; Il peut y avoir plusieurs dates de points dans une étude
- temps de participation ou temps de suivi : délai entre la date d'entrée et la date des dernières nouvelles (ou la date de point) si cette date est antérieure à la date de point.
- recul : délai entre le début de l'étude et la date de point; pour l'étude on parle de recul maximum qui est different du recul pour l'individu

Censures : cas de la censure à droite

- Existence d'observations incomplètes : une des caractéristiques des données de survie
- Censure : phénomène rencontré lorsque l'on recueil des données de survie
- censure à droite : une durée de vie (survie) est dite censurée à droite si l'individu n'a pas subi l'évènement à sa date de dernière nouvelles.
  - cas 1 "exclus vivant" : individus qui n'ont pas subi l'évènement à la date de point (censure administrative).
  - cas 2 "perdues de vue" : individu qui ont quitté l'étude à une date à laquelle il n'avait pas encore subit l'évènement

Censures : cas de la censure à droite

- on parle de censure non informative : hypothèse d'indépendance entre la cause de la censure et l'évènement étudié.
  - Exemple : perdus de vue pour des raisons telles qu'un déménagement, un refus de continuer à participer à l'étude.
- biais de censure informative : hypothèse d'indépendance entre la cause de la censure et l'évènement étudié non vérifiée; Ce biais impacte sur l'estimation du risque de présenter l'évènement
  - Exemple : perdus de vue car leur état de santé s'est gravement dégradé pour une cause liée à l'évènement étudié. On enlèverait de l'étude les personnes les plus à risque: risque de surestimation de leur probablité de survivre de la maladie étudiée.

Censures : cas de la censure à gauche

- censure à gauche : un délai de survie (durée de vie) est dit censuré à gauche si l'individu a déjà subi l'évènement avant que l'on commence à l'observer dans une étude
- On ne connait pas toujours la date exacte d'entrée dans l'étude
- Exemple : Un épidémiologiste étudie la durée d'apparition des symptomes dus à la Covid-19. La durée est une variable aléatoire T et C est la censure administrative de l'étude (1er Mai). Pour les patients qui ont déjà des symptômes de la covid-19 et diagnostiqués à la PCR, on a une censure à gauche car pour eux le délai est inconnu et inférieur à la date de censure administrative de l'étude
- Cas non fréquent dans la litterature : les critères d'inclusion exigent dans ces cas de ne pas inclure des sujets qui ont déjà subi l'évènement au moment de l'inclusion

Censures : cas de la censure par intervalles et troncature

- censure par intervalles : une durée de vie (survie) est dite censurée par intervalle si au lieu de l'observer avec exactitude, on a uniquement l'information que l'évènement a eu lieu entre deux deux dates connues.
- Troncature : une durée de vie (survie) est dite tronquée si elle est conditionnelle à un autre évènement.
  - La troncature à gauche est la plus fréquente en analyse de survie : implique d'un individu n'est observable que si sa durée de vie est supérieure à une certaine valeur.

Censures : cas de la censure par intervalles et troncature

- Exemple : durée de survie étudiée à partir d'une cohorte tirée au sort dans la population. Seuls les sujets vivants à la date de l'enquête sont étudiés et inclus dans l'enquête.
- troncature à gauche différente de la censure à gauche : certains sujets ne sont pas observables et seulement un sous-échantillon est étudié dans le cas de la troncature.

Censures : cas de la censure par intervalles et troncature

NB: Dans le cas de la censure à gauche : Informations incomplètes pour certains individus bien qu'ils soient suivis dans l'échantillon

- Dans le cas de la censure par intervalle si au lieu de l'observer avec exactitude, on a uniquement l'information que l'évènement a eu lieu entre deux deux dates connues.
- Troncature par intervalles : troncature à gauche et à droite
- Remarques : pour aller loin, voir http://www.numdam.org/article/JSFS_1994___135_4_3_0.pdf

Section 3

Distribution de survies et fonctions associées aux distributions de survie

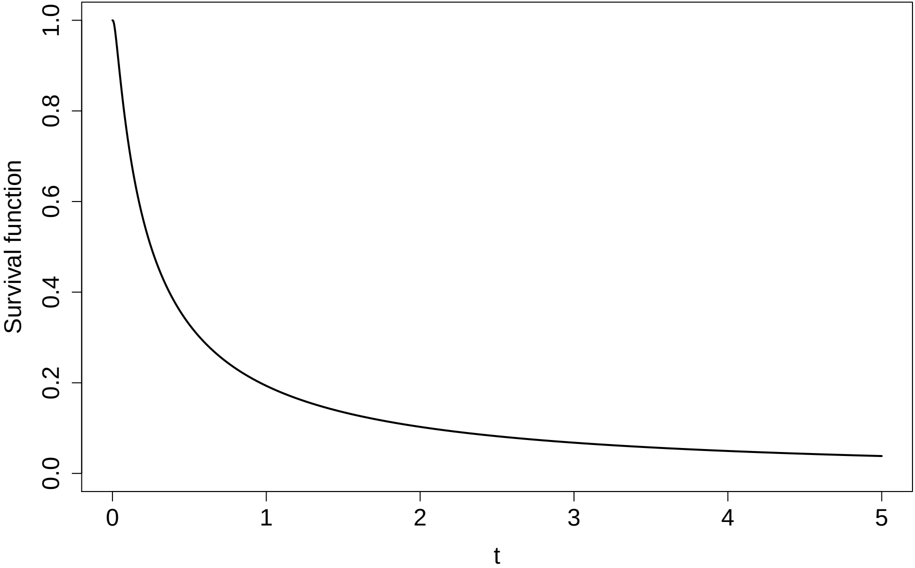
Fonction de survie

- Supposons T une variable continue dicrète
- Nous allons identifier quelques fonctions définies sur $R+$ et qui peuvent être utilisées pour représenter la loi de probabilité de T.
- la fonction de survie $S(t)$: probabilité d'être en vie jusq'au temps t. C'est une fonction décroisante de 1 à 0.

Sa forme analytique s'écrit comme suit :

$$S(t) = P(T > t)$$

Fonction de survie



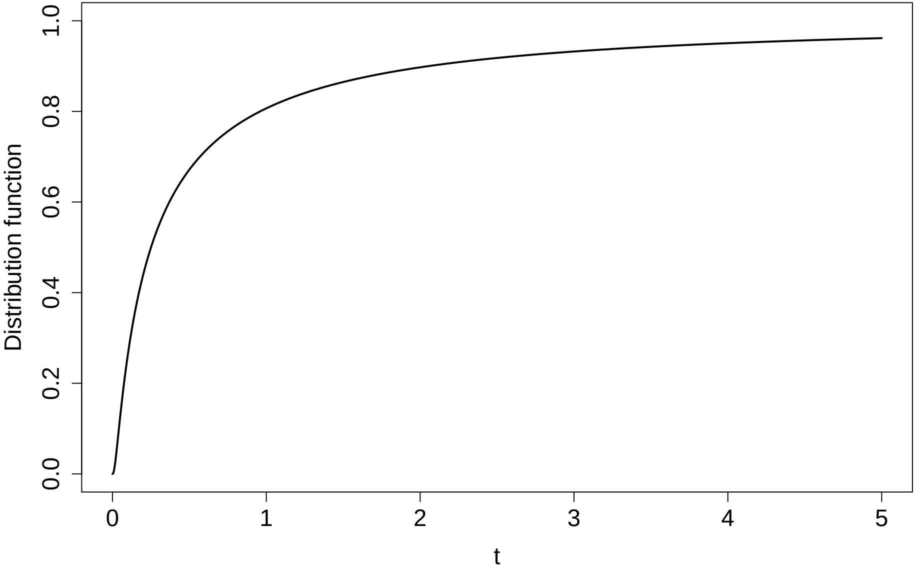
Fonction de répartition

- Supposons T une variable continue dicrète
- la fonction de répartition $F(t)$: probabilité que l'évènement se réalise entre de 0 à t.

Sa forme analytique s'écrit comme suit :

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

Fonction de répartition



Fonction de risque cumulé

- Supposons T une variable continue dicrète
- la fonction de risque cumulé $\Lambda(t)$: probabilité que l'évènement se réalise entre de 0 à t.

Sa forme analytique s'écrit comme suit :

$$\Lambda(t) = -log(S(t)) = -\int_0^t \lambda(u)du$$

avec λ la fonction de risque, i.e. la probabilité que l'évènement se réalise entre t et Δ_t

$$\lambda(t) = lim_{\Delta_t \to 0^+} = \frac{P(T \leq t \leq t + \Delta_t | t + \Delta_t \leq t)}{\Delta_t}$$

Fonction de risque cumulé : remarques

Il exite plusieurs fonctions de distribition de survie :

- La distribion exponentielle
- La distribution de Weibull
- La distribution de Weibull généralisée
- L'exponentiel de la distribution de Weibull (https://rstudio-pubs-static.s3.amazonaws.com/291890_72feb2b9a2cf4ec1bfb88b0bb1b80b34.html)
- Pour aller loin : https://www.r-bloggers.com/2019/06/parametric-survival-modeling/

Section 4

Méthode de Kaplan-Meier

Estimateur de Kaplan-Meier : Introduction

- Estimateur non paramétrique : estimateur du produit limite (1958)
- Sa définition résulte d'un raisonnement simple : ne pas avoir un évènement réalisé au temps t signifie que l'évènement ne s'est pas réalisé
- Supposons :
  - d_i : nombre d'individus pour lequel l'évènement d'intérêt se réalise au temps t_i
  - ni: nombre d'individus à risque de l'évènement d'intérêt au temps t_i
  - pour calculé la survie $S()$ au temps t_i : on calcul la probabilité que l'évènement ne se realise pas au temps t_{i-1} et la probabilité que l'évènement ne se realise pas au temps t_i sachant quel ne s'était pas réalisé au temps t_{i-1}.

Estimateur de Kaplan-Meier : calcul

$$S(t_i) = P_i * P_{i-1} * ... * P_1$$

avec $\hat{P}_i = \frac{ni-di}{ni}$

- Estimation au temps t_{i+ 1} :

$$\hat{S}(t_{i+1}) = \hat{S}(t_i) * (1 - \frac{d_{i+1}}{n_{i+1}})$$

- La médiane estimée de survie : délai T_M tel que $\hat{S}(T_M) = 0.5$

Estimateur de Kaplan-Meier : Variance et intervalle de confiance

- $\hat{S}(t)$ a asymptotiquement une distribution normale (Anderson et al. 1993) : car on utilise des proportions comme estimateurs des probabilités pour calculer $\hat{S}(t)$
- $\forall\ t \in [0; t_k[$, on a asymptotiquement $\hat{S}(t) \sim N(S(t); \hat{\sigma}_t^2$ où $\hat{\sigma}_t^2$ est la variance de $\hat{S}(t)$ estimée par la formule de Greenwood :

$$\hat{\sigma}_t^2 = [\hat{S}(t)]^2 * \sum_{j:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- On a donc l'intervalle de confiance de la survie :

$$IC_{95\%} = [\hat{S}(t) \pm 1.96 * \hat{\sigma}_t]$$

Estimateur de Kaplan-Meier : application (1)

- Cas où l'évènement est le décès
- 100 individus à risque au temps t
- 20 décès
- P_i la probabilité conditionelle de ne pas décéder à ce temps est :

$$\frac{100-20}{100} = 4/5$$

- La probabilité conditionnelle de décès à ce temps est :

$$\frac{20}{100} = 1/5$$

Estimateur de Kaplan-Meier : application (2)

Tableau : Estimation de survie (Kaplan-Meier) de 16 femmes de l'étude Paquid (délai depuis l'entrée en institution)

temps décès	Nombre décès	effectifs à risque	prob condi	Survie estimée
t_i	d_i	n_i	((ni-di)/ni)	S(t)
0	0	16	1.00	1.00
0.32	1	16	0.94	0.94
0.70	1	15	0.93	0.88
2.01	1	14	0.93	0.81
2.31	1	12	0.92	0.74
2.95	1	11	0.91	0.68
3.26	1	10	0.90	0.61

Estimateur de Kaplan-Meier : application (2)

Détail du calcul dans l'étude paquid:

- $S(t = 0) = 1$
- $S(t = 0.32) = S(t = 0) * 0.94 = 0.94$
- $S(t = 0.70) = S(t = 0.32) * 0.93 = 0.8742$
- $S(t = 2.01) = S(t = 0.70) * 0.93 = 0.813006$

Section 5

Test de Logrank

Hypothèses

- Test basé sur les rangs des temps d'évènements
- Vise à tester une hypothèse nulle (H0) contre une hypothèse alternative (H1).
- Formulation dans le cadre de comparaison de deux courbes A et B de survie :
    - Hypothèse nulle (H0) : les courbes de survie ne sont pas différentes soit $SA(t) = SB(t)$ pour tout $t > 0$.
    - Hypothèse alternative (H1) : les courbes de survie sont différentes soit $S_A(t) \neq S_B(t)$

Conditions d'application

- Peu ou pas d'ex-aequos
- Indépendance des temps d'évènements

Paramètre et loi statistique du test

- Test statistique est basé sur la statistique du Khi-deux.
- Sous l'hypothèse H0 on calculera le nombre de décès (ou événements) « attendus » dans chacun des groupes (nombre de décès estimés).
- On comparera ensuite le nombre de décès (ou événements) observés aux nombre de décès attendus.

On définit :

- d_{li} : nombre d'évènements observés en t_i dans deux groupes $l = 1, 2$
- d_i : nombre total d'évènements observés en t_i avec $d_i = d_{1i} + d_{2i}$
- n_{li} : nombre de sujets à risque en t_i (temps observés d'évènements) dans le groupe l avec $n_i = n_{1i} + n_{2i}$

Statistique de test (1)

On peut présenter les observations dans un tableau de contingence :

```
-----------------------------------------
         |Evenements | Censures     |
------------------------------------------
Groupe 1 |d_{1i}     |n_{1i}-d_{1i} | n_{1i}
Groupe 2 |d_{2i}     |n_{2i}-d_{2i} | n_{2i}
------------------------------------------
         |d_i        |n_{i}-d_{i}   | n_i
```

Statistique de test (2)

Sous HO , le nombre d'évènements attendus pour le groupe l, au temps t_i, est une variable aléatoire distribué selon une loi hypergéométrique d'espérance :

$$e_{li} = n_{li}\frac{di}{ni}$$

et de variance identique pour les deux groupes

$$v_i = \frac{d_i(n_i - d_i)n_{1i}n_{2i}}{n_i^2(n_i - 1)}$$

La statistique de test du logrank suit assymptotiquement , sous H0 une loi de Khi-deux (χ^2) à 1 degré de liberté :

$$X^2 = \frac{(\sum_{i=1}^{k} d_{1i} - \sum_{i=1}^{k} e_{1i})^2}{\sum_{i=1}^{k} v_i}$$

Section 6

Quelques références

Quelques références

Commenges, D., & Jacqmin-Gadda, H. (2015). Modèles biostatistiques pour l'épidémiologie. De Boeck Superieur.

Hill Catherine, Com-Nougué, Catherine, & Kramar, Andrew. (1990). Analyse statistique des données de survie.

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In Modeling survival data: extending the Cox model (pp. 39-77). Springer, New York, NY.

Collett, D. (2015). Modelling survival data in medical research. CRC press.

Moore, D. F. (2016). Applied survival analysis using R. Switzerland: Springer.