Analyse de données censurées : Introduction

Bibliographie

- Hill, C., Com-Nougué, C., & Kramar, A. (1990). Analyse statistique des données de survie. INSERM; Médecine-Sciences-Flammarion.
- Collett, D. (2003). Modelling Survival Data in Medical Research. Chapman et Hall/CRC.

Types de données censurées - la censure à droite

Données de suivi de patients

Deux domaines principaux d'application :

- médecine, biologie
- actuariat, banque, assurance

On s'intéresse souvent à des durées :

1. Durée de survie de patients ayant eu un infarctus
2. Durée de rémission d'une leucémie aigüe
3. Durée de séropositivité sans symptôme de patients infectés par le VIH
4. Durée de fièvre chez un patient atteint de pneumonie
5. Durée de chomâge
6. Durée avant de faire jouer son assurance.

On distingue **l'évènement d'intérêt**

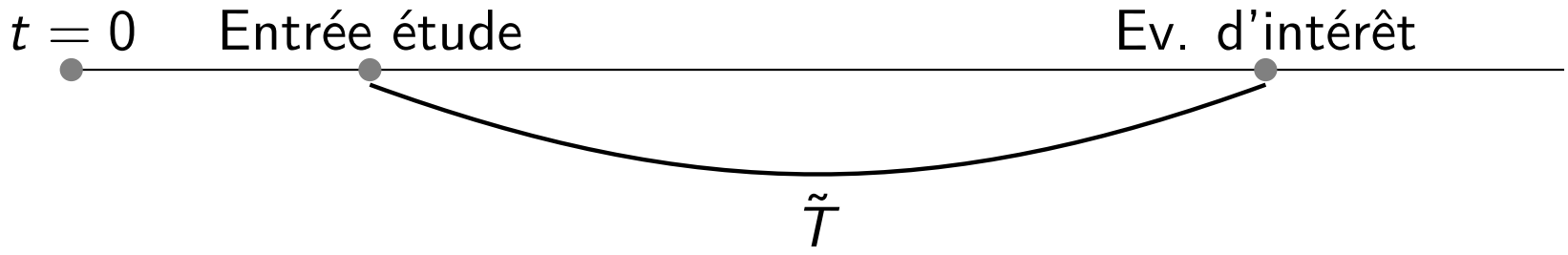
1. Décès du patient après l'infarctus
2. Fin de la rémission
3. Début des symptômes chez un patient séropositif
4. Fin de la fièvre chez un patient atteint de pneumonie,
5. Nouveau travail
6. Moment où l'on fait jouer son assurance

de la **variable à expliquer** ou **variable d'intérêt** : durée avant l'apparition de l'évènement d'intérêt

1. Temps écoulé avant le décès
2. Temps écoulé avant la fin de la rémission
3. Temps écoulé sans symptôme
4. Temps écoulé avant la fin de la fièvre,
5. Temps écoulé avant de retrouver un travail
6. Temps écoulé avant de faire jouer son assurance.

D'un point de vue statistique on note $\tilde{T}$ la variable aléatoire d'intérêt.

- Elle représente un **temps**, c'est donc une variable aléatoire **continue** et **positive**.
- $\tilde{T}$ représente le temps écoulé depuis l'entrée dans l'étude et l'apparition de l'évènement d'intérêt.

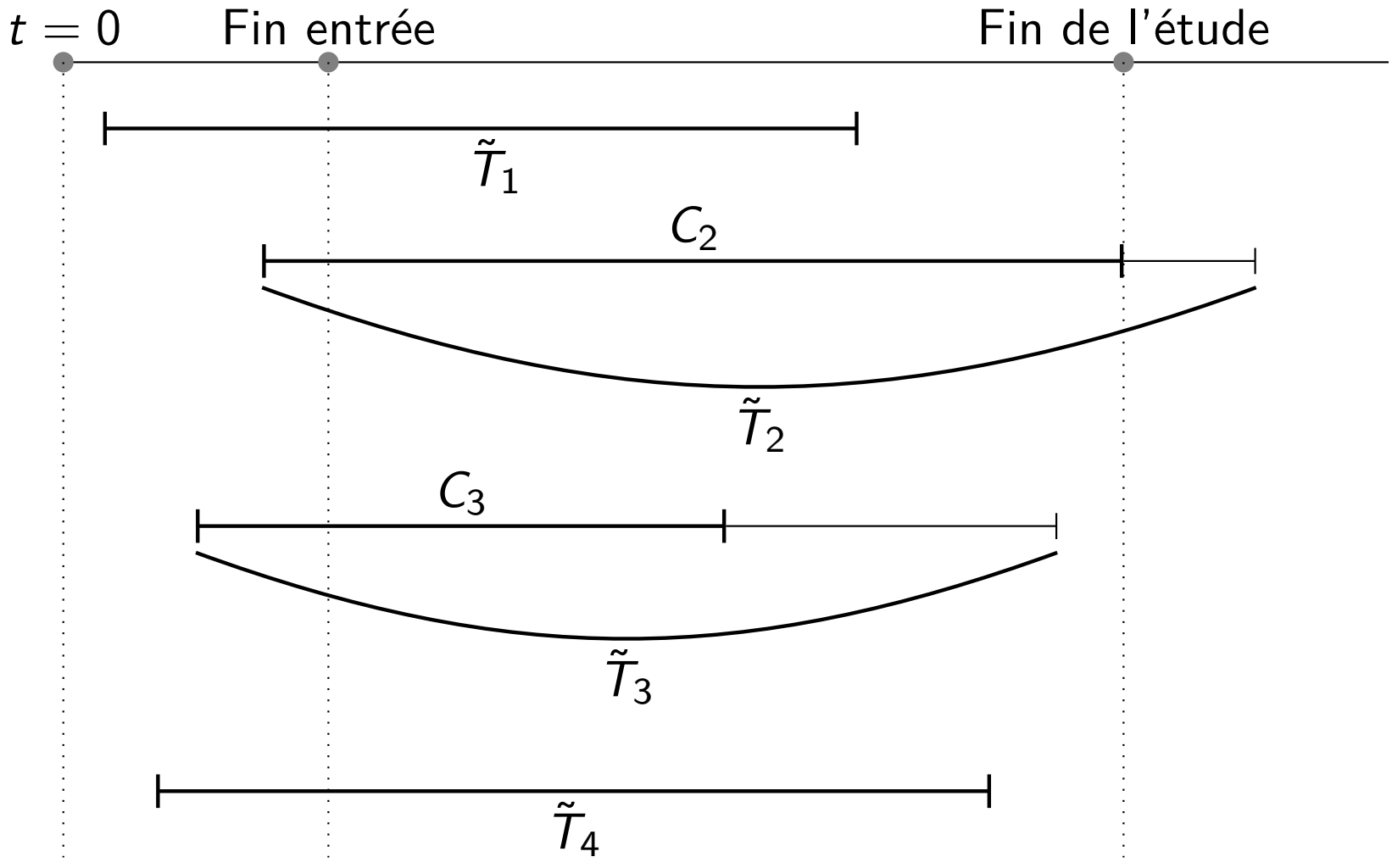


- $t = 0$ représente le début de l'étude.

La censure à droite

- Dans une étude où l'on suit des individus au cours du temps ("follow-up study'' en anglais), les individus rentrent à des instants différents et l'évènement d'intérêt se produit à des instants différents.
- Chez certains individus, on n'observe pas le moment où se produit l'évènement d'intérêt à cause de la **censure à droite** :
  - personne perdue de vue (elle quitte l'étude spontanément, elle déménage ...)
  - l'étude s'est terminée avant que l'évènement ne se produise (censure administrative).

La censure à droite : un exemple sur quatre patients



La censure à droite : un exemple sur quatre patients

Dans cet exemple,

- pour l'individu 1, la variable d'intérêt $\tilde{T}_1$ est observée.
- pour l'individu 2, la variable d'intérêt $\tilde{T}_2$ n'est pas observée. Seule la variable C_2 est observée ! C'est un cas de **censure administrative**.
- pour l'individu 3, la variable d'intérêt $\tilde{T}_3$ n'est pas observée (individu **perdu de vue**). Seule la variable C_3 est observée !
- pour l'individu 4, la variable d'intérêt $\tilde{T}_4$ est observée.

La censure à droite

- On est en présence de censure à droite quand **l'évènement d'intérêt n'est pas toujours observé**.
- Pour les individus censurés, on observe une durée **plus petite** que la variable d'intérêt.

D'une manière générale, on note les observations, pour $i = 1, \ldots, n$, de la façon suivante :

$$\begin{cases} T_i = \min(\tilde{T}_i, C_i) \\ \Delta_i = I(\tilde{T}_i \leq C_i). \end{cases}$$

Exemple 1 : les données de Freireich (1963)

Durees de remission (en semaines) obtenue par des stéroïdes chez des patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP).

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

Le signe $+$ correspond à des patients qui ont quitté l'étude à la date considérée. Pour ces individus les durées sont donc censurées à droite. Par exemple le 4 ieme patient est perdu de vue au bout de 6 semaines de traitement avec le 6-MP : il a donc une durée de rémission supérieure à 6 semaines.

Exemple 2 : données (fictives) d'abonnement téléphonique

Durées d'abonnement (en mois) à un forfait mobile

Durée observée	Fournisseur téléphonique	Durée observée	Fournisseur téléphonique
8	A	220	N
8	N	365+	N
13	A	632	N
18	A	700	N
23	A	852+	N
52	A	1296	N
63	A	1296+	N
63	A	1328+	N
70	N	1460+	N
76	N	1976+	N
180	N	1990+	N
195	N	2240+	N
210	N		

Les données avec + sont censurées. A et N : deux fournisseurs différents.

Les données censurées requièrent un traitement particulier

Si on enlève les données censurées on perd de l'information !

- Dans l'exemple 1, si on enlève les données censurées, on ne tient pas compte des durées de rémission les plus longues et on sous évalue l'effet du traitement 6-MP.
- Dans l'exemple 2, si on enlève les données censurées, c'est à dire les 8 valeurs censurées on ne tient pas compte des clients qui ont justement les durées d'abonnement les plus longues.

Si l'on ne prend pas en compte la censure, en faisant comme si une donnée censurée est égale à notre variable d'intérêt on aura tendance à sous évaluer les durées !

Les fonctions types utilisées en analyse de survie - Introduction et propriétés

Cinq fonctions essentielles ($t \geq 0$)

- la fonction de répartition : $F(t) = \mathbb{P}[\tilde{T} \leq t]$.
- la fonction de survie : $S(t) = \mathbb{P}[\tilde{T} > t]$.
- la densité :

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t]}{\Delta t}$$

- le risque instantané ou “hazard rate’ ’ en anglais :

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t]}{\Delta t}$$

- le risque cumulé $H(t)$:

$$H(t) = \int_0^t h(s) ds.$$

Cinq fonctions essentielles

- F est une fonction **croissante** telle que $F(0) = 0$ et $\lim_{t\to\infty} F(t) = 1$.
- S est une fonction **décroissante** telle que $S(0) = 1$ et $\lim_{t\to\infty} S(t) = 0$. On a la relation : $S(t) = 1 - F(t)$.
- la densité mesure la probabilité que l'évènement d'intérêt se produise sur un intervalle de temps infinitésimal. On a les relations : $F(t) = \int_0^t f(x)dx$ et $F'(t) = f(t)$.
- le risque instantané mesure la probabilité que l'évènement d'intérêt se produise sur un intervalle de temps infinitésimal ($[t, t + \Delta t[$), conditonnellement au fait d'**être à risque** que l'évènement se produise ($\tilde{T} \geq t$).

Si le temps est en jours : $h(t)$ est la probabilité de décéder le jour $t + 1$ sachant que l'on est encore en vie au jour t.

Relations entre risque instantané et survie (1)

D'après la formule de Bayes :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t, \tilde{T} \geq t]}{\Delta t} \times \frac{1}{P[\tilde{T} \geq t]} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Puisque $S'(t) = (1 - F(t))' = -f(t)$, on a également :

$$h(t) = -\frac{S'(t)}{S(t)}.$$

Relations entre risque instantané et survie (2)

Par ailleurs $-S'(t)/S(t) = [-\log(S(t))]'$ et donc

$$h(t) = [-\log(S(t))]'$$
$$\int_0^t h(u)du = [-\log(S(t))]$$
$$\exp\left(-\int_0^t h(u)du\right) = S(t).$$

La f.d.r, la fonction de survie, la densité, le risque instantané et le risque cumulé caractérisent tous la loi de $\tilde{T}$, de telle sorte que si l'on connaît l'une de ces cinq fonctions on peut retrouver les quatre autres !

Exemples de risques instantanés

- le risque instantané **constant** : $h(t) = a$. On a alors $S(t) = \exp(-at)$, $\tilde{T} \sim \mathcal{E}(a)$. C'est une loi dite sans mémoire ! Loi typique pour étudier la durée de vie de composants électroniques, d'une ampoule etc.
- le risque instantané **croissant** : il y a "veillissement' ' ! C'est le risque le plus classique. Par exemple, pour modéliser la durée de vie de patients à partir du moment qu'ils ont développé une maladie pulmonaire chronique, qu'ils ont eu un certain cancer etc. Egalement, la durée de vie d'une voiture à partir du premier accident ...
- le risque instantané **décroissant** : assez rare en pratique, révèle généralement de l'hétérogénéité dans la survie des patients. Un exemple classique est la mortalité infantile jusqu'à un an dans un pays "pauvre' '.

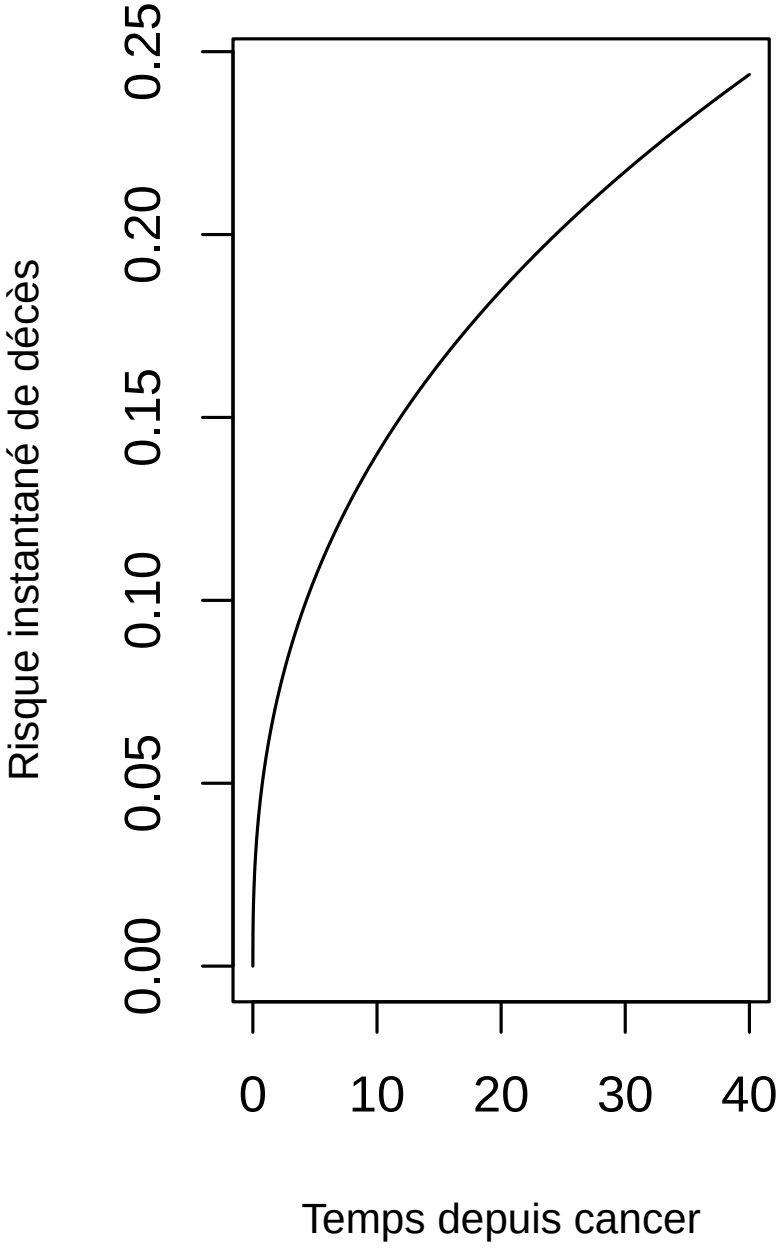
Exemples de risque instantané croissant/décroissant

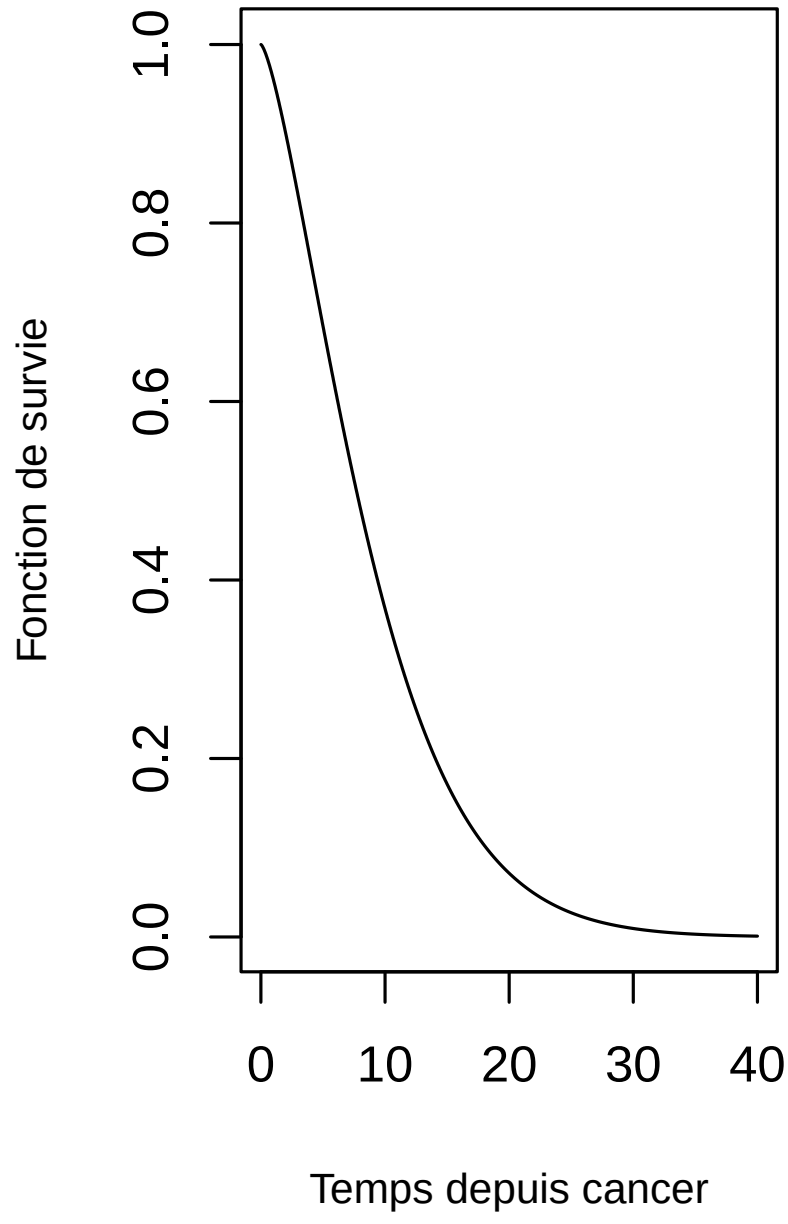
On peut modéliser un risque instantané monotone croissant/décroissant par une loi de Weibull de paramètre de forme a ($a > 1$ risque croissant, $a < 1$ risque décroissant) et de paramètre d'échelle b.

$$h(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1}$$

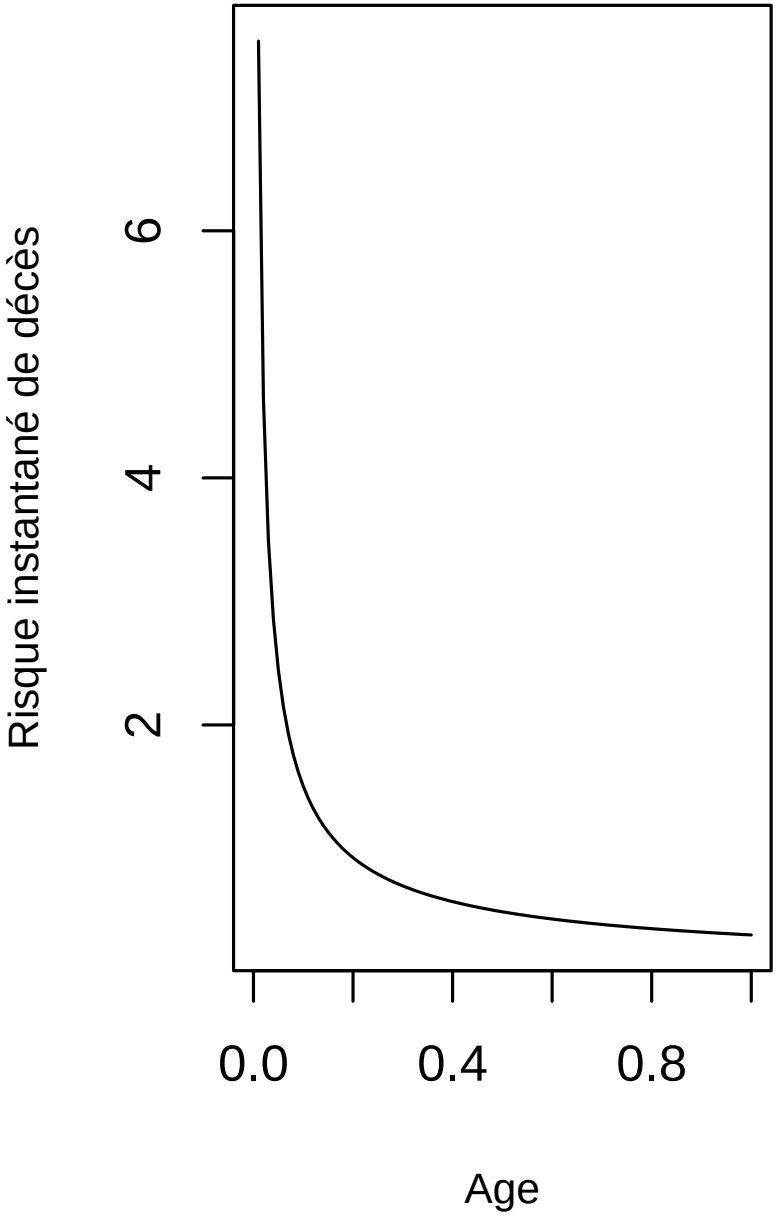
$$S(t) = \exp\left(-\left(\frac{t}{b}\right)^{a}\right).$$

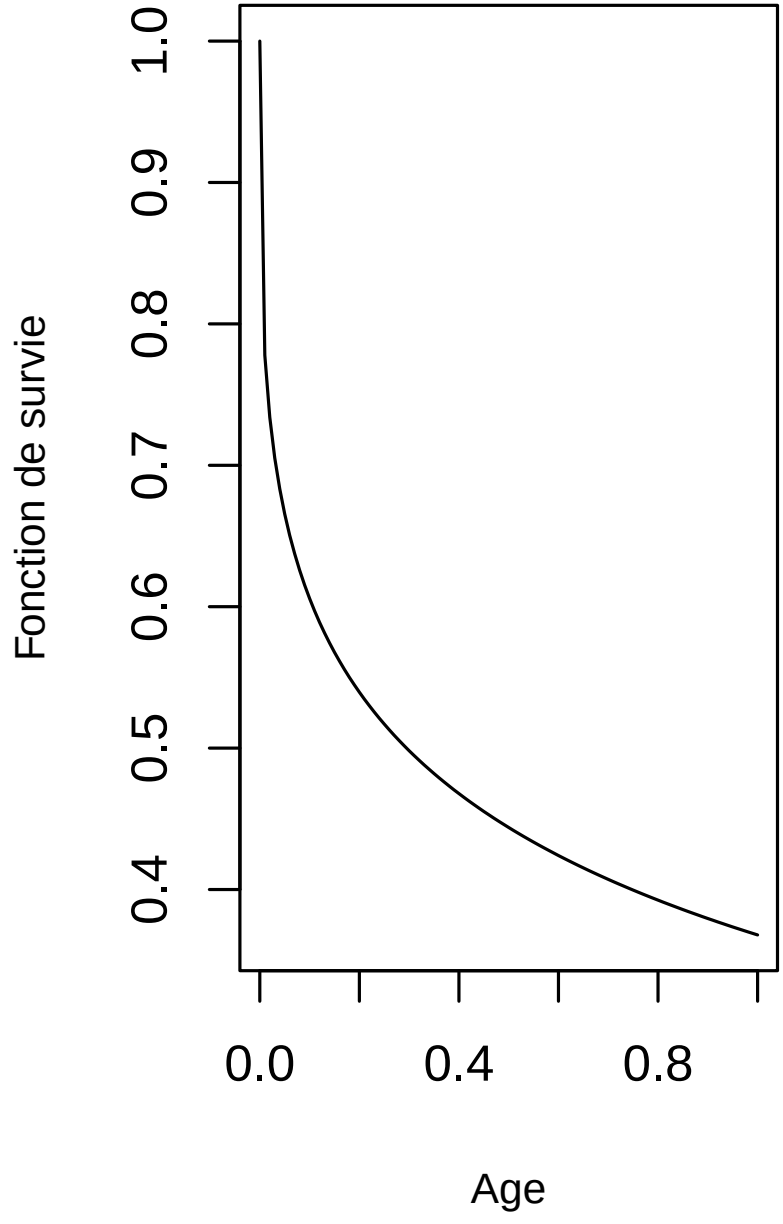
Exemples de risque instantané croissant ($a = 1.4, b = 10$)





Exemples de risque instantané décroissant ($a = 0.3, b = 1$)





Exemples de risques instantanés

- Il existe aussi des risques instantanés en forme de $\cup$ (“bathtub’ ’ en anglais) : risque diminue en début, se stabilise ensuite puis risque élevé en fin de vie.
- Enfin, il existe des risques instantanés en forme de $\cap$: risque élevé en début, se stabilise ensuite puis risque faible en fin de vie.

Quelques propriétés sur la fonction de survie

- Dans le chapitre suivant on verra comment estimer le risque instantané, sans introduire de **biais**, à partir des observations :

$$\begin{cases} T_i = \min(\tilde{T}_i, C_i) \\ \Delta_i = I(\tilde{T}_i \leq C_i). \end{cases}$$

- Une fois le risque instantané estimé on pourra estimer la fonction de survie.
- A partir de cet estimateur on essaiera de retrouver des quantités d'intérêt comme la médiane de $\tilde{T}$, les quartiles, la moyenne, la variance etc.

Quelques propriétés sur la fonction de survie

- Pour l'espérance et la variance, on peut montrer qu'on a :

$$\mathbb{E}[\tilde{T}] = \int_0^\infty S(t)dt$$
$$\mathbb{V}[\tilde{T}] = 2\int_0^\infty S(t)dt - \left(\int_0^\infty S(t)dt\right)^2.$$

- Le quantile d'ordre p de $\tilde{T}$, $Q(p)$, est égal à :

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

Exemple d'étude biaisée qui ne prend pas en compte la censure

Etude sur la durée de vie des gauchers

Les droitiers vivraient 9 ans de plus en moyenne que les gauchers d'après une étude chez Nature (1980) et une autre étude chez Science (1990) ...

http://www.bbc.com/news/magazine-23988352