MATH 60604
Modélisation statistique
§ 7b - Vraisemblance et analyse de survie

HEC Montréal
Département de sciences de la décision

Fonctions de survie et de risque

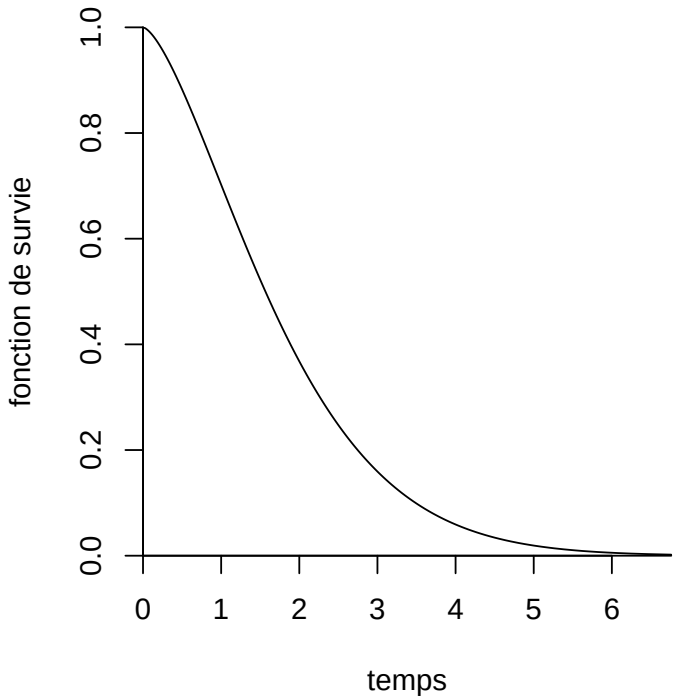
Soit T la variable aléatoire représentant le temps de survie

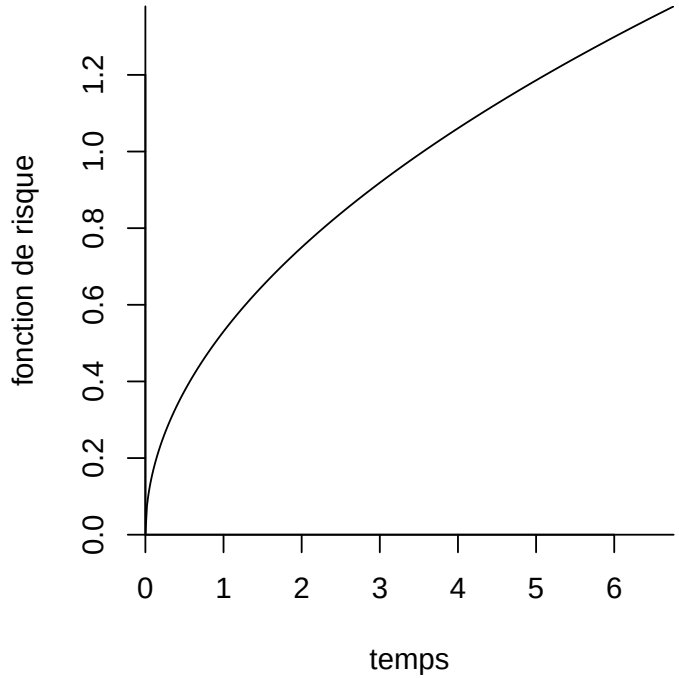
- La fonction de survie, $S(t) = \mathrm{P}\left(T > t\right)$, caractérise complètement la loi de T.
- On veut souvent savoir quelles périodes présentent un plus fort taux de défaillance. La fonction de risque (taux de défaillance, taux de risque) de T est

$$\begin{aligned} h(t) &= \lim_{\delta \to 0} \frac{\mathrm{P}\left(t < T < t + \delta \mid T > t\right)}{\delta} \\ &= \lim_{\delta \to 0} \frac{1}{\delta} \frac{\mathrm{P}\left(t < T < t + \delta\right)}{\mathrm{P}\left(T > t\right)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

On peut interpréter la fonction de risque comme étant la probabilité instantanée de “mourir” au temps t, compte tenu de la survie jusqu'au temps t.

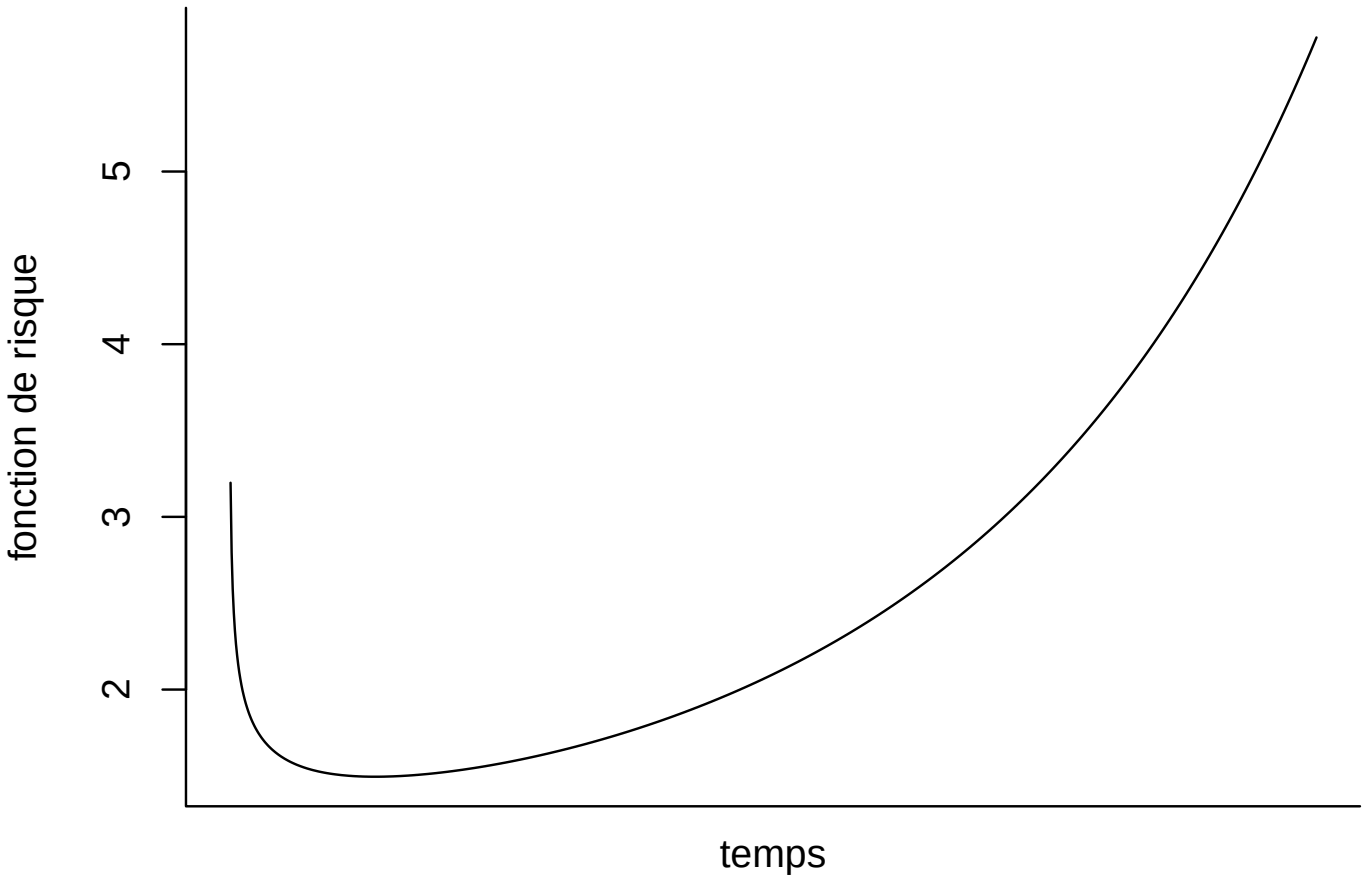
Fonctions de survie et de risque





La fonction de survie décroît de $S(0) = 1$ de manière monotone. Plus le risque $h(t)$ est élevé, plus la survie décroît rapidement.

Fonction de risque en forme de bain



Le risque initial est fort (mortalité infantile, défaut de fabrication) initialement, puis décroît et se stabilise. Au fil du temps, le risque augmente de nouveau (défaillance accrue avec l'âge).

Censure aléatoire et vraisemblance

On observe $T_i = \min\{T_i^0, C_i\}$. Si une observation est censurée à droite au temps c, on sait que $S(c) = \mathrm{P}\left(T_i^0 > c\right)$

- en d'autres mots, le temps de survie excède c.

Si on a de la censure aléatoire, la base de données contient un indicateur δ_i où

$$T_i = \begin{cases} T_i^0, & \delta_i = 1 \text{ (événement observé)} \\ C_i, & \delta_i = 0 \text{ (censure à droite)} \end{cases}$$

Contribution à la vraisemblance

Soit $S(t;\boldsymbol{\theta}) = \mathrm{P}\left(T_i^0 > t\right)$ la fonction de survie de T_i^0. Avec T_i^0 indépendant de C_i, chaque observation contribue

$$L_i(\boldsymbol{\theta}) = \begin{cases} f(t_i;\boldsymbol{\theta}), & \delta_i = 1 \text{ (événement observé)} \\ S(t_i;\boldsymbol{\theta}), & \delta_i = 0 \text{ (censure à droite)} \end{cases}$$

à la vraisemblance. La log vraisemblance s'écrit

$$\ell(\boldsymbol{\theta}) \equiv \sum_{i:\delta_i=1} \ln f(t_i;\boldsymbol{\theta}) + \sum_{i:\delta_i=0} \ln S(t_i;\boldsymbol{\theta})$$

Approches pour l'inférence

Plusieurs approches s'offrent à nous pour estimer la fonction de survie (ou la fonction de risque).

- paramétrique: choisir une famille de lois (Weibull, log normale, Gompertz, exponentielle) pour T.
  - \+ permet d'incorporer des variables explicatives aisément
  - \+ estimés continus, permet d'extrapoler la courbe
  - − notre modèle peut être mal spécifié
  - − peu flexible: la loi peut mal s'ajuster aux données
- nonparamétrique: aucune distribution assumée.
  - − pas de variables explicatives
  - \+ hypothèse minimales, garanties théoriques quand la taille de l'échantillon n est grande.
  - \+ flexible.
  - − estimés discontinus,
  - − on ne peux extrapoler au delà du plus grand temps de défaillance observé.

Modèles paramétriques pour la survie: loi exponentielle

Soit $T_i \overset{\text{iid}}{\sim} \mathrm{E}(\lambda)$ des variables exponentielles d'espérance λ^{-1}.

- La fonction de survie de T est $S(T) = \exp(-\lambda t)$ et
- la fonction de risque $h(t) = \lambda$ est **constante**.

La log vraisemblance pour un échantillon aléatoire de taille n s'écrit

$$\ell(\lambda) = \sum_{i=1}^{n} \{\delta_i \ln \lambda - \lambda T_i\}.$$

Le maximum de vraisemblance est $\widehat{\lambda} = \sum_{i=1}^n \delta_i / \sum_{i=1}^n T_i$.

- L'estimé du temps de survie est infini si aucune défaillance n'est observée.
- On obtient les erreurs-types à l'aide de la matrice d'information observée $j(\widehat{\lambda}) = \sum_{i=1}^n \delta_i / \widehat{\lambda}^2$; les données censurées ne contribuent pas d'information.