Données longitudinales et modèles de survie

6. Modèles paramétriques

André Berchtold

Département des sciences économiques, Université de Genève

Cours de Master

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES
ÉCONOMIQUES ET SOCIALES
Département des sciences
économiques

Plan du cours

Plan du cours

Paramétrique ou semi-paramétrique ?

- Le modèle de Cox est un modèle semi-paramétrique, car la distribution exacte du risque ou de la survie n'est jamais connue, même si les coefficients du modèle ont pu être estimés à partir d'un ensemble de données. De même, le risque de base et la fonction de survie de base ne sont pas spécifiés.
- Par opposition, un modèle paramétrique est un modèle dans lequel les temps de survie sont supposés être distribués selon une loi parfaitement connue.
- Les distributions les plus couramment utilisées sont : Exponentielle, Weibull, Gompertz, Log-logistique, Pareto.

Avantages de l'approche paramétrique

- Théoriquement, une courbe de survie est une fonction valant 1 au temps 0 et 0 à l'infini :

$$S(0) = 1 \text{ et } S(\infty) = 0$$

- Lorsque cette courbe est approximée soit par des méthodes de type Kaplan-Meier, soit par un modèle semi-paramétrique, le résultat peut être un peu différent pour deux raisons :
  1. S'il y a peu de données à disposition, la "courbe" aura plutôt une forme en escalier.
  2. S'il reste des sujets à risque de subir l'événement en fin d'étude, alors la courbe de survie n'atteindra pas son minimum de zéro.
- Ces problèmes n'existent pas avec les modèles paramétriques.

Désavantages de l'approche paramétrique

- Pour utiliser l'approche paramétrique, il faut avoir de bonnes raisons de penser que les temps de survie sont approximativement distribués selon une certaine loi connue plutôt qu'une autre.

Densité, hasard et survie (1)

- Dans les modèles paramétriques, le temps de survie est distribué selon une loi dont la fonction de densité, $f(t)$, s'exprime en fonction de paramètres.
- Une fois que ces paramètres sont connus (estimés), il est possible de calculer à la fois la fonction de hasard instantané, $h(t)$, et celle de survie, $S(t)$.
- En pratique, la connaissance de l'une des trois fonctions permet d'obtenir les deux autres à l'aide de relations mathématiques.

Densité, hasard et survie (2)

- Survie en fonction de la densité :

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du$$

- Survie en fonction du hasard :

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

- Densité en fonction du hasard et de la survie :

$$f(t) = h(t)S(t)$$

Hypothèses PH et AFT (1)

- Le modèle de Cox est un modèle à hasard proportionnel (Proportional Hazard : PH), c'est-à-dire que le rapport des risques pour deux individus est constant et indépendant du temps.
- Les modèles paramétriques n'ont pas tous besoin de cette hypothèse. En revanche, ils peuvent reposer sur une autre hypothèse : AFT (Acceleration Failure Time model).
- Certains modèles paramétriques comme le modèle exponentiel et celui de Weibull peuvent même s'adapter indifféremment à l'une ou l'autre de ces hypothèses. En revanche, le modèle log-logistique par exemple ne peut pas respecter l'hypothèse PH, et le modèle de Gompertz est PH, mais pas AFT.

Hypothèses PH et AFT (2)

- L'hypothèse PH signifie en pratique que l'effet des variables explicatives est multiplicatif par rapport au hasard.
- En revanche, l'hypothèse AFT signifie que l'effet des variables explicatives est multiplicatif par rapport au temps de survie.
- Le respect de l'une ou l'autre des hypothèses implique une interprétation différente des paramètres.

Facteur d'accélération (1)

- L'hypothèse AFT est plus facile à comprendre à partir de la notion de facteur d'accélération. Supposons que dans une population, les non-fumeurs aient une durée de vie en moyenne 1.1 fois supérieure à celle des fumeurs. Nous pouvons écrire

$$S_{nf}(t) = 1.1 S_f(t) = \gamma S_f(t)$$

- Le paramètre γ est le facteur d'accélération. Il s'interprète comme un facteur de dilatation ou de contraction du temps : La probabilité qu'un fumeur survive 75 ans est égale à la probabilité qu'un non-fumeur survive $75 \cdot 1.1 = 82.5$ années.

Facteur d'accélération (2)

- Formellement, on peut définir γ comme le ratio constant des temps de survie de deux groupes de personnes pour n'importe quel valeur $S(t)$.
- Par exemple, si $\gamma = 1.1$, alors le temps médian de survie du groupe des non-fumeurs est 1.1 fois celui du groupe des fumeurs.

Plan du cours

Distribution exponentielle

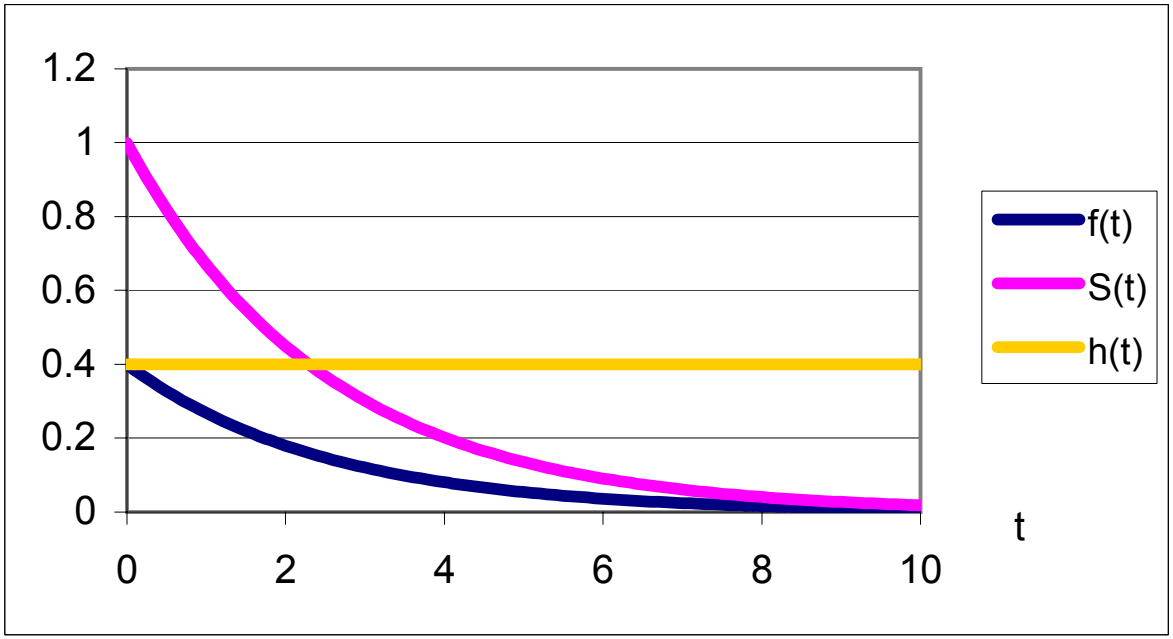
- Distribution du temps d'attente entre deux événements qui surviennent indépendamment de façon purement aléatoire (processus de Poisson).
- $\Rightarrow$ risque instantané constant

$$\begin{aligned} h(t) &= \lambda\,, \quad \lambda > 0 \\ S(t) &= \exp(-\lambda t) \\ f(t) &= \lambda \exp(-\lambda t) \end{aligned}$$

$$\mathrm{E}(T) = \frac{1}{\lambda} \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}$$

- Plus le risque λ est grand, plus l'espérance de survie est faible.

Forme de la distribution exponentielle



$\lambda = 0.4$,

$\mathrm{E}(T) = 1/0.4 = 2.5$, $\mathrm{Var}(T) = (1/0.4)^2 = 6.25$, $\sigma_T = 2.5$

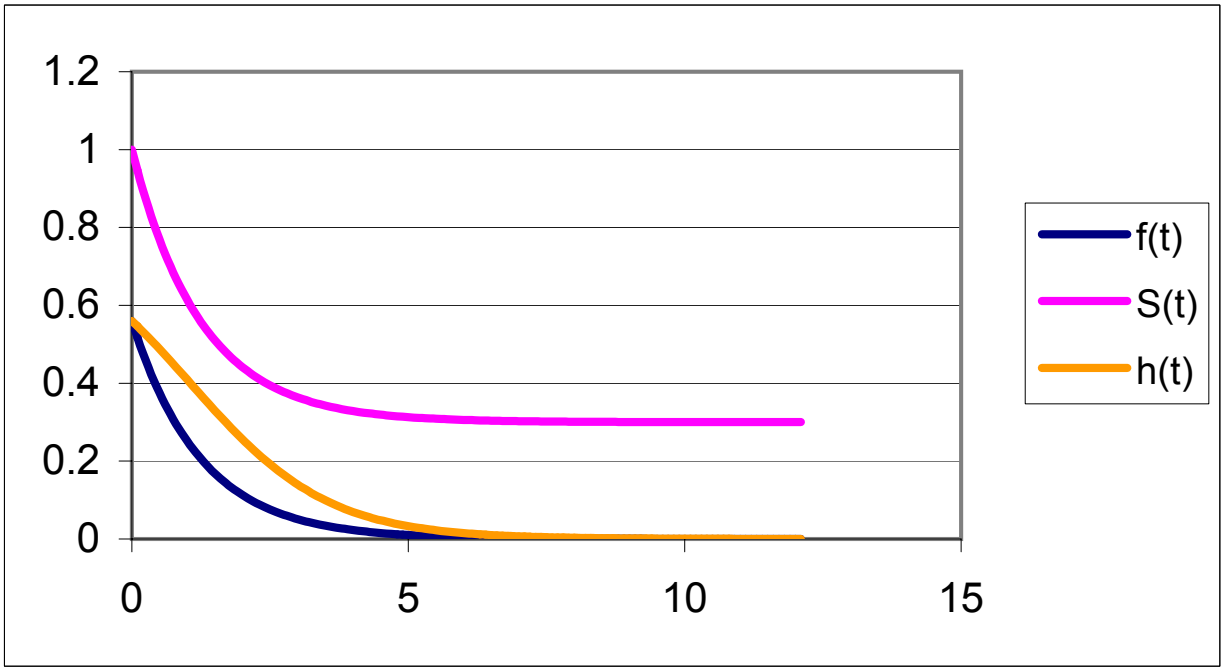
Mélange de deux distributions exponentielles

- Problème appelé “migrant-sédentaire” (“mover-stayer’).
- Soit deux sous-populations avec risque instantané constant : $h_1(t) = \lambda, \forall t$; $h_2(t) = 0, \forall t$. Alors,

	population 1	population 2	en tout
proportion	w	$1-w$	1
densité	$\lambda \exp(-\lambda t)$	0	$w\lambda \exp(-\lambda t)$
survie	$\exp(-\lambda t)$	1	$(1-w) + w \exp(-\lambda t)$
risque instantané	$h_1(t) = \lambda$	$h_2(t) = 0$	$\frac{\lambda}{1 + \frac{1-w}{w} \exp(\lambda t)}$
espérance $E(T)$	$1/\lambda$	∞	w/λ

- Le risque instantané du mélange est non constant.

Forme du mélange de deux distributions exponentielles



$\lambda_1 = 0.8$
$\lambda_2 = 0$
$w = 0.7$
$E(T) = 0.875$

Distribution de Weibull (1)

- C'est une généralisation de la loi exponentielle qui correspond au cas $\alpha = 1$.
- Le risque h croît ou décroît de façon monotone.
- Le paramètre α modifie la forme de la distribution.

$$
\begin{aligned}
h(t) &= \lambda\alpha(\lambda t)^{\alpha-1}\ , \quad t > 0,\ \ \lambda > 0,\ \ \alpha > 0 \\
S(t) &= \exp(-(\lambda t)^{\alpha}) \\
f(t) &= \lambda\alpha(\lambda t)^{\alpha-1}\ \exp(-(\lambda t)^{\alpha})
\end{aligned}
$$

Distribution de Weibull (2)

- Caractéristiques :

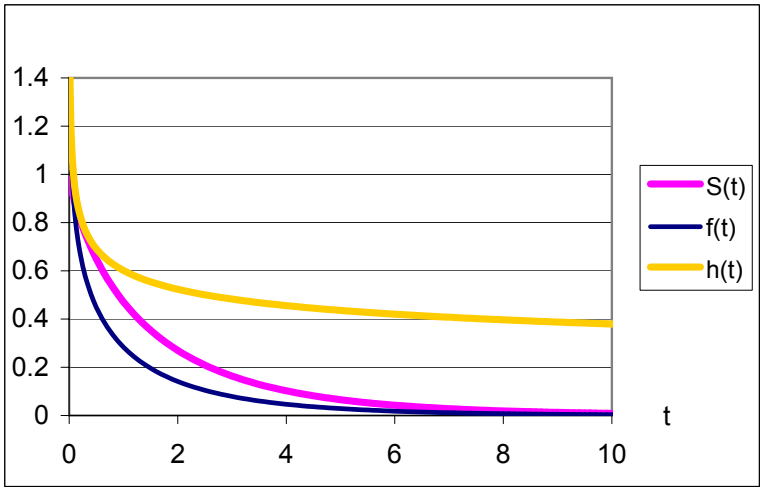
$$\mathrm{E}(T) = \Gamma\left(\frac{1+\alpha}{\alpha}\right)/\lambda$$

$$\mathrm{Var}(T) = \left(\Gamma\left(\frac{2+\alpha}{\alpha}\right) - \Gamma\left(\left(\frac{1+\alpha}{\alpha}\right)\right)^2\right)/\lambda^2$$

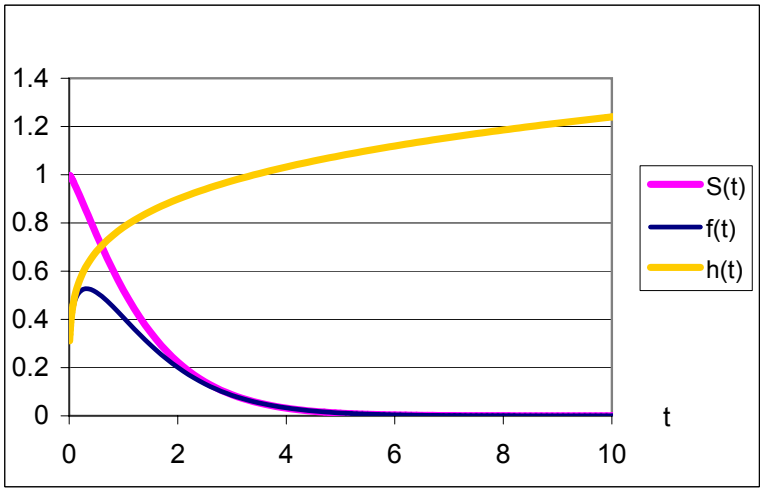
- $\Gamma(y)$ est la fonction gamma :

$$\Gamma(y) = (y-1)\Gamma(y-1) = \int_0^\infty x^{y-1}e^{-x}dx$$

Forme de la distribution de Weibull



$\lambda = 0.7$
$\alpha = 0.8$
$\mathrm{E}(T) = 1.62$



$\lambda = 0.7$
$\alpha = 1.2$
$\mathrm{E}(T) = 1.34$

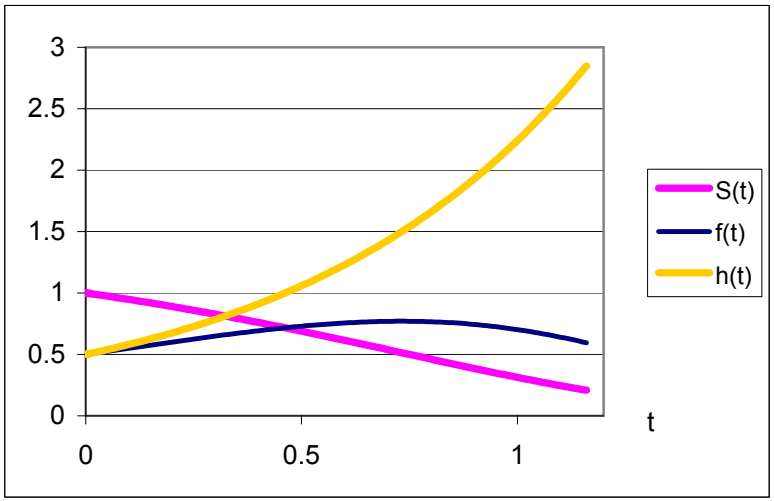
Distribution de Gompertz

- Cette distribution s'obtient lorsque le risque varie de façon proportionnelle à sa valeur.
- Elle est très utilisée pour la distribution du taux de mortalité.
- $T \sim$ Gompertz :

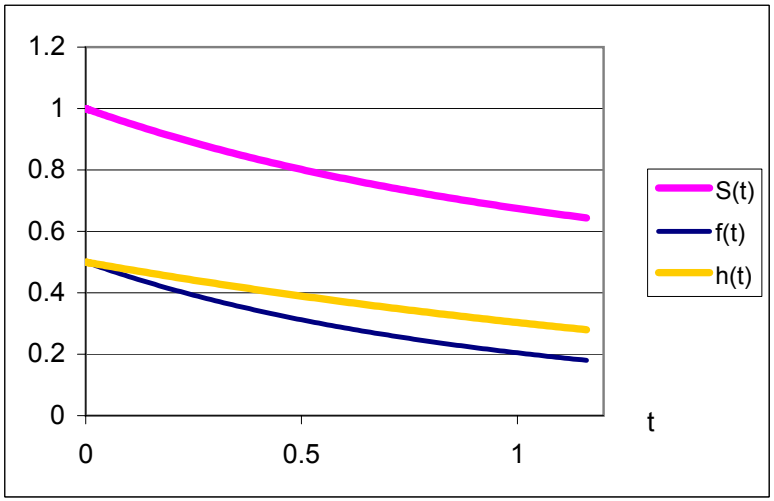
$$h(t) = \lambda \exp(\gamma t)$$

- λ : mortalité de base
- γ : influence de l'âge

Forme de la distribution de Gompertz



$$\lambda = 0.5$$
$$\gamma = 1.5$$



$$\lambda = 0.5$$
$$\gamma = -0.5$$

Distribution de Gompertz-Makeham

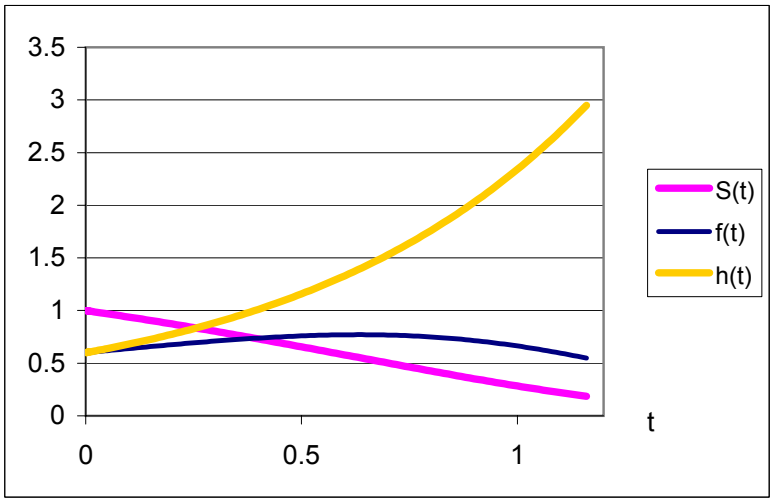
- Makeham ajoute un paramètre supplémentaire $\alpha > 0$ pour tenir compte de la mortalité accidentèle.

$$h(t) = \alpha + \lambda \exp(\gamma t) \qquad (t \geq 0)$$

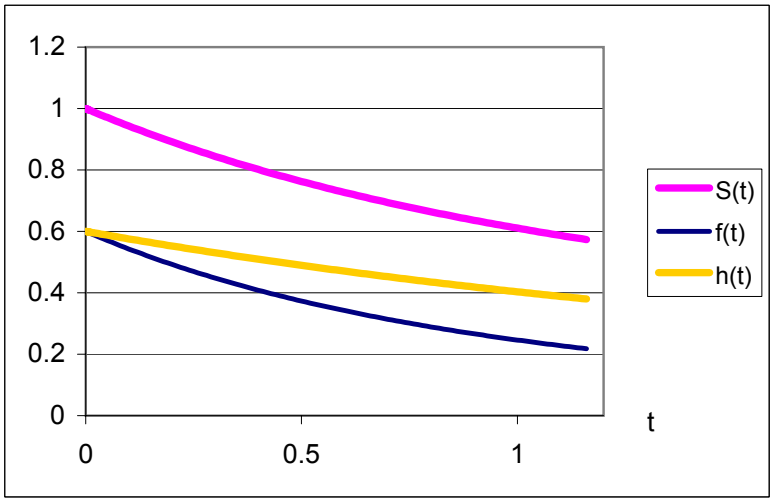
$$S(t) = \exp\left(-\alpha t - \frac{\lambda}{\gamma}(\exp(\gamma t) - 1)\right)$$

$$f(t) = \left(\alpha + \lambda \exp(\gamma t)\right) \exp\left(-\alpha t - \frac{\lambda}{\gamma}(\exp(\gamma t) - 1)\right)$$

Forme de la distribution de Gompertz-Makeham



$$\lambda = 0.5$$
$$\gamma = 1.5$$
$$\alpha = 0.1$$



$$\lambda = 0.5$$
$$\gamma = -0.5$$
$$\alpha = 0.1$$

Distribution Log-logistique (1)

- Le risque instantané $h(t)$ est une fonction non-monotone de t.
- Le paramètre α est un paramètre de forme :
  - $\alpha \leq 1$: hasard décroissant au fil du temps
  - $\alpha > 1$: hasard croissant, puis décroissant au fil du temps (d'où une distribution unimodale)
- *Remarque : La loi log-normale est une distribution similaire pour laquelle il n'y a cependant pas de forme explicite pour $h(t)$.*

Distribution Log-logistique (2)

- Soit

$$\ln T = -\ln \lambda + U/\alpha \Leftrightarrow T = \frac{1}{\lambda} \exp(U/\alpha)$$

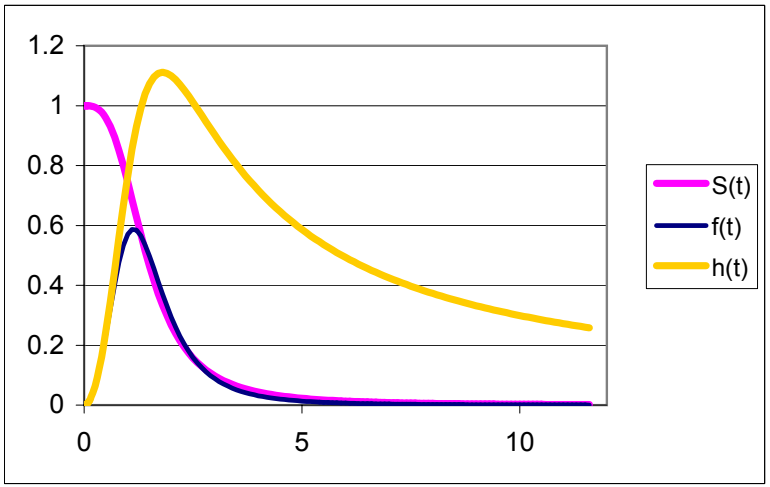
avec $U \sim$ logistique, alors

$$T \sim \text{Log-logistique}(\lambda, \alpha)$$

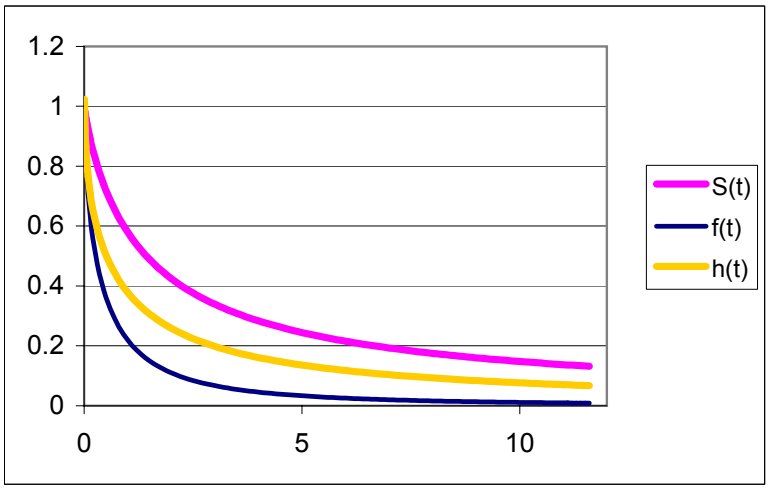
- Caractéristiques :

$$\begin{aligned} h(t) &= \frac{\lambda\alpha(\lambda t)^{\alpha-1}}{1+(\lambda t)^{\alpha}}, \quad t>0, \ \lambda>0, \ \alpha>0 \\ S(t) &= \frac{1}{1+(\lambda t)^{\alpha-1}} \\ f(t) &= \lambda\alpha(\lambda t)^{\alpha-1}\left[1+(\lambda t)^{\alpha}\right]^{-2} \end{aligned}$$

Forme de la distribution de Log-logistique



$\lambda = 0.7$
$\alpha = 3$



$\lambda = 0.7$
$\alpha = 0.9$

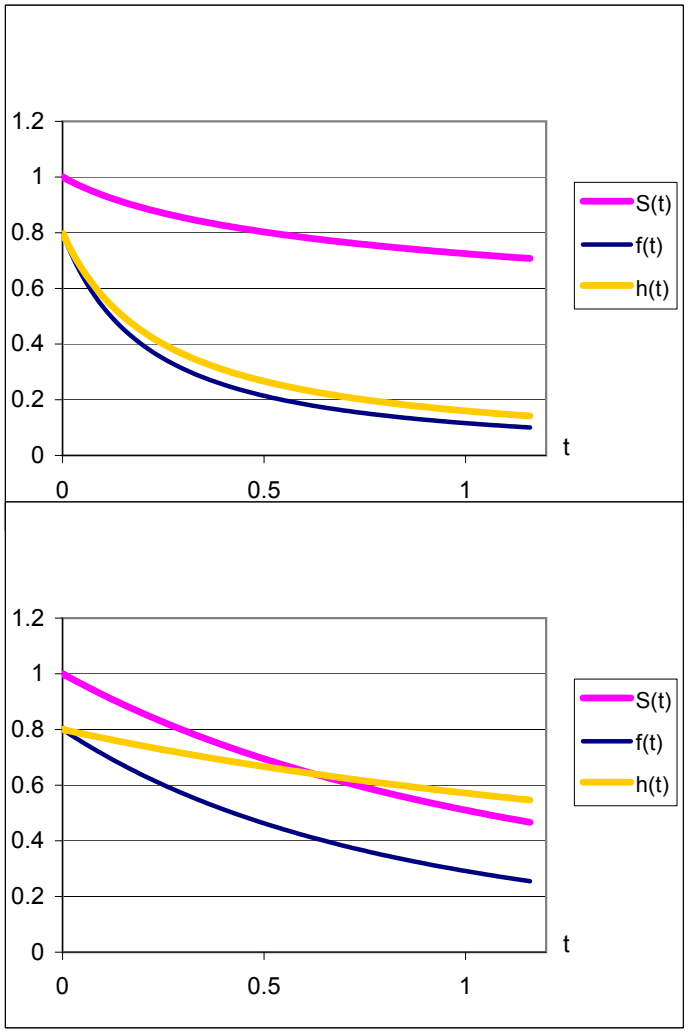
Distribution de Pareto

- Risque instantané λ constant pour chaque individu mais variant entre individus selon une loi gamma $\gamma(a, p)$ de moyenne $\lambda_0 = a/p$, dont la densité est

$$f(\lambda) = \frac{p^a}{\Gamma(a)} \lambda^{(a-1)} \exp(-p\lambda) = \frac{(a/\lambda_0)^a}{\Gamma(a)} \lambda^{(a-1)} \exp(-a\lambda/\lambda_0)$$

$$\begin{aligned} h(t) &= a\left(t + \frac{a}{\lambda_0}\right)^{-1} \\ S(t) &= \left(\frac{a}{\lambda_0}\right)^a \left(t + \frac{a}{\lambda_0}\right)^{-a} \\ f(t) &= a\left(\frac{a}{\lambda_0}\right)^a \left(t + \frac{a}{\lambda_0}\right)^{-(a+1)} \end{aligned}$$

Forme de la distribution de Pareto



$\lambda_0 = 0.8$
$a = 0.2$

$\lambda_0 = 0.8$
$a = 2$

Plan du cours

Deux approches

- Il y a deux approches principales permettant de choisir le modèle théorique le plus adapté aux données :
  1. L'approche graphique consiste à comparer la distribution effective du risque ou de la survie avec la distribution théorique suggérée par les différentes lois et à choisir le modèle dont on est le plus proche.
  2. L'approche par ajustement consiste à estimer les différents modèles et à choisir celui qui s'ajuste le mieux aux données sur la base du coefficient d'information d'Akaike (AIC).

Choix entre distributions théoriques (1)

modèle	fonction	propriété
Exponentiel	$h(t)$ $H(t)$	indépendante de t linéaire en t
Weibull	$\ln(-\ln S(t))$	linéaire en $\ln t$
Log-logistique	$\ln(1/S(t)-1)$	linéaire en $\ln t$
Gompertz	$\ln h(t)$ $\ln(\ln[\Delta S(t)])$	linéaire en t linéaire en t
Migrant-sédentaire	$\ln[\Delta S(t)]$	linéaire en t
Pareto	$\frac{1}{h(t)}$	linéaire en t

Choix entre distributions théoriques (2)

- En pratique, on fait des présentations graphiques des transformées des estimations $\hat{S}_k$ et $\hat{h}_k$ en fonction de t et on compare avec les propriétés attendues des différentes lois.
- Si les intervalles $[t_{k-1}, t_k)$ sont de longueur variable (en particulier avec Kaplan-Meier)
  $\Rightarrow$ ajuster les $\hat{h}_k$ et $\Delta\hat{S}_k$ pour obtenir des valeurs se rapportant à une unité de temps :

$$\hat{h}(t_k) = \frac{2\hat{h}_k}{(t_{k+1} - t_k)(2 - \hat{h}_k)}$$

$$\Delta\hat{S}(t_k) = \hat{f}(t_k) = \frac{\hat{S}_{k-1} - \hat{S}_k}{t_k - t_{k-1}}$$

Critère d'information d'Akaike

- Le critère d'information d'Akaike est défini comme

$$\text{AIC} = -2\text{LL}(M) + 2k$$

où LL est la log-vraisemblance et k est le nombre de paramètres du modèle M.

- Le modèle minimisant ce critère est le modèle offrant le meilleur ajustement aux données.
- *Remarque : Etant donné que les vraisemblances qui sont maximisées pour l'obtention des paramètres d'un modèle de Cox et d'un modèle paramétrique sont différentes, il n'est pas possible de comparer ces deux types de modèles à l'aide d'AIC.*

Plan du cours

Exemple : Données biographiques allemandes

- Données extraites de l'enquête biographique allemande réalisée entre 1981 et 1983 (Mayer & Brückner, 1989) et utilisées notamment par (Blossfeld & Rohwer, 2002).
- Trois cohortes de naissance : 1929-1931 (coho1), 1939-1941 (coho2), 1949-1951 (coho3).
- Echantillon de n=600 emplois.
- Comment le niveau d'éducation (*edu*), l'expérience sur le marché du travail (*lfx*), le nombre d'emplois précédents (*pnoj*) et le prestige de l'emploi (*pres*) influencent-ils
  - le risque de terminer un emploi ?
  - la durée de l'emploi ?
- On souhaite aussi contrôler les effets par cohorte.

Exponentiel : AFT

```
Call:
survreg(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3, dist = "exponential")
              Value Std. Error     z        p
(Intercept)  4.48944   0.279500 16.06 4.68e-58
edu         -0.07730   0.024703 -3.13 1.75e-03
lfx          0.00318   0.000938  3.39 6.99e-04
pnoj        -0.05964   0.044153 -1.35 1.77e-01
pres         0.02801   0.005530  5.06 4.10e-07
coho2       -0.60804   0.113551 -5.35 8.57e-08
coho3       -0.61080   0.118542 -5.15 2.57e-07

Scale fixed at 1

Exponential distribution
Loglik(model)= -2466   Loglik(intercept only)= -2514
        Chisq= 96.07 on 6 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 600
```

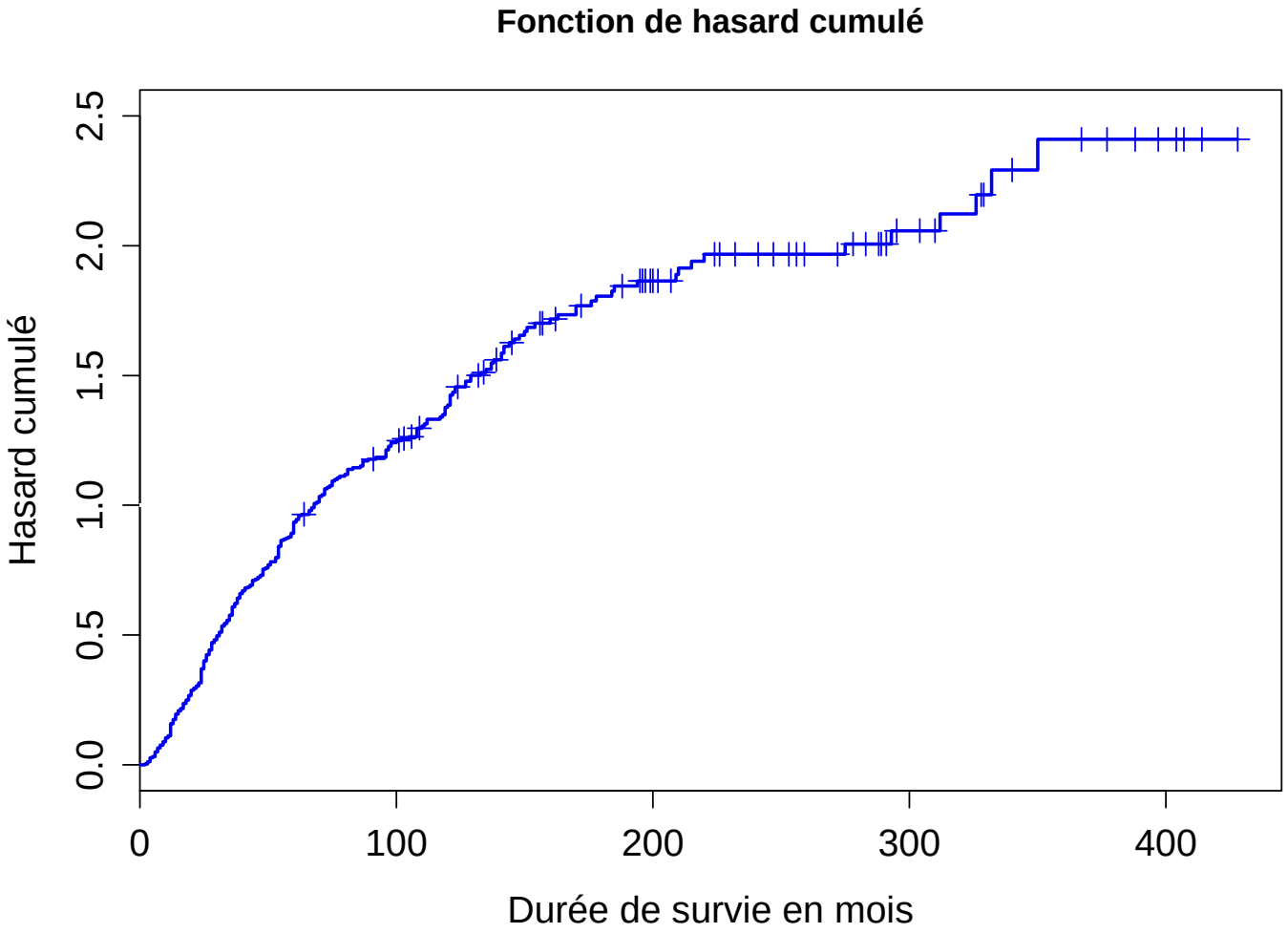
Weibull : AFT

```
Call:
survreg(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3, dist = "weibull")
               Value Std. Error     z        p
(Intercept)  4.40631    0.30600 14.40 5.20e-47
edu         -0.07793    0.02701 -2.88 3.92e-03
lfx          0.00364    0.00104  3.50 4.58e-04
pnoj        -0.06375    0.04835 -1.32 1.87e-01
pres         0.02916    0.00603  4.84 1.30e-06
coho2       -0.60620    0.12417 -4.88 1.05e-06
coho3       -0.57768    0.13035 -4.43 9.34e-06
Log(scale)   0.09028    0.03636  2.48 1.30e-02

Scale= 1.09

Weibull distribution
Loglik(model)= -2462.8   Loglik(intercept only)= -2504.9
        Chisq= 84.28 on 6 degrees of freedom, p= 4.4e-16
Number of Newton-Raphson Iterations: 5
n= 600
```

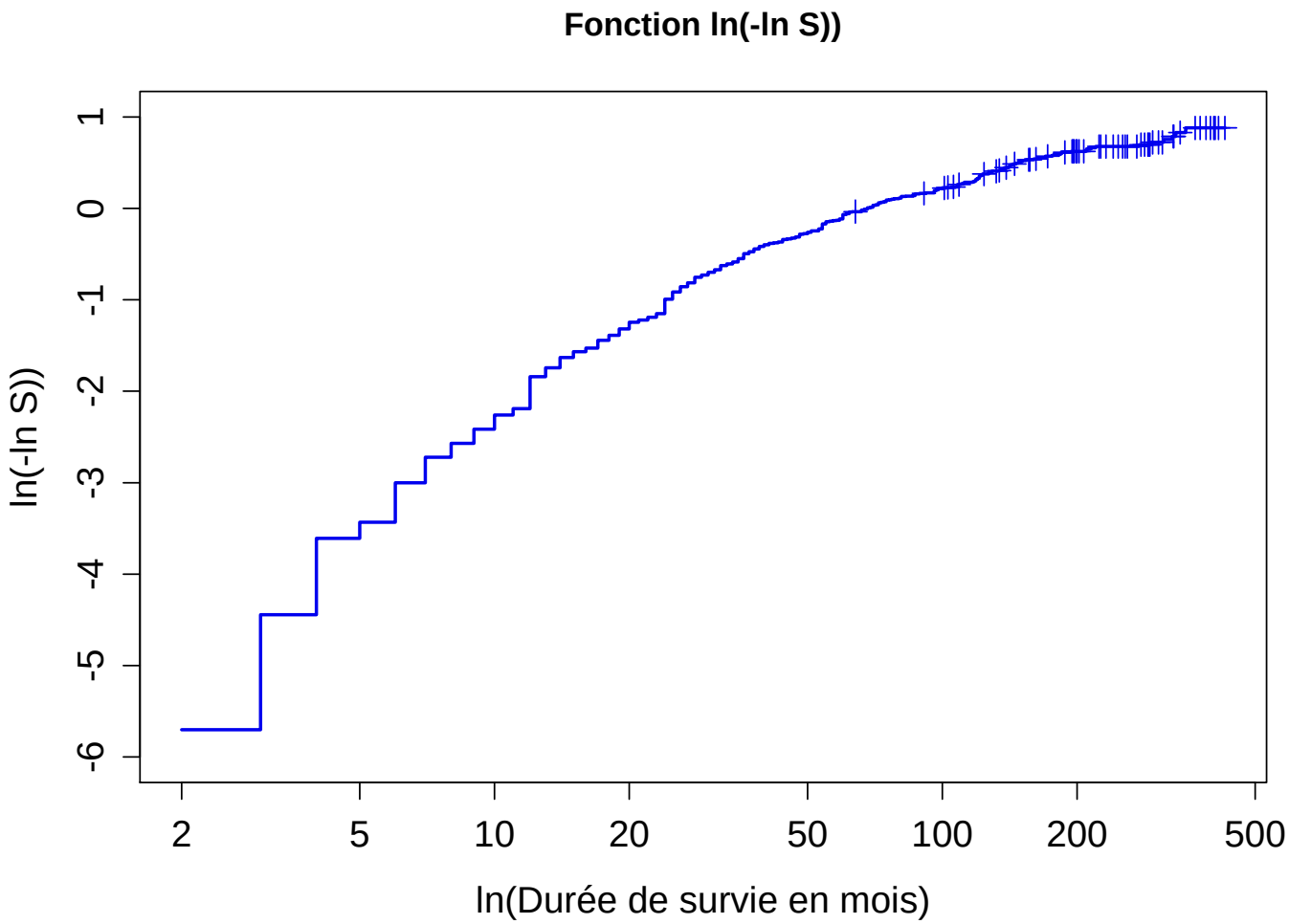
Log-logistique : AFT

```
Call:
survreg(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3, dist = "loglogistic")
               Value Std. Error     z        p
(Intercept)  3.77846   0.292578 12.91 3.74e-38
edu         -0.08187   0.026856 -3.05 2.30e-03
lfx          0.00425   0.000913  4.66 3.21e-06
pnoj        -0.09704   0.047640 -2.04 4.17e-02
pres         0.02967   0.005609  5.29 1.22e-07
coho2       -0.53382   0.124941 -4.27 1.93e-05
coho3       -0.40048   0.131818 -3.04 2.38e-03
Log(scale)  -0.36148   0.038563 -9.37 7.01e-21

Scale= 0.697

Log logistic distribution
Loglik(model)= -2418.9   Loglik(intercept only)= -2460.5
        Chisq= 83.16 on 6 degrees of freedom, p= 7.8e-16
Number of Newton-Raphson Iterations: 4
n= 600
```

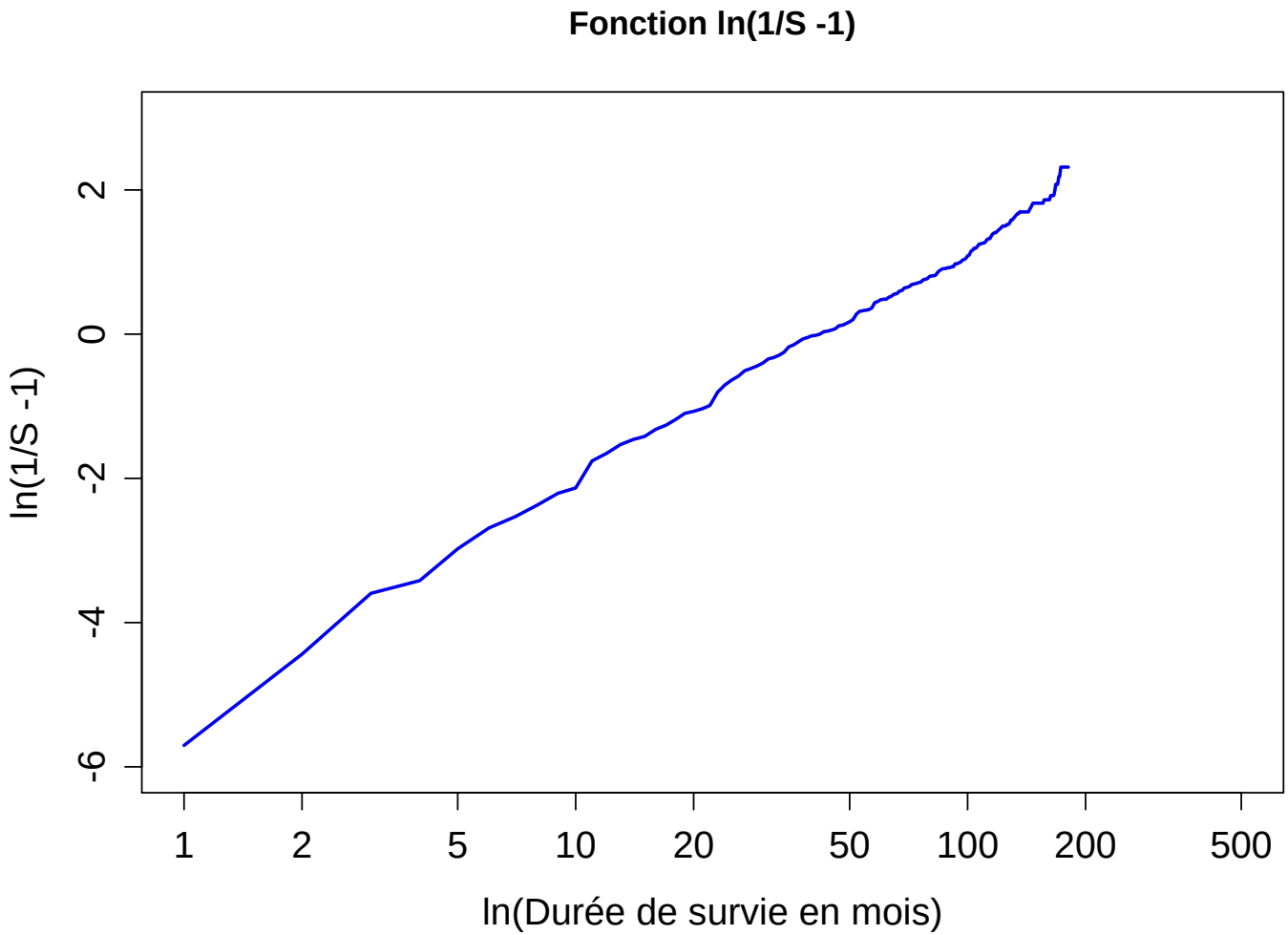
Graphique de $H(t)$



Graphique de $\ln(-\ln S(t))$



Graphique de $\ln(1/S(t) - 1)$



AIC

Modèle	LL	k	AIC
Exponentiel	-2466	7	4932
Weibull	-2462.8	8	4941.6
Log-logistique	-2418.9	8	4853.8

Bibliographie

- Blossfeld HP, Rohwer G (2002) *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ : Lawrence Erlbaum.
- Mayer KU, Brückner E (1989) *Lebensverläufe und Wohlfahrtsentwicklung. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929-1931, 1939-1941, 1959-1951.* Materialen aus der Bildungsforschung. Berlin : Max-Planck Institut für Bildungsforschung.