Estimation indirecte en sciences humaines :

une méthode bayésienne

Henri Caussinus, Institut de Mathématiques de Toulouse,

en collaboration avec Daniel Courgeau, INED
Isabelle Séguy, INED
Luc Buchet, CNRS

Un exemple en paléodémographie

Estimer une loi de mortalité à partir de restes osseux, c'est-à-dire :

- estimer l'âge d'un individu à partir de caractéristiques d'un os (synostose crânienne),

- ou plutôt estimer la répartition des âges au décès dans une population à partir d'un échantillon de squelettes.

(noter que les deux problèmes sont passablement différents)

C’est un problème de statistique traditionnel

C’est un problème de statistique traditionnel

Connaissant la longueur du bateau, le nombre de hublots, le nombre de passagers, etc.

quel est l’âge du capitaine ?

Même chose en plus « historique », et on peut ajouter le hauteur du grand mât…

âge = fonction (longueur, etc.) à peu près

âge = fonction (longueur, etc.) à peu près

↑

données

de référence

↑

liste de bateaux pour lesquels on connaît
la longueur, etc. et aussi l’âge du capitaine

âge = fonction (longueur, etc.) à peu près

↑ ↑

données précision

de référence attendue

Un exemple en paléodémographie

Objectif : estimer la loi de mortalité d'une population du passé, c'est-à-dire la distribution des âges au décès

au moyen de données

- historiques
- peu nombreuses
- peu fiables

Mais quelques informations a priori (en dehors des données) sont éventuellement disponibles.

données

- Observations sur un site cible : état osseux (stade) sur m squelettes.

- Données « de référence » :

  croisent âge et stade

  sur un échantillon pour lequel les deux informations sont disponibles, en espérant que la distribution des âges à stade osseux donné est identique dans le nouveau site considéré.

Ages et stades sont ici supposés discrétisés, répartis en classes

Exemple de données de référence

âges

Stage	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80+
A	14	39	36	13	14	16	20	12	6	3	5	2	4	2
B	2	10	8	19	14	13	15	13	14	5	3	5	4	11
C	1	4	5	1	5	9	10	9	9	7	9	11	7	13
D	1	2	3	5	3	4	6	11	10	10	8	8	9	15
E	0	2	0	5	4	2	2	6	10	7	13	7	9	28
total	18	57	52	43	40	44	53	51	49	32	38	33	33	69

Exemples de données « de site »

	Antibes	Maubuisson
A	m_1=21	3
B	m_2=14	5
C	m_3=12	9
D	m_4=16	6
E	m_5=10	14
total	m =73	37

Modélisation

Paramètres :

probabilités conditionnelles : $p_{i/j}$ matrice P

probabilités des classes d'âge : p_j vecteur p

avec i = 1,…l et j = 1,…,c

D'où les probabilités des stades : $\pi_i = \Sigma_j p_j P_{i/j}$

ou matriciellement $\pi = P\ p$

Les données de site (m_i) fournissent une information sur π

Les données de référence fournissent une information sur P.

Ce qui devrait permettre d'estimer p.

Modèles utilisés :

- **régression classique supposant les $p_{i/j}$ parfaitement connus ou régression sur variables entachées d'erreur en admettant le caractère aléatoire des estimations des $p_{i/j}$ (nécessite $l \geq c$).**

- **les m_i sont multinomiaux de probabilités π_i, les $p_{i/j}$ étant supposés parfaitement connus au moyen des données de référence connus (nécessite $l \geq c$).**

- **même chose mais on tient compte du fait que les $p_{i/j}$ sont eux-mêmes estimés au moyen d'échantillons multinomiaux.**

Les résultats sont catastrophiques
problème très instable
espace paramétrique trop grand.

On peut chercher à introduire le fait que les p_j ne sont pas n'importe quoi

Première voie :

réduire convenablement l'espace paramétrique.

On trouve deux façons de faire dans la littérature :

- Utiliser un modèle de mortalité classique, comme la loi de Gompertz, qui fait dépendre les p_j de seulement deux paramètres.

- Considérer $\pi = P\ p$ comme une régression mais, au lieu de chercher le vecteur p estimé dans un espace vectoriel, le chercher dans (l'enveloppe convexe d') un ensemble fini de vecteurs censés représenter les lois de mortalité « possibles » (Bocquet-Appel et Bacro, 2008).

Une autre voie :

Méthode bayésienne

- Densité conjointe de M, P, p

$$f(M,P,p) = C.g(p)G(P)\prod_i \left(\sum_j p_j p_{i/j}\right)^{m_i}$$

- Densité conditionnelle de p sachant M

$$\frac{\int f(M,P,p)dP}{\iint f(M,P,p)dpdP}$$

Espérance a posteriori d'une fonction φ(p)

$$\iint \varphi(p)\, f(M,P,p)\,dPdp \Big/ \iint f(M,P,p)\,dPdp$$

égale à

$$E(\varphi(X)\prod_i(\sum_j X_j Y_{ij})^{m_i}) \Big/ E(\prod_i(\sum_j X_j Y_{ij})^{m_i})$$

où X est un vecteur de loi g

Y est un vecteur de loi G

Avec $\varphi(p) = p_j$ on a la moyenne a posteriori de p_j

Avec $\varphi(p) = \mathbf{1}_{[pi < x]}$ on a la fonction de répartition a posteriori de p_j qui permet de calculer des intervalles de crédibilité

etc.

Lois a priori

- G est déduite des données de référence : en l'absence d'autre information, on considère une a priori uniforme pour les c vecteurs de probabilités conditionnelles $p_{i/j}$ ce qui, combiné avec le modèle multinomial, conduit à c lois de Dirichlet indépendantes.

- Pour g, plusieurs arguments peuvent être mis en avant.

  Les paléodémographes ont défini un « standard préindustriel » qui est une espèce de loi de mortalité moyenne pour une large période historique. Une loi de Dirichlet dont les moyennes correspondent à ce standard incorpore cette information de façon naturelle. On peut fignoler en modifiant les paramètres de cette loi pour tenir compte d'autres faits connus. Plusieurs lois ont cependant mêmes moyennes, selon la valeur retenue pour la somme S des paramètres individuels exprimant plus ou moins de confiance dans la loi a priori. Un certain nombre de simulations ont conduit à une proposition simple (S = c), mais d'autres valeurs peuvent être envisagées.

Exemple des religieuses de Maubuisson

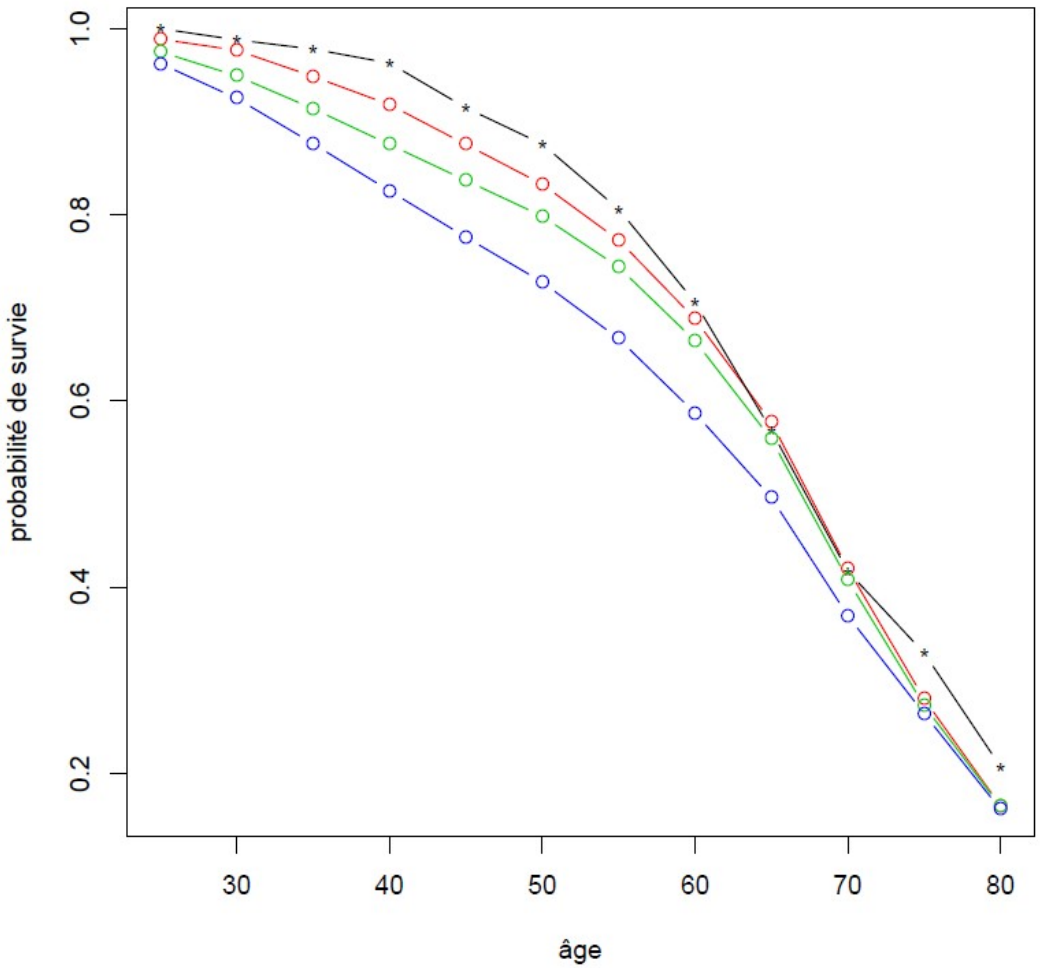
âge en classes quinquennales de 20 à 80 ans plus une classe ouverte (≥ 80)

(donc 13 classes au total)

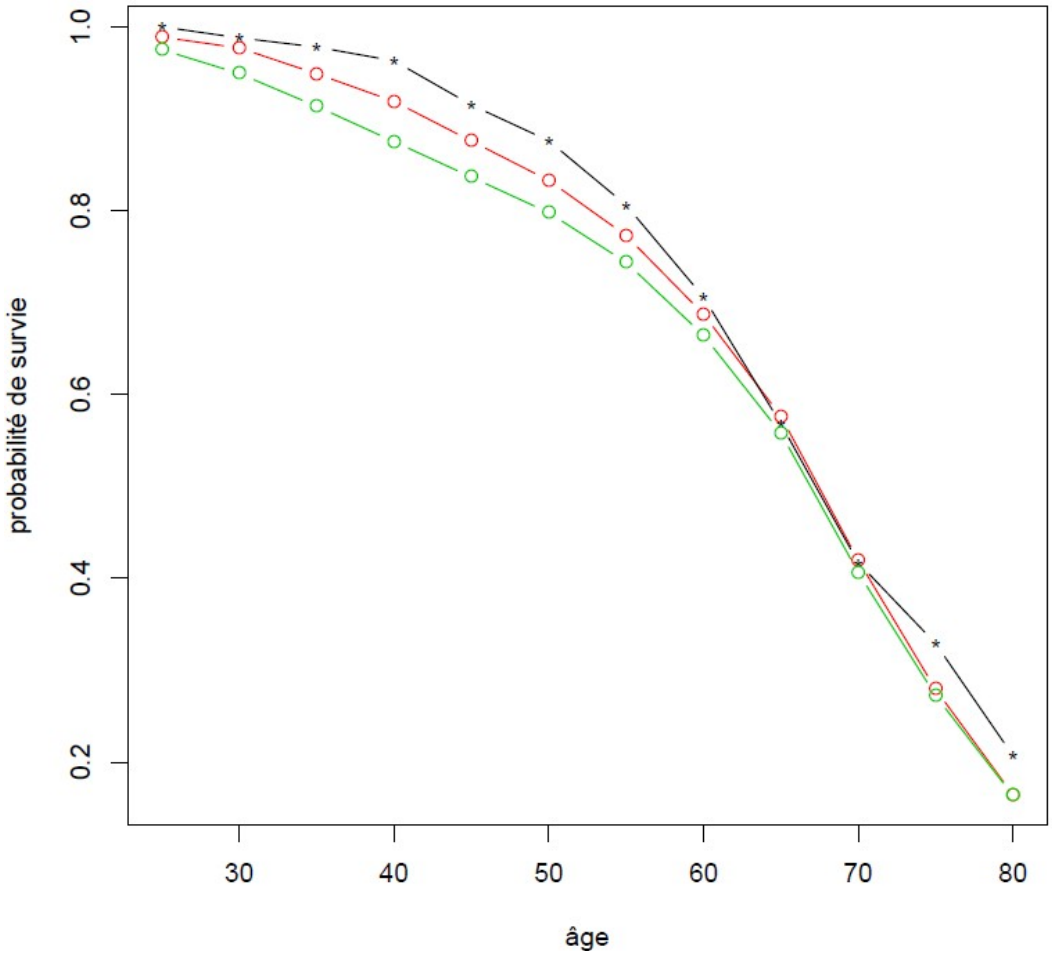
5 stades osseux (synostose crânienne)

Registres et trois estimations selon l'a priori

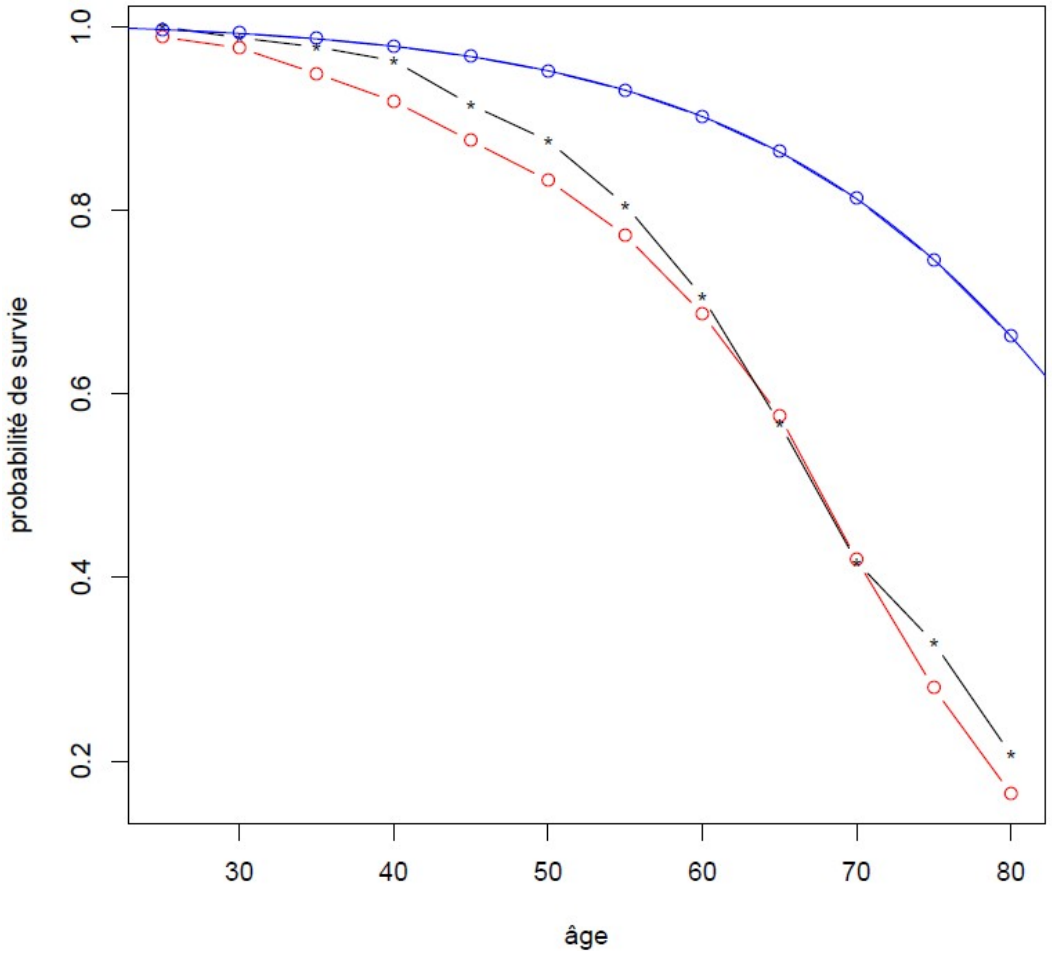
uniforme standard standard modifié



Registres a priori standard a priori standard modifié

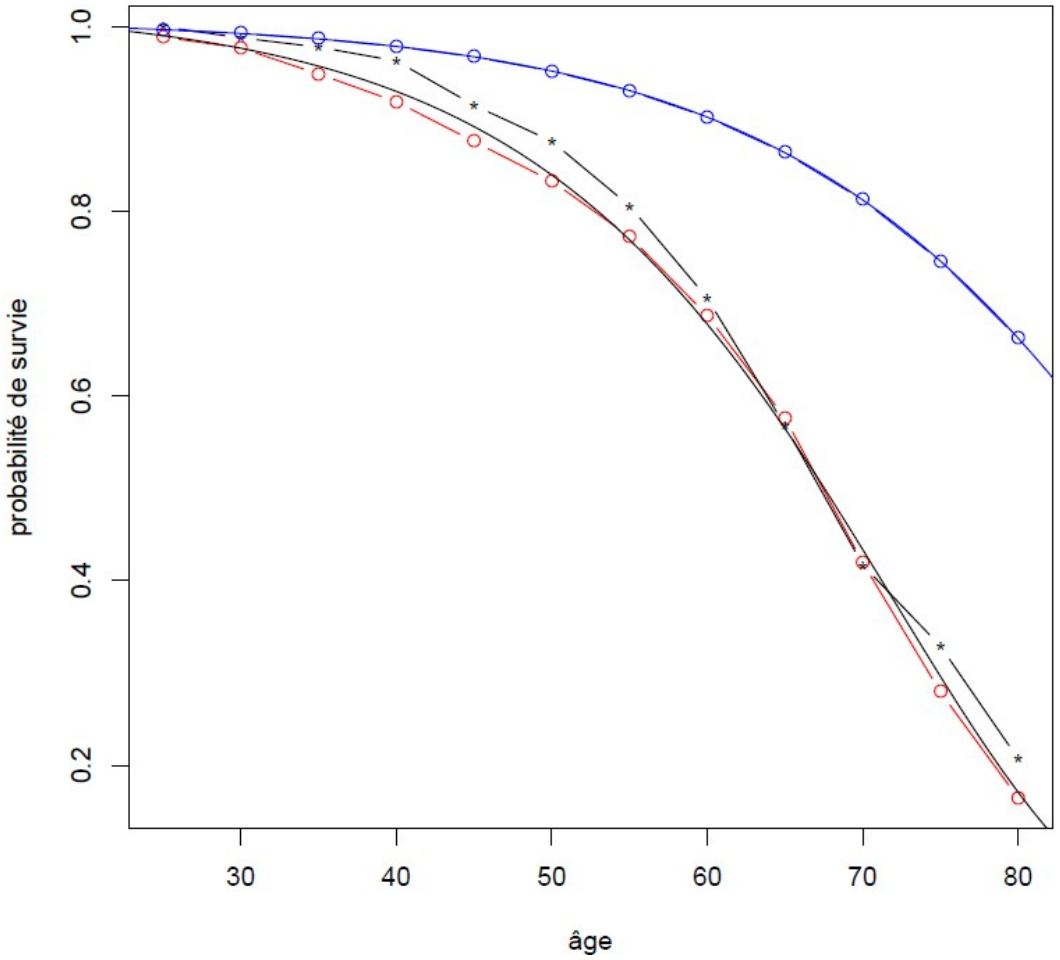


Registres Bayes maximum de vraisemblance
avec modèle de Gompertz



Bayes et modèle de Gompertz ajusté

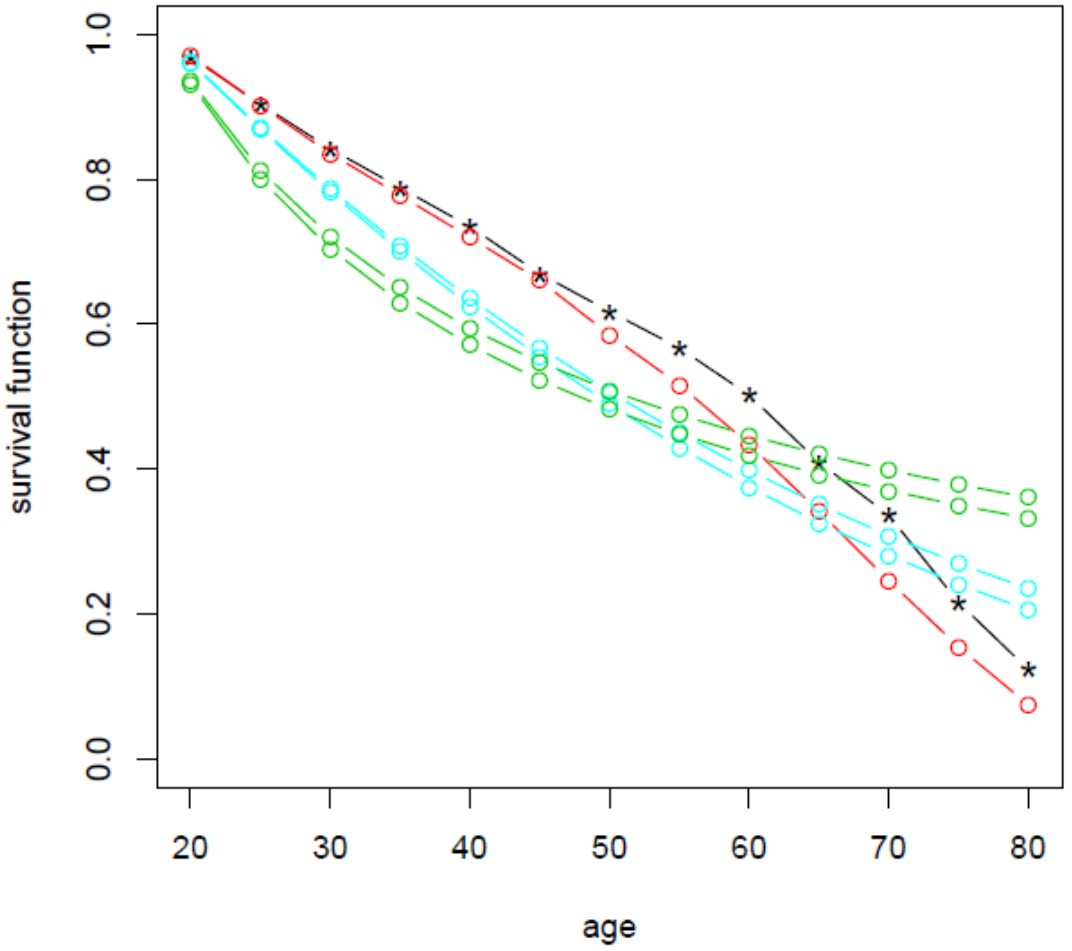
maximum de vraisemblance
avec modèle de Gompertz



Exemple d’Antibes

Là encore on a des registres (XIXème siècle) permettant d’évaluer la méthode.

Registres Bayes Gompertz Weibull



Exemple de Frénouville

2 sites voisins

Site gallo-romain (fin du 3ème siècle à moitié du 5ème) m=69

Site mérovingien (6ème - 7ème siècles) m=200

5 stages osseux

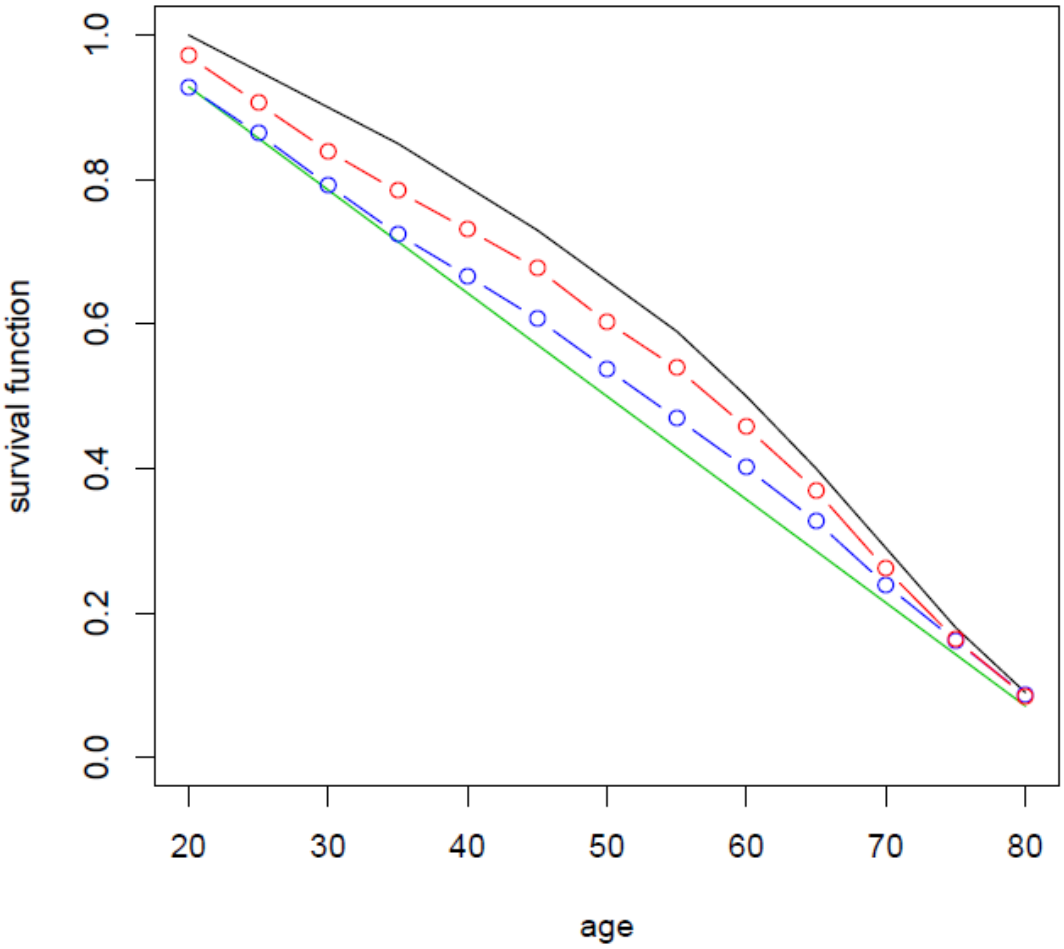
14 classes d'âge : 12 classes quinquennales encadrées des classes (18-19 ans) et (≥80)

Deux études comparatives :

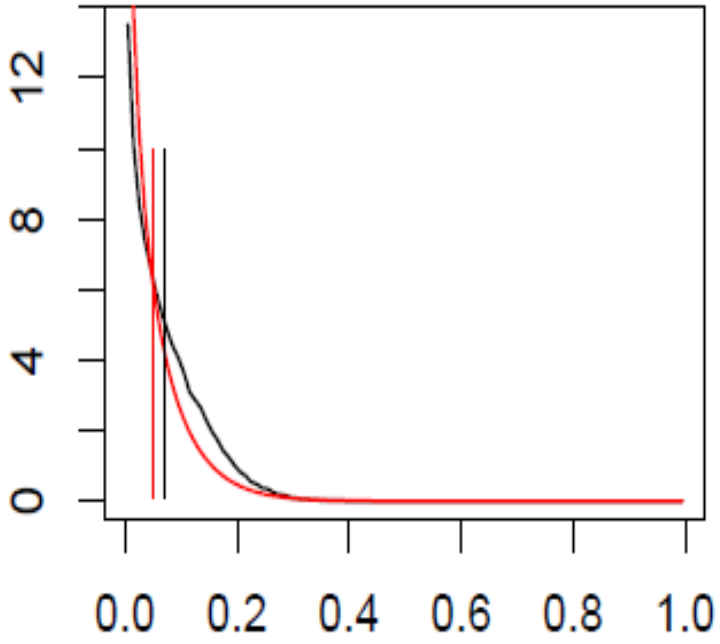
par sexe

par période

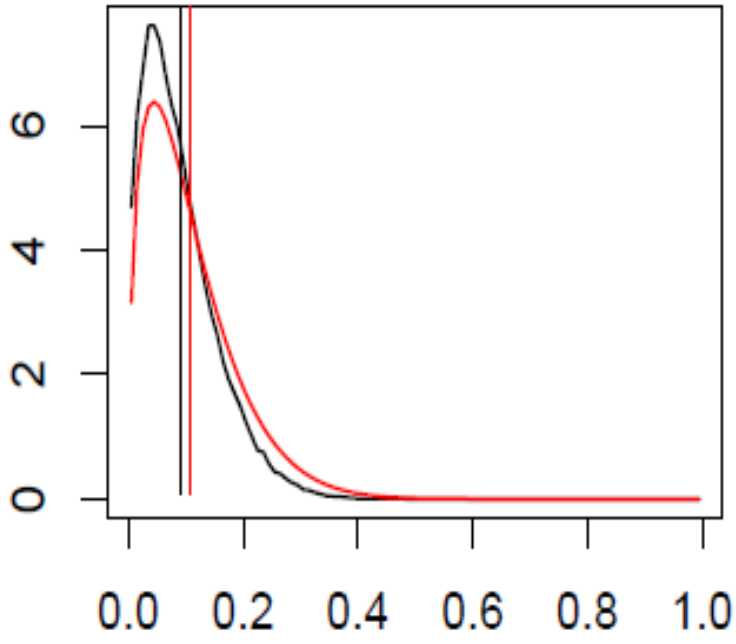
Période gallo-romaine



Frénouville gallo-romain
densités a priori et a posteriori

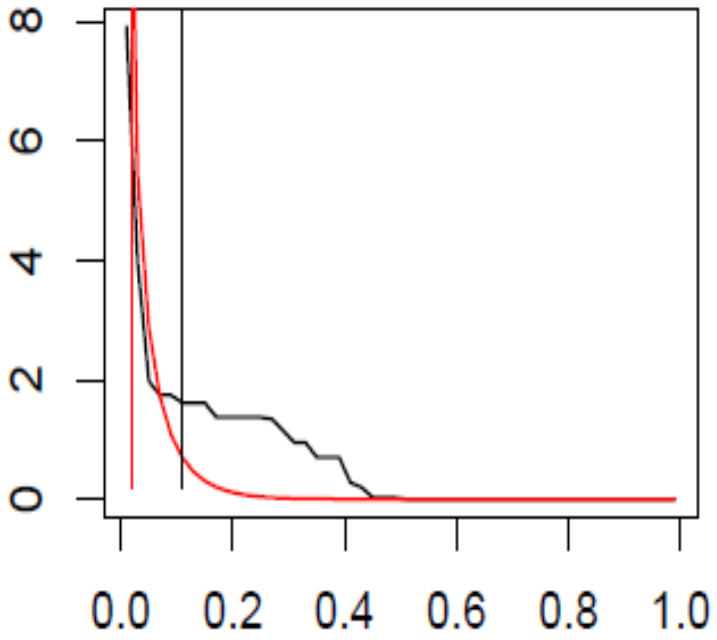


20 – 24 ans

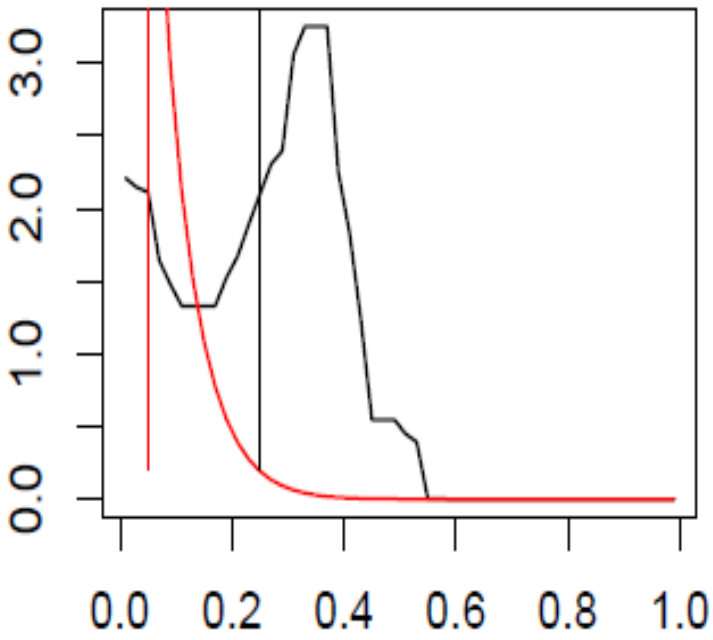


70 – 74 ans

Frénouville Mérovingien
densités a priori et a posteriori



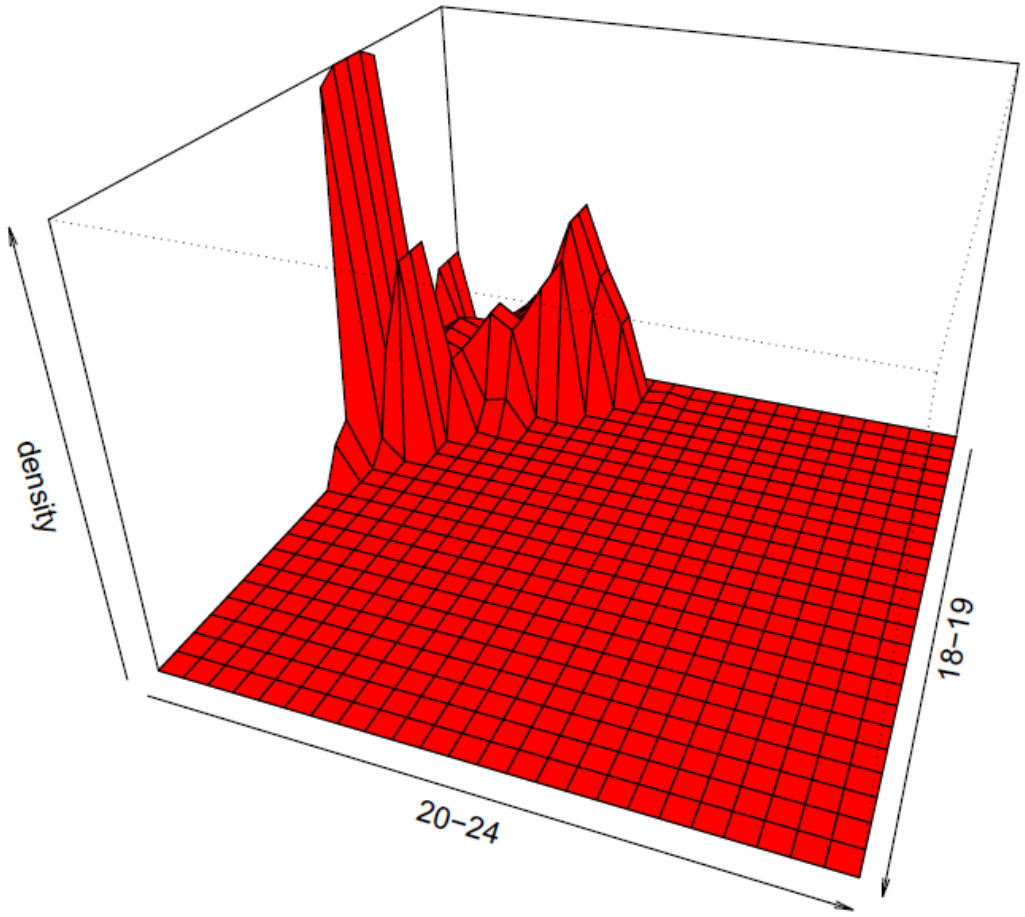
18 – 19 ans



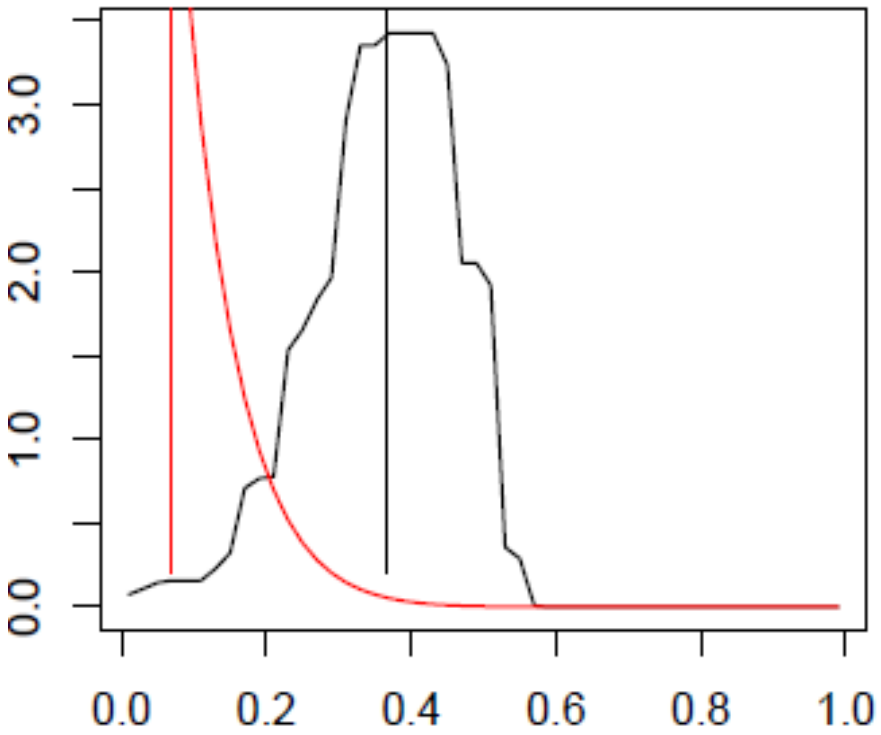
20 – 24 ans

Pour les deux premières classes d’âge, forte corrélation négative (- 0.778)

Pour les deux premières classes d'âge, forte corrélation négative et densité conjointe bimodale

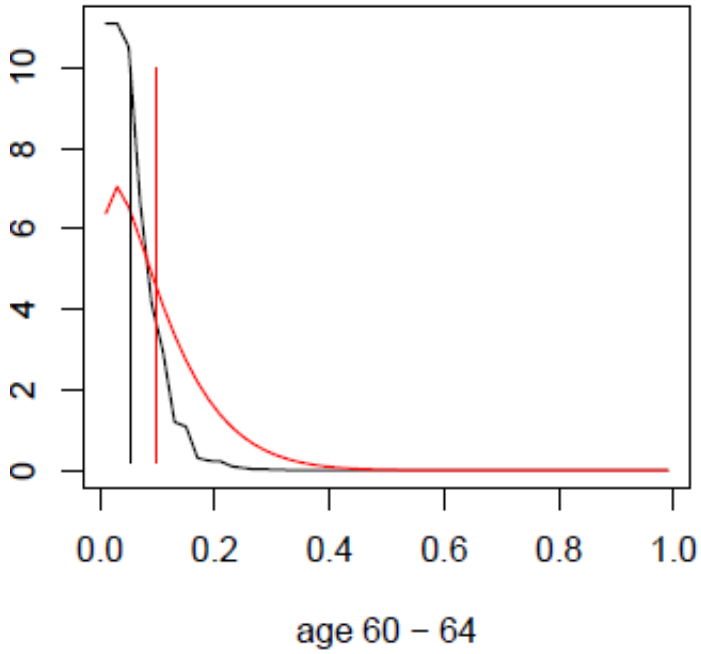


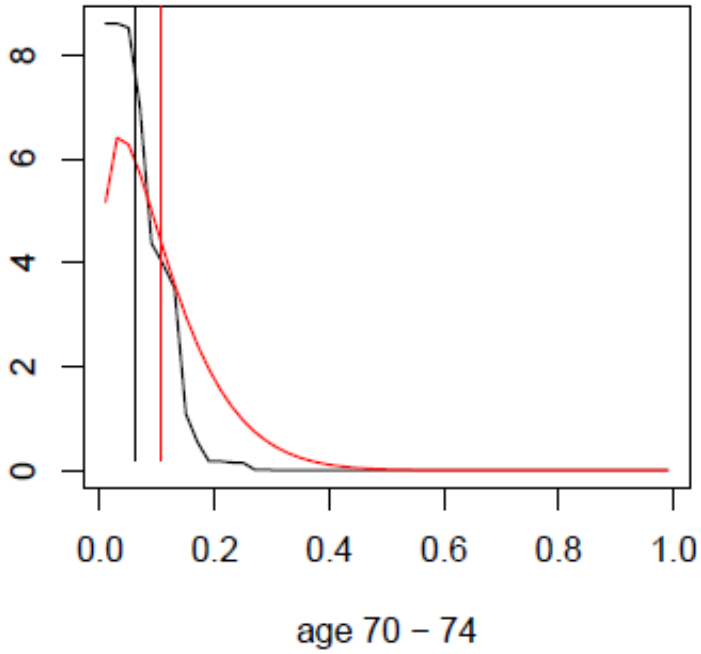
Difficile d’avoir un résultat avec une bonne précision pour les deux premières classes. On va donc les regrouper.



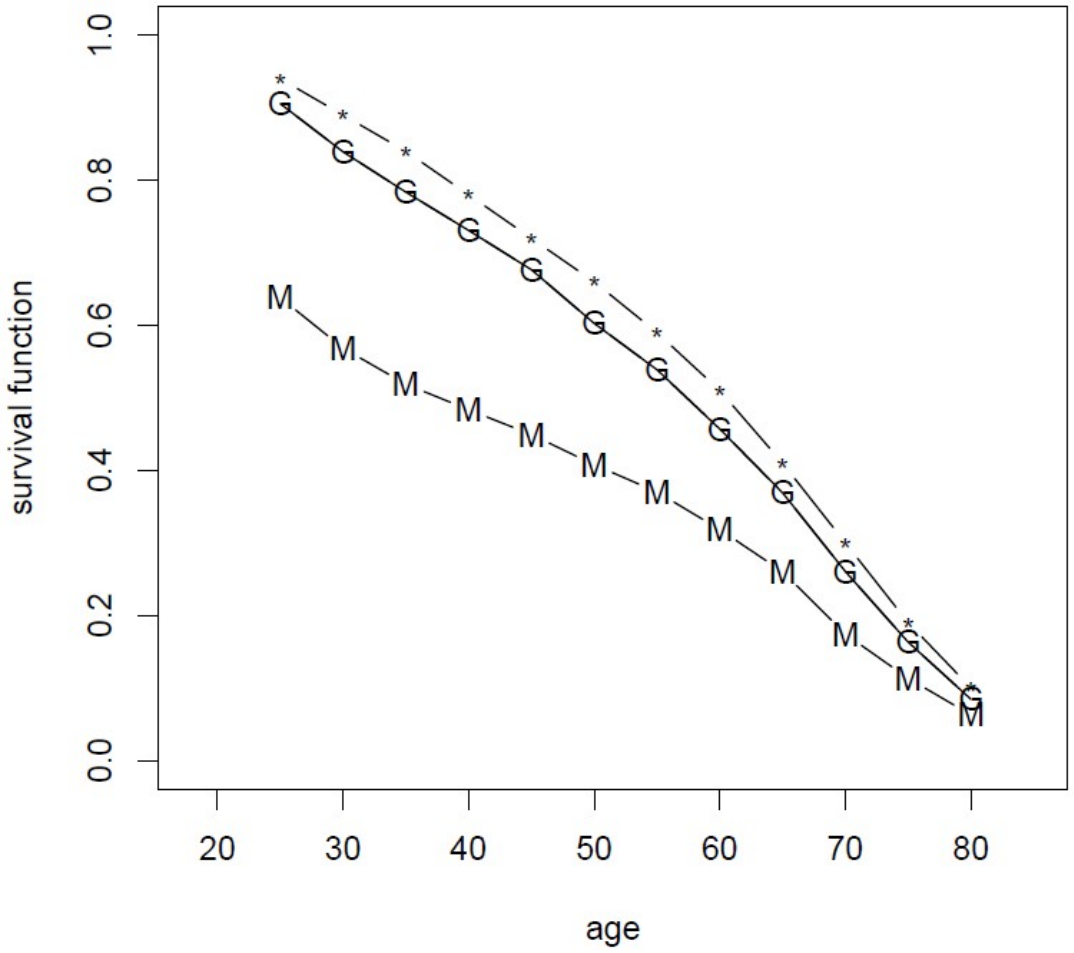
18 – 24 ans

Deux autres classes d’âge

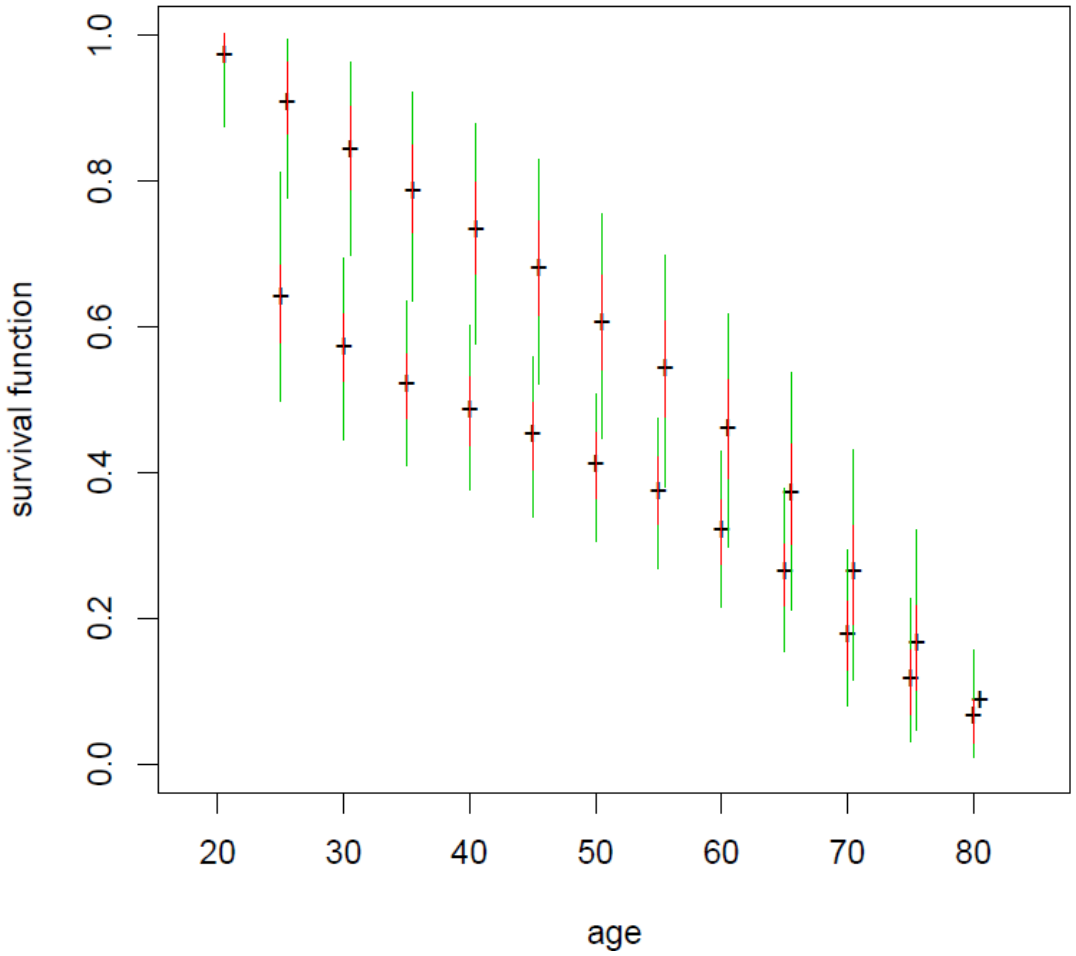




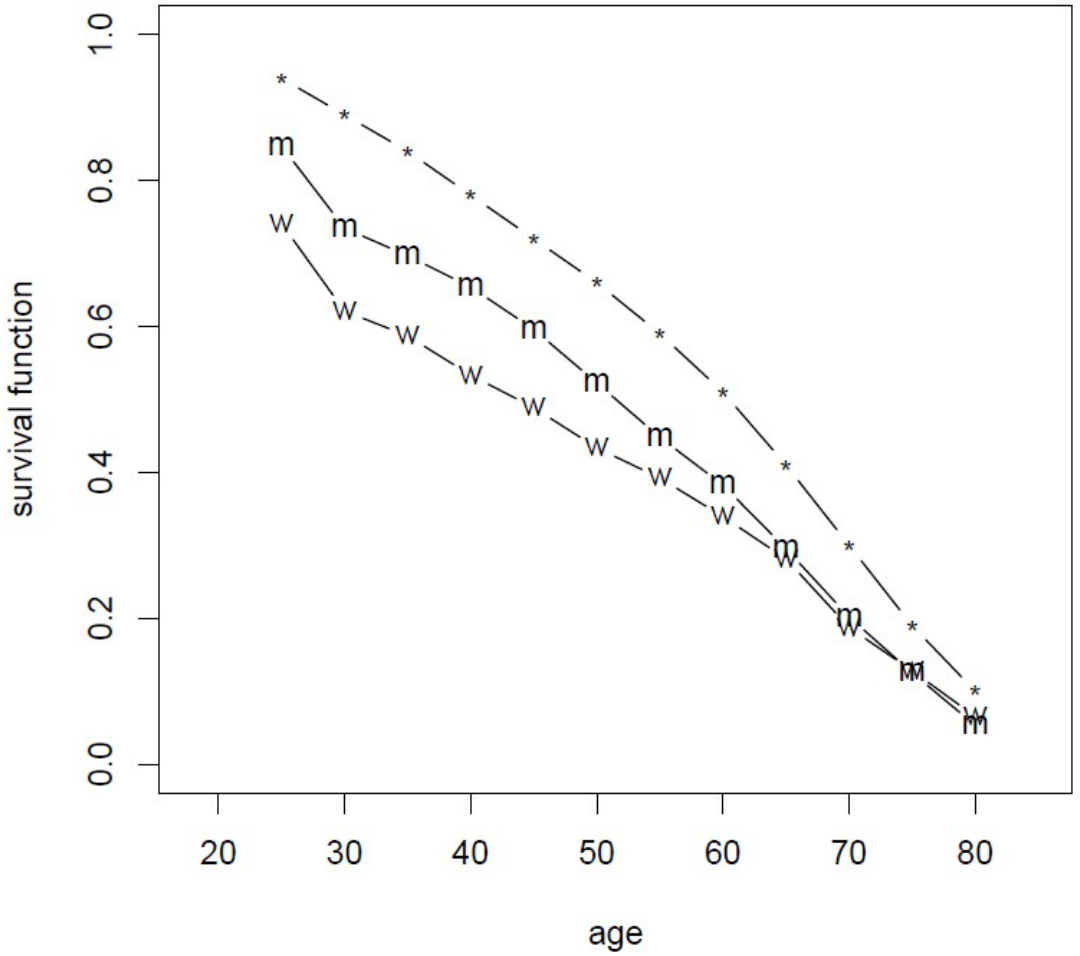
Comparaison des époques gallo-romaine et mérovingienne sur les survies



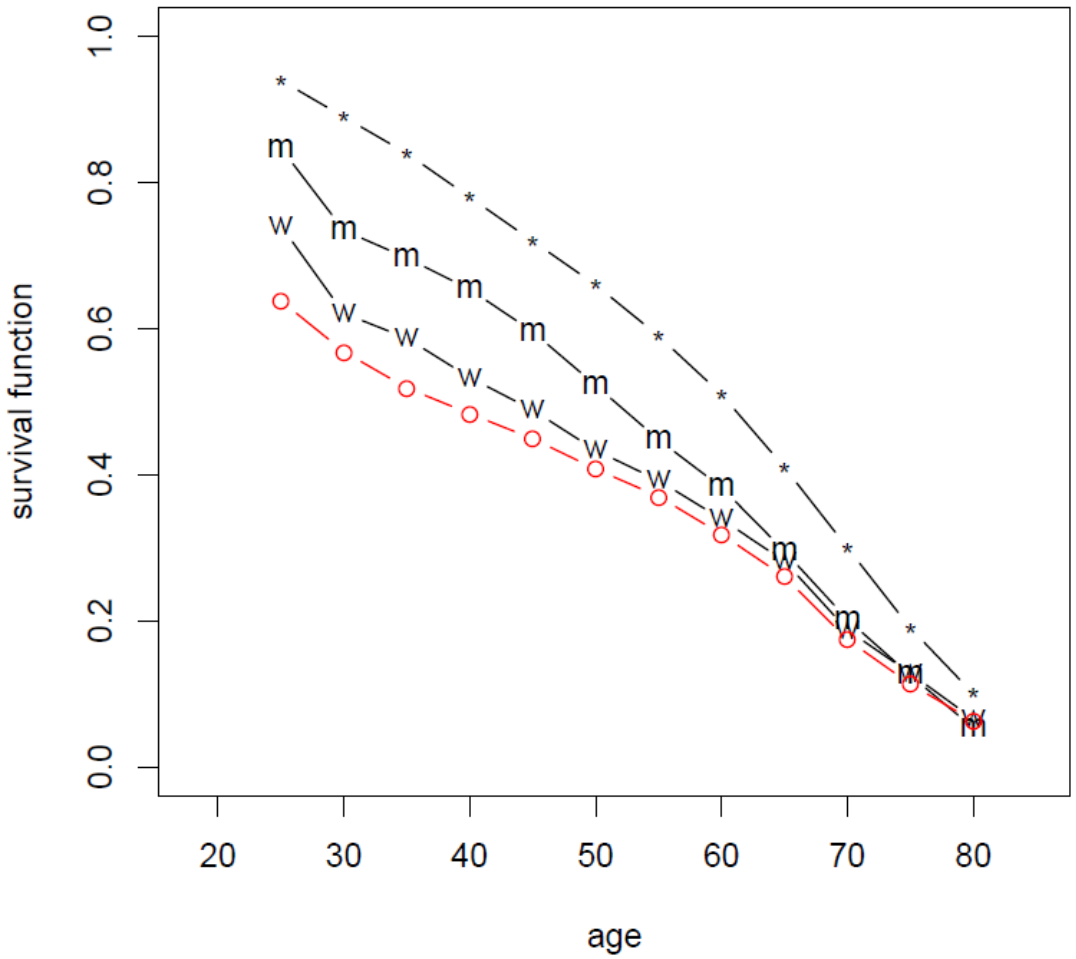
Avec des intervalles de crédibilité à 50 et 90 %



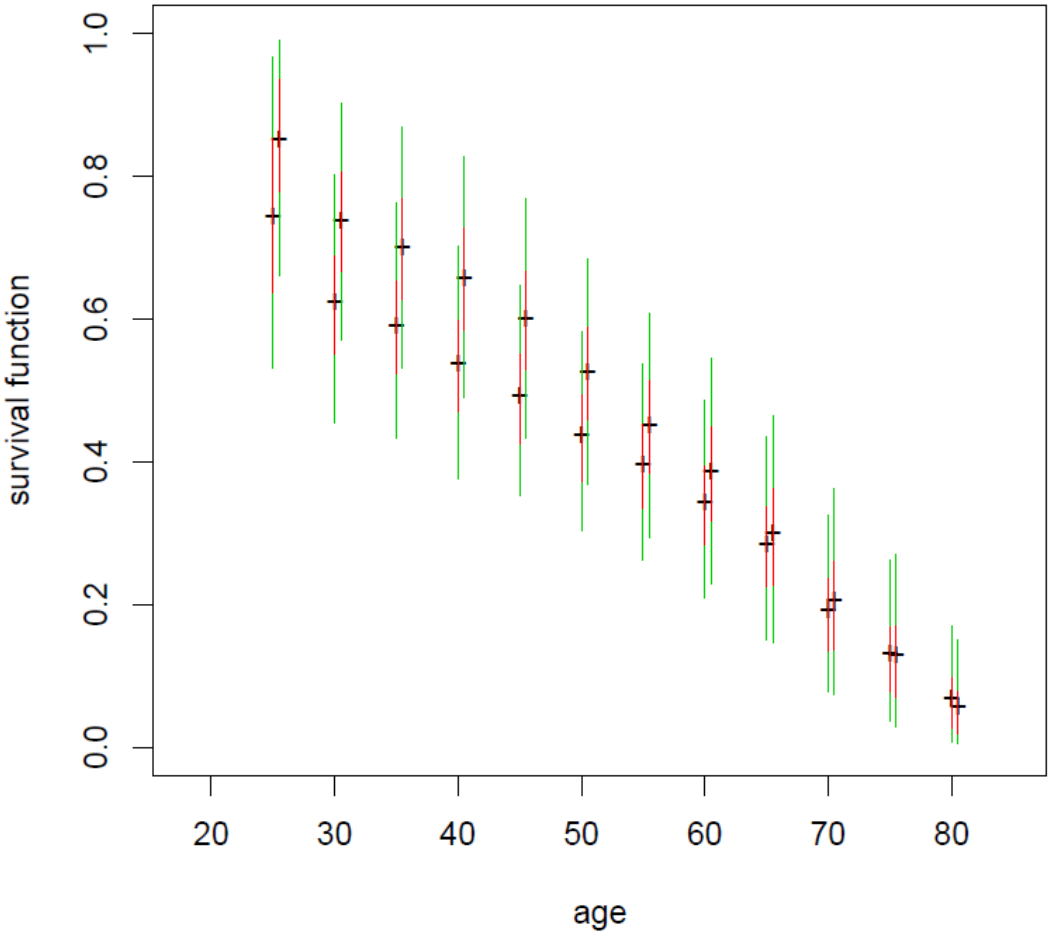
Comparaison par sexes pour l’époque mérovingienne



Comparaison par sexes pour l’époque mérovingienne



Comparaison par sexes – intervalles de crédibilité (époque mérovingienne)



Une autre loi a priori ?

Le standard préindustriel qui donne la moyenne de la loi a priori de Dirichlet a été obtenu comme moyenne de tables de mortalité relevées dans la littérature (167 tables pour les sexes réunis). Peut-on utiliser ces tables au-delà de leur moyenne pour avoir une idée de la distribution a priori dans son ensemble ?

La distribution empirique de ces tables est-elle une « bonne » loi a priori ?

Pour le contrôler :

- Reprendre certaines estimations avec une telle loi a priori, en particulier des cas tests où existent des registres,
- Examiner de près la distribution des tables compilées,
- En déduire les questions à poser aux paléodémographes pour essayer de conclure.

Reprise de l'exemple d'Antibes

A priori de Dirichlet :

âge	18 - 19	20 - 24	25 - 29	30 - 34	…
estimation	0.029	0.068	0.068	0.062	...
écart-type	0.045	0.064	0.067	0.064	…

A priori uniforme sur les lois compilées :

âge	18 - 19	20 – 24	25 – 29	30 – 34	…
estimation	0.020	0.048	0.050	0.052	…
écart-type	0.007	0.015	0.016	0.015	…

Reprise de l'exemple d'Antibes

A priori de Dirichlet :

écart-type 0.045 0.064 0.067 0.064 …

A priori uniforme sur les lois compilées :

écart-type 0.007 0.015 0.016 0.015 …

Reprise de l'exemple de Maubuisson

A priori de Dirichlet :

âge	18 - 19	20 - 24	25 - 29	30 - 34	…
estimation	0.025	0.026	0.035	0.037	...
écart-type	0.030	0.030	0.039	0.042	…

A priori uniforme sur les lois compilées :

âge	18 - 19	20 – 24	25 – 29	30 – 34	…
estimation	0.039	0.042	0.044	0.046	…
écart-type	0.009	0.008	0.008	0.007	…

Reprise de l'exemple de Maubuisson

A priori de Dirichlet :

écart-type 0.030 0.030 0.039 0.042 …

A priori uniforme sur les lois compilées :

écart-type 0.009 0.008 0.008 0.007 …

Comparaison aux registres

Pour Antibes

légèrement favorable à la nouvelle loi a priori

Pour Maubuisson

nettement défavorable à la nouvelle loi à cause de l'estimation très surévaluée de la mortalité des jeunes :

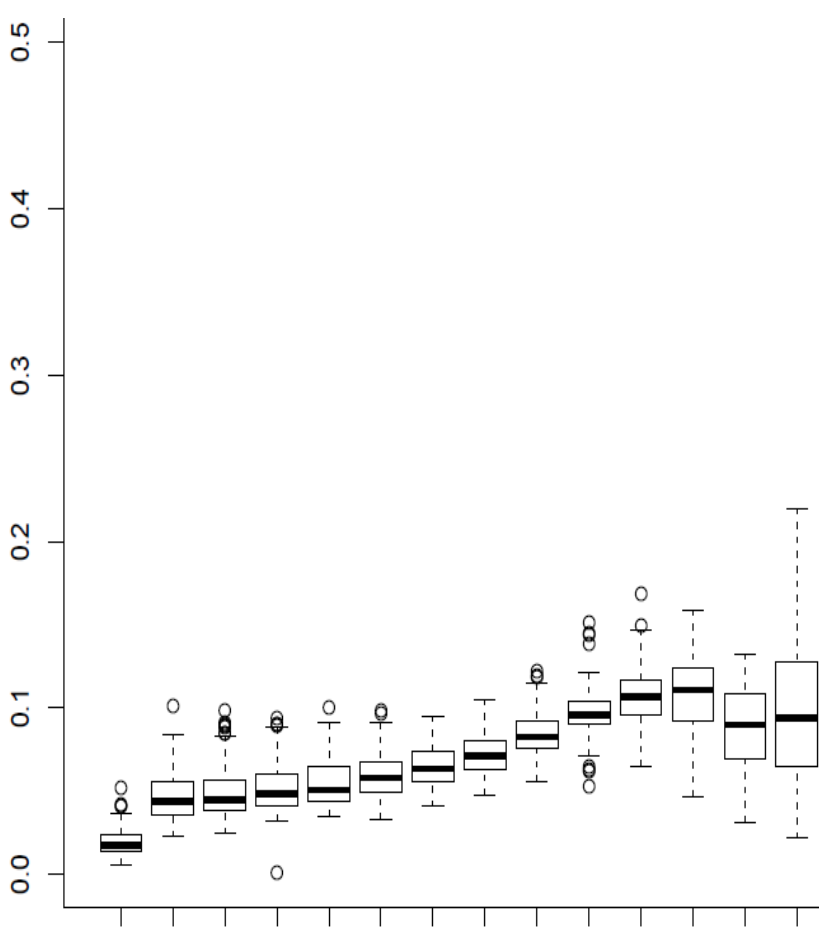
	20-24 ans	25-30 ans	30-34 ans
registres	0	.012	.010
Dirichlet	.025 (.011)	.026 (.012)	.035 (.028)
Compil.	.039	.042	.044

La nouvelle loi resserre beaucoup l'éventail des possibilités a priori

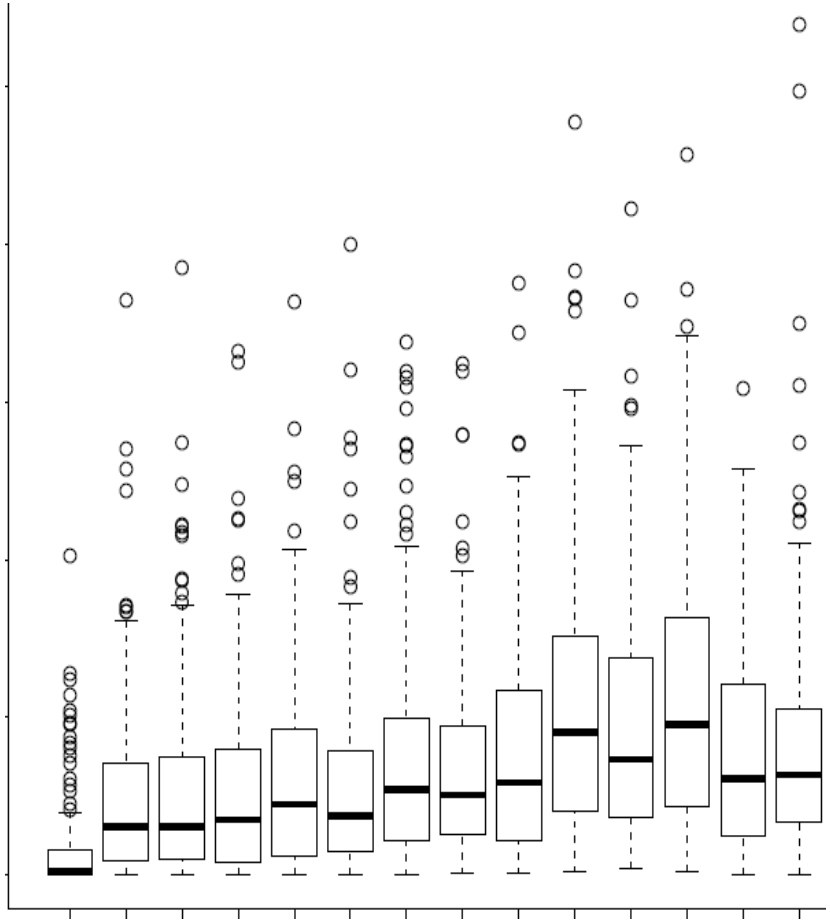
Par exemple dans les tables compilées

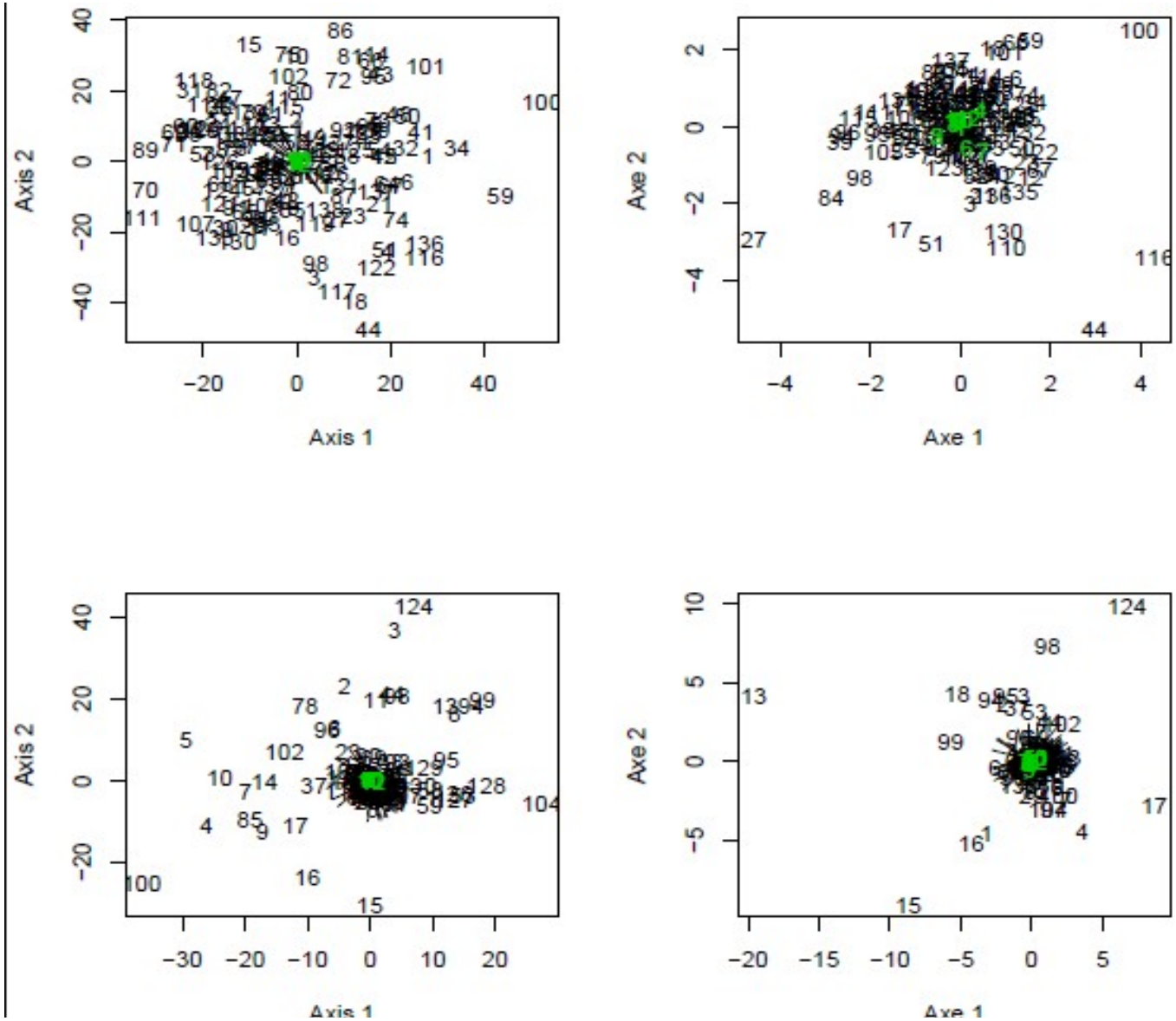
	↓ probabilité minimum	↓ probabilité maximum	Registres Maubuisson
20-24 ans	.0098	.0628	0
80 ans et +	.0208	.2149	.207

Lois compilées



Loi de Dirichlet




Axis 2
Axis 1
Axe 2
Axe 1
Axis 2
Axis 1
Axe 2
Axe 1

Frénouville époque mérovingienne

Classe d’âge 18 - 24 ans

	estimation	écart-type
A priori Dirichlet	0.354	0.090
A priori compil.	0.130	0.015

.

Pour cette classe la mortalité maximale observée parmi les lois compilées est 0.152

Frénouville époque mérovingienne

Classe d’âge 18 - 24 ans

	estimation	écart-type
A priori Dirichlet	0.354	0.090
A priori compil.	0.130	0.015
A priori « mixte »	0.355	0.083

Evaluer l'âge d'un individu présentant le stade i

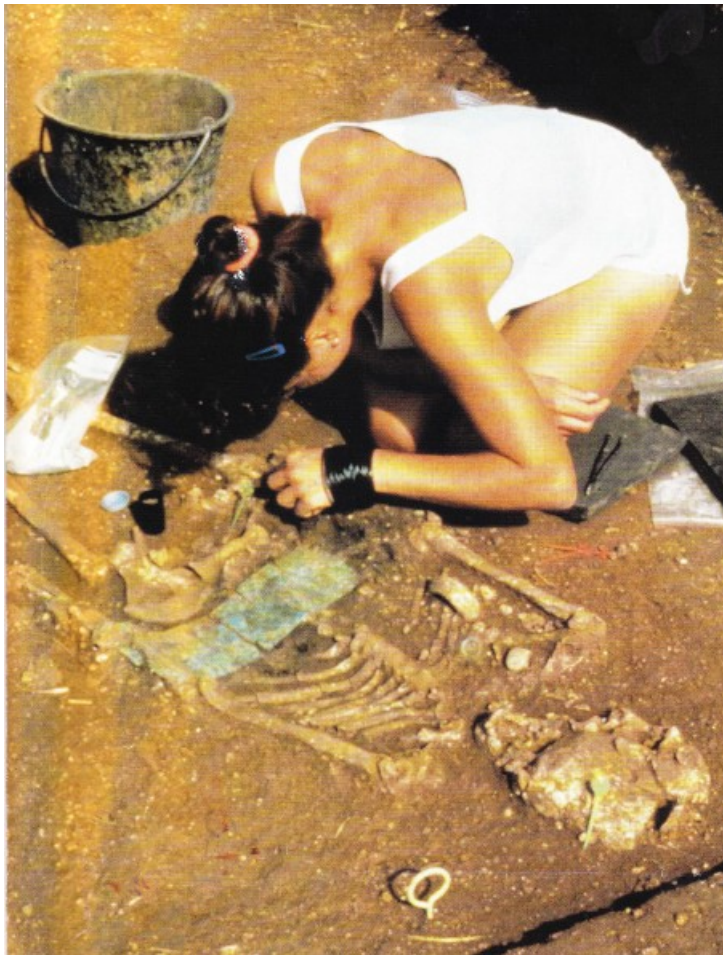
La probabilité de l'âge j connaissant le stade i est

$$p_{j/i} = \frac{p_j p_{i/j}}{\sum_h p_h p_{i/h}}$$

où

- les $p_{j/i}$ sont analogues aux probabilités de référence,
- il faut une information sur les p_j .

Merci pour votre attention



Loi de Gompertz

Si X est la durée de vie

$$S(t) = P(X>t) = \exp(\lambda(1 - e^{\rho t}))$$

Quotient instantané de mortalité :

$$- S'(t) / S(t) = \lambda \rho e^{\rho t}$$

Survie conditionnelle (pour $a < t$) :

$$P(X>t / X>a) = \exp(\lambda(e^{\rho a} - e^{\rho t}))$$

Loi de Dirichlet

$$x = (x_1, \ldots, x_k) \in D \quad \Leftrightarrow \quad x_i > 0 \ \ pour\ tout \ \ i = 1, \ldots, k \quad et \quad \sum_{i=1}^{k} x_i = 1$$

$$d(x) = \begin{cases} \dfrac{\Gamma(a_.)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} x_i^{a_i - 1} & pour\ x \in D \\ 0 & pour\ x \notin D \end{cases}$$

$$E(X_i) = \frac{a_i}{a_.} \qquad Var(X_i) = \frac{a_i (a_. - a_i)}{a_.^2 (a_. + 1)} \qquad Cov(X_i X_j) = -\frac{a_i\ a_j}{a_.^2 (a_. + 1)}$$

Peut-on « corriger » les données de référence ?

Stage	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80+
A	14	39	36	13	14	16	20	12	6	3	5	2	4	2
B	2	10	8	19	14	13	15	13	14	5	3	5	4	11
C	1	4	5	1	5	9	10	9	9	7	9	11	7	13
D	1	2	3	5	3	4	6	11	10	10	8	8	9	15
E	0	2	0	5	4	2	2	6	10	7	13	7	9	28
total	18	57	52	43	40	44	53	51	49	32	38	33	33	69

Modèle de Goodman
(J.A.S.A. 1991, 86, 416, 1085-1138)

$$p_{ij} = \alpha_i \beta_j e^{\varphi \mu_i \nu_j}$$ à l'ordre 1

$$p_{ij} = \alpha_i \beta_j \exp\left(\sum_{k=1}^{K} \varphi_k \mu_{ik} \nu_{jk}\right)$$ cas général

Le modèle d'ordre 1 dit « en gros » qu'il y a une loi bi-normale sous-jacente, les paramètres μ et ν donnant des scores qui représentent les abscisses des marges correspondantes.

Ajustement du modèle de Goodman sur le cas 5 x 8

Le choix de dimension au moyen d'un critère tels que AIC ou BIC conduit à l'ordre 1. On a pour estimation

des μ_i :

-1.359 -0.206 0.418 0.988 1.348

des ν_j :

-1.877 -1.385 -0.516 -0.384 0.195 1.032 1.033 1.504

Le coefficient de corrélation est estimé à 0.40

ce qui n'est pas très gros !