Méthode Bayésienne d'estimation du seuil optimal d'un biomarqueur et de son intervalle de crédibilité

Fabien Subtil & Muriel Rabilloud

Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique Santé
Service de Biostatistique, Hospices Civils de Lyon

Journée AppliBUGS - Juin 2012

Exemples

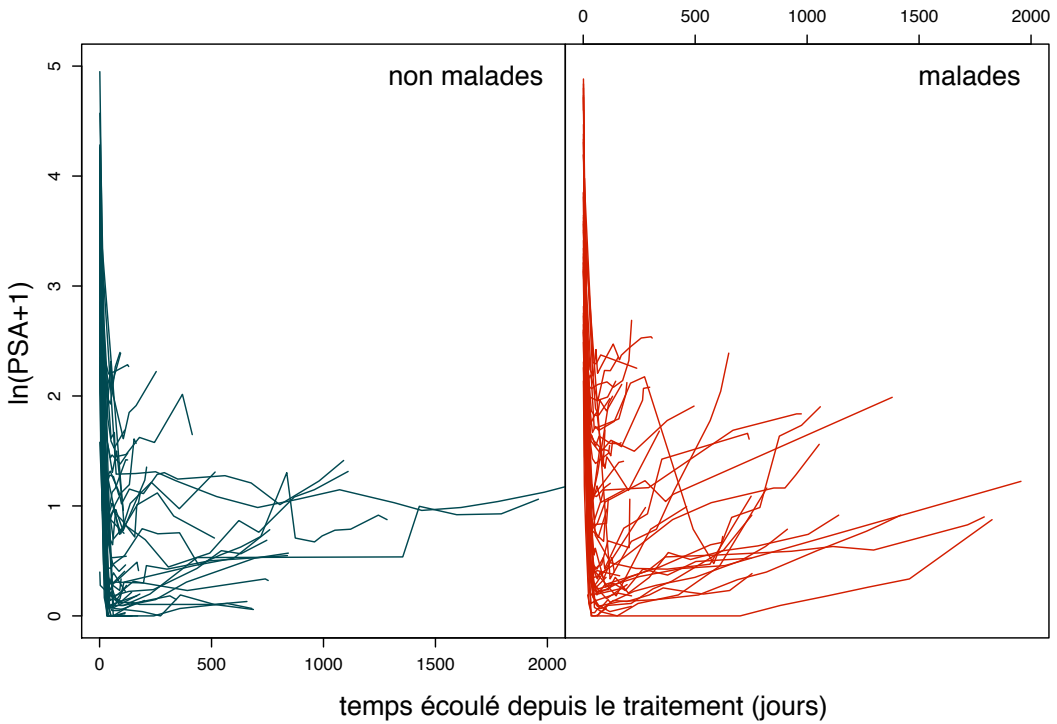
Seuil optimal d'un biomarqueur

Biomarqueurs avec valeurs extrêmes

Mélanges de loi

Cancer de la prostate et nadir de PSA

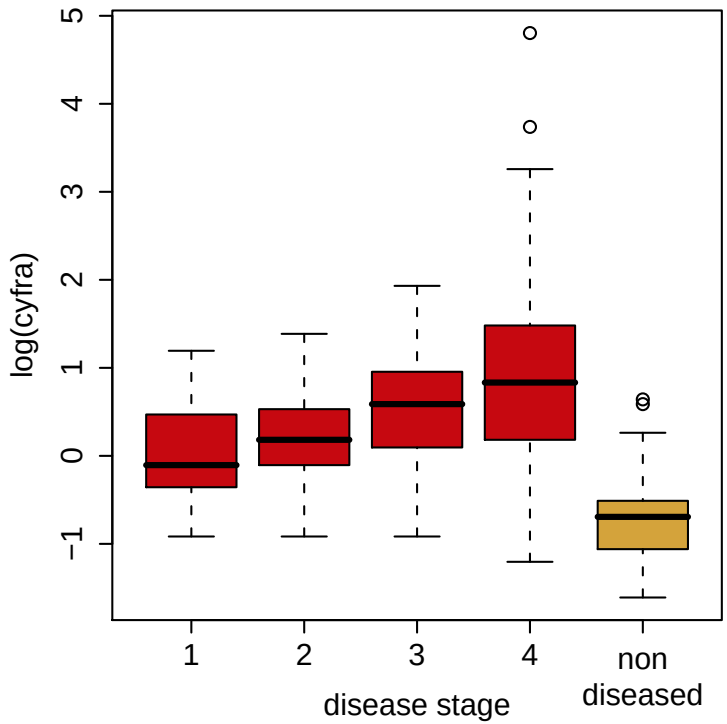
Utilisation des mesures de PSA effectuées après traitement par ultrasons pour détecter une récidive du cancer de la prostate.
Données de 150 patients ayant présenté une récidive, et 139 sans récidive (preuve histologique).



Utilisation du nadir de PSA pour détecter une récidive.
Quel seuil utiliser ?

Cancer des voies aérodigestives supérieures

300 patients atteints d'un cancer des voies aérodigestives supérieures (71 contrôles).
Dosage du marqueur tumoral Cyfra 21-1.



A partir de quelle valeur de Cyfra 21-1 faut-il affirmer la présence d'un cancer des voies aérodigestives supérieures ?

Qu'est ce qu'un seuil optimal ? (1/2)

Un seuil optimal est une valeur du biomarqueur maximisant une fonction d'utilité.

Exemple : nombre moyen de patients correctement classés par le test.

$$U(c) = N[\underbrace{\text{Sen}(c) \times \pi}_{\text{vrais positifs}} + \underbrace{\text{Spe}(c) \times (1-\pi)}_{\text{vrais négatifs}}]$$

N : nombre de patients ; π : prévalence de la pathologie.

Qu'est ce qu'un seuil optimal ? (2/2)

Fonction d'utilité définie en termes d'état de santé

Introduction des bénéfices et coûts des classements corrects et des erreurs.

$$U(c) = N[\underbrace{\text{Sen}(c) \times \pi \times U_{VP}}_{\text{vrais positifs}} + \underbrace{(1 - \text{Sen}(c)) \times \pi \times U_{FN}}_{\text{faux négatifs}} + \underbrace{(1 - \text{Spe}(c)) \times (1 - \pi) \times U_{FP}}_{\text{faux positifs}} + \underbrace{\text{Spe}(c) \times (1 - \pi) \times U_{VN}}_{\text{vrais négatifs}}]$$

U_{VP}, U_{FN}, U_{FP} et U_{VN} : utilités associées à un état de santé.

Objectif

Estimer c qui maximise

$$\tilde{U}(c) = \text{Sen}(c) + \text{Spe}(c) \times \frac{U_{VN} - U_{FP}}{U_{VP} - U_{FN}} \times \frac{1 - \pi}{\pi} = \text{Sen}(c) + \text{Spe}(c) \times R$$

Méthode non paramétrique d'estimation du seuil

Difficulté : estimer Sen(c) et Spe(c) quel que soit c.

Utiliser la fonction de répartition empirique du biomarqueur dans les deux groupes :

- pas d'hypothèse sur la distribution du biomarqueur ;
- méthode peu précise ;
- pas d'intervalle de confiance, sauf par bootstrap ; probabilité de couverture pas toujours acceptable (Schisterman et Perkins, 2007).

Méthode paramétrique d'estimation du seuil (1)

Modéliser la distribution du biomarqueur dans les deux groupes et utiliser les paramètres de distribution pour estimer le seuil optimal.

Exemple : biomarqueur dont la distribution peut être reproduite par une loi normale dans les deux groupes $\mathcal{N}(\mu_0, \sigma_0^2)$ et $\mathcal{N}(\mu_1, \sigma_1^2)$.

Seuil optimal :

$$c = \frac{\mu_0(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\sigma_0^2 \ln(b^2 R^2)}}{b^2 - 1}$$

$a = \mu_1 - \mu_0$, $b = \sigma_1/\sigma_0$.

- Estimation ponctuelle : méthode du plugin.
- Intervalle de confiance : méthode Delta.

Méthode paramétrique d'estimation du seuil (2)

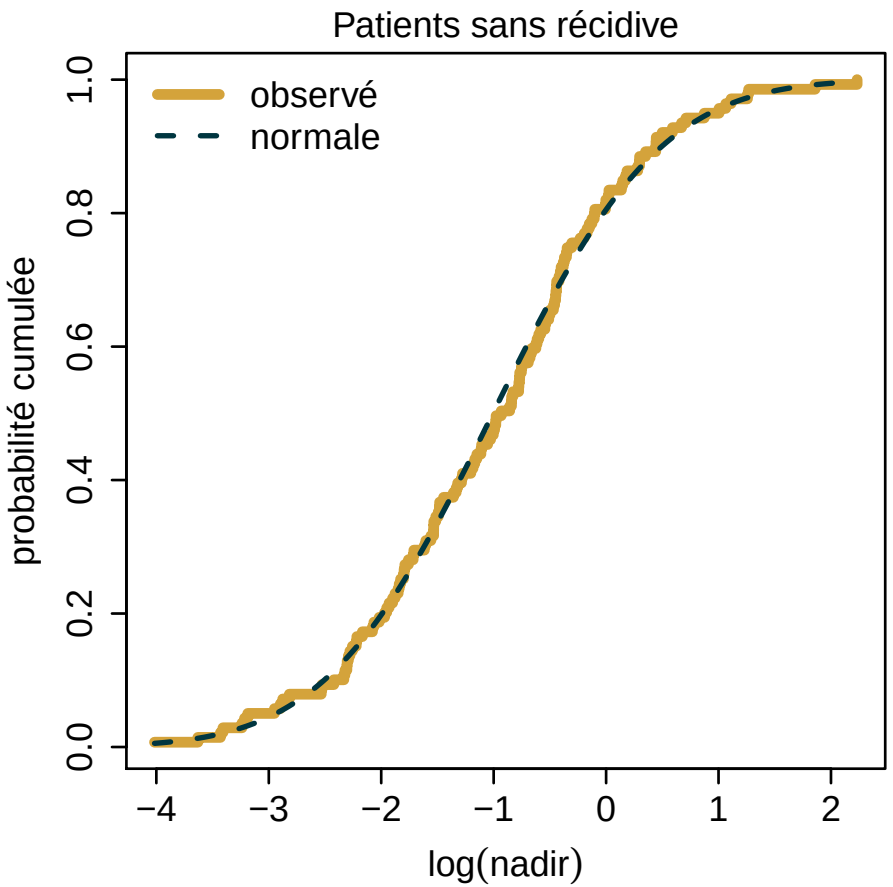
Méthode plus précise.

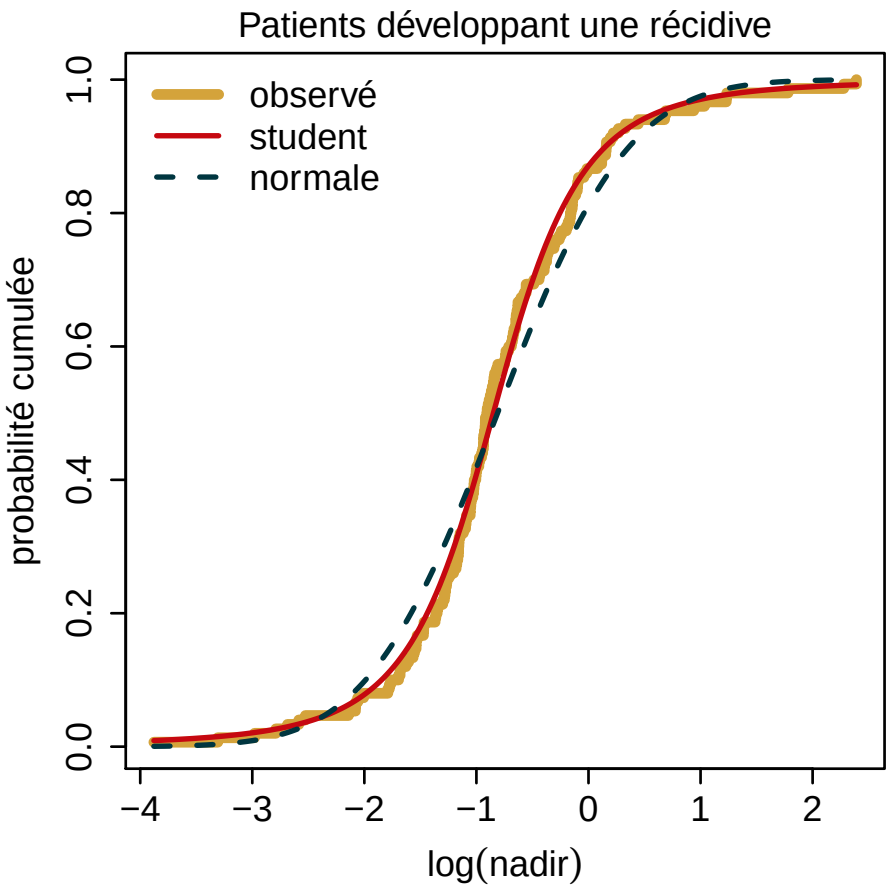
Solution explicite au problème d'optimisation dans le cas de lois normales, log normales, ou gamma.

En l'absence de formule explicite du seuil optimal :

- estimation ponctuelle : méthode numérique de maximisation de fonction (Newton Raphson) ;
- intervalle de confiance : bootstrap.

Distribution du logarithme du nadir de PSA





Inférence et méthodes MCMC

Inférence à partir d'un échantillon de valeurs représentatives de la distribution a posteriori d'un paramètre :

- estimation ponctuelle
  - moyenne ;
  - médiane ;
  - mode.
- intervalle de crédibilité à 95 %
  - quantiles 2.5 et 97.5 % ;
  - région de plus haute densité de probabilité (HPD).

Soit $\theta_1, \ldots, \theta_n$ un ensemble de paramètres, et $\theta^\star = f(\theta_1, \ldots, \theta_n)$.
Un échantillon de valeurs représentatives de la distribution a posteriori de $\theta^\star$ est obtenu à partir des distributions a posteriori de $\theta_1, \ldots, \theta_n$.

Estimation Bayésienne du seuil optimal

Paramètres de distribution			Fonction d'utilité		Seuil optimal
θ_{01}	θ_{11}	$\rightarrow$	$\text{Sen}(c\|\theta_{11}) + \text{Spe}(c\|\theta_{01}) \times R$	$\rightarrow$	c_1
θ_{02}	θ_{12}		$\text{Sen}(c\|\theta_{12}) + \text{Spe}(c\|\theta_{02}) \times R$		c_2
$\vdots$	$\vdots$		$\vdots$		$\vdots$
θ_{0K}	θ_{1K}		$\text{Sen}(c\|\theta_{1K}) + \text{Spe}(c\|\theta_{0K}) \times R$		c_K

Simulations

Design

Valeurs de biomarqueur générées pour 100 malades et 100 non malades ; 5 000 répétitions.

Non malades :

$$Y_0 \hookrightarrow \mathcal{N}(-0.30, 0.05^2)$$

Malades :

$$Y_1 \hookrightarrow t_\nu(-0.25, 0.05)$$

Valeurs de paramètres

$\nu = \{1, 4, 8, 12\}$.

Critères d'intérêt

- biais relatif de l'estimation du seuil ;
- probabilité de couverture de l'intervalle de crédibilité à 95 %.

Résultats

	Biais relatif*		Probabilité de couverture†	
ν	Gauss	Student	Gauss	Student
1	0.2929	0.0082	0.000	0.932
4	0.0435	0.0036	0.532	0.955
8	0.0148	0.0017	0.868	0.952
12	0.0086	0.0011	0.920	0.955

* moyenne a posteriori † méthode HPD.

Gauss : lois normales pour les deux groupes.
Student : lois de Student pour les deux groupes.

Diagnostic de récidive du cancer de la prostate

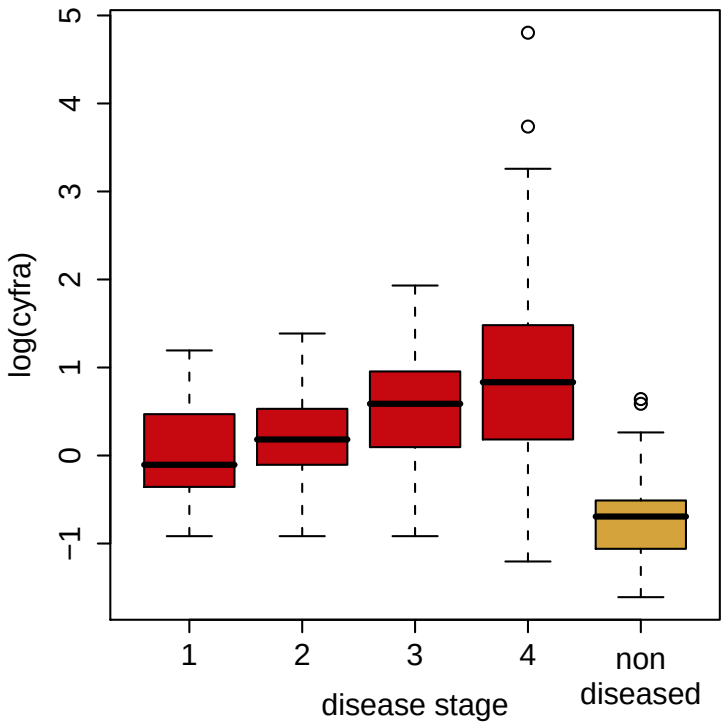
Prévalence de 52 %.

$\frac{U_{VP}-U_{FN}}{U_{VN}-U_{FP}}$	Student		Normale	
0.43	11.620	[0.329, 19.400]	12.180	[8.560,18.000]
0.67	0.412	[0.236, 0.397]	0.484	[0.215, 0.847]
1	0.215	[1.165, 0.261]	0.177	[0.120, 0.246]
1.5	0.153	[0.105, 0.208]	0.096	[0.046, 0.149]
2.33	0.084	[0.000, 0.138]	0.060	[0.023, 0.107]
4	0.011	[0.000, 0.074]	0.038	[0.007, 0.068]

Mélanges de loi selon la sévérité de la maladie

Cyfra 21-1 :

- patients sans cancer : loi log-normale assez adaptée ;
- patients avec cancer : mélange de distributions suivant la sévérité de la maladie.



Loi de Dirichlet (1)

Evénement binaire

Soit θ la probabilité d'une maladie spécifique. Sur un échantillon de taille n, x_1 malades ont été observés.

Vraisemblance :

$$X_1 \hookrightarrow \mathcal{B}(n, \theta)$$

A priori :

$$\theta \hookrightarrow \text{Beta}(a_1, a_2)$$

A posteriori :

$$\theta | y_1 \hookrightarrow \text{Beta}(a_1 + x_1, a_2 + n - x_1)$$

La loi Beta est conjuguée à la vraisemblance binomiale.

La loi de Dirichlet (2)

Variable catégorielle à plus de deux modalités

$$\left.\begin{array}{l} X_1 \rightarrow \theta_1 \\ X_2 \rightarrow \theta_2 \\ X_3 \rightarrow \theta_3 \\ X_4 \rightarrow \theta_4 \\ X_5 \rightarrow \theta_5 \\ X_6 \rightarrow \theta_6 \end{array}\right\} \text{Dirichlet}(a_1, \ldots, a_6)$$

Vraisemblance :

$$(x_1, \ldots, x_6) \hookrightarrow \text{Multinomial}(\theta_1, \ldots, \theta_6)$$

A posteriori :

$$(\theta_1, \ldots, \theta_6) \hookrightarrow \text{Dirichlet}(a_1 + x_1, \ldots, a_6 + x_6)$$

La loi de Dirichlet est conjuguée à la vraisemblance multinomiale.

La loi de Dirichlet (3)

Autre paramétrisation :

$$(\theta_1, \ldots, \theta_6) \hookrightarrow \text{Dirichlet}(a_1, \ldots, a_6)$$

$$(\theta_1, \ldots, \theta_6) \hookrightarrow \text{Dirichlet}\left(n\left(\frac{a_1}{n}, \ldots, \frac{a_6}{n}\right)\right)$$

$$(\theta_1, \ldots, \theta_6) \hookrightarrow \text{Dirichlet}(n(p_1, \ldots, p_6))$$

n : mesure de précision ; $p_1, \ldots, p_6$: probabilité a priori des différents événements.

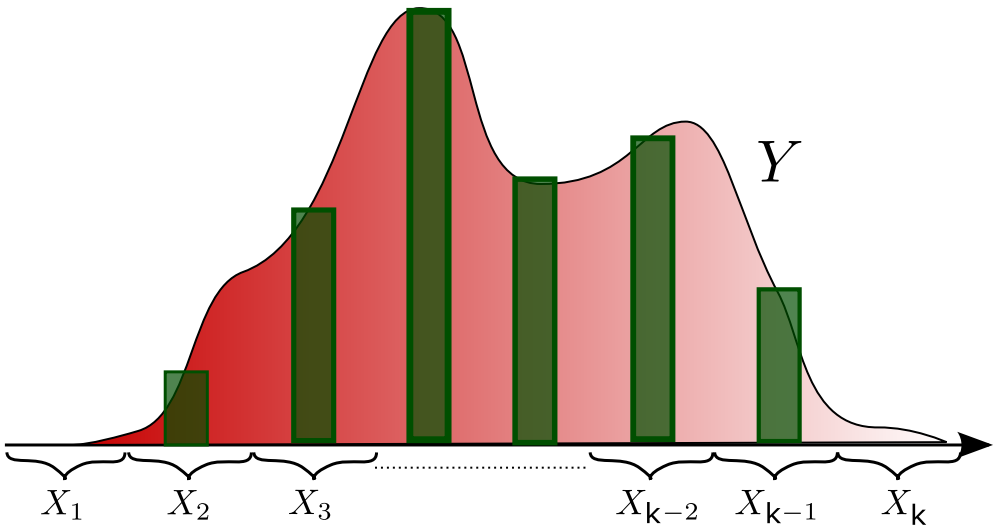
Processus de Dirichlet

Soit une variable aléatoire Y. L'espace des valeurs possibles pour Y peut-être divisé en une infinité de sous-espaces X_j, $j = 1, \ldots, k$.
Soit $\theta_j = P(Y = X_j)$. L'ensemble des θ

- suit une loi de Dirichlet ;
- forme une distribution de probabilité pour Y, (G).

$$\begin{aligned} Y|G &\hookrightarrow G \\ G &\hookrightarrow \mathcal{DP}(M.G_0) \end{aligned}$$

Le processus de Dirichlet est une simple extension de la loi de Dirichlet au cas de données continues.



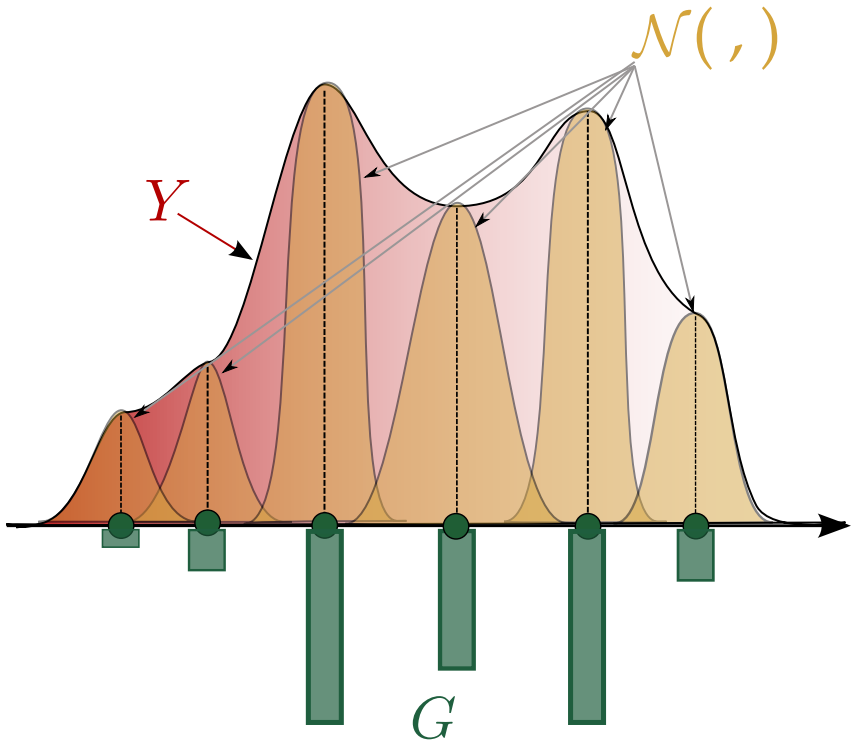
Mélange de processus de Dirichlet

$$
\begin{aligned}
y_i|\beta_i,\sigma_i^2 &\hookrightarrow \mathcal{N}(\beta_i,\sigma_i^2)\\
\beta_i,\sigma_i^2|G &\hookrightarrow G\\
G &\hookrightarrow \mathcal{DP}(M.G_0)
\end{aligned}
$$

Hyper paramètres :

$$
\begin{aligned}
G_0 &\hookrightarrow \mathcal{N}(\mu_0,\sigma_0^2)\\
&\quad \times \text{Inv}-\text{Gamma}(\delta_0,\delta_1)\\
M &\hookrightarrow \text{Gamma}(\alpha_0,\alpha_1)
\end{aligned}
$$

G : mass point distribution.
G_0 : a priori pour G.
M : degré de certitude concernant G_0.



Simulations

Design

Valeurs de biomarqueurs générées pour $N/2$ malades et $N/2$ non malades ; 5 000 répétitions.

Non malades :

$$Y_0 \hookrightarrow \mathcal{N}(-0.3, 0.07^2)$$

Malades :

$$Y_1 \hookrightarrow 0.5 \times \mathcal{N}(0.05, \sigma_1^2) + 0.5 \times \mathcal{N}(-0.25, \sigma_2^2)$$

Valeurs de paramètres

- $N = \{200, 400\}$;
- $\sigma_1 = \{0.07, 0.08, 0.10\}$;
- $\sigma_2 = \{0.05, 0.07\}$.

Critères d'évaluation

- biais relatif de l'estimation du seuil ;
- probabilité de couverture de l'intervalle de crédibilité à 95 %.

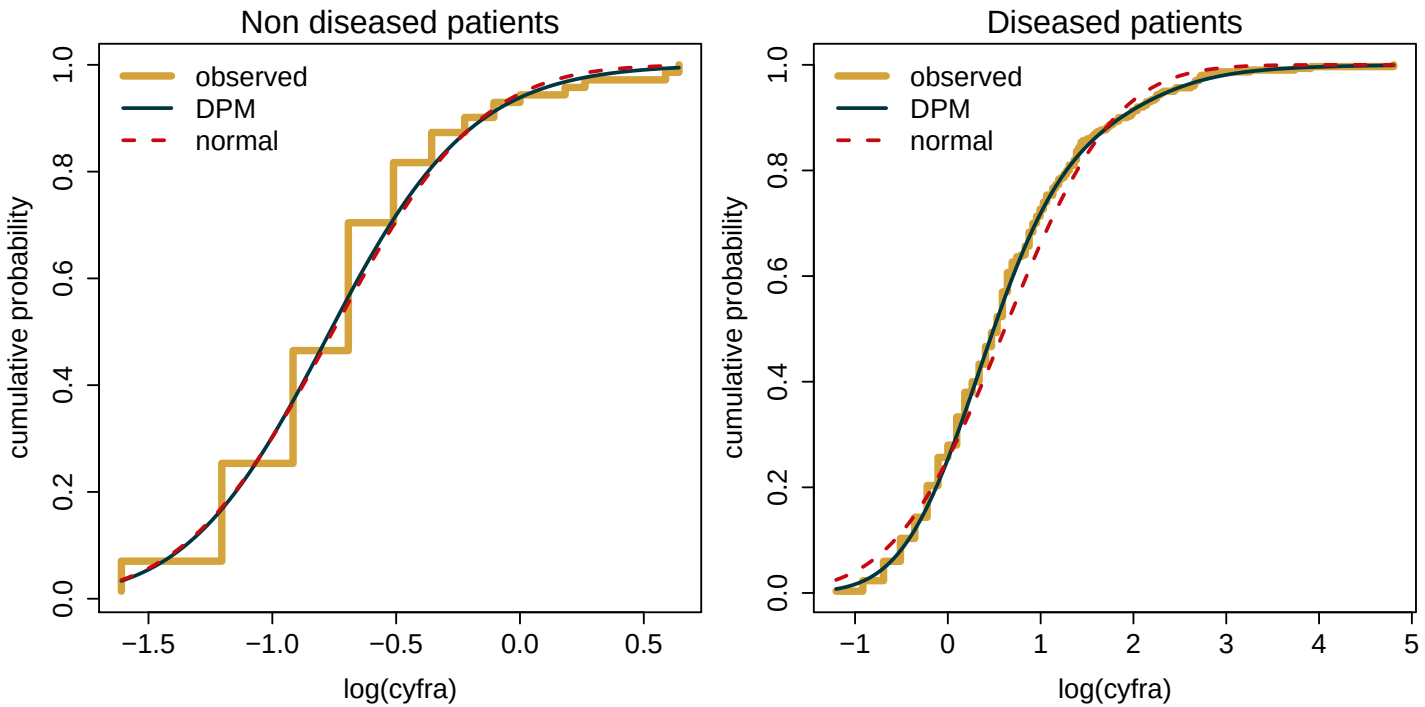
Résultats

			Biais relatif*		Probabilité de couverture†	
N	σ_1	σ_2	Gauss	MPD	Gauss	MPD
400	0.07	0.07	0.040	-0.018	0.623	0.947
400	0.08	0.05	0.166	0.008	0.000	0.956
400	0.10	0.05	0.172	0.002	0.000	0.959
200	0.07	0.07	0.040	-0.022	0.792	0.944
200	0.08	0.05	0.166	0.011	0.000	0.960
200	0.10	0.05	0.172	0.004	0.000	0.961

* moyenne a posteriori du seuil optimal † méthode HPD.

Gauss : lois normales pour les deux groupes ;
MPD : mélange de processus de Dirichlet pour les deux groupes.

Application



Seuil optimal $((U_{VP} - U_{FN})/(U_{VN} - U_{FP}) = 1)$:

Estimation ponctuelle

- mode : 2.19 ;
- médiane : 2.46 ;
- moyenne : 2.68.

Intervalle de crédibilité à 95 %

- HPD : [1.66, 4.33] ;
- quantile : [1.80, 4.89].

Conclusion

- Méthode unifiée d'estimation du seuil optimal d'un biomarqueur, quelle que soit la distribution du biomarqueur.
- Plus de contrainte dans le choix des lois.
- Difficultés :
  - déterminer la distribution du biomarqueur dans les deux groupes ;
  - échantillonner dans la distribution a posteriori des paramètres.
- Espoir : ne plus utiliser le seuil qui conduit à la meilleure combinaison sensibilité / spécificité.