Modèle de mélange sur distributions binomiales : un exemple réel

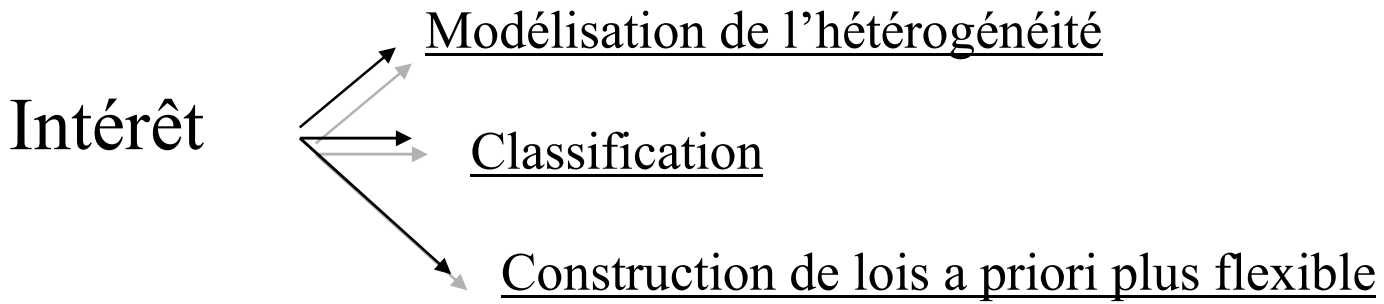
Isabelle Albert (INRA-MIA, Paris), Jean-Baptiste Denis (INRA-MIA, Jouy)

PLAN

- Introduction aux modèles de mélange
- Les données
- Modèle binomial
- Modèle hiérarchique
- Modèle de mélange
- Conclusion

INTRODUCTION

- Regain d'intérêt autour des modèles de mélange de distributions de probabilités
  - Variétés des techniques d'estimation (ex: bayésienne, EM)
  - Outils disponibles (MIXMOD, WINBUGS)



Modèles de mélange

$$y_i \sim f(y) = \sum_{j=1}^{k} p_j f\left(y|\theta_j\right), \quad i = 1, \ldots n$$

Observations indépendantes

Densités de probabilités

Poids des composantes du mélange $(p_j \geq 0 \quad \text{avec} \quad \sum_{j=1}^{k} p_j = 1)$

- Estimation des poids et des paramètres des distributions

⇨ <u>*k* connu ou inconnu</u>

Modèles de mélange

– Mélange quelconque : *f , familles de lois quelconques*

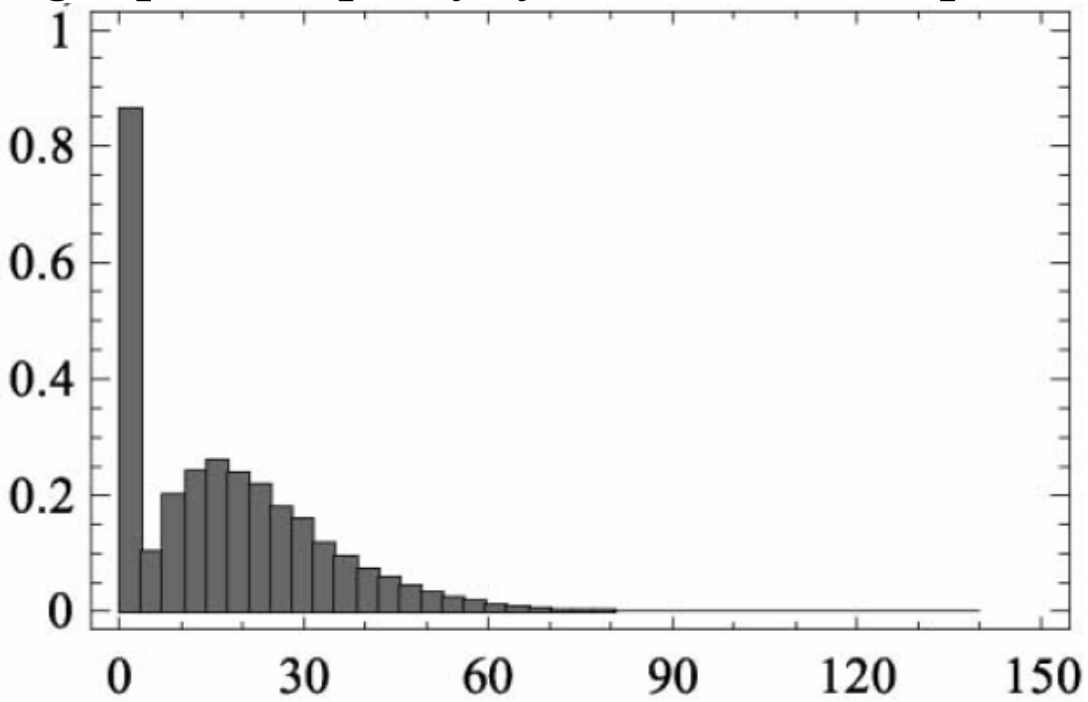


Fig. 4. Typical consumption histogram.

Modèles de mélange

– Mélange gaussien, poissonien,…

- Gaussien :

$$y_i \sim f(y) = \sum_{j=1}^{k} p_j f\left(y|\theta_j\right), \quad i = 1,\ldots n$$

$$\text{avec} \quad f\left(y|\theta_j\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}\left(y-\mu_j\right)^2\right\}$$

Modèles de mélange

– Mélange gaussien

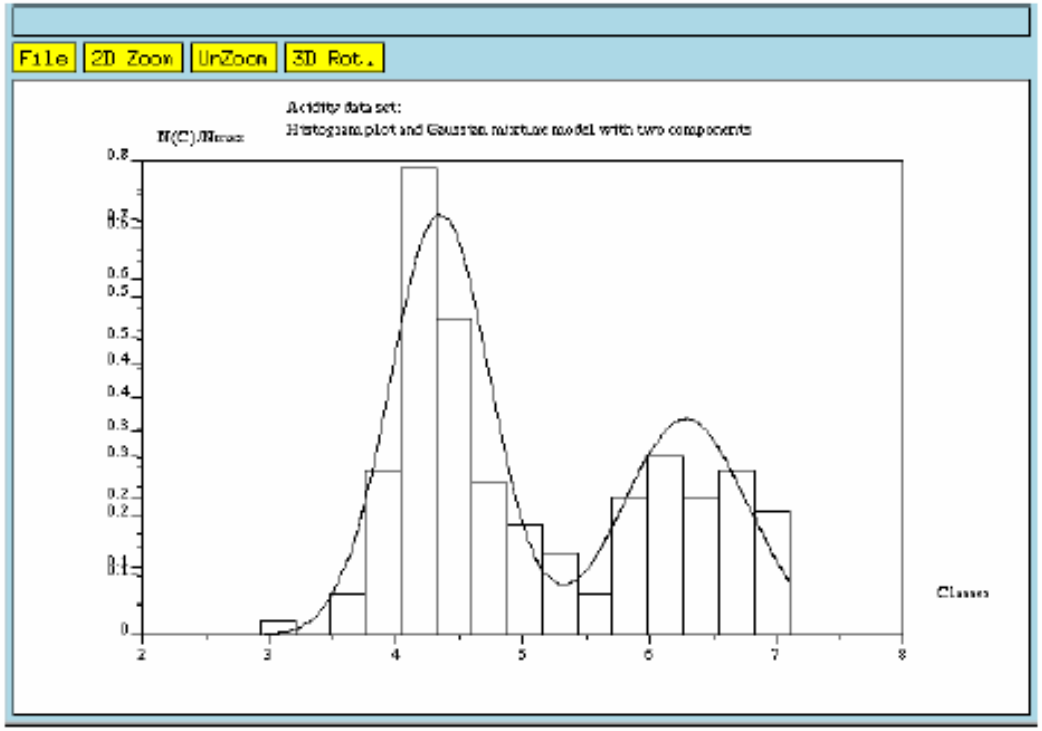


Figure 1.1: Plot of fitted two-component normal mixture density for acidity data set

Modèles de mélange

– Mélange binomiale

$X_1, \ldots, X_n$ i.i.d.

$$P(X = k) = \sum_{j=1}^{J} w_j \binom{n}{k} p_j^k (1 - p_j)^{n-k}, \quad k = 0,1,\ldots,n$$

- Mélange Beta-Binomiale

$$P(k|n, \alpha_\omega, \beta_w) = \int_0^1 dp \, P_{\text{Bin}}(k|n, p) \, P_{\text{Beta}}(p|\alpha_\omega, \beta_w)$$

Illustration

- Estimation de la prévalence de la contamination par *L. monocytogenes* dans le lait cru.
  - Données bibliographiques internationales :
    - 91 études répertoriées (issues de 2 rapports de synthèse)

      Dans chaque étude i, n_i prélèvements réalisés et r_i résultats positifs (n_i de 14 à 1 227 053)

      Par étude, l'estimation de la prévalence est donnée par :

$$\hat{p}_i = \frac{r_i}{n_i}.$$

#	[a]	[b]	Référence	Origine	n	r	$IC_{95\%,\text{inf}}$ (%)	$\widehat{p}$ (%)	$IC_{95\%,sup}$ (%)
1		*	Sutherland, 1989	Australie	600	0	0.0	0.0	0.5
2		*	Hartung, 2000	Allemagne	415	0	0.0	0.0	0.7
3		*	Quaglio et al., 1992	Italie	276	0	0.0	0.0	1.1
4	*		Arimi et al, 1997	Costa-Rica	220	0	0.0	0.0	1.4
5	*	*	Ibrahim and Mc Rea, 1991	Australie	150	0	0.0	0.0	2.0
6	*		Soncini and Piantoni, 1993	Italie	142	0	0.0	0.0	2.1
7	*		Takai et al., 1990	Japon	120	0	0.0	0.0	2.5
8	*		Luisjan-Morales et al., 1995	Mexique	100	0	0.0	0.0	3.0
9	*	*	Lovett et al., 1987	USA	100	0	0.0	0.0	3.0
10	*		Gelosa, 1990	Italie	85	0	0.0	0.0	3.5
11	*	*	Patterson et al., 1989	USA	84	0	0.0	0.0	3.5
12	*	*	Stone, 1987	N-Zélande	71	0	0.0	0.0	4.1
13		*	Desmeures et al., 1997	France	69	0	0.0	0.0	4.2
14	*	*	Vessau, 1988	Canada	55	0	0.0	0.0	5.3
15	*		Tiscione et al., 1994	Italie	50	0	0.0	0.0	5.8
16	*	*	Doyle and Schoeni, 1986	USA	50	0	0.0	0.0	5.8
17	*		Masse et al., 1990	Italie	40	0	0.0	0.0	7.2
18	*		Franzin, 1992	Italie	32	0	0.0	0.0	8.9
19	*	*	Casarotti et al., 1994 ; Destro et al., 1991	Brésil	20	0	0.0	0.0	13.9
20	*		Prentice, 1994	Danemark	1227053	278	0.0	0.0	0.0
21	*		Gospèarovic et al., 1989	Allemagne	635	2	0.0	0.3	1.1
22	*		Bachman and Spahr, 1995	Suisse	4046	14	0.2	0.3	0.6
23	*		Bachman and Spahr, 1996	Suisse	340	2	0.1	0.6	2.1
24	*		Eyles, 1992	Australie	169	1	0.0	0.6	3.3
25		*	Hartung, 2000	Allemagne	964	9	0.4	0.9	1.8
26		*	Breer and Schopfer, 1989	Suisse	317	4	0.3	1.3	3.2
27	*	*	Farber et al., 1988	Canada	445	6	0.5	1.3	2.9
28	*	*	Kwiatek et al., 1992 ; Rola et al., 1994	Pologne	134	2	0.2	1.5	5.3
29	*	*	Davidson et al., 1989	Canada	256	4	0.4	1.6	4.0
30	*		Tiwari and Aldenrath, 1990	Canada	252	4	0.4	1.6	4.0
31	*	*	Donnely et al., 1987	USA	939	15	0.9	1.6	2.6
32		*	Hartung, 2000	Allemagne	187	3	0.3	1.6	4.6
33	*		Prentice, 1996	Finlande	59	1	0.0	1.7	9.1
34	*		Fedio and Jackson, 1990	Canada	426	8	0.8	1.9	3.7
35	*	*	Lieven and Paulz, 1988	USA	100	2	0.2	2.0	7.0
36	*	*	Lovett et al., 1987	USA	100	2	0.2	2.0	7.0
37	*	*	Lovett et al., 1987	USA	100	2	0.2	2.0	7.0
38	*		Legnani et al., 1995	Italie	98	2	0.2	2.0	7.2
39	*		Rezavi-Rohani and Hedaiatinia, 1990	Iran	190	4	0.6	2.1	5.3
40	*		Fenlon and Wilson, 1989	R.-U.	560	14	1.4	2.5	4.2
41	*	*	Fenlon and Wilson, 1990	Ecosse	540	14	1.4	2.6	4.3
42		*	Steele et al., 1997	Canada	1720	47	2.0	2.7	3.6
43	*		D'Errico et al., 1990	Italie	290	8	1.2	2.8	5.4
44	*		El-Lebouby and Fayed, 1992	Egypte	236	7	1.2	3.0	6.0
45	*	*	Lund et al, 1991	USA	300	9	1.4	3.0	5.6

a : cité par [64] ; b : cité par [60]

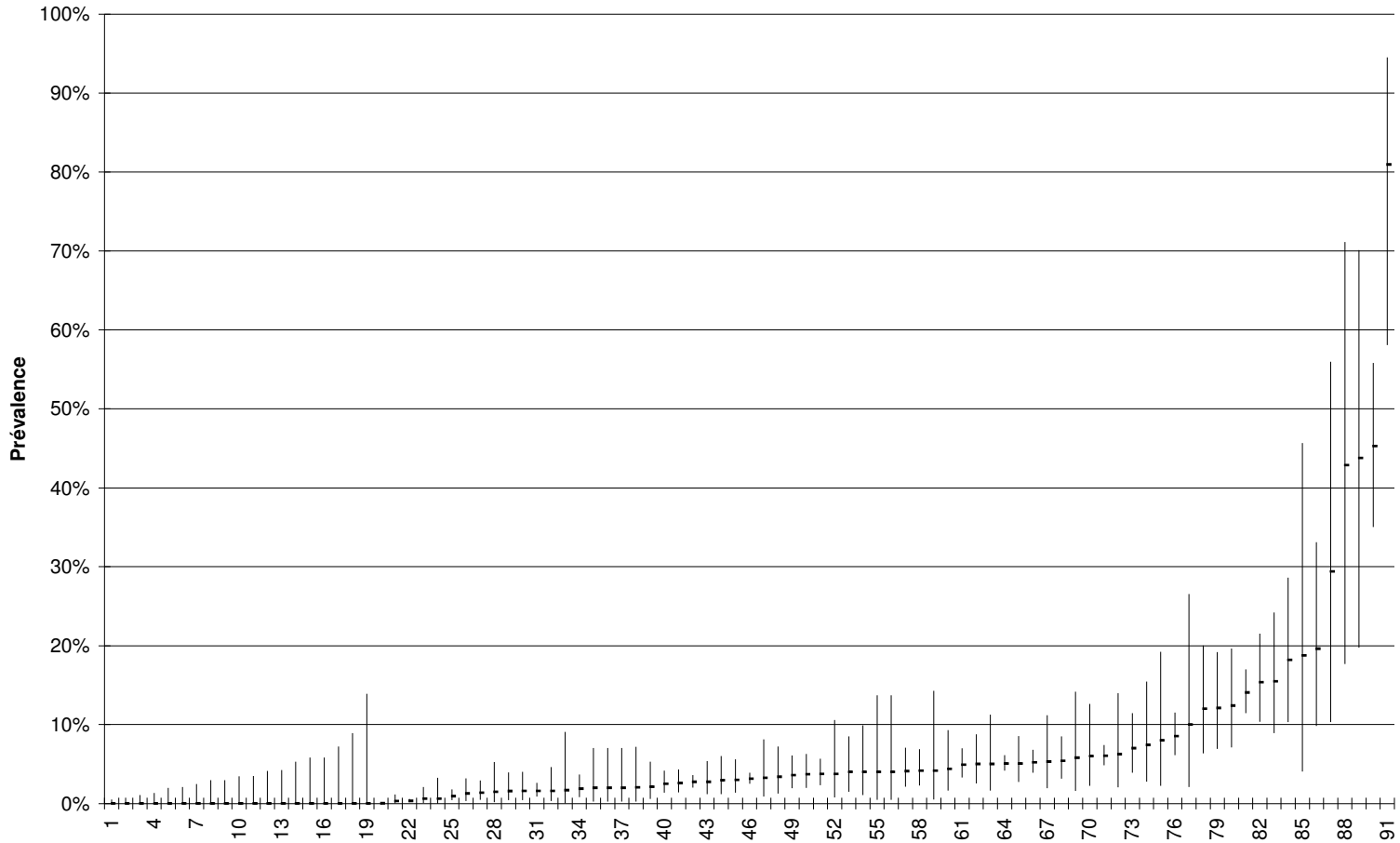
TAB. 4.14 – Données de prévalence issues de la bibliographie (début)

#	a	b	Référence	Origine	n	r	$IC_{95\%,\text{inf}}$ (%)	$\widehat{p}$ (%)	$IC_{95\%,\text{sup}}$ (%)
46	*	*	Doores and Amelang, 1990	USA	2511	79	2.5	3.1	3.9
47	*		Mickova and Koneeny, 1990	Tchécoslovaquie	123	4	0.9	3.3	8.1
48	*		Mickova, 1991	Tchécoslovaquie	177	6	1.3	3.4	7.2
49	*	*	Greenwood et al., 1991	R.-U.	361	13	1.9	3.6	6.1
50	*		Gilbert, 1990	R.-U.	350	13	2.0	3.7	6.3
51	*	*	Anonymous, 1990 ; Pitt et al., 1999	France	561	21	2.3	3.7	5.7
52	*	*	Rodler and Wérbler, 1988	Hongrie	80	3	0.8	3.8	10.6
53	*		Satio et al., 1991	Japon	150	6	1.5	4.0	8.5
54	*	*	Lovett et al., 1987	USA	100	4	1.1	4.0	9.9
55	*	*	Rodler and Wérbler, 1989	Hongrie	50	2	0.5	4.0	13.7
56	*	*	Lovett et al., 1987	USA	50	2	0.5	4.0	13.7
57	*	*	Rohrbach et al., 1992	USA	292	12	2.1	4.1	7.1
58	*	*	Glebel, 1986	France	337	14	2.3	4.2	6.9
59		*	Greenwood et al., 1991	R.-U.	48	2	0.5	4.2	14.3
60	*	*	Anonymous, 1988, Beckers et al., 1987	Pays-Bas	137	6	1.6	4.4	9.3
61	*	*	Rea et al., 1992	Irlande	589	29	3.3	4.9	7.0
62	*	*	Moura et al., 1993	Brésil	220	11	2.5	5.0	8.8
63	*	*	Lovett et al., 1987	USA	100	5	1.6	5.0	11.3
64	*		O'Donnell, 1995	R.-U.	2009	102	4.2	5.1	6.1
65	*		Prentice, 1995	Finlande	256	13	2.7	5.1	8.5
66	*	*	Wnorowski, 1991	Afrique du Sud	961	50	3.9	5.2	6.8
67	*	*	Gilmour and Harvey, 1990	Irlande	113	6	2.0	5.3	11.2
68	*	*	Slade et al., 1988	Canada	315	17	3.2	5.4	8.5
69		*	Desmeures et al., 1998	France	69	4	1.6	5.8	14.2
70	*	*	Lieven and Paulz, 1988	USA	100	6	2.2	6.0	12.6
71	*	*	Anonymous, 1989 ;Pitt et al., 1999	France	1409	85	4.8	6.0	7.4
72	*		Cheng et al., 1993	Taiwan	80	5	2.1	6.3	14.0
73		*	Kozak et al., 1996	USA	200	14	3.9	7.0	11.5
74	*		Rola et al., 1994	Pologne	81	6	2.8	7.4	15.4
75	*	*	IDF, 1989 :Kozak et al., 1996	Irlande	50	4	2.2	8.0	19.2
76		*	Anonymous, 1990	Danemark	445	38	6.1	8.5	11.5
77	*	*	El-Marrakchi et al., 1993	Maroc	30	3	2.1	10.0	26.5
78	*	*	Lovett et al., 1987	USA	100	12	6.4	12.0	20.0
79		*	Fleming et al., 1995	USA	124	15	6.9	12.1	19.2
80	*		Hayes et al., 1996	USA	121	15	7.1	12.4	19.6
81	*	*	Fenlon et al., 1995	R.-U.	640	90	11.5	14.1	17.0
82	*		Harvey and Gilmour, 1992	R.-U.	176	27	10.4	15.3	21.5
83		*	Rola et al., 1994	Pologne	97	15	8.9	15.5	24.2
84	*	*	Sharif and Tansil, 1991	Turquie	77	14	10.3	18.2	28.6
85		*	Harvey and Gilmour, 1992	Irlande	16	3	4.0	18.8	45.6
86	*	*	Glebel, 1987 ; Kozak et al., 1996	France	51	10	9.8	19.6	33.1
87		*	Harvey and Gilmour, 1992	Irlande	17	5	10.3	29.4	56.0
88		*	Harvey and Gilmour, 1992	Irlande	14	6	17.7	42.9	71.1
89		*	Harvey and Gilmour, 1992	Irlande	16	7	19.8	43.8	70.1
90		*	Dominguez-Rodriguez et al., 1985	Espagne	95	43	35.0	45.3	55.8
91	*	*	Wnorowski, 1990	Afrique du Sud	21	17	58.1	81.0	94.6

a : cité par [64] ; b : cité par [60].

TAB. 4.15 – Données de prévalence issues de la bibliographie (fin)

Illustration



Illustration

⇨ Supposons qu'il y ait une prévalence mondiale unique de *L. monocytogenes* dans le lait cru, p.

- En statistique bayésienne, p suit une distribution dite a priori car p est incertain.
- On évalue à partir des données cette incertitude (par la distribution dite a posteriori de p).

$$p \sim Bêta(1,1), \quad r_i \sim Bino(p, n_i)$$

$$p|r_i, n_i \sim Bêta\left(1+\sum_i r_i, 1+\sum_i (n_i - r_i)\right),$$

$$Application \quad p|r_i, n_i \sim Bêta(1329, 1255727),$$

$$E(p) = 0.0011, \quad \sigma(p) = 0.00003$$

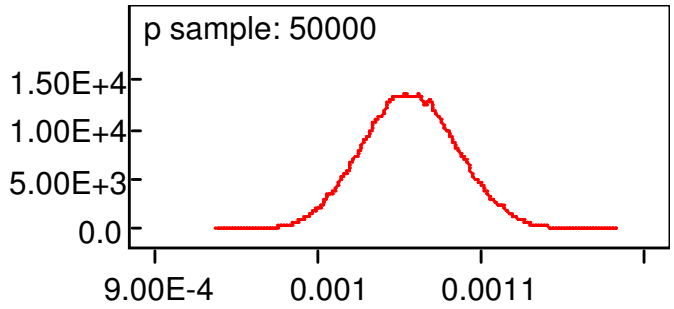


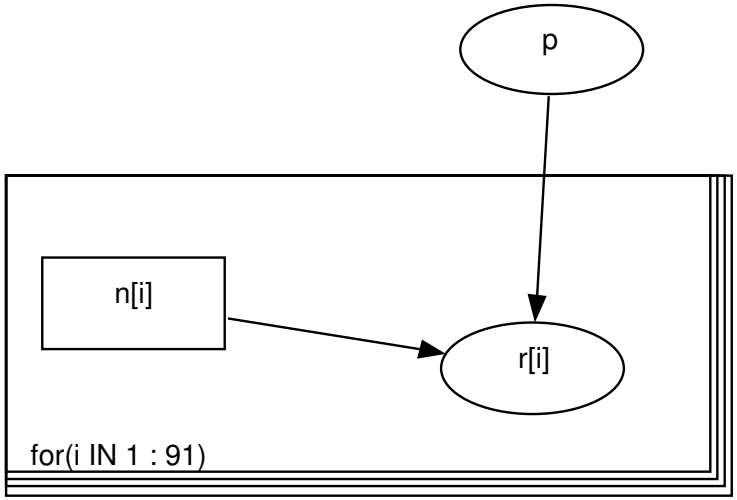
Illustration : graphe/code

```
model {

        for( i in 1 : N ) {
            r[i] ~ dbin(p,n[i])
             }

         p~ dbeta(1,1)

        }
```




```
list(N=91,
n=c(600,415,276,220,150,142,120,100,100,85,84,71,69,55,50,50,40
,32,20,1227053,635,4046,340,169,964,317,445,134,256,252,939,187
,59,426,100,100,100,98,190,560,540,1720,290,236,300,2511,123,17
7,361,350,561,80,150,100,50,50,292,337,48,137,589,220,100,2009,
256,961,113,315,69,100,1409,80,200,81,50,445,30,100,124,121,640,176,97,77,16,51,17,14,16,95,21),
r=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,278,2,14,2,1,9,4,6,2,4,4,15,3,1,8,2,2,2,2,4,14,14,47,8,7,9,79,4,6,13,13,21,3,6,4,2
,12,14,2,6,29,11,5,102,13,50,6,17,4,6,85,5,14,6,4,38,3,12,15,15,90,27,15,14,3,10,5,6,7,43,17))
```

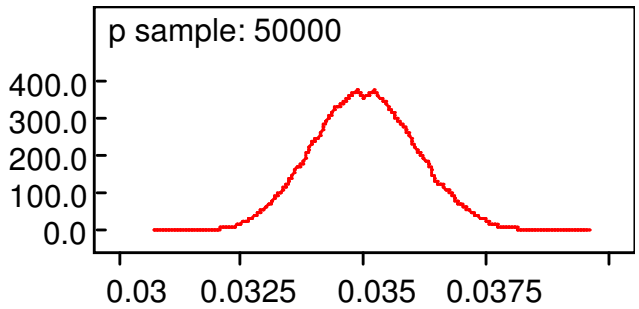
Illustration

⇨ Supposons qu'il y ait une prévalence mondiale unique de *L. monocytogenes* dans le lait cru, p.

– Retrait de l'étude 20 (n_i = 1 255 726 et r_i = 278)

$$p|r_i, n_i \sim Bêta(1051, 28922),$$

$$E(p) = 0.03506, \quad \sigma(p) = 0.001057$$



Illustration

⇨ <u>Supposons qu'il y ait une variabilité, entre les études, de la prévalence mondiale de *L. mono.* dans le lait cru.</u>

Prise en compte des variations extra-binomiales

On va supposer les études « échangeables ».

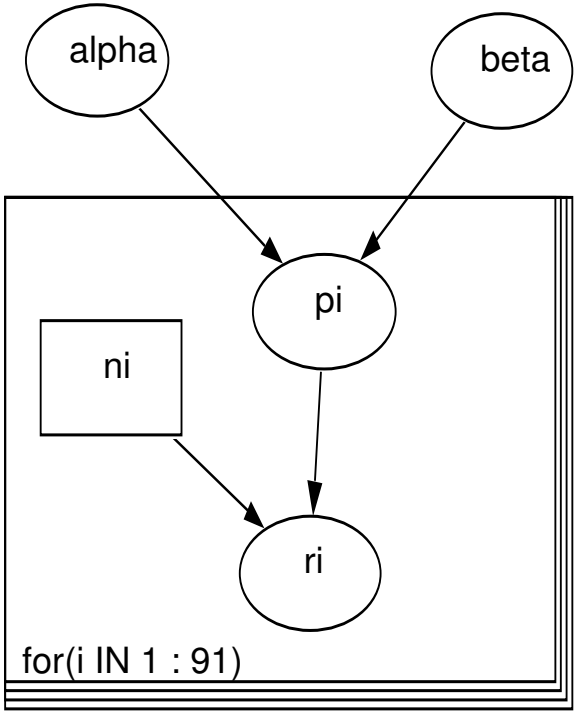
Modèle à effets aléatoires/modèle hiérarchique

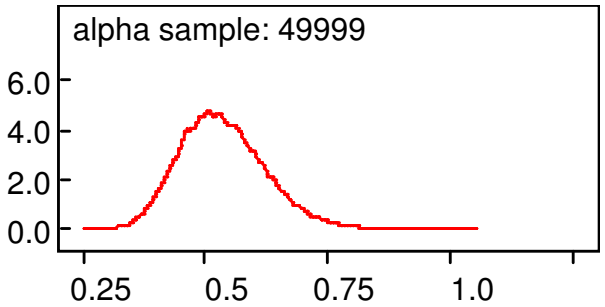
2 niveaux hiérarchiques :

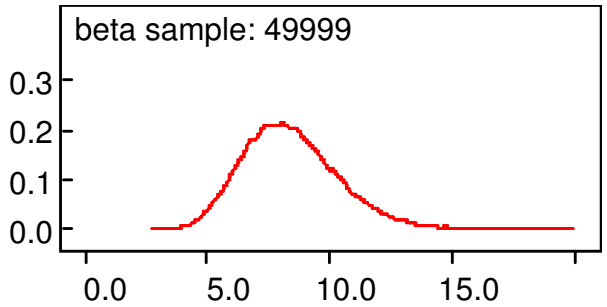
- L'étude : $(r_i | n_i, p_i) \sim Bin(n_i, p_i)$
- L'ensemble des étude : $p_i \sim Bêta(\alpha, \beta)$

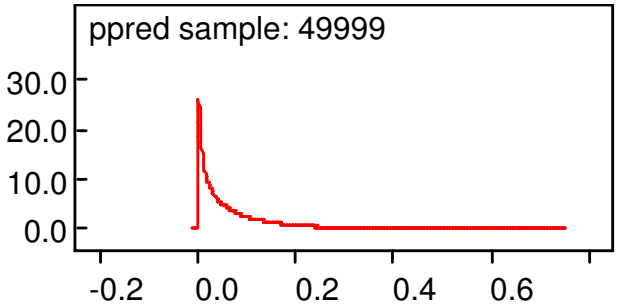
En statistique bayésienne, les hyperparamètres (α, β) étant eux-mêmes aléatoires comme tout paramètre. Ils possèdent **des distributions a priori** qui mesure l'incertitude sur la prévalence mondiale de la contamination.

Illustration









node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	0.535	0.08774	8.645E-4	0.383	0.528	0.7264	20001	49999
beta	8.399	1.947	0.01881	5.119	8.225	12.71	20001	49999
ppred	0.06154	0.07951	3.59E-4	8.242E-5	0.0311	0.2899	20001	49999

Illustration : code

```
model {

    for( i in 1 : N ) {
        p[i] ~ dbeta(alpha,beta)
        r[i] ~ dbin(p[i],n[i])
            }

    ppred ~ dbeta(alpha,beta)
    alpha ~ dnorm(0,0.001)I(0,)
    beta ~ dnorm(0,0.001)I(0,)
}
```

Illustration

Supposons qu'il y ait une variabilité de la prévalence mondiale de *L. mono.* dans le lait cru avec deux sous-populations

Même modèle que précédemment mais la prévalence suit un mélange de deux lois Bêta (études échangeables dans 2 sous-pop.)

L'attribution des différentes observations à l'une ou l'autre des distributions n'est pas connue

$$(r_i | n_i, p_i) \sim Bin(n_i, p_i)$$

$$(p_i | w_i) \sim Bêta(\alpha_{w_i}, \beta_{w_i})$$

$$w_i \sim Cat(P) \quad \text{(variable latente d'allocation} = 1 \text{ ou } 2)$$

$$P \sim Dirichlet(\mathrm{A})$$

Illustration

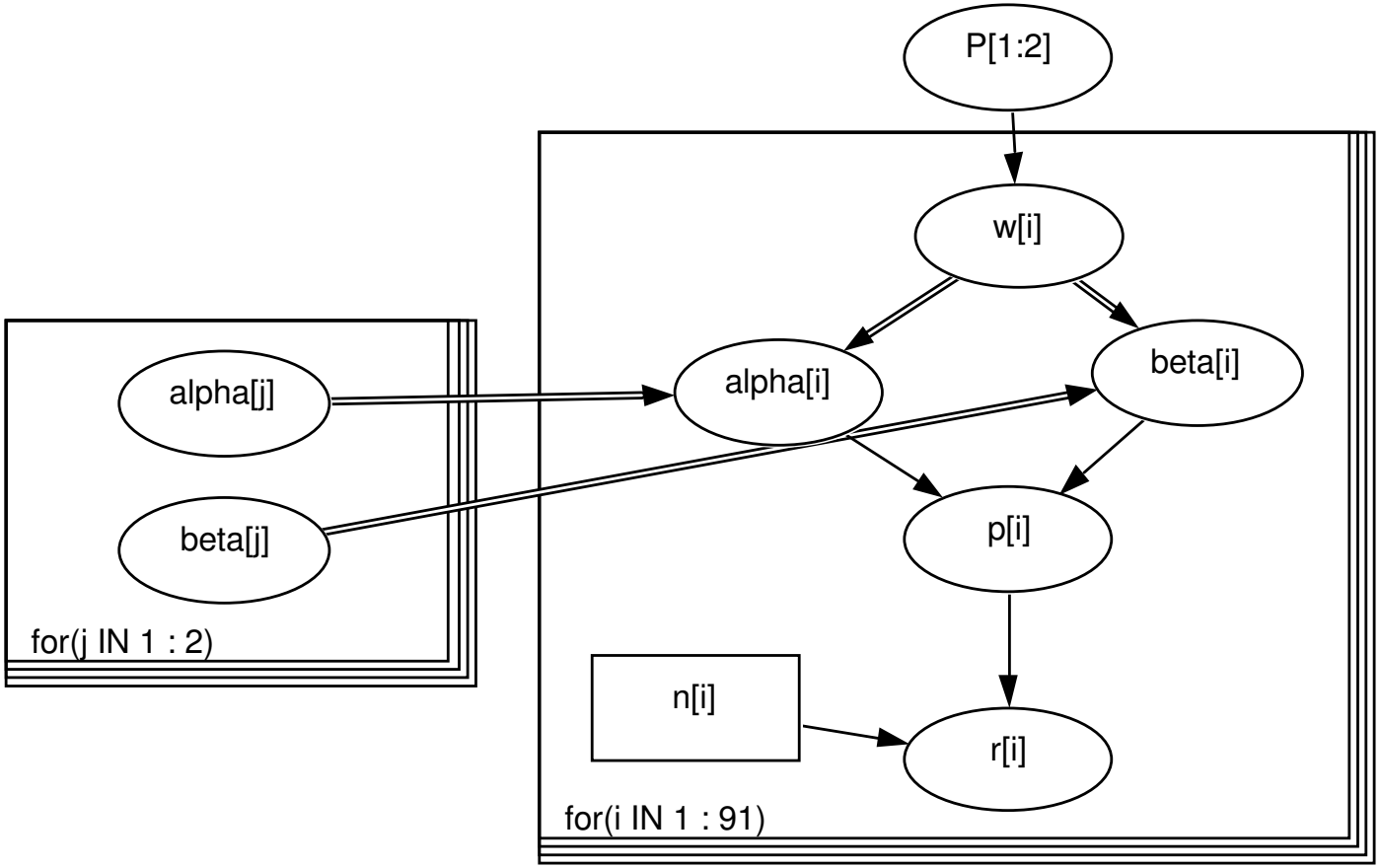


Illustration : code

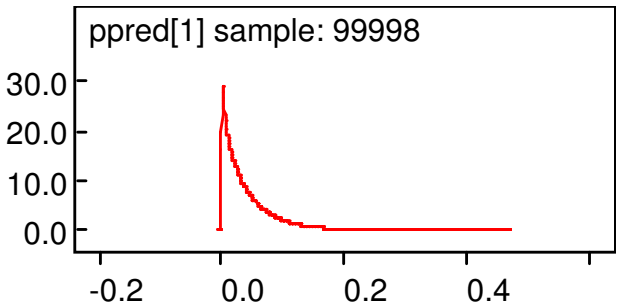
```
model;
{
   P[1:2] ~ ddirch(A[1:2])
   for( i in 1 : N ) {
      w[i] ~ dcat(P[1:2])
      p[i] ~ dbeta(alpha[w[i]],beta[w[i]])
      r[i] ~ dbin(p[i],n[i])
   }
   for( j in 1 : 2 ) {
      alpha[j] ~ dnorm(0.0,0.001)I(0,)
      beta[j] ~ dnorm(0.0,0.001)I(0,)
      ppred[j] ~ dbeta(alpha[j],beta[j])
   }
}
```

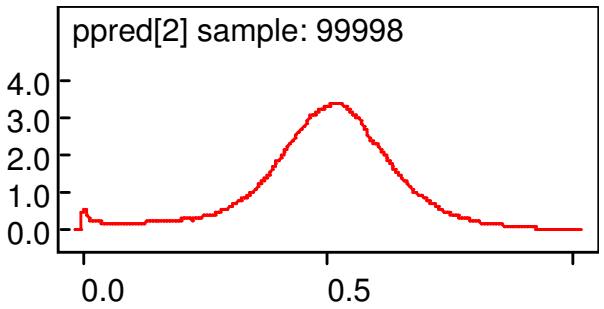
```
list(N=91,A=c(1,1),n=c(600,415,276,220,150,142,120,100,100,85,84
,71,69,55,50,50,40,32,20,1227053,635,4046,340,169,964,317,
445,134,256,252,939,187,59,426,100,100,100,98,190,560,540,1720,
290,236,300,2511,123,177,361,350,561,80,150,100,50,50,292,337,4
8,137,589,220,100,2009,256,961,113,315,69,100,1409,80,200,81,
50,445,30,100,124,121,640,176,97,77,16,51,17,14,16,95,21),
r=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,278,2,14,2,1,9,4,6,2,
4,4,15,3,1,8,2,2,2,2,4,14,14,47,8,7,9,79,4,6,13,13,21,3,6,4,2,2
,12,14,2,6,29,11,5,102,13,50,6,17,4,6,85,5,14,6,4,38,3,12,15,15
,90,27,15,14,3,10,5,6,7,43,17),w=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,1,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,2)
```

Illustration

Résultats :

$$E(ppred1) = 0.041,$$
$$\sigma(ppred1) = 0.05$$
$$E(ppred2) = 0.51,$$
$$\sigma(ppred2) = 0.14$$





Illustration

Résultats :

$$E(P_1) = 0.94$$
$$E(P_2) = 0.06$$

node	mean	sd	MC error	2.5%	median	97.5%	sample
P[1]	0.9363	0.05078	0.001494	0.8022	0.9464	0.9859	49999
P[2]	0.06373	0.05078	0.001494	0.01409	0.05358	0.1978	49999

Peu d'observations sont modélisées selon la seconde loi bêta

Illustration

Résultats :

node	mean	sd	MC error	2.5%	median	97.5%	sample
w[77]	1.016	0.1258	0.002043	1.0	1.0	1.0	59998
w[78]	1.026	0.1584	0.003247	1.0	1.0	2.0	59998
w[79]	1.025	0.1564	0.003037	1.0	1.0	2.0	59998
w[80]	1.027	0.1617	0.003274	1.0	1.0	2.0	59998
w[81]	1.039	0.1926	0.004388	1.0	1.0	2.0	59998
w[82]	1.045	0.2074	0.004762	1.0	1.0	2.0	59998
w[83]	1.044	0.2055	0.004549	1.0	1.0	2.0	59998
w[84]	1.065	0.2464	0.005581	1.0	1.0	2.0	59998
w[85]	1.068	0.252	0.004844	1.0	1.0	2.0	59998
w[86]	1.083	0.2761	0.005529	1.0	1.0	2.0	59998
w[87]	1.356	0.4788	0.01077	1.0	1.0	2.0	59998
w[88]	1.79	0.4074	0.008035	1.0	2.0	2.0	59998
w[89]	1.85	0.3575	0.007052	1.0	2.0	2.0	59998
w[90]	1.953	0.2125	0.003504	1.0	2.0	2.0	59998

Seules les publications 78 à 91 sont classés dans la seconde catégorie dans plus de 2.5% des itérations. Seules les observations 88 à 91 sont classées dans la seconde catégorie dans plus de 50% des itérations.

Conclusions

- Nous avons retenu uniquement la première loi bêta pour représenter la situation française

R. Pouillot, I. Albert et J.B. Denis. Caractérisation quantitative de la variabilité de l'écologie de *Listeria monocytogenes* par méta-analyse statistique, en vue de l'analyse du risque de listériose lié à la consommation de lait cru. Rapport Technique, INSERM/CNRS/MIRE, 128 pages, 2001

Conclusions

- Extensions à un nombre inconnu de classe -> Reversible Jump (Richardson and Green, On bayesian analysis of mixtures with an unknown number of components, J.R.S.S.B, 1997, 59,731-292.)

  A suite of reversible jump MCMC components for WinBUGS http://www.winbugs-development.org.uk/

  ou comparaison de modèle avec un nombre de classe fixé (Facteur de Bayes, **PAS** DIC)

Conclusions

– Extensions « plus de deux catégories » -> loi multinomiale et distribution de Dirichlet sur les p_i dont les paramètres alpha appartiennent à différentes classes de mélange

Conclusions

- Mélange sur lois a priori : possible en Winbugs si mélange de même type de loi
  - Exemple :

```
model{
        for (i in 1:N){
                r[i]~dbin(p,n[i])
                }
        w~dcat(Pi[])
        p ~dbeta(alpha[w],beta[w])
        alpha[1]<-8
        alpha[2]<-1
        beta[1]<-28
        beta[2]<-1
        }
```

Références

- McLachlan et Peel (2000). *Finite mixture Models.* John Wiley & Sons, Inc.
- Marin, Jean-Michel and Mengersen, Kerrie and Robert, Christian (2005) *Bayesian modelling and inference on mixtures of distributions*, in Dey, D. and Rao, C.R., Eds. *Handbook of Statistics Volume 25*. Elsevier Sciences. http://www.ceremade.dauphine.fr/~xian/publications.html
- Richardson and Green, On bayesian analysis of mixtures with an unknown number of components, J.R.S.S.B, 1997, 59,731-292.
- http://www.winbugs-development.org.uk/
- Exemple Winbugs : Eyes: Normal Mixture Model