



Mathématiques et sciences humaines

Mathematics and social sciences

179 | Automne 2007

Varia

Inférence statistique, échangeabilité et approche mutiniveau

Statistical inference, exchangeability and multilevel approach

Daniel Courgeau



Édition électronique

URL : <http://journals.openedition.org/msh/6793>

DOI : 10.4000/msh.6793

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 1 septembre 2007

Pagination : 5-19

ISSN : 0987-6936

Référence électronique

Daniel Courgeau, « Inférence statistique, échangeabilité et approche mutiniveau », *Mathématiques et sciences humaines* [En ligne], 179 | Automne 2007, mis en ligne le 21 décembre 2007, consulté le 19 avril 2019. URL : <http://journals.openedition.org/msh/6793> ; DOI : 10.4000/msh.6793

INFÉR[^] NC[^] STATISTIQU[^], ÉCHANG[^] ABILITÉ [^] T APPROCH[^] MULTINIV[^] AU

Daniel COURG[^] AU¹

RÉSUMÉ – *Cet article examine les problèmes posés par une inférence statistique en sciences sociales. Pour pouvoir passer d'une population à un nouvel individu sur qui on veut effectuer l'inférence, il est utile d'utiliser le concept d'échangeabilité, mis en évidence par de Finetti. Cela est montré pour un modèle logit simple avec deux groupes et pour un modèle multiniveau où l'on observe un plus grand nombre de groupes. Dans ces deux cas le paradoxe de Simpson vient jouer et peut fournir des résultats inverses selon que l'on travaille sur les données agrégées ou décomposées par groupe. Le concept d'échangeabilité permet, en utilisant les probabilités appropriées, de résoudre un certain nombre de problèmes posé par cette inférence. Mais il est nécessaire d'utiliser à la fois les données sur la population et des informations obtenues par d'autres moyens sur le sujet étudié, pour pouvoir réaliser cette inférence.*

MOTS-CLÉS – Approche multiEiveau, ÉchaEgeabilité, IEféreEce statistique, Paradoxe de SimpsOE, Probabilité épistémique, Probabilité objectiviste, ScieEces sociales

SUMMARY – Statistical iEféreEce, exchaEgeability aEd multilevel approach
This paper is concerned with the problems of statistical inference in social sciences. In order to pass from a population to a new individual by inference, de Finetti's idea of exchangeability is useful. This is shown in a simple logit model and in a multilevel model, with a larger number of observed groups, where the Simpson's paradox arises depending on whether you work on aggregate data or on group data. The concept of exchangeability permits, by using the appropriate probabilities, to solve a number of problems arising in statistical inference. But it is necessary to use not only the data but also information inferred by other means in order to make the final inference.

KEY WORDS – Epistemic probabilities, ExchaEgeability, Multilevel approach, Objective probability, SimpsOE paradox, Social scieEces, Statistical iEféreEce

1. INTRODUCTION

Cet article reprend et étend les discussions passionnantes que nous avons eues avec Henry Rouanet et ses collègues au sujet d'un article publié dans cette revue sur *Régression et analyse géométrique des données* [Rouanet *et al.*, 2002]. Nous utiliserons en particulier le même exemple qu'eux sur les résultats d'un examen dans une ville, mais bien sûr notre propos est différent du leur : ils proposaient de conjuguer les procédures des deux analyses ; nous proposons d'examiner les conditions sous lesquelles on peut effectuer une inférence statistique à partir de ces données.

¹ Institut National d'Études Démographiques (IN[^] D), 133 boulevard Davout Paris 75980 Cedex 20, daniel.courageau@wanadoo.fr

Nous considérons ici l'inférence statistique comme un processus qui permet de passer de données concernant un ensemble d'unités à des assertions au sujet d'une nouvelle unité [Johnson, 1932 ; Lindley et Novick, 1961 ; Rouanet *et al.*, 1998]. En démographie et plus généralement en sciences humaines, ces unités ont été, dès les origines de ces disciplines au XVII^e siècle, les individus, considérés ici en grand nombre, dont on étudie le comportement. Mais plus récemment l'approche multiniveau est venue compléter cette hypothèse et a mis en évidence une pluralité de niveaux qu'il paraît nécessaire d'étudier simultanément en montrant comment ils vont s'articuler entre eux, sans en privilégier aucun [Greenland, 2000 ; Courgeau, 2004(b), 2007]. Nous allons essayer ici de montrer les problèmes que cette approche va poser pour l'inférence statistique et essayer d'en résoudre certains.

Pour ce faire nous aurons d'abord à examiner un paradoxe classique en statistique, prolongé ensuite sous l'approche multiniveau : le paradoxe de Simpson. Nous verrons également comment des prédictions sont possibles à partir de ces exemples et en utilisant le concept d'échangeabilité introduit par de Finetti [1974], ainsi que d'autres hypothèses sur le comportement de la population étudiée.

2. LE PARADOXE DE SIMPSON

Il nous semble en premier lieu utile de revenir sur l'histoire de ce paradoxe. On a l'habitude de l'appeler le « paradoxe de Simpson », du nom de l'auteur qui l'a traité de façon détaillée en 1951, même si Cohen et Nagel en parlaient déjà en 1934 et si Chung l'avait déjà exposé en 1942, en utilisant les notions intéressantes d'événements favorables, indifférents ou défavorables, montrant qu'elles conduisaient à diverses conclusions dont le paradoxe de Simpson et la non-transitivité de ces notions. C'est aussi ce que Rouanet *et al.* [2002] appellent le « paradoxe des lycées ». Également Lindley et Novick [1981], dans un article sur le rôle de l'échangeabilité en inférence statistique, reprennent ce paradoxe avec le concept de sous-population chez Fisher ou le concept d'unités échangeables en bayésien. Je m'inspirerai de ces réflexions pour proposer des prédictions en accord avec les données et indiquer les raisons pour lesquelles il est délicat de faire des prédictions sans de nouvelles hypothèses.

Présentons le à partir d'un exemple simple donné par Rouanet *et al.* [2002] qui avait déjà été discuté dans cette revue pour analyser les effets de structure intervenant dans ces données.

On suppose que les résultats à un examen des élèves d'une ville se présentent d'abord sous la forme du Tableau 1.

Genre	Réussites	Échecs	Total
Garçons	24	36	60
Filles	36	24	60

TABLEAU 1. Lycées réunis

Il en résulte que les taux de succès des garçons (40 %) est très largement inférieur à celui des filles (60 %), montrant que les filles réussissent mieux que les garçons. Cependant la ville comportant deux lycées on peut également observer pour chaque lycée les effectifs de garçons et de filles avec les succès ou les échecs à l'examen (Tableau 2).

Genre	Réussites	Échecs	Total
	Lycée 0		
Garçons	15	35	50
Filles	1	9	10
Lycée 1			
Garçons	9	1	10
Filles	35	15	50

TABLEAU 2. Lycées séparés

On constate cette fois-ci un résultat a priori surprenant : dans chaque lycée le pourcentage de réussites des garçons est maintenant supérieur à celui des filles. Dans le premier lycée le taux de succès des garçons est de 30 % comparé à 10 % pour les filles ; dans le second lycée le taux de succès des garçons est de 90 % comparé à 70 % pour les filles. On dit dans ce cas que le lycée est un facteur de confusion, lorsque l'on étudie l'effet du sexe sur la réussite à l'épreuve. Pour lever la contradiction apparente entre ces résultats il est donc nécessaire d'effectuer une analyse plus approfondie de ces données.

En premier lieu nous pouvons nous contenter de travailler pour le moment sur l'*approche objectiviste* ou *fréquentiste* des probabilités [Courgeau, 2004(a)] : cela est justifié ici par le fait que nous observons des effectifs importants, comme nous l'avons dit dans l'introduction. La fréquence d'un événement A , $P(A)$, est le rapport du nombre de cas favorables à cet événement au nombre total de cas observés. De plus, plutôt que d'appliquer à ces données un modèle de régression linéaire, comme dans l'article de Rouanet *et al.*, qui ne convient pas pour des variables dépendantes dichotomiques (cf. par exemple [Gourieroux, 1989, p. 9-11]), il est préférable dans ce cas d'utiliser un modèle logit, probit ou log-log complémentaire, qui évitent tous les problèmes que poseraient l'utilisation d'une régression linéaire dans ce cas. Je prendrai ici le cas du modèle logit, mais les autres modèles conduisent à des résultats très proches, sinon identiques.

Ce modèle s'écrit, en notant de façon simple les trois caractéristiques binaires observées (R la réussite à l'examen, 1 pour la réussite et 0 pour l'échec, S le sexe, 1 pour les filles et 0 pour les garçons, L le lycée, 0 ou 1) :

$$P(R = 1 | S, L) = \left\{ 1 + \exp \left[-(a_0 + a_1 S + a_2 L) \right] \right\}^{-1}$$

soit encore, en notant pour simplifier $P = P(R = 1 | S, L)$:

$$\log \left(\frac{P}{1 - P} \right) = a_0 + a_1 S + a_2 L$$

Avec les données précédentes on estime ces paramètres :

$$a_0 = -0,85$$

$$a_1 = -1,35$$

$$a_2 = 3,04$$

qui permettent de reconstruire l'ensemble des probabilités issues du Tableau 2 :

$$\begin{aligned}
P(R = 1 | S = 0, L = 0) &= \{1 + \exp(-a_0)\}^{-1} = 0,3 \\
P(R = 1 | S = 1, L = 0) &= \{1 + \exp(-a_0 - a_1)\}^{-1} = 0,1 \\
P(R = 1 | S = 0, L = 1) &= \{1 + \exp(-a_0 - a_2)\}^{-1} = 0,9 \\
P(R = 1 | S = 1, L = 1) &= \{1 + \exp(-a_0 - a_1 - a_2)\}^{-1} = 0,7
\end{aligned}$$

Ces résultats montrent clairement deux effets. En premier lieu l'effet du sexe est identique pour les deux lycées : être une fille réduit la probabilité de 0,2 points par rapport à celle des garçons de façon identique dans les deux lycées. Cet effet du sexe semble donc être parfaitement indépendant du lycée dans lequel se trouve l'élève. En second lieu l'effet de la qualité de l'enseignement dans chaque lycée est très clair : le lycée 0 a des résultats inférieurs à ceux du lycée 1 de 0,6 points, quel que soit le sexe de l'élève. Cet effet du lycée semble donc être parfaitement indépendant du sexe de l'enfant et ne dépendre que de l'enseignement qui y est donné.

On a également la possibilité d'estimer la matrice des variances-covariances de ces divers paramètres [McCullag, Nelder, 1983].

Cet ensemble permet enfin de calculer les probabilités des événements donnés dans le tableau 1. Ainsi par exemple pour les garçons :

$$\begin{aligned}
P(R = 1 | S = 0) &= P(R = 1 | S = 0, L = 0)P(L = 0 | S = 0) + P(R = 1 | S = 0, L = 1)P(L = 1 | S = 0) \\
&= 0,3 \frac{5}{6} + 0,9 \frac{1}{6} = 0,4
\end{aligned}$$

et de façon semblable pour les filles :

$$P(R = 1 | S = 1) = 0,1 \frac{1}{6} + 0,7 \frac{5}{6} = 0,6$$

Mais du fait que le lycée est un facteur de confusion ces résultats sont de peu d'intérêt, car ils mélangent des résultats obtenus dans des lycées différents qui peuvent conduire, comme ici, au paradoxe de Simpson.

3. INFÉRENCES À PARTIR DE CES RÉSULTATS

Nous allons maintenant, pour faire des prédictions, utiliser l'*approche épistémique* des probabilités, car l'inférence va porter ici sur un seul nouveau cas : la probabilité épistémique² d'un événement A , celui que peut connaître le nouveau cas, sachant qu'un autre événement B s'est produit, ceux qu'ont connus les précédents élèves : $p(A | B)$, sera le degré de crédibilité d'un événement, compte tenu de ce que l'on sait des autres circonstances.

Également, avant d'essayer de faire ces prédictions, il nous faut définir ce que l'on appelle l'échangeabilité, notion introduite par de Finetti [1974], qui va les permettre. Soient X et Y deux variables aléatoires mesurées sur un certain nombre d'unités. Pour la

² Nous utilisons ici le terme épistémique dans le sens proposé par Hacking [2002] : cette probabilité concerne l'évaluation des degrés auxquels il est raisonnable de croire en la vérité de propositions.

$i^{\text{ème}}$ unité, ces variables ont les valeurs X_i et Y_i qui sont inconnues du chercheur. Aussitôt qu'il les observe, ces variables deviennent des nombres x_i et y_i qui ne sont plus des variables aléatoires.

Un nombre n d'unités est dit échangeable pour X si la distribution de probabilité conjointe $p(X_1, \dots, X_n)$ est invariante sous quelque permutation que ce soit des unités. Une unité supplémentaire est échangeable pour X si l'ensemble des $(n + 1)$ unités sont aussi échangeables. Ainsi on pourrait juger échangeables pour la réussite à l'examen les 60 garçons du tableau 1 : il s'agit d'un jugement personnel, qui nous l'avons vu ne tient pas, car cette réussite dépend aussi du lycée.

Un nombre n d'unités est dit échangeable pour X , sachant que $Y = y$, si la distribution conditionnelle jointe, $p(X_1, \dots, X_n | Y_i = y, \forall i)$, est invariante sous quelque permutation que ce soit des unités. Ainsi on peut juger échangeables pour la réussite au test les 50 garçons du lycée 0, conditionnellement à leur établissement : il s'agit à nouveau d'un jugement personnel, qui nous l'avons vu tient mieux que le précédent.

Il est dès lors possible de montrer que, si n est important, alors la probabilité épistémique que l'événement A arrive dans le $(n + 1)^{\text{ème}}$ cas est simplement la fréquence avec laquelle l'événement s'est produit dans les n cas précédents. Une preuve rigoureuse de ce résultat, bien qu'il paraisse assez intuitif, requiert en fait le théorème de de Finetti [1974] sur la structure des séquences échangeables. Ainsi, on peut écrire plus simplement ce résultat sous la forme $p(A) = P(A)$, où $p(A)$ est la probabilité épistémique de l'événement pour le $(n + 1)^{\text{ème}}$ individu et $P(A)$ est la fréquence estimée pour les n individus précédents. De même, si X est conditionnellement échangeable étant donné Y alors on peut dire que : $p(A | Y = y) = P(A | Y = y)$ où, par exemple, l'événement A est la réussite à l'examen d'un garçon et y le lycée.

Pour pouvoir prédire les chances de réussite lors d'une épreuve postérieure d'un nouvel élève, il nous faudra dès lors faire des hypothèses sur l'échangeabilité de certaines caractéristiques.

Par exemple, si l'on cherche à prédire les chances d'un nouveau garçon, élève du lycée 0, on peut faire sans trop de problèmes l'hypothèse que l'ensemble des garçons, élèves du lycée 0, est échangeable pour les chances de réussite, et que ce nouvel élève l'est également avec eux. Dans ce cas on pourra prédire que ce garçon a trois chances sur dix de réussir son test :

$$p(R = 1 | S = 0, L = 0) = P(R = 1 | S = 0, L = 0) = 0,3$$

Et une fille aura une chance sur dix.

Comme autre exemple, si l'on cherche à prédire la probabilité de réussite d'un nouvel élève dont on ignore le lycée de provenance, mais dont on connaît le sexe (par exemple, masculin, soit $S = 0$), on peut écrire :

$$p(R = 1 | S = 0) = p(R = 1 | S = 0, L = 0)p(L = 0 | S = 0) + p(R = 1 | S = 0, L = 1)p(L = 1 | S = 0)$$

D'où, compte tenu du cas précédent :

$$p(R = 1 | S = 0) = 0,3p(L = 0 | S = 0) + 0,9p(L = 1 | S = 0)$$

Mais dans ce cas, contrairement à ce que l'on observait dans l'échantillon initial, on ne connaît plus les probabilités épistémiques de tirer le lycée 0 ou 1, sachant que l'on a tiré un garçon. On voit donc que l'échangeabilité pour le lycée, le sexe étant donné, n'est pas plus vérifiée pour ce nouvel élève qu'elle ne l'était pour les 60 autres. Cependant dans la mesure où l'on a vu que l'effet du lycée était essentiellement dû à son enseignement quelque soit sexe de l'élève, donc que le sexe du nouvel élève tiré et le lycée sont ici indépendants, on peut estimer que le nouvel individu tiré est échangeable selon le lycée avec le reste de la population, où la probabilité de tirer un lycée est égale à $\frac{60}{120}$. Alors $p(L | S) = 0,5$ et :

$$p(R = 1 | S = 0) = 0,3 * 0,5 + 0,9 * 0,5 = 0,6$$

On voit que ce nouvel élève aura plus de chances de réussir au baccalauréat qu'une fille, pour laquelle cette probabilité sera 0,4. Ce résultat montre que l'utilisation du tableau n° 1 est à déconseiller dans ce cas, pour prédire ces chances, car il donne un résultat inverse, lié à la non-échangeabilité des observations. Mais bien entendu cette prévision n'est valable que sous l'hypothèse faite ici, que l'effet du lycée est dû à son enseignement, qu'il faudrait vérifier avec plus de précision. Nous verrons comment cela est possible en observant un plus grand nombre de lycées.

Ce problème a également été traité par Lindley et Novick [1981] en considérant des variables aléatoires plus générales que des variables binaires.

4. ANALYSE MULTINIVEAU

Voyons maintenant comment généraliser ces résultats à un nombre élevé n de lycées. Une première solution est d'effectuer des analyses logit séparées sur chacun des lycées, i , avec comme seule variable prédictrice le sexe :

$$P_i(R = 1 | S) = \left\{ 1 + \exp \left[- (a_{0_i} + a_{1_i} S) \right] \right\}^{-1}$$

Dans ce cas certaines écoles pourront avoir des effectifs très faibles et l'écart-type des paramètres a_{0_i} et a_{1_i} pourra être très important, rendant les résultats non-interprétables.

Si l'on veut un résultat plus général, valable pour une population plus large de lycées (l'ensemble des lycées français par exemple), alors on peut considérer les lycées choisis pour nous donner une information plus générale sur l'ensemble des lycées français. De la même façon que l'on tire un échantillon d'individus pour en déduire des estimations sur la population, on peut tirer un échantillon de lycées pour en déduire des informations sur l'ensemble des lycées. On utilise ici, comme plus haut, l'approche *objectiviste* ou *fréquentiste* des probabilités. Dans ce cas on peut considérer les paramètres a_{0_i} et a_{1_i} comme des variables aléatoires que l'on peut écrire :

$$a_{0_i} = \alpha_0 + u_{0_i} \quad \text{et} \quad a_{1_i} = \alpha_1 + u_{1_i}$$

où u_{0_i} et u_{1_i} sont des variables aléatoires de moyenne nulle et de variances égales à :

$$\text{var}(u_{0_i}) = \sigma_{u_0}^2 \quad \text{var}(u_{1_i}) = \sigma_{u_1}^2 \quad \text{et} \quad \text{cov}(u_{0_i}, u_{1_i}) = \sigma_{u_0 u_1}$$

On a alors un modèle logit multiniveau :

$$P(R = 1 | S) = \left\{ 1 + \exp \left[- \left(\alpha_0 + u_{0_i} + \langle \alpha_1 + u_{1_i} \rangle S \right) \right] \right\}^{-1}$$

On sait estimer les paramètres et la matrice des variances-covariances de ce modèle, ainsi que les valeurs pour chaque lycée des u_{0_i} et u_{1_i} [Goldstein, 2003].

5. EXEMPLE D'ANALYSE MULTINIVEAU

Prenons un exemple qui prolonge celui que nous avons étudié plus haut, en favorisant maintenant les filles. Soient 10 lycées dans lesquels on obtient d'abord les résultats globaux donnés dans le Tableau 3.

Genre	Réussites	Échecs	Total
Garçons	915	1085	2000
Filles	685	1315	2000

TABLEAU 3. Résultats globaux obtenus dans 10 lycées

On voit que pour l'ensemble des 10 lycées le taux de succès des filles est inférieur à celui des garçons (34,25 % contre 45,75 %), soit de plus de 10 %.

Portons maintenant les résultats obtenus dans chaque lycée dans le Tableau 4.

Lycée	Garçons			Filles			Total
	Réussite	Échec	Total	Réussite	Échec	Total	
1	3	17	20	76	304	380	400
2	12	48	60	85	255	340	400
3	25	75	100	90	210	300	400
4	42	98	140	91	169	260	400
5	63	117	180	88	132	220	400
6	88	132	220	81	99	180	400
7	117	143	260	70	70	140	400
8	150	150	300	55	45	100	400
9	187	153	340	36	24	60	400
10	228	152	380	13	7	20	400
Ensemble	915	1085	2000	685	1315	2000	4000

TABLEAU 4. Résultats obtenus par lycée

On voit facilement que dans chaque lycée le taux de succès des filles est toujours supérieur à celui des garçons de 5 % : le paradoxe de Simpson est donc toujours vérifié pour ces lycées.

Réalisons une analyse multiniveau sur ces lycées, avec comme variable prédictrice le sexe (0 pour les garçons, 1 pour les filles). On constate d'abord que l'introduction de la variable « Sexe » dans les aléatoires donne une variance nulle ($\sigma_{u_1}^2 = 0$) : nous la supprimerons donc de ces aléatoires par la suite. On obtient alors les résultats du Tableau 5.

Paramètres	Estimateurs (écart-type)		
	Modèle 1	Modèle 2	Modèle 3
<i>Fixes :</i>			
Constante (α_0)	-0,5378 (0,2031)	-0,5378 (0,2030)	-1,6410 (0,1102)
Sexe (α_1)	0,1913 (0,0807)	0,1913 (0,0807)	0,2207 (0,0816)
Pourcentage de garçons (α_2)			2,1860 (0,1466)
<i>Aléatoires :</i>			
$\sigma_{u_0}^2$	0,3844 (0,1775)	0,3841 (0,177)	0
$\sigma_{e_0}^2$	1	0,999 (0,0224)	1

TABLEAU 5. Résultats de modèles multiniveaux sur les 10 lycées

Le modèle 1 montre bien que les garçons ont un plus faible taux de succès que les filles, quelque soit le lycée où ils se trouvent (36,87 % contre 41,42 %). En revanche on a un aléa au niveau lycée, qui montre des résultats différents selon les lycées. On peut calculer les estimateurs de ces aléas u_{0j} pour chaque lycée et comparer les taux de succès estimés avec ceux observés (Tableau 6).

Lycée	u_{0j}	Taux de succès estimé		Taux de succès observé	
		Garçons	Filles	Garçons	Filles
1	-1,0196	17,4	20,3	15	20
2	-0,7509	21,6	25,0	20	25
3	-0,5088	26,0	29,8	25	30
4	-0,2845	30,5	34,7	30	35
5	-0,0725	35,2	39,6	35	40
6	0,1313	40,0	44,6	40	45
7	0,3302	44,8	49,6	45	50
8	0,5267	49,7	54,5	50	55
9	0,7240	54,6	59,3	55	60
10	0,9239	59,5	64,0	60	65

TABLEAU 6. Comparaison des taux de succès observés et estimés dans les 10 lycées, par le modèle 1

La comparaison est tout à fait satisfaisante, montrant un taux de succès des filles supérieur de 5 % approximativement au taux de succès des garçons dans chaque lycée.

Le modèle 2 lève la condition d'une variance binomiale au niveau individuel (égale à 1) et montre que cette condition est en fait parfaitement vérifiée, la variance étant égale à 0,999 avec un écart type de 0,02237. Les autres paramètres estimés sont d'ailleurs inchangés. Nous pouvons donc considérer par la suite le modèle 1 comme parfaitement vérifié.

Le modèle 3 fait intervenir la variable contextuelle, la proportion de garçons dans chaque lycée. On voit qu'elle explique parfaitement l'effet aléatoire au niveau lycée, sa variance tombant à une valeur égale à 0. Il suffit donc d'un modèle logit simple pour expliquer l'ensemble des comportements observés. On obtient pour chaque lycée et pour chaque sexe les estimations suivantes des taux de succès :

Pourcentage de garçons	Taux de succès estimé des garçons	Taux de succès estimé des filles
5 %	17,8	21,2
15 %	21,2	25,1
25 %	25,1	29,4
35 %	29,4	34,2
45 %	34,1	39,3
55 %	39,2	44,6
65 %	44,5	50,0
75 %	50,0	55,5
85 %	55,5	60,8
95 %	60,7	65,8

TABLEAU 7. Comparaison des taux de succès estimés des garçons et des filles par le modèle 3

On voit qu'à nouveau ces estimations sont très correctes. Cependant, bien que les taux de succès estimés par le modèle 1 et le modèle 3 soient très proches des taux de succès observés, les hypothèses sous-jacentes à ces estimations sont en fait très différentes. Voyons les plus en détail.

Sous le modèle 1, *purement multiniveau*, l'hypothèse sous-jacente est que les taux de succès sont liés au lycée et à son personnel enseignant, car il ne comporte aucune variable contextuelle permettant de faire intervenir des caractéristiques propres aux élèves. L'aléa au niveau lycée est important et vient assurer le raisonnement précédent.

Sous le modèle 3, *multiniveau contextuel*, on fait intervenir une variable contextuelle propre aux élèves de chaque lycée : le pourcentage des élève de sexe masculin. Dans ce cas il apparaît que ce pourcentage va jouer un rôle fondamental et permet par lui seul de distinguer les lycées entre eux. En revanche, dans ce modèle l'aléa au niveau lycée devient nul. Cela montre que l'explication des différences est entièrement due aux proportions des deux sexes et non plus au personnel enseignant, comme l'analyse précédente semblait le montrer.

Bien entendu ce modèle est un peu artificiel et conduit à ce résultat extrême. Dans les cas réels observés la solution est moins tranchée, mais conduit cependant à des situations semblables. Ainsi dans le cas de la migration des agriculteurs norvégiens comparés aux autres professions, traité en détail dans un article [Courgeau, 2002] et repris dans l'ouvrage plus général [Courgeau, 2004(b), 2007] un premier modèle multiniveau pur conduit à distinguer les comportements selon la région d'origine, alors qu'un second modèle multiniveau contextuel permet de nuancer ce résultat en montrant l'effet du pourcentage d'agriculteurs présents dans chaque région, qui réduit fortement l'aléa régional, sans cependant l'annuler complètement.

Lorsque nous essayerons de faire des prédictions il nous faudra tenir compte de ces nouveaux éléments.

6. ANALYSE MULTINIVEAU BAYÉSIENNE

Nous avons la possibilité d'estimer des modèles bayésiens en utilisant les méthodes MCMC (Markov Chain Monte Carlo) et Bootstrap, avec le logiciel Mlwin [Goldstein, 2003 ; Browne, 2004]. La méthode MCMC tire son nom de l'idée que, pour produire des approximations acceptables d'intégrales et d'autres fonctions dépendant d'une loi d'intérêt, il suffit de générer une chaîne de Markov de loi limite, la loi d'intérêt [Robert, 2006]. La méthode Bootstrap consiste à construire une série d'échantillons semblables à l'échantillon observé et à utiliser cette série pour estimer les paramètres d'intérêt de cette loi. Nous avons supposé ici qu'un modèle binomial était bien adapté, comme nous l'avons montré en non-bayésien. Il est facile de le vérifier ici, dans le cas bayésien.

Pour l'option MCMC, nous utilisons la méthode d'échantillonnage de Metropolis Hastings qui est applicable aux données binaires, comme c'est le cas ici. Cet algorithme permet la construction d'un mécanisme de génération pour produire les chaînes de Markov convergeant vers la loi d'intérêt [Robert, 2006]. Nous avons également deux possibilités pour la distribution *a priori* : soit une distribution gamma, soit une distribution uniforme. Nous utiliserons ces deux possibilités. Les estimations ont été faites avec 20.000 itérations.

Pour l'option Bootstrap, nous avons choisi d'effectuer 500 tirages dans 10 ensembles : un premier essai sur 100 tirages dans 5 ensembles nous donnait des estimations bi-modales pour la densité du noyau de certains paramètres. Nous avons utilisé d'une part l'option paramétrique en partant des estimateurs des moindres carrés. L'option non-paramétrique conduit à des résultats instables, la valeur des paramètres augmentant avec le nombre d'itérations : nous ne l'avons pas reportée ici.

Cela fournit les estimations des paramètres et des aléatoires, avec le rappel de l'estimation non-bayésienne, cf. le Tableau 8.

Si les paramètres fixes estimés changent peu selon la méthode utilisée, l'aléa régional est beaucoup moins stable. L'utilisation d'un modèle MCMC conduit à une variance plus élevée de cet aléa non significativement différente de zéro, en particulier lorsque la distribution *a priori* est uniforme. L'utilisation d'un bootstrap paramétrique conduit en revanche à une estimation de cet aléa très proche de celle du modèle non bayésien et tout à fait significative. Les densité du noyau et divers diagnostics de précision de ces estimateurs sont donnés (Raftery-Lewis sur les quantiles et Brooks-

Drapeer sur la moyenne) : ils ne fournissent aucun élément nouveau par rapport au modèle non-bayésien.

Paramètres	Estimateurs (écart-type)			
	Non-bayésien	MCMC (gamma)	MCMC (uniforme)	Bootstrap (paramétrique)
<i>Fixes :</i>				
Constante (α_0)	-0,538 (0,203)	-0,534 (0,246)	-0,545 (0,255)	-0,512 (0,191)
Sexe (α_1)	0,191 (0,081)	0,193 (0,083)	0,198 (0,081)	0,197 (0,081)
<i>Aléatoires :</i>				
$\sigma_{u_0}^2$	0,384 (0,177)	0,548 (0,354)	0,750 (0,560)	0,439 (0,184)
$\sigma_{e_0}^2$	1	1	1	1

TABLEAU 8. Résultats des modèles multiniveaux non-bayésiens et bayésiens, sans l'intervention du pourcentage de garçons

L'introduction du pourcentage de garçons ne permet plus d'estimer aucun modèle MCMC : cela doit être dû au fait que l'échantillon construit n'introduit plus d'aléas et il faudrait élaborer un échantillon moins strict pour éviter cela. En revanche il permet d'estimer un modèle bootstrap (Tableau 9).

Paramètres	Estimateurs (écart-type)	
	Non-bayésien	Bootstrap (paramétrique)
<i>Fixes :</i>		
Constante (α_0)	-1,641 (0,110)	-1,642 (0,108)
Sexe (α_1)	0,221 (0,082)	0,223 (0,081)
Pourcentage de garçons (α_2)	2,186 (0,147)	2,187 (0,145)
<i>Aléatoires :</i>		
$\sigma_{u_0}^2$	0	0
$\sigma_{e_0}^2$	1	1

TABLEAU 9. Résultats des modèles multiniveaux non-bayésien et bayésiens, avec l'intervention du pourcentage de garçons

Comme on peut le voir les deux modèles donnent des estimateurs pratiquement identiques, l'aléa au niveau école étant non significativement différent de zéro. Les trois caractéristiques fixes expliquent correctement les résultats au baccalauréat, sans faire intervenir d'aléa au niveau école.

On peut conclure de cet examen que l'approche multiniveau bayésienne n'introduit pas de résultats vraiment nouveaux dans le cas étudié ici, où les effectifs observés sont importants, et conduit à des prédictions identiques à celle du modèle non-bayésien.

7. INFÉRENCES À PARTIR DES RÉSULTATS MULTINIVEAUX

On peut à nouveau faire des prédictions à partir de ces résultats. Si l'on tire un nouvel individu, peut-on prédire sa réussite à l'aide des paramètres $\alpha_0, \alpha_1, \alpha_2, u_{0i}$, estimés à partir de la population initialement observée ? Comme précédemment cette prédiction sera une probabilité épistémique. Examinons divers cas, car nous avons vu que selon que l'on fait une analyse *purement multiniveau* ou une analyse *multiniveau contextuelle* les hypothèses vont changer de façon importante.

Voyons d'abord les résultats obtenus lorsque l'on a fait une analyse *purement multiniveau*. Rappelons sous une autre forme ce que nous avons dit précédemment : le jugement d'échangeabilité va porter dans ce cas sur la réussite de l'élève, étant donné son lycée et son sexe, ayant ici des effets indépendants.

Plaçons nous d'abord dans le cas où l'on observe un nouvel élève de sexe donné, sachant le lycée où il se trouve. Ainsi s'il s'agit d'un nouvel élève de sexe masculin soumis à un examen de même type dans le lycée n° 7, son taux de succès peut être estimé comme précédemment par le modèle 1 du Tableau 5 et par les estimations du Tableau 6 :

$$p(R = 1 | S = 0, L = 7) = P(R = 1 | S = 0, L = 7) = 0,448$$

soit 44,8 %. Le résultat pour une fille du même lycée serait 49,6 % par le même modèle.

Si l'on ignore son lycée on peut alors écrire :

$$p(R = 1 | S = 0) = P(R = 1 | S = 0, L = 1)p(L = 1 | S = 0) + \dots + P(R = 1 | S = 0, L = 10)p(L = 10 | S = 0)$$

On connaît les probabilités de réussite des garçons dans chacun des lycées, par les pourcentages observés, mais on ignore les probabilités épistémiques des lycées selon le sexe. Pour le modèle 1, on pourra faire l'hypothèse, comme dans l'exemple précédent des deux lycées, que chacune de ces probabilités sont égales entre elles, soit égales à un dixième, car l'effet du lycée semble être ici indépendant du sexe de l'élève. Dans ce cas on obtient : $p(R = 1 | S = 0) = 0,379$. Un raisonnement semblable chez les filles conduit à : $p(R = 1 | S = 1) = 0,421$, soit une meilleure probabilité de réussite d'une nouvelle fille que d'un nouveau garçon. Ces résultats sont contraires à ce que donne le Tableau 3, tous lycées réunis.

Si le nouvel individu est maintenant issu d'un lycée hors de l'échantillon, il nous faut tenir un raisonnement légèrement différent. Ce lycée étant hors de l'échantillon on ne peut pas estimer l'aléa qui lui correspond et on doit donc le poser comme égal à la moyenne des lycées observés, donc nul. Ce qui conduit à :

$$p(R = 1 | S = 0) = [1 + \exp(0,5378)]^{-1} = 0,369 .$$

On voit que ce résultat est très proche de celui obtenu lorsque l'on ignorait le lycée, alors qu'il était situé dans l'échantillon.

Voyons maintenant les résultats obtenus lorsque l'on fait une analyse *multiniveau contextuelle*. Dans ce le jugement d'échangeabilité porte sur la réussite de l'élève et

l'effet du lycée, mesuré par son pourcentage de garçons, étant donné le sexe du nouvel élève.

Plaçons nous dans le premier cas précédent où il s'agit d'un nouvel élève de sexe masculin soumis à un examen de même type dans le lycée n° 7. Si l'on ne dispose d'aucune information sur le pourcentage de garçons lors de cette épreuve on peut le considérer comme inchangé par rapport à la précédente et le modèle 3 du Tableau 5 nous donne :

$$p(R = 1 | S = 0, L = 7) = P(R = 1 | S = 0, L = 7) = 0,445$$

et pour les filles du même lycée 0,500, soit un résultat très proche de ce que donnait le modèle 1.

Cependant, si pour cet examen le pourcentage de garçons de ce lycée a changé, ce que l'on pourrait savoir *a priori*, le modèle 3 fournira une autre estimation de ce taux de succès. En effet, si ce pourcentage tombe à 55 % au lieu de 65 % son taux de succès prévu tombera à 39,2 %, car la caractéristique pertinente est maintenant la part de garçons dans le lycée et non le lycée lui-même, alors que le modèle 1 donnera toujours 44,8 %. Dans ce cas les prédictions faites par les deux modèles risquent de différer fortement entre elles.

Si l'on ignore maintenant le lycée où se trouve l'élève, on pourra toujours écrire la même formule, que dans le cas précédent. Mais l'hypothèse faite pour le modèle 1 ne vaut plus pour le modèle 2, car on voit que c'est en fait le pourcentage de garçons dans le lycée qui introduit les différences de résultats. Si l'on n'a pas de nouvelle information sur les pourcentages de garçons pour cette nouvelle épreuve, les pourcentages les plus pertinents à utiliser seront ceux observés pour l'épreuve antérieure et cela conduit aux probabilités de succès pour un nouveau garçon : $p(R = 1 | S = 0) = 0,4575$ et de : $p(R = 1 | S = 1) = 0,3425$ pour une nouvelle fille, soit un résultat correspondant à celui des lycées réunis et inverse de ce que fournissait le modèle 1. Si l'on connaît les nouveaux pourcentages de garçons dans chacun des lycées pour cette nouvelle épreuve, on doit alors les appliquer pour calculer la probabilité de succès d'un nouveau garçon. Ainsi, si pour ce nouvel examen les pourcentages de garçons sont égaux à ceux des filles dans tous les lycées, alors le résultat sera proche de celui obtenu par le modèle 1, car on peut calculer, à l'aide des résultats du Tableau 7 : $p(R = 1 | S = 0) = 0,377$ pour les garçons, à comparer avec 0,379 par le modèle 1, et $p(R = 1 | S = 1) = 0,496$ pour les filles, toujours à comparer avec 0,421. En revanche chacune des répartitions possibles donnera maintenant un résultat différent, dépendant de cette répartition.

Voyons enfin le cas où le lycée est hors de l'échantillon. Si l'on suppose qu'on connaît son pourcentage de garçons alors on pourra avoir une estimation plus satisfaisante que dans le cas d'une analyse multiniveau simple. Par exemple si son pourcentage de garçons est égal à 70 % alors la probabilité de succès prévue pour ce nouveau garçon sera de 0,427, valeur interpolée entre les résultats des lycées avec 65 % et 70 % de garçons. Cependant si ce pourcentage est inconnu au moment de l'estimation on devra se contenter de la valeur

$$P(R = 1 | S = 0) = [1 + \exp(1,641 - 0,5 * 2,186)]^{-1} = 0,366$$

valeur proche de ce que donnerait le modèle multiniveau simple.

Bien entendu nous avons développé ici ces exemples en détail pour montrer les différences de résultats selon le jugement d'échangeabilité choisi comme le plus pertinent. Mais dans le cas présent il est évident que le modèle 3 doit être choisi comme le plus pertinent car il utilise toute l'information fournie par les Tableaux de données 3 et 4. Également l'utilisation d'un modèle bayésien ne change en rien ces résultats.

8. CONCLUSIONS

Cet article a permis de montrer comment le paradoxe de Simpson pouvait être correctement résolu en utilisant un modèle multiniveau adéquat [Wunsch, 2007] et comment l'utilisation du concept d'échangeabilité permettait une inférence statistique à partir de l'observation d'un échantillon. Cependant le lien entre le nouvel individu pour lequel on effectue cette inférence et l'échantillon initial doit être formulé de façon très précise si l'on désire réaliser une prédiction correcte. Ce lien nécessite des hypothèses sous-jacentes aux données, qu'il est nécessaire de combiner à celles-ci pour poser les conditions exactes d'échangeabilité et réaliser l'inférence finale. Ainsi, pour le premier modèle *purement multiniveau*, on estime que l'échangeabilité porte sur le résultat conditionné par le lycée et le sexe alors que, pour le second modèle *multiniveau contextuel*, l'échangeabilité porte sur le résultat et le lycée conditionnés seulement par le sexe de l'élève. Dans ce dernier cas on ne considère que deux sous populations : les garçons et les filles. Le fait qu'un individu fasse partie d'un lycée ou d'un autre n'intervient plus que par le pourcentage de garçons qu'il a, et non par son personnel enseignant.

Bien entendu d'autres approches sont possibles pour traiter ce problème d'inférence. Ainsi les concepts de *population* et de *sous-population* utilisés par Fisher [1956] peuvent remplacer ceux d'*échangeabilité* et d'*échangeabilité conditionnelle* [Lindley, Novick, 1981 ; Courgeau, 2004(a)]. Mais le fait que l'échangeabilité se réfère explicitement à des variables aléatoires, alors que le concept de population ne peut s'y référer, enlève une partie de l'intérêt de l'approche suivie par Fisher. Ainsi il est très difficile d'indiquer comment un individu peut se rattacher à une population donnée alors que le critère d'échangeabilité, qui travaille sur des unités, permet plus clairement de rattacher des individus entre eux. C'est la raison pour laquelle nous n'avons pas développé ici cette approche fréquentiste et avons préféré une approche bayésienne de l'inférence.

Également, nous avons supposé ici travailler sur un échantillon important sur lequel on pouvait estimer les probabilités utilisées pour réaliser l'inférence, par leur fréquence. D'ailleurs l'estimation bayésienne tentée sur cet échantillon, pour le cas multiniveau, donne des résultats proches de l'estimation fréquentiste. Cependant, si l'on travaille sur un faible échantillon des complexités vont survenir, car l'estimation des probabilités de l'événement étudié doit dans ce cas se faire en utilisant des méthodes bayésiennes. Ceci entraîne des considérations sur l'échangeabilité d'un petit nombre d'individus que nous ne développerons pas ici. On peut cependant penser que cela n'apportera qu'une très faible information sur l'inférence à effectuer, rendant ce problème de peu d'intérêt.

Enfin certains auteurs [Draper *et al.*, 1993] considèrent un concept d'échangeabilité à la fois plus étendu que celui utilisé par de Finetti et également plus primitif que la notion de probabilité. Nous indiquons seulement ici cette possibilité sans entrer dans les discussions entraînées par cet article et données à sa suite, en particulier

celle de Lindley (p. 29-30), qui refuse de considérer l'échangeabilité comme un concept plus fondamental que la probabilité.

Pour conclure nous pouvons dire que les données ne fournissent pas une information suffisante pour pouvoir réaliser une inférence correcte. Il faut également utiliser d'autres informations, soit tirées de l'analyse faite à partir de ces données, soit tirées de l'expérience pratique du chercheur sur le sujet, qui vont pouvoir rendre cette inférence réalisable.

BIBLIOGRAPHIE

- BROWNE W. J., *MCMC estimation in MLwin* [Version 2.0], London, Institute of Education, 2004.
- CHUNG K.-L., "On mutually favorable events", *The Annals of Statistics* 13, 1985, p. 338-349.
- COHEN M. R., NAGEL E., *An introduction to logic and scientific method*, New-York, Harcourt Brace, 1934.
- COURGEAU D., « Évolution ou révolutions dans la pensée démographique ? », *Mathématiques et Sciences humaines* 160, 2002, p. 49-76.
- COURGEAU D., « Probabilités, démographie et sciences sociales », *Mathématiques et Sciences humaines* 167, 2004(a), p. 27-50.
- COURGEAU D., *Du groupe à l'individu. Synthèse multiniveau*, Paris, éditions de l'Institut National d'Études Démographiques, 2004(b).
- COURGEAU D., "Multilevel synthesis", *The Springer Series on Demographic Methods and Population Analysis*, Dordrecht, Springer, 2007.
- DE FINETTI B., *Theory of probability*, 2 vol., London, Wiley, 1974.
- DRAPER D., HODGES J. S., MALLOWS C. L., PREGIBON D., "Exchangeability and data analysis", *Journal of the Royal Statistical Society A*, 156, 1993, p. 9-37.
- FISHER R. A., *Statistical methods and scientific inference*, Edinburgh, Oliver and Boyd, 1956.
- GOLDSTEIN H., *Multilevel statistical models*, London, Edward Arnold, 2003.
- GOURIEROUX, C., *Économétrie des variables qualitatives*, Paris, Economica, 1989.
- GREENLAND S., "Principles of multilevel modelling", *International Journal of Epidemiology*, 29, 2000, p. 158-167.
- HACKING I., *L'émergence de la probabilité*, Paris, Seuil, 2002.
- JOHNSON W.E., "Probability: the deductive and inductive problem", *Mind* 41, 1932, p. 409-423.
- LINDLEY, D. V., NOVICK M. R., "The role of exchangeability in inference", *The Annals of Statistics* 9, 1981, p. 45-58.
- McCULLAGH P., NELDER J. A., *Generalized Linear Models*, London, Chapman and Hall, 1983.
- ROUANET H., BERNARD J.-M., BERT M.-C., LECOUTRE B., LECOUTRE M.-P., LE ROUX B., *New ways in statistical methodology*, Bern, Peter Lang, 1998.
- ROBERT C. P., *Le choix bayésien*, Paris, Springer-Verlag, 2006.
- ROUANET H., LEBARON F., LE HAY V., ACKERMANN W., LE ROUX B., « Régression et analyse géométrique des données : réflexions et suggestions », *Mathématiques et Sciences humaines* 160, 2002, p. 13-45.
- SIMPSON E. H., "The interpretation of interaction in contingency tables", *Journal of the Royal Statistical Society Series B*, 13, 1951, p. 238-241.
- WUNSCH G., "Confounding and control", *Demographic Research* 16, 2007, p. 97-120.