ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 4 — 1986

Comment évaluer un nombre de catégories par échantillonnage

Jacques ROUAULT, Pierre CAPY *

RÉSUMÉ. – Cet article décrit une méthode d'estimation du nombre de catégories présentes dans une urne de composition inconnue. A partir d'une composition *a priori* du nombre et des fréquences des catégories, il est possible de calculer une erreur d'échantillonnage et d'estimer le nombre d'individus dont il faudrait disposer pour que cette erreur devienne inférieure à un risque fixé. Des tirages successifs permettent de converger vers une composition représentative de l'urne au risque fixé.

Estimating the Number of Categories by Sampling

ABSTRACT. – This paper describes a method for estimating the number of categories in an urn of unknown composition. From *a priori* values for the number and frequencies of categories, it is possible to calculate a sampling error and to estimate the number of individuals to draw which would be necessary for decreasing this error beyond a given risk. Successive assays lead to a convergence towards a convenient representation of the urn, in agreement with the fixed risk.

* J. ROUAULT, P. CAPY : Laboratoire de Biologie et Génétique évolutives CNRS, 91190 Gif-sur-Yvette. Les auteurs remercient vivement MM. J.-M. GOUX (Université Paris-VII) et G. ROY (INSEE), ainsi que les deux lecteurs anonymes pour leurs remarques pertinentes et constructives qui ont permis d'améliorer significativement la qualité de ce texte.

1 Introduction

De nombreuses études statistiques supposent connue *a priori* la composition exacte de l'urne où sont effectués les tirages. Par composition, il faut entendre le nombre de catégories (couleurs, classes, espèces, ...) et leurs fréquences respectives. L'application de ces méthodes aux domaines économique ou biologique pose problème dans la mesure où la composition des urnes faisant l'objet des prélèvements n'est en général pas connue avec certitude. Celle-ci peut cependant être approchée avec une erreur plus ou moins grande selon le nombre d'individus qui y sont prélevés. La question qui fait l'objet de ce travail consiste à déterminer le nombre minimal d'individus qu'il est nécessaire de prélever pour que l'erreur d'échantillonnage puisse être considérée comme négligeable devant un risque fixé.

La théorie des sondages (DESABIE [1966], GOURIEROUX [1981]) permet dans certains cas de fixer la taille de l'échantillon quand d'une part, le nombre de classes est défini *a priori*, et d'autre part, que les effectifs de chaque classe peuvent être considérés comme très grands. Quand certaines classes représentent un effectif trop faible, elles peuvent être regroupées dans une classe dénommée « autres » ou « divers ». De même, si seules les classes les plus fréquentes sont connues *a priori*, il sera défini une classe rebut regroupant les autres catégories. Quand les effectifs des différentes catégories ne peuvent plus être considérés comme très grands, le modèle gaussien utilisé dans la théorie des sondages peut être malgré cela appliqué à condition d'introduire une correction (AMEGANDJIN [1970]). Ce procédé est classique en statistique mathématique où l'estimation de la variance subit un biais fonction de la taille de l'échantillon.

Ce type de problème a été très peu abordé dans le domaine des sciences biologiques. Au plus pouvons-nous citer une étude consacrée à la notion de « richesse spécifique » qui offre un estimateur du nombre d'espèces capturées en fonction de la taille de l'échantillon (HURLBERT [1971]). Cependant, cette méthode n'est pas opératoire du fait de son caractère absolu (non probabiliste) (PEET [1974]). Cette démarche traduit la diminution de l'erreur d'échantillonnage avec l'accroissement de l'effectif. Trivialement, l'erreur devient nulle pour un effectif infini. Le défaut majeur de ce type d'approche est de ne pas prendre en considération le *coût* que représente l'échantillonnage d'une unité : coût en travail, en moyens, en temps, mais également coût écologique dans le cas de prélèvements destructifs. De l'équilibre entre les deux paramètres de l'échantillonnage que sont d'une part l'*erreur* (liée à la fraction non prélevée) et d'autre part le *coût* (de la fraction prélevée) résulte le concept d'*optimisation* d'une série de prélèvements.

Dans cet article, nous allons nous efforcer de répondre à la question initialement posée dans le cadre le plus général qu'il est possible d'envisager : composition (nombre de classes et fréquences de celles-ci) totalement inconnue *a priori*, effectifs de taille quelconque (y compris très petits). Il sera alors possible de déterminer le nombre minimal d'individus à échantillonner pour un risque donné aussi bien dans une enquête (cas de codage des réponses à une question ouverte) qu'en écologie (dénombrement d'espèces) ou en génétique (dénombrement d'allèles).

2 Probalité associée à un échantillon

Nous supposerons dans un premier temps que l'échantillonnage est réalisé au sein d'une urne où sont prélevées des boules de couleurs (ou catégories, ou classes) distinguables sans ambiguïté. La réalisation d'une épreuve d'échantillonnage de taille nt consistera à tirer un ensemble de nt boules.

Dans ce paragraphe, nous supposerons que la composition de l'urne de référence est connue *a priori* : celle-ci se compose de nc couleurs de fréquences respectives p_i ($1 \leqslant i \leqslant nc$). Nous définirons la probabilité de tirage associée à un échantillon de nt boules comme étant la *probabilité d'observer dans cet échantillon au moins une fois chacune des nc couleurs présentes dans l'urne.*

Dans le cas usuel de tirages équiprobables sans remise dans une urne de taille finie, on est conduit à utiliser des lois hypergéométriques. De plus, comme les tirages sont effectués sans remise, les fréquences p_i sont modifiées après chaque tirage.

Si nous supposons que la taille nt du tirage est négligeable devant la taille de l'urne, la probabilité de ne pas observer une couleur de fréquence p_i dans l'échantillon est alors égale à l'expression binomiale $(1-p_i)^{nt}$. Donc la probabilité d'observer cette couleur au moins une fois est $1-(1-p_i)^{nt}$.

La probabilité de tirer a_i fois (avec i variant de 1 à nc) chacune des nc couleurs de l'urne suit une loi multinomiale de paramètres nt (somme de a_i) et p_i. La somme $\mathrm{P}t(nc)$ des probabilités de tirage, qui correspond à l'ensemble des événements $(a_1, ..., a_i, ..., a_{nc})$ avec $a_i \neq 0$, représente la probabilité de tirer au moins une fois chacune des nc couleurs de l'urne.

La quantité $\mathrm{P}t(nc)$ croît vers l'unité quand la taille nt de l'effectif augmente. Nous définirons l'*erreur d'échantillonnage* re comme le complément à 1 de la probabilité de tirage : $re = 1 - \mathrm{P}t(nc)$.

3. Calcul de la probabilité

3.1. Deux couleurs

Le calcul exhaustif de toutes les probabilités multinomiales des événements correspondant à $a_i \neq 0$, $\forall i$ devient très vite extrêmement fastidieux quand nt et/ou nc dépassent quelques unités, même quand ce calcul est programmé sur ordinateur. Nous avons recherché une expression équivalente de la probabilité $\mathrm{P}e(nt)$ dont l'expression analytique soit uniquement fonction de nt, nc et des p_i (les a_i n'apparaissant plus).

Pour $nc=2$, on obtient immédiatement :

$$Pe(2)=1-p_1^{nt}-p_2^{nt}$$

La figure 1 montre les variations de $Pe(2)$ en fonction de nt pour différentes valeurs des probabilités de tirage p_1 et p_2.

3.2. Trois couleurs

Pour $nc=3$, la probalité $Pe(3)$ recherchée est égale à 1 moins la somme des probabilités trinomiales correspondant à au moins un des a_i nul. Si nous notons $\text{trino}(a_1, a_2, a_3)$ la valeur de la probabilité trinomiale de paramètres $nt=a_1+a_2+a_3$ et p_1, p_2, p_3, nous aurons :

$$Pe(3)=1-\sum_{a_1=1}^{nt}\text{trino}(a_1, 0, nt-a_1)$$
$$-\sum_{a_2=1}^{nt}\text{trino}(nt-a_2, a_2, 0)-\sum_{a_3=1}^{nt}\text{trino}(0, nt-a_3, a_3)$$

Chaque somme représente le développement d'un binome de la forme $(p_1+p_2)^{nt}$ au premier terme près, donc :

$$Pe(3)=1-(p_1+p_2)^{nt}-(p_1+p_3)^{nt}$$
$$-(p_2+p_3)^{nt}+p_1^{nt}+p_2^{nt}+p_3^{nt}$$

La figure 2 montre les variations de $Pe(3)$ en fonction de nt pour différentes valeurs des probabilités de tirage p_1, p_2 et p_3.

3.3. Cas général

Une démonstration de la formule analytique pour un nombre quelconque de catégories peut être effectué ainsi :

Soit A_i l'événement $A_i=(a_i\neq 0)$. On a :

$$Pe(nc)=P\left(\bigcap_{i=1}^{nc}A_i\right)=1-P\left(\bigcup_{i=1}^{nc}\overline{A_i}\right).$$

Cette dernière expression peut être calculée par le théorème de Poincaré :

$$P\left(\bigcup_i\overline{A_i}\right)=\sum_i P(\overline{A_i})-\sum\sum_{i>j}P(\overline{A_i}\cap\overline{A_j})+\ldots$$

En remarquant que :

$$P(\overline{A_i})=(1-p_i)^{nt}$$
$$P(\overline{A_i}\cap\overline{A_j})=(1-p_i-p_j)^{nt},\qquad\ldots$$

On voit ainsi apparaitre la structure générale de l'expression analytique.

FIGURE 1

Variation de la probabilité d'échantillonnage Pe(2) (en ordonnée) en fonction du nombre NT de boules prélevés (en abscisse) pour des urnes contenant NC = 2 couleurs.

Les différentes courbes correspondent à des couleurs de fréquences respectives 10 : 10 ($p_1 = p_2 = 0{,}5$), 11 : 9 ($p_1 = 0{,}55$; $p_2 = 0{,}45$), ..., 19 : 1 ($p_1 = 0{,}95$; $p_2 = 0{,}05$), 39 : 1 ($p_1 = 0{,}975$; $p_2 = 0{,}025$), 79 : 1 ($p_1 = 0{,}987\,5$; $p_2 = 0{,}012\,5$) et 159 : 1 ($p_1 = 0{,}993\,75$; $p_2 = 0{,}006\,25$).

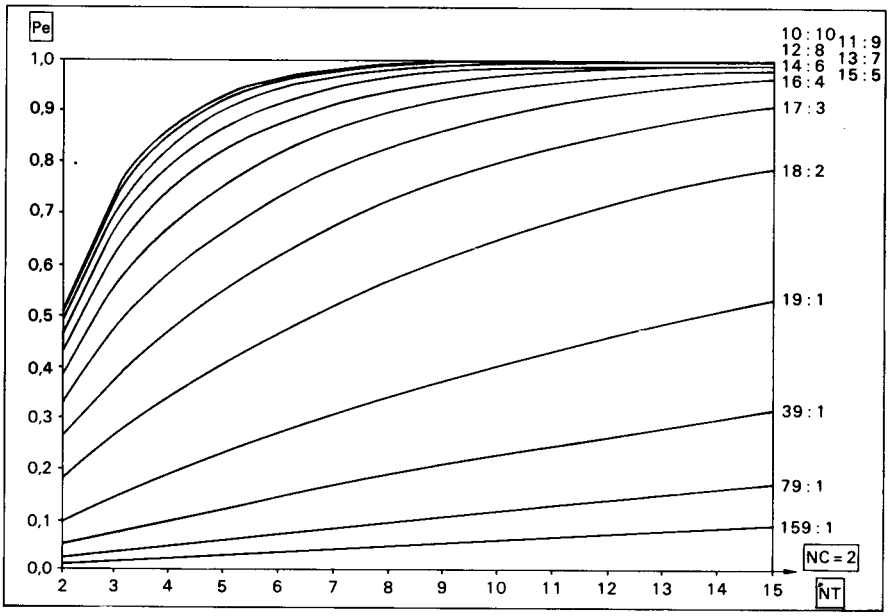
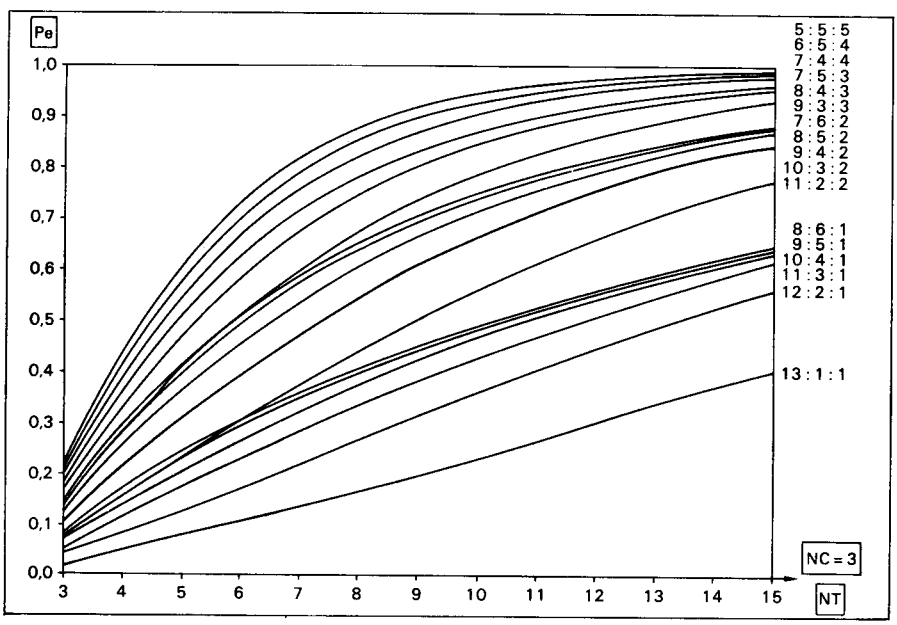


FIGURE 2

Variation de la probabilité d'échantillonnage Pe(3) (en ordonnée) en fonction du nombre NT de boules prélevés (en abscisse) pour les urnes contenant NC = 3 couleurs.

Les différentes courbes correspondent à des couleurs de fréquences respectives 5 : 5 : 5 ($p_1 = p_2 = p_3 = 0{,}333\,33$), 6 : 5 : 4 ($p_1 = 0{,}4$; $p_2 = 0{,}33\,333$; $p_3 = 0{,}26\,666$), et 13 : 1 : 1 ($p_1 = 0{,}866\,66$; $p_2 = p_3 = 0{,}066\,66$).



3.4. Exemple

L'expression générale de la probabilité recherchée peut s'exprimer par la formule suivante :

$$Pe(nc) = \sum_{z=0}^{nc-1} (-1)^z S_{nc,\ z}$$

Où $S_{nc,\ z}$ représente la somme des puissances nt-ièmes de toutes les sommes de combinaisons de nc-z valeurs p_i parmi nc, soit :

$$S_{nc,\ z} = \sum_{k=1}^{C_{nc}^{nc-3}} T_{nc,\ z,\ k}^{nt}$$

Où $T_{nc,\ z,\ k}$ est une somme de $nc-z$ valeurs p_i.

Cette procédure de calcul peut être réalisée simplement grâce à l'emploi de la récursivité. Les auteurs tiennent à la disposition des lecteurs qui en feront la demande un exemple de programme rédigé en Pascal.

Pour $nc=4$, nous aurons :

$T_{4,\ 0,\ 1} = p_1+p_2+p_3+p_4 = 1$

$S_{4,\ 0} = 1$

$T_{4,\ 1,\ 1} = p_1+p_2+p_3$, $T_{4,\ 1,\ 2} = p_1+p_2+p_4$,

$T_{4,\ 1,\ 3} = p_1+p_3+p_4$ $T_{4,\ 1,\ 4} = p_2+p_3+p_4$

$S_{4,\ 1} = (p_1+p_2+p_3)^{nt}+(p_1+p_2+p_4)^{nt}+(p_1+p_3+p_4)^{nt}+(p_2+p_3+p_4)^{nt}$

$T_{4,\ 2,\ 1} = p_1+p_2$, $T_{4,\ 2,\ 2} = p_1+p_3$, $T_{4,\ 2,\ 3} = p_1+p_4$

$T_{4,\ 2,\ 4} = p_2+p_3$, $T_{4,\ 2,\ 5} = p_2+p_4$, $T_{4,\ 2,\ 6} = p_3+p_4$

$S_{4,\ 2} = (p_1+p_2)^{nt}+(p_1+p_3)^{nt}+(p_1+p_4)^{nt}$
$+(p_2+p_3)^{nt}+(p_2+p_4)^{nt}+(p_3+p_4)^{nt}$

$T_{4,\ 3,\ 1} = p_1$, $T_{4,\ 3,\ 2} = p_2$, $T_{4,\ 3,\ 3} = p_3$, $T_{4,\ 3,\ 4} = p_4$

$S_{4,\ 3} = p_1^{nt}+p_2^{nt}+p_3^{nt}+p_4^{nt}$

D'où nous obtenons :

$$\begin{aligned} Pe(4) = 1 &- (p_1+p_2+p_3)^{nt} - (p_1+p_2+p_4)^{nt} - (p_1+p_3+p_4)^{nt} \\ &- (p_2+p_3+p_4)^{nt} + (p_1+p_2)^{nt} + (p_1+p_3)^{nt} + (p_1+p_4)^{nt} \\ &+ (p_2+p_3)^{nt} + (p_2+p_4)^{nt} + (p_3+p_4)^{nt} - p_1^{nt} - p_2^{nt} - p_3^{nt} - p_4^{nt} \end{aligned}$$

4. Algorithme d'échantillonnage

A présent, nous allons supposer que nous ignorons la composition de l'urne, mais que nous disposons déjà d'une estimation *a priori* de celle-ci basée sur un tirage de *nt* boules. Nous allons effectuer un raisonnement inférentiel en fixant la valeur *r* du risque de première espèce, c'est-à-dire le risque de rejeter à tort une hypothèse vraie, ce qui se traduit ici par une évaluation du nombre de couleurs strictement inférieure à la valeur réelle. C'est le risque de ne pas détecter une ou plusieurs couleurs présentes dans l'urne.

L'estimation dont nous disposons pour l'urne permet d'appliquer le raisonnement probabiliste du paragraphe précédant. Nous calculons la probabilité $Pe(nc)$ associée à l'échantillon *nt*, et l'erreur d'échantillonnage associée *re*. La règle de décision résulte de la comparaison de *re* à *r* :

$re \leqslant r$: L'erreur d'échantillonnage étant inférieure au risque de première espèce, nous pouvons conclure que l'échantillon d'effectif *nt* est représentatif de l'urne au risque *re*, et *a fortiori r*.

$re > r$: l'erreur d'échantillonnage est supérieure au risque de première espèce. Par conséquent, l'échantillon de taille *nt* ne peut pas être considéré comme représentatif de l'urne au risque *r*. Cependant, nous allons utiliser la composition *a priori* de l'urne dont nous disposons pour estimer le nombre de tirages $nt' > nt$ dont nous aurions dû disposer pour que l'erreur d'échantillonnage correspondante *re'* soit inférieure à *r*.

La quantité $dt = nt' - nt$ représente la taille du tirage complémentaire à effectuer. Ce n'est que dans la mesure où chaque tirage est indépendant du précédant (ou du suivant) que ce protocole peut être appliqué. Dans le cas contraire, *nt'* boules (au lieu de *dt*) doivent être retirées.

Effectuer un tirage complémentaire a pour effet de modifier l'estimation de la composition de l'urne : les fréquences des couleurs déjà tirées se trouvent modifiées, et de nouvelles couleurs peuvent apparaitre. A la suite du tirage complémentaire, nous disposons d'une estimation *a posteriori* de la composition de l'urne.

Nous nous trouvons alors dans une situation comparable à celle exposée au début de ce paragraphe : il suffit simplement de considérer les fréquences observées *a posteriori* comme des probabilités de tirage *a priori*.

Ce protocole définit un algorithme d'échantillonnage par navettes (ou tirages complémentaires) successives. Pour chaque navette, la composition *a priori* de l'urne est définie par la composition *a prosteriori* de la navette précédante. La convergence de l'agorithme résulte de la croissance du nombre de catégories estimé et du fait que ce nombre est borné par la vraie valeur. Il est cependant évident qu'en présence d'une catégorie de fréquence très faible (inférieure au risque *r* et a $1/nc$), cette catégorie ne sera pas représentée dans l'échantillon, et en voie de conséquence, celui-ci sera biaisé.

Il est intuitif que le choix d'une valeur plus faible pour le risque r réduira ce biais, mais induira corrélativement un accroissement de la taille de l'échantillon.

Il est évident que, plus le risque de première espèce r est faible, plus le nombre nv de navettes, comme le nombre nt de tirages seront grands. A la différence des travaux de HURLERT [1971] et de PEET [1974], cette liaison fonctionnelle a été probabilisée, ce qui a pour corrolaire la formulation d'une règle de décision.

Une fois la nature de la règle de récurrence définie, il convient de préciser comment déterminer la première composition supposée de l'urne. Face à une problématique donnée, le praticien a en général une idée de la composition de l'urne. Il peut pour cela faire référence à des prélèvements antérieurs ou analogues qui lui offrent un bon point de départ.

Si cela n'est pas le cas, une remarque s'impose : le premier tirage ne doit pas être d'effectif trop faible. Par exemple, pour une composition ($p_1 = 0{,}9$, $p_2 = 0{,}1$), une première valeur nt trop faible conduira souvent à une estimation complètement erronnée (c'est-à-dire la présence d'une seule couleur). Une attitude raisonnable serait de considérer une couleur de plus que le nombre attendu *a priori*. Quant au choix des fréquences, le praticien pourra s'en remettre en partie au *principe d'équirépartition de l'ignorance* de Bayes-Laplace. Plus précisément, si nous désirons éviter de ne pas détecter une $nc+1$-ième classe de fréquence f, nous choisirons comme composition *a priori* nc classes de fréquence $(1-f)/nc$ (supérieure à f) et une $nc+1$-ième classe de fréquence f.

5. Exemple d'application de l'algorithme

Considérons une urne composée de 4 couleurs de fréquences respectives 0,5; 0,3; 0,1 et 0,1. Nous fixons un risque de première espèce $r = 0{,}05$.

1. Dans l'ignorance de la composition de l'urne, nous supposons *a priori* l'existence de 2 couleurs équiprobables. Nous devons alors tirer $nt_1 = 6$ boules pour atteindre la probabilité $Pe_1 = 0{,}969$ de tirer les 2 couleurs présupposées (figure 3, courbe 1 : 1). Nous tirons 4 : 1 : 1 boules correspondant à 3 couleurs, la probabilité associée est $Pe_1^* = 0{,}417$ ($re^* = 0{,}583 > r$). Nous devons donc envisager un tirage complémentaire.

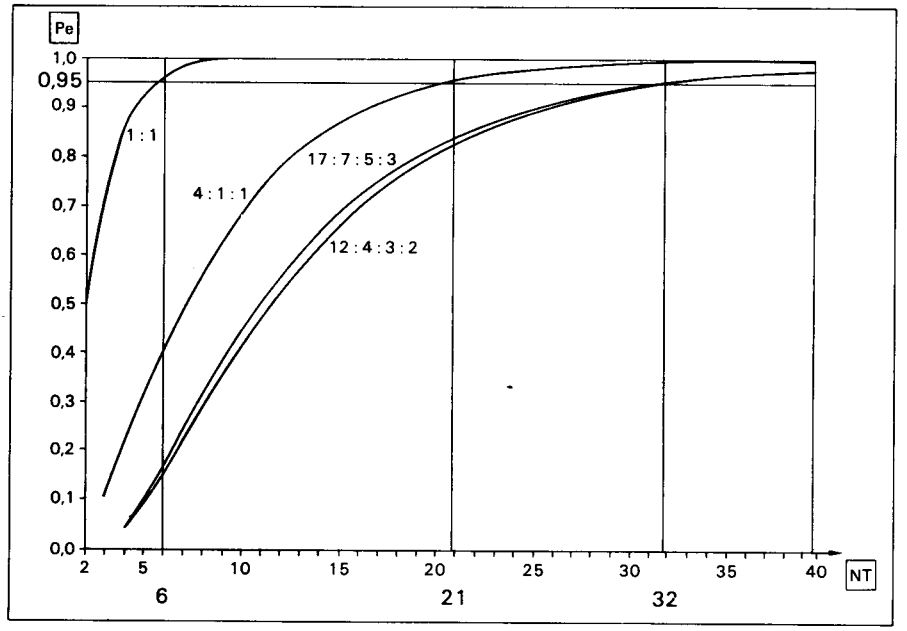
2. D'après la nouvelle composition *a priori* de l'urne (courbe 4 : 1 : 1), nous devons disposer de $nt_2 = 21$ boules pour avoir $Pe_2 = 0{,}957$. Le tirage complémentaire devra donc comporter $21 - 6 = 15$ boules. La seconde navette donne 8 : 3 : 2 : 2 boules, ce qui conduit à une composition 12 : 4 : 3 : 2 correspondant à $Pe_2^* = 0{,}831$ ($re^* = 0{,}169 > r$).

3. Suivant cette nouvelle composition, nous devrons disposer de 32 boules pour avoir $Pe_3 = 0.951$. Le tirage complémentaire de 11 boules donne

FIGURE 3

Exemple d'application de l'algorithme. Représentation des variations de la probabilité d'échantillonnage Pe (en ordonnées) en fonction du nombre NT de boules prélevées (en abscisse) pour les différentes estimations successives de la composition de l'urne : 1:1, 4:1:1, 12:4:3:2 et 17:7:5:3.

Les 3 traits verticaux visualisent le nombre total de boules NT' qu'aurait du contenir l'échantillon pour que la probabilité associée Pe dépasse 0,95.*



5 : 3 : 2 : 1, ce qui, additionné aux prélèvements précédants donne une composition de 17 : 7 : 5 : 3. La probabilité $Pe_3^* = 0{,}953$ donne une erreur d'échantillonnage *re** de 0,047, inférieure à 0,05. Nous considèrons alors que l'algorithme a convergé après avoir effectué 3 navettes. (Les effectifs observés ont été obtenus par simulation.)

6. Simulations

Afin de valider l'algorithme proposé et de vérifier la vraisemblance des résultats, nous avons procédé à 4 séries de simulations correspondant respectivement à des risque de première espèce de 0,05; 0,01; 0,005 et 0,001. Dans chaque cas, environ 6 000 simulations ont été effectuées.

Ces simulations concernent une urne théorique contenant des boules de 2 couleurs aux fréquences respectives 0,9 et 0,1. La composition initiale supposée de l'urne est 0,5 : 0,5. Cependant, le processus force le tirage d'au moins 2 couleurs dans le premier échantillon, afin d'éliminer des effets de bord indésirables (c'est-à-dire la convergence immédiate de l'algorithme si une seule couleur est observée).

La figure 4 présente la distribution du nombre total de boules tirées jusqu'à l'obtention de la convergence du processus. On remarquera l'aspect très discontinu de la distribution. L'observation essentielle est que la distribution se déplace vers la droite et s'étale de plus en plus quand la valeur du risque r est de plus en plus petite. Ce résultat est logique, mais les 4 distributions indiquent qu'il est utopique de vouloir déterminer un modèle prédictif (même statistiquement) de nt en fonction de r. Quoi qu'il en soit, toute approximation analytique (et en particulier gaussienne) est à proscrire.

La distribution des valeurs observées pour la classe de fréquence 0,9 fait l'objet de la figure 5. La valeur théorique 0,9 correspond bien à la position centrale de la distribution pour les 4 valeurs du risque. Quand la valeur du risque r diminue, les valeurs observées se resserrent de plus en plus autour de la valeur théorique. Nous avons vu que cela correspondait à l'accroissement du nombre de boules tirées. La distribution observée (particulièrement pour $r=0{,}001$) pourrait être modélisée par une loi beta.

Le nombre nv de navettes effectuées croît évidemment quand la valeur du risque diminue (figure 6). Le nombre de cas de convergence au cours de la première navette est très faible pour $r=0{,}5$ et 0,01, et devient négligeable pour r plus faible. Pour la moitié des simulations effectuées, on observe une convergence du processus en 2 navettes. Le nombre moyen de navettes est respectivement 2,55; 3,03; 3,27 et 3,15. Il est logique que le nombre nv de navettes et le nombre nt de boules tirées croissent quand le risque r diminue. Pour les mêmes raisons, la qualité des résultats obtenus augmente avec le nombre de navettes, mais il s'agit d'un phénomène discret qui est de nature à expliquer la répartition discontinue des observations.

7. Discussion et conclusion

Nous avons pu fournir une réponse à la question initialement posée par le développement d'une méthodologie qui s'adapte à tous les contextes possibles, y compris quand les classes ne sont pas prédéfinies et/ou les effectifs observés dans certaines classes sont très petits. La notion de coût est liée à la valeur du risque de première espèce : une diminution du risque entraîne une augmentation du coût de l'échantillonnage, mais induit corrélativement une amélioration sensible de la qualité des estimations.

Dans la pratique, la méthode développée ci-dessus peut être appliquée dans deux contextes différents :

1. Pour effectuer un échantillonnage. On fixe la valeur du risque de première espèce, la composition *a priori* de l'urne, et on lance le processus d'échantillonnage par navettes jusqu'à la convergence. C'est ce qui a été effectué dans l'exemple présenté plus haut.

2. Pour qualifier un échantillon ou une série d'échantillons. A chaque échantillon, nous pouvons associer une probabilité de tirage $Pe(nc)$ et un risque re. Les tests statistiques ultérieurement appliqués à cet échantillon devront intégrer l'erreur d'échantillonnage ainsi estimée. En particulier, l'erreur d'échantillonnage devra être négligeable devant la valeur du risque

FIGURE 4

Simulation de l'algorithme d'échantillonnage à partir d'une urne de composition 0,9 : 0,1. Représentation de la distribution des répétitions (en ordonnée) du nombre NT de boules tirées (en abscisse) pour un risque R = 0,05; 0,01; 0,005; 0,001 et environ NS = 6 000 répétitions

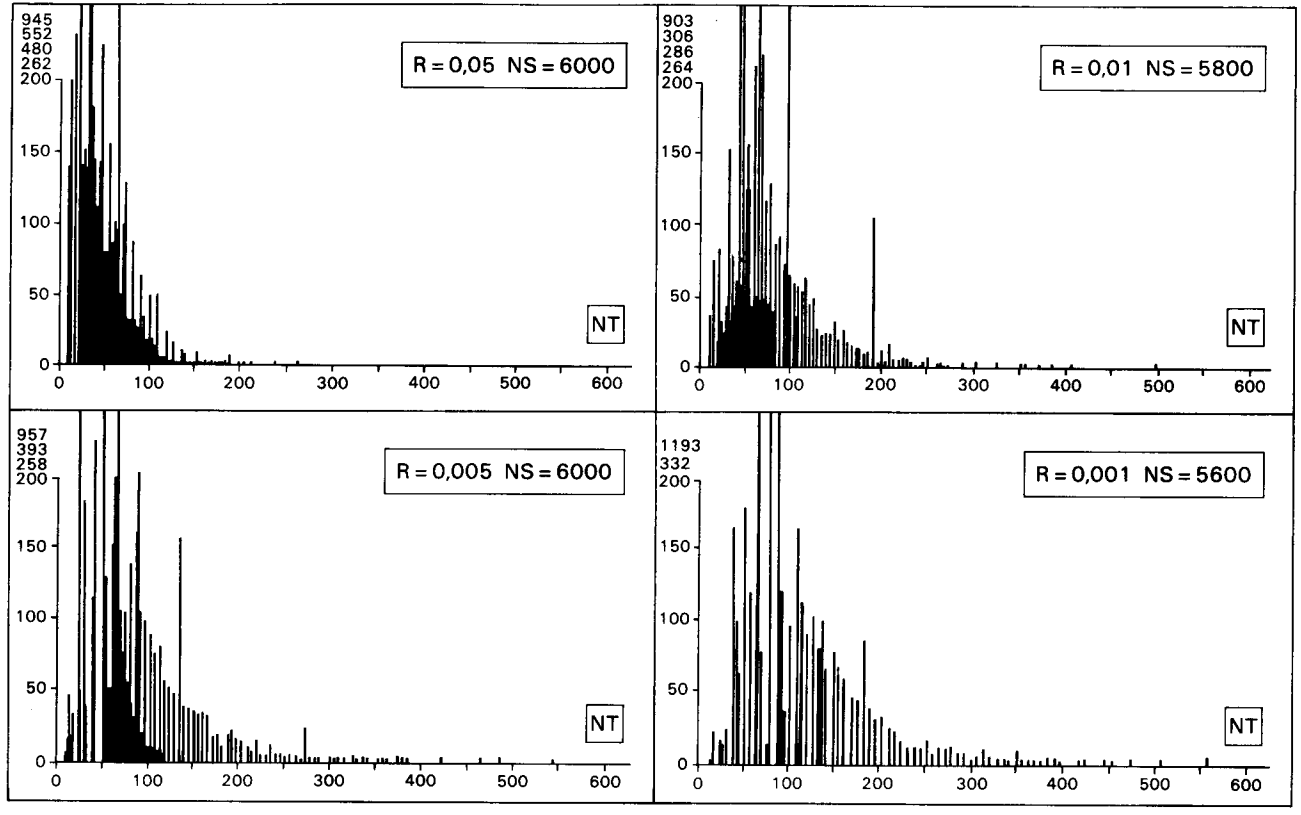


FIGURE 5

Simulation de l'algorithme d'échantillonnage à partir d'une urne de composition 0,9 : 0,1. Représentation de la distribution des répétitions (en ordonnée) des valeurs observées pour la couleur de fréquence théorique F 1 = 0,9 (en abscisse) pour un risque R = 0,05; 0,01; 0,005; 0,001 et environ NS = 6 000 répétitions

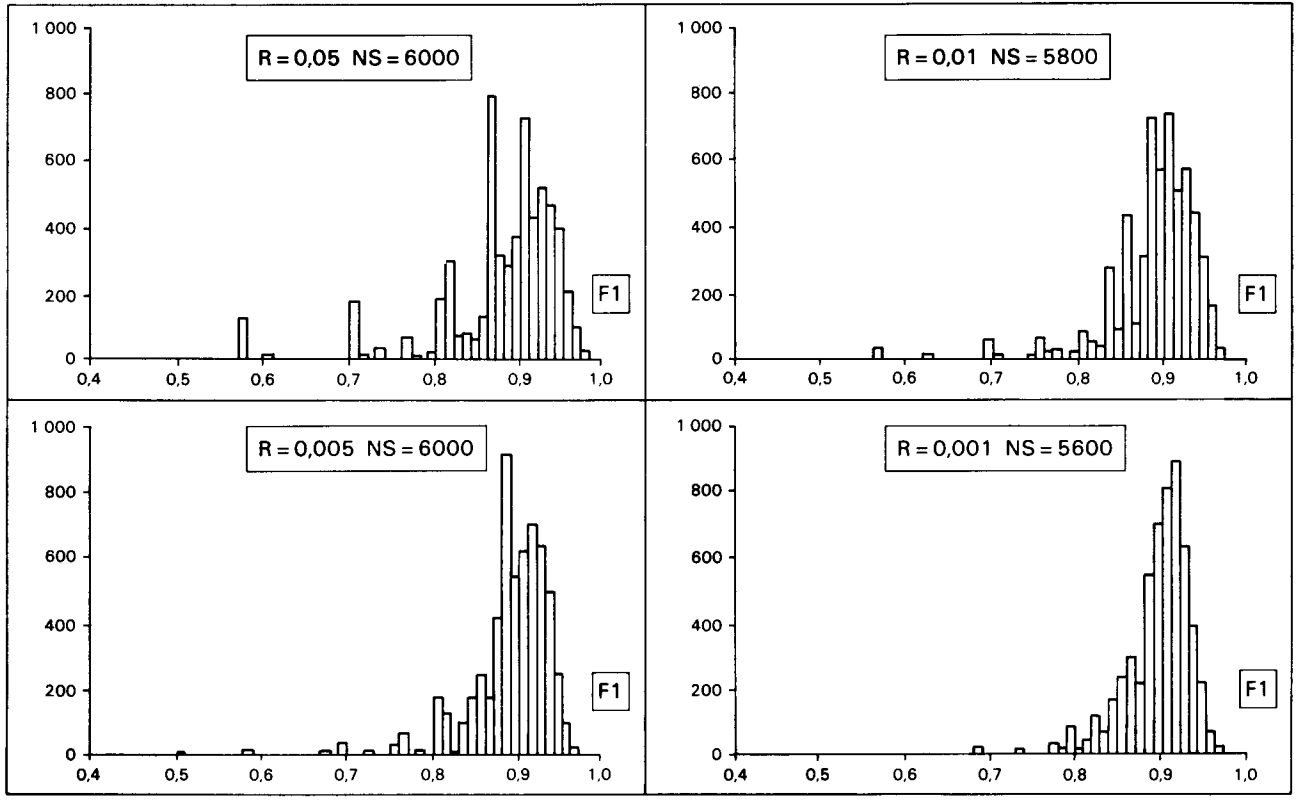
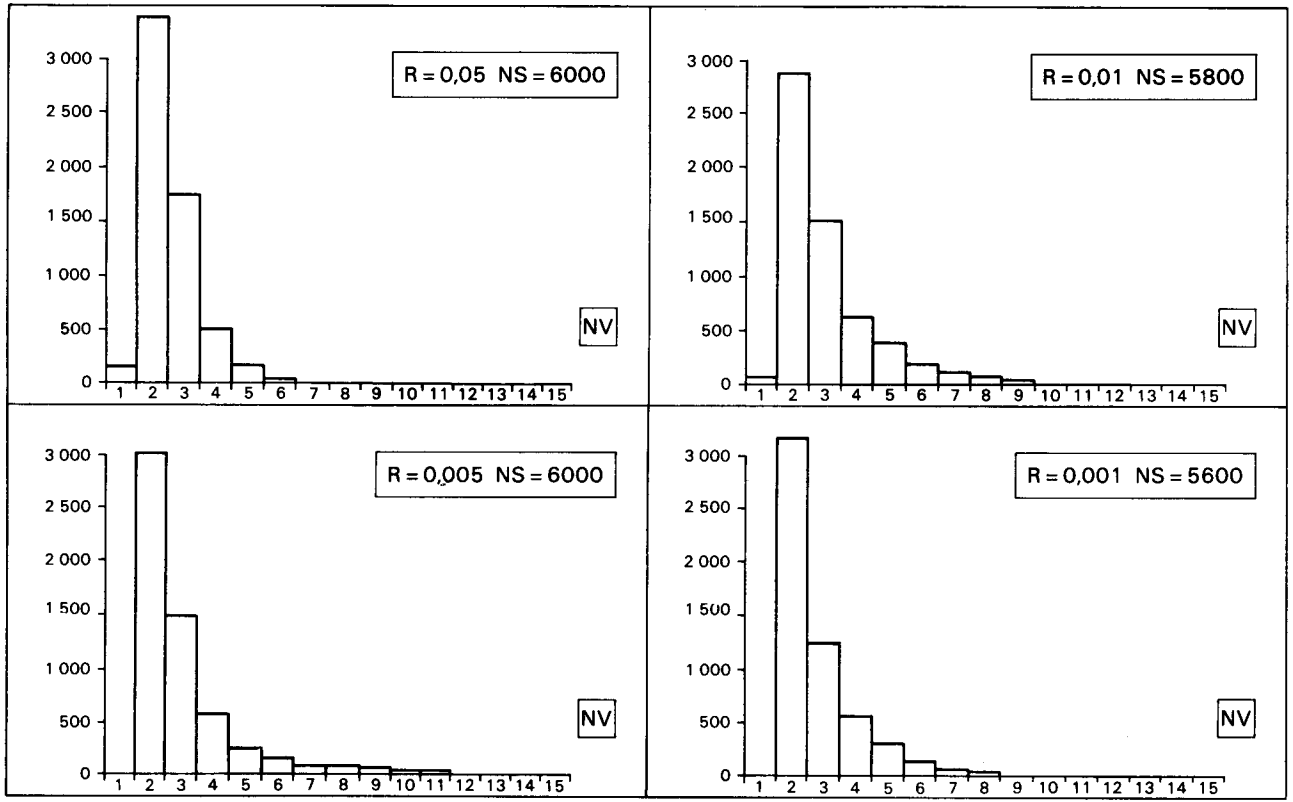


FIGURE 6

Simulation de l'algorithme d'échantillonnage à partir d'une urne de composition 0,9 : 0,1. Représentation de la distribution des répétitions (en ordonnée) du nombre NV de navettes (en abscisse) effectuées pour un risque R = 0,05; 0,01; 0,005; 0,001 et environ NS = 6 000 répétitions



de première espèce considérée dans le test statistique. L'aspect arbitraire du choix du risque de première espèce dans les tests statistiques disparaît alors pour faire place au concept de *valeur plancher du risque de première espèce.* Dans le cas de plusieurs échantillons, deux stratégies sont possibles :

a. Évaluer l'erreur d'échantillonnage relative à chaque échantillon. L'erreur d'échantillonnage relative à l'ensemble sera la plus grande valeur obtenue.

b. Fixer une valeur *a priori* de l'erreur et écarter tout échantillon dont l'erreur d'échantillonnage est supérieure à la valeur fixée.

Un grand nombre de variantes à la méthode exposée peuvent être ainsi construites. Cependant elles reprendront les mêmes caractéristiques, en particulier le principe des navettes successives. Selon les choix effectués, certaines variantes serait plus exigeantes alors que d'autres paraitront plus permissives.

Nous avons employé la notion de « risque de première espèce » sans faire explicitement référence à un test statistique. Bien qu'il s'agisse ici d'un problème d'estimation, cette notion conserve un sens dans le contexte particulier de l'estimation d'un paramètre discret. En effet, le choix de l'estimateur retenu entraîne que les réponses sont forcément inférieures ou égales à la valeur recherchée, et la fonction de perte sous-jacente au problème ne prend que les valeurs 0 et 1. Ces conditions permettent d'assimiler ce problème d'estimation très particulier à un problème de test, et autorisent la comparaison d'une « erreur d'échantillonnage » à un « risque de première espèce ».

D'un point de vue inférentiel, la méthode proposée s'apparente à la logique néo-bayésienne. Tout d'abord, le principe d'équirépartition de l'ignorance est appliqué pour définir une première composition *a priori* de l'urne; ensuite, les fréquences observées *a posteriori* après un tirage deviennent les probabilités de tirage *a priori* de la navette suivante. De nombreuses critiques ont été formulées à l'encontre des raisonnements néo-bayésiens (MATALON [1968]; HAMAKER [1977]). Cependant, le fait que ce raisonnement soit répété jusqu'à convergence de l'algorithme permet de neutraliser l'arbitraire du choix de la composition initiale.

● Références bibliographiques

AMEGANDJIN, J. (1970). — « L'admissibilité des estimateurs et la statistique d'ordre dans la théorie de l'échantillonnage sur des populations finies », *Thèse de 3e cycle*, Paris.

DESABIE, J. (1966). — *Théorie et pratique des sondages*, Dunod.

GOURIEROUX, C. (1981). — *Théorie des sondages*, Economica.

HAMAKER, H. C. (1977). — « Bayesianism; a Threat to the Statistician Profession? » *Intenational Statistical Review*, 45/2, p. 111-115.

HURLBERT, S. H. (1971). — « The Nonconcept of Species Diversity : a Critique and Alternative Parameters », *Ecology*, 52, p. 577-86.

MALATON, B. (1967). — « Epistémologie des probabilités », In *Logique et connaissance scientifique* sous la direction de Jean PIAGET, Encyclopédie de la Pléiade, p. 526-553.

PEET, R. K. (1974). — « The Measurement of Species Diversity », *Annual Review of Ecology and Systematics*, p. 295-306.