Sur Certains Problemes D'Estimation Dans les Cas de Double Echantillonnage

M. Lamotte et M. Schutzenberger

Biometrics, Vol. 7, No. 3 (Sep., 1951), 275-282.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195109%297%3A3%3C275%3ASCPDDL%3E2.0.CO%3B2-1

Biometrics is currently published by International Biometric Society.

SUR CERTAINS PROBLÈMES D'ESTIMATION DANS LES CAS DE DOUBLE ÉCHANTILLONNAGE.

M. LAMOTTE ET M. SCHUTZENBERGER

Laboratoire de Génétique, Faculté des Sciences, Paris
et Centre de Génétique, Hopital St. Louis, Paris

I. INTRODUCTION

LORSQU'UNE espèce polymorphe pour un certain caractère est répartie en colonies plus ou moins isolées, la fréquence du caractère varie d'une colonie à l'autre. Il est intéressant de déterminer la forme de la distribution de ces fréquences, mais cette détermination pose un problème statistique complexe. On ne peut étudier, en effet, qu'un nombre limité de colonies et chacune d'elles n'est connue qu'imparfaitement par l'analyse d'un échantillon prélevé sur son effectif total. Il y a donc deux échantillonnages successifs: l'un au niveau de chaque colonie, l'autre au niveau de l'ensemble de toutes les colonies de l'espèce.

Un tel problème de double échantillonnage se pose d'ailleurs dans d'autres domaines que l'étude des populations naturelles. On le rencontre notamment en Physiologie ou en Psychologie expérimentale, lorsque chaque individu d'un groupe est caractérisé par la probabilité qu'il a de répondre d'une certaine manière à un stimulus donné. Pour étudier le groupe, on ne peut évidemment analyser le comportement que d'un nombre limité d'individus, et chacun d'eux n'est soumis qu'à un nombre fini de tests.

D'autres exemples pourraient être donnés et c'est seulement pour des raisons de commodité que nous utiliserons ici le langage correspondant au cas de la distribution d'un caractère qualitatif dans les colonies d'une espèce.

Nous ferons les hypothèses suivantes:

(a) l'effectif N des colonies est le même pour toutes;

(b) l'effectif n de l'échantillon prélevé dans chacune d'elles est également partout le même.

II. LA DISTRIBUTION D'ÉCHANTILLONNAGE

Soit $f(p)$ la fonction de probabilité de la fréquence p du caractère dans l'ensemble des colonies et soit $f^*(r)$ la loi de distribution de la fréquence r du caractère dans les ν échantillons. Alors que $f(p)$ est définie pour toutes les valeurs 0, $1/N$, $2/N$, i/N, $(N-1)/N$, 1,—ou pour toute valeur de p entre 0 et 1 si N peut être considéré comme infini—, $f^*(r)$ n'est définie que pour les valeurs 0, $1/n$, $2/n$, $\cdots$ k/n, $\cdots$ $(n-1)/n$, 1; $f^*(r)$ est donc toujours groupée en classes plus larges que $f(p)$.

Soit $g(r, p)$ la probabilité pour que l'on observe une fréquence $r(=k/n)$ quand l'échantillon est prélevé dans une colonie où la fréquence est $p = i/N$. Le théorème des probabilités composées permet d'écrire:

$$\text{(II, Ia)} \qquad f^*(r) = \sum_{i=0}^{N} g(r, p) f(i/N)$$

ou

$$\text{(II, Ib)} \qquad f^*(r) = \int_0^1 g(r, p) f(p)\, dp$$

selon que N est fini ou infini.

En pratique, on n'aura à envisager pour la loi $g(r, p)$ que les cas suivants:

Cas hypergéométrique, si N n'est pas très grand par rapport à n:

$$\text{(II, 2a)} \qquad g(k/n, i/N) = \frac{(N-n)!n!i!(N-i)!}{N!k!(n-k)!(i-k)!(N-i-n+k)!}$$

Cas binomial, si N est très grand par rapport à n:

$$\text{(II, 2b)} \qquad g(k/n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Il est intéressant de remarquer que les valeurs de $f^*(0)$ et de $f^*(I)$ sont toujours au moins égales aux valeurs respectives $f(0)$ et $f(1)$, puisque l'on a toujours $g(0, 0) = 1$ et $g(1, 1) = 1$, quelle que soit la loi de probabilité $g(r, p)$. On a:

$$f^*(0) = f(0) + \sum_{i=1}^{N} g(0, i/N) f(i/N)$$

$$\text{et} \qquad f^*(1) = f(1) + \sum_{i=0}^{N-1} g(1, i/N) f(i/N)$$

On notera également que, si $f(p)$ est symétrique par rapport à la valeur $\frac{1}{2}$, il en est de même pour $f^*(r)$.

Fonction caractéristique et moments de la distribution d'échantillonnage.

La fonction caractéristique de la distribution de $f^*(r)$ est définie par:

$$\varphi_r^*(t) = \sum_{k=0}^{n} e^{(k/n)t} f(k/n)$$

On a donc:

$$\varphi_r^*(t) = \sum_{k=0}^{n} e^{k/n} \sum_{i=0}^{N} g(k/n, i/N) f(i/N)$$

(II, 3)

$$= \sum_{i=0}^{N} f(i/N) \sum_{k=0}^{n} e^{(k/n)t} g(k/n, i/N) = \sum_{i=0}^{N} \psi_{r,p}(t) f(i/N),$$

c'est-à-dire que la fonction caractéristique de $f^*(r)$ est égale à la moyenne, étendus à toutes les colonies, de la fonction caractéristique $\psi_{r,p}$ attachée à la loi du deuxième échantillonnage.

De cette relation générale entre les fonctions caractéristiques $\varphi(t)$ et $\psi(t)$ on déduit facilement l'expression des moments autour de zéro de $f^*(r)$ en fonction des mêmes moments de $f(p)$:

Cas hypergéométrique:

$$\mu_1^* = \mu_1$$

$$\mu_2^* = \frac{N(n-1)\mu_2 + (N-n)\mu_1}{n(N-1)}$$

$$\mu_3^* = \frac{N^2(n-1)(n-2)\mu_3 + 3(N-n)N(n-1)\mu_2}{n^2(N-1)(N-2)} + \frac{(N-n)(N-2n)\mu_1}{n^2(N-1)(N-2)}$$

(II, 4a)

$$\mu_4^* = \frac{N^3(n-1)(n-2)(n-3)\mu_4 + 6(N-n)N^2(n-1)(n-2)\mu_3}{n^3(N-1)(N-2)(N-3)} + \frac{N(N-n)(n-1)[7(N-2n)+3n+1]\mu_2}{n^3(N-1)(N-2)(N-3)} + \frac{[(N-n)(N-2n)(N-3n) - N(n-1)]\mu_1}{n^3(N-1)(N-2)(N-3)}$$

.....

Cas binomial:

$$\mu_1^* = \mu_1$$

$$\mu_2^* = \frac{(n-1)\mu_2 + \mu_1}{n}$$

$$\mu_3^* = \frac{(n-1)(n-2)\mu_3 + 3(n-1)\mu_2 + \mu_1}{n^2}$$

(II, 4b)

$$\mu_4^* = \frac{(n-1)(n-2)(n-3)\mu_4 + 6(n-1)(n-2)\mu_3}{n^3} + \frac{7(n-1)\mu_2 + \mu_1}{n^3}$$

.....

$$\mu_h^* = \frac{\sum_{j=1}^{h} \mathfrak{s}_h^j \frac{n!}{(n-j)!}\mu_j}{n^h},$$

où les $\mathfrak{s}_h^j$ sont les nombres de stirling de seconde espèce.

Pour $h \leq n$, les formules (II, 4a) et (II, 4b) donnent les valeurs de μ_h^* en fonction des moments μ_j d'ordre $j \leq h$. Pour $h > n$, elles donnent les valeurs de μ_h^* en fonction des moments μ_j d'ordre $j \leq n$ (avec $n < h$) seulement. En effet, seuls les n premiers moments de $f^*(r)$ sont indépendants, et ses moments d'ordre supérieur à n peuvent être exprimés en fonction des n premiers, puisque $f^*(r)$ ne prend que $n + 1$ valeurs.

Inversement, on peut aussi connaître l'expression des moments autour de zéro de la loi $f(p)$ en fonction des mêmes moments de $f^*(r)$:

Cas hypergéométrique:

$$\mu_1 = \mu_1^*$$

$$\mu_2 = \frac{n(N-1)\mu_2^* - (N-n)\mu_1^*}{N(n-1)}$$

$$\mu_3 = \frac{n^2(N-1)(N-2)\mu_3^* - 3n(N-1)(N-n)\mu_2^*}{N^2(n-1)(n-2)} + \frac{(N-n)(2N-n)\mu_1^*}{N^2(n-1)(n-2)}$$

(II, 5a)

$$\mu_4 = \frac{n^3(N-1)(N-2)(N-3)\mu_4^* - 6n^2(N-1)(N-2)(N-n)\mu_3^*}{N^3(n-1)(n-2)(n-3)}$$

$$+ \frac{n(N-1)(N-n)(11N-7n-1)\mu_2^*}{N^3(n-1)(n-2)(n-3)}$$

$$- \frac{[(N-n)(2N-n)(3N-n)-(N-n)+(n-1)N]\mu_1^*}{N^3(n-1)(n-2)(n-3)}$$

.....

Cas binomial:

$$\mu_1 = \mu_1^*$$

$$\mu_2 = \frac{n\mu_2^* - \mu_1^*}{n-1}$$

(II, 5b)
$$\mu_3 = \frac{n^2\mu_3^* - 3n\mu_2^* + 2\mu_1^*}{(n-1)(n-2)}$$

$$\mu_4 = \frac{n^3\mu_4^* - 6n^2\mu_3^* + 11n\mu_2^* - 6\mu_1^*}{(n-1)(n-2)(n-3)}$$

.....

$$\mu_h = \frac{(n-h)!}{n!} \sum_{i=1}^{h} n^i S_h^i \mu_i^*$$

où les S_h^i sont les nombres de stirling de première espèce.

Les formules (II, 5a) et (II, 5b) donnent les valeurs des μ_h en fonction des μ_i^* d'ordre $j \leq h$ lorsque $h \leq n$. Pour $h > n$, ces formules deviennent illusoires: la connaissance de la distribution $f^*(r)$ ne permet pas d'obtenir la suite complète des moments de $f(p)$, mais seulement ceux d'ordre au plus égal à n.

III. ESTIMATION DES CARACTÉRISTIQUES DE LA DISTRIBUTION DES COLONIES

L'observation fournit directement les valeurs de $f^*(r)$ pour les diverses valeurs $0, 1/n, \cdots (n-1)/n, 1$ des fréquences dans les ν échantillons de n individus. On peut alors calculer les valeurs m_h^* des moments, jusqu'à l'ordre n, de la distribution expérimentale. Par définition, ces m_h^* sont des estimations correctes (unbiased) des μ_h^* et, par conséquent, les valeurs m_h obtenues à partir des m_h^* par les formules II,5, linéaires,

seront aussi des estimations correctes des moments μ_h de la fonction de probabilité. On remarque que ces estimations sont indépendantes du nombre ν des colonies prospectées.

Il n'en serait plus de même pour des caractéristiques dont l'expression ne serait pas une fonction linéaire des μ_h^*. Mais on pourra remarquer alors que les formules II,I montrent que le double échantillonnage est équivalent à un échantillonnage simple où:

(i) la loi $f(p)$ est remplacée par la loi $f^*(r)$, et

(ii) chaque échantillon ne compte que pour *une* observation.

Il sera donc possible, en utilisant les formules classiques, de construire des estimations correctes des caractéristiques de $f(p)$.

IV. AJUSTEMENT D'UNE COURBE DE FRÉQUENCES À DES RÉSULTATS EXPÉRIMENTAUX

Pour remonter des résultats expérimentaux à la loi de probabilité $f(i/N)$, on dispose de $n + 1$ valeurs de $f^*(r)$ pour $r = 0, 1/n, \cdots (n - 1)/n$ et 1, d'où l'on déduit les moments m_0^*, m_1^*, $\cdots$ m_n^* et les estimations m_0, m_1, $\cdots$ m_n.

Il est possible, à partir de ces $n + 1$ quantités, d'ajuster une courbe $f(i/N)$ de forme donnée a priori, et dont l'expression comprenne moins de $n + 1$ paramètres.

Si l'on recherche une approximation de $f(i/N)$ par une suite de polynomes f_h de degré h (inférieur à n), on imposera naturellement, pour déterminer le polynome f_h, les $h + 1$ conditions:

$$\sum_{i=0}^{N} \left(\frac{i}{N}\right)^j f(i/N) = m_j \qquad \text{pour } j = 0, 1, 2, \cdots, n$$

ou, dans le cas binomial où N peut être considéré comme infini:

$$\int_0^1 p^j f_h(p) = m_j \qquad \text{pour } j = 0, 1, 2, \cdots, n.$$

Dans ce dernier cas, l'emploi des polynomes de legendre,—où la variable x habituelle est remplacée par $2p - 1$—, facilitera beaucoup les calculs. En effet, si l'on veut ajuster une courbe de degré h, on prendra:

$$f_h(p) = 1 + c_1P_1(p) + c_2P_2(p) + \cdots + c_iP_i(p) + \cdots + c_hP_h(p),$$

où les P_i sont les polynomes de legendre avec variable transformée $p = (1 + x)/2$ et les c_i des coefficients, dépendant des résultats expérimentaux. Le tableau qui suit donne les expressions des polynomes P_i et des coefficients c_i en fonction des moments m_i^* de la distribution observée des échantillons, pour $j \leq 6$:

$$P_1 = 2p - 1$$

$$P_2 = 6p^2 - 6p + 1$$

$$P_3 = 20p^3 - 30p^2 + 12p - 1$$

$$P_4 = 70p^4 - 140p^3 + 90p^2 - 20p + 1$$

$$P_5 = 252p^5 - 630p^4 + 560p^3 - 210p^2 + 30p - 1$$

$$P_6 = 924p^6 - 2772p^5 + 3150p^4 - 1680p^3 + 420p^2 - 42p + 1$$

$$c_1/3 = 2m_1^* - 1$$

$$c_2/5 = (6nm_2^* - 6nm_1^*)(n-1)^{-1} + 1$$

$$c_3/7 = [20n^2m_3^* - 30n^2m_2^* + (12n^2 - 6n + 4)m_1^*] \cdot [(n-1)(n-2)]^{-1} - 1$$

$$c_4/9 = [70n^3m_4^* - 140n^3m_3^* + (90n^3 - 30n^2 + 50n)m_2^* - (20n^3 - 30n^2 + 50n)m_1^*][(n-1)(n-2)(n-3)]^{-1} + 1$$

$$c_5/11 = [252n^4m_5^* - 630n^4m_4^* + (560n^4 - 140n^3 + 420n^2)m_3^* - (210n^4 - 210n^3 + 630n^2)m_2^* + (30n^4 - 90n^3 + 280n^2 - 100n + 48)m_1^*] \cdot [(n-1)(n-2)(n-3)(n-4)]^{-1} - 1$$

$$c_6/13 = [924n^5m_6^* - 2772n^5m_5^* + (3150n^5 - 630n^4 + 2940n^3)m_4^* - (1680n^5 - 1260n^4 + 5880n^3)m_3^* + (420n^5 - 840n^4 + 3990n^3 - 1050n^2 + 1176n)m_2^* - (42n^5 - 210n^4 + 1050n^3 - 1050n^2 + 1176n)m_1^*] \cdot [(n-1)(n-2)(n-3)(n-4)(n-5)]^{-1} + 1$$

V. SOMMAIRE

Nous avons été amenés, dans l'étude d'une population répartie en groupes distincts dont on ne peut observer qu'un nombre limité, et chacun seulement par un échantillon, à résoudre certains problèmes de double échantillonnage. Nous avons montré comment obtenir, à partir des résultats d'observation, des estimations correctes des moments de

la loi de distribution de la fréquence d'un caractère qualitatif dans la population.

La connaissance de ces estimations permet d'ajuster aux résultats expérimentaux une courbe de fréquences dont l'expression comporte un nombre de paramètres au plus égal à l'effectif des échantillons.

VI. SUMMARY

In a study of a population distributed in distinct groups of which only a limited number can be observed, each by a sample, it has been necessary to solve certain problems of double sampling.

A demonstration has been made of how to obtain from the results of these observations unbiased estimates of the moments of the frequency function of a character in the population.

Knowing these estimates, it is possible to fit to the experimental results a frequency curve the equation of which includes at the most a number of parameters equal to the size of a sample.