Ann. Biol. anim. Bioch. Biophys., 1963, **3** (4), 427-432

NOTE SUR UNE ESTIMATION DE FRÉQUENCE GÉNIQUE DANS UNE ÉTUDE DE GROUPES SANGUINS

F. GROSCLAUDE et L. OLLIVIER

Station centrale de Génétique animale,
Centre national de Recherches zootechniques, Jouy-en-Josas (Seine-et-Oise)

SOMMAIRE

Dans leur étude des groupes sanguins de la race bovine *Montbéliarde*, GROSCLAUDE et MILLOT (1962) ont été amenés à poser, entre autres, le problème suivant : soit, dans une population, représentée par un échantillon de N femelles, un locus A, caractérisé par deux allèles A^A et A^a, le premier étant dominant, de sorte que le phénotype [A] représente les deux catégories génotypiques A^A/A^A et A^A/A^a ; on se propose d'estimer la fréquence q_a du gène A^a dans la population, et la variance de cette estimation, sans poser l'hypothèse de panmixie mais sachant que chaque femelle, croisée avec un mâle hétérozygote, a donné un produit de phénotype connu.

La présente note a pour but de justifier la solution qu'ont donnée de ce problème GROSCLAUDE et MILLOT sans démonstration.

Le principe de la méthode de résolution est d'utiliser à la fois l'information fournie par le nombre de mères homozygotes récessives, et celle qu'apporte le nombre de mères de phénotype [A] ayant donné un produit homozygote récessif.

Si n et r sont les nombres respectifs de mères formant ces deux catégories, l'application de la méthode du maximum de vraisemblance conduit aux résultats suivants :

$$\hat{q}_a = \frac{n + 2r}{N} \qquad Var(\hat{q}_a) = \frac{1}{N^3}\left[n(N-n) + 4r(N-r-n)\right]$$

L'hypothèse de panmixie est ensuite testée, en comparant les nombres observés de mères des trois catégories, n, r, $N-r-n$, aux valeurs théoriques correspondantes :

$$N\hat{q}^2, \qquad \frac{2}{N}\hat{q}(1-\hat{q}), \qquad N\left[(1-\hat{q})^2 + \frac{3}{2}(1-\hat{q})\right]$$

INTRODUCTION

Sauf dans certains cas comme celui du locus FV, l'estimation de la fréquence des gènes de groupes sanguins dans une population bovine ne peut se faire par simple comptage des gènes, car ceux-ci ne sont pas régulièrement identifiables, un même phénotype pouvant résulter de diverses combinaisons génotypiques.

Les auteurs qui se limitent, pour résoudre le problème, à l'échantillon de référence, sont amenés à poser l'hypothèse de panmixie. Au contraire, dans leur étude des groupes sanguins de la race *Montbéliarde*, GROSCLAUDE et MILLOT (1962) utilisent l'information supplémentaire apportée par les produits des mères composant l'échantillon.

En pratique, la méthode suivante a été utilisée : l'échantillon étudié est constitué par un certain nombre de vaches pour lesquelles il a été possible de trouver un descendant issu d'un taureau homozygote récessif pour le maximum de loci et en particulier pour le locus B.

Cependant, le taureau retenu n'était pas homozygote récessif à tous les loci, il était en particulier hétérozygote à trois loci peu complexes ; dans ces trois cas, le problème d'estimation se pose en termes identiques ; dans le cas du locus A, il s'énonce comme suit :

Soit un échantillon de N femelles prises au hasard dans la population. Les techniques sérologiques permettent de les classer en deux catégories : celles dont le phénotype est [A], et dont le génotype est alors soit A^A/A^A, soit A^A/A^a, et celles dont le phénotype est [*a*], donc le génotype A^a/A^a. Soit *n* le nombre de femelles [*a*]. Les N femelles, croisées avec un mâle hétérozygote, ont donné chacune un produit ; on connait le phénotype de ces produits. Soit *r* le nombre de mères [A] ayant donné un produit [*a*].

On se propose d'estimer la fréquence q_a du gène A^a dans la population d'origine et d'estimer la variance de cette estimation.

On veut ensuite tester l'hypothèse de panmixie.

SOLUTION DU PROBLÈME

a) *Estimation de la fréquence* q_a

L'échantillon observé est composé de trois catégories d'animaux :

— *n* femelles A^a/A^a

— *r* femelles [A] qui, accouplées à un mâle A^A/A^a ont donné un produit A^a/A^a

— N-*n*-*r* femelles [A] qui, accouplées à un mâle A^A/A^a ont donné un produit [A].

Si l'on désigne par R, 4 Q et P les probabilités respectives des trois génotypes A^a/A^a, A^A/A^a et A^A/A^A, on voit que les trois catégories observées ont des probabilités théoriques d'apparition de R, Q, 1-Q-R. Remarquons que l'échantillon obéit en fait à une loi trinomiale. La probabilité pour un gamète de la population de porter A^a est :

$$q = R + 2Q$$

L'estimation optimum conjointe de R et Q est celle qui rend maximum l'équation de vraisemblance. Le logarithme de cette équation s'écrit dans le cas présent :

$$\log L = n \log R + r \log Q + (N\text{-}n\text{-}r) \log (1\text{-}R\text{-}Q).$$

Les deux équations d'estimation de R et Q sont :

$$\frac{\partial \log L}{\partial R} = \frac{n}{R} - \frac{N\text{-}n\text{-}r}{1\text{-}R\text{-}Q} = 0$$

$$\frac{\partial \log L}{\partial Q} = \frac{r}{Q} - \frac{N\text{-}n\text{-}r}{1\text{-}R\text{-}Q} = 0$$

qui peuvent s'écrire :

$$\frac{N\text{-}n\text{-}r}{1\text{-}R\text{-}Q} = \frac{n}{R} = \frac{r}{Q} = \frac{N}{1}$$

$$\hat{Q} = \frac{r}{N} \qquad \hat{R} = \frac{n}{N}$$

L'estimation optimum de q est :

$$\hat{q} = \hat{R} + 2\hat{Q} = \frac{n + 2r}{N}.$$

Remarquons que l'on retrouve ce résultat en faisant le raisonnement suivant : la fréquence du gène A^a est égale à la somme, rapportée à l'effectif total, du nombre des homozygotes A^a/A^a et de la moitié du nombre des hétérozygotes ; or, puisqu'une mère hétérozygote donne un produit homozygote récessif avec la probabilité 1/4, on peut estimer à $4r$ le nombre de mères hétérozygotes ; donc $\hat{q} = \frac{n + 2r}{N}$.

b) *Variance de l'estimation de q_a*

La variance de $\hat{q}$ se déduit de celles de $\hat{Q}$ et $\hat{R}$, et de leur covariance :

$$\text{Var}\,(\hat{q}) = 4\,\text{Var}\,(\hat{Q}) + \text{Var}\,(\hat{R}) + 4\,\text{Cov}\,(\hat{Q}, \hat{R})$$

Dans la méthode du maximum de vraisemblance, les variances et covariances s'obtiennent à partir des dérivées secondes de l'équation de vraisemblance.

Dans le cas d'un seul paramètre estimé θ, la variance de l'estimation est donnée par l'expression suivante (le symbole E désignant l'espérance mathématique) :

$$-\frac{1}{E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)}$$

Dans le cas présent, avec deux paramètres, nous avons 4 dérivées secondes, qui peuvent se disposer en une matrice de 2×2. Les variances et les covariances sont les espérances mathématiques des éléments de la matrice inverse changés de signe.

Les calculs donnent : $\text{Var.}\,(\hat{R}) = \frac{R\,(1\text{-}R)}{N}$ $\text{Var.}\,(\hat{Q}) = \frac{Q\,(1\text{-}Q)}{N}$ $\text{Cov.}\,(\hat{Q}, \hat{R}) = \frac{QR}{N}$.

Ces variances et covariances peuvent aussi se déduire plus simplement des propriétés bien connues de la loi multinomiale (KEMPTHORNE, 1957).

En remplacant Q et R dans les expressions des variances et des covariances par $\hat{Q}$ et $\hat{R}$, on obtient une valeur approchée de l'estimation de q :

$$\text{Var}\,(\hat{q}) = \frac{1}{N^3}\left[n\,(N\text{-}n) + 4r\,(N\text{-}r\text{-}n)\right].$$

c) *Test de l'hypothèse de panmixie*

La connaissance de q permet maintenant de tester l'hypothèse de panmixie. Pour cela il faut déterminer si les probabilités P, 4Q et R peuvent être regardées

comme étant de la forme p^2, $2pq$, q^2, ce qui exige que l'on compare par un test de χ^2

$$N\text{-}r\text{-}n \text{ à } N\left[(1-\hat{q})^2 + \frac{3}{2}\hat{q}(1-\hat{q})\right]$$

$$r \text{ à } \frac{N}{2}\hat{q}(1-\hat{q})$$

$$n \text{ à } N\hat{q}^2$$

Le nombre de degrés de liberté est 1 puisque l'on observe 3 catégories d'individus et que 2 relations doivent être vérifiées, à savoir :

$$\text{Effectif total} = N$$
$$N\hat{q} = n + 2r$$

d) *Calculs numériques*

Les valeurs numériques obtenues par GROSCLAUDE et MILLOT sont les suivantes :

$$N = 400 \qquad n = 124 \qquad r = 48$$

On trouve : $\hat{q} = 0{,}550$

$$\sqrt{\text{Var}(\hat{q})} = 0{,}035$$

Effectifs observés	Effectifs calculés
228	229,5
48	49,5
124	121
$\chi^2 = 0{,}13$	$0{,}50 < P(\chi^2) < 0{,}75$

L'hypothèse de panmixie est donc parfaitement recevable.

DISCUSSION

Une méthode fréquemment utilisée pour les estimations de fréquences géniques consiste à prendre la racine carrée de la fréquence des homozygotes récessifs. C'est le deuxième cas envisagé par COTTERMAN (1954) dans sa revue des divers problèmes d'estimation que pose l'étude des populations soumises à échantillonnage.

Dans le cas présent, cette estimation conduirait à :

$$\hat{q}_a = \sqrt{\frac{n}{N}}$$

avec

$$\text{Var}(\hat{q}_a) = \frac{1 - \hat{q}_a^2}{4N}$$

Cependant l'emploi de cette méthode est critiquable dans le problème qui nous intéresse. D'une part, elle n'utilise pas l'information dont nous disposons sur les produits des mères. D'autre part, elle pose, a priori, l'hypothèse de panmixie. Si cette hypothèse n'est pas vérifiée, l'estimation de la fréquence sera biaisée. Nous

présentons, en annexe, une solution valable dans l'hypothèse de panmixie, qui tient compte de toute l'information disponible.

Dans beaucoup d'études de groupes sanguins, les auteurs testent d'abord l'hypothèse de panmixie au locus FV, où les fréquences géniques peuvent être déterminées par simple comptage ; l'hypothèse étant vérifiée à ce locus, on suppose qu'elle l'est également à tous les autres. Cependant, ce mode d'extrapolation est purement arbitraire et il semble préférable, dès que cela est possible, d'utiliser d'autres méthodes, dont celle que nous venons de développer est un exemple.

Notons que si nous avons raisonné dans le cas d'un locus de groupes sanguins, cette méthode est naturellement applicable à toutes les situations où l'on considère un locus biallélique avec dominance d'un allèle ; elle s'applique aussi, que les unités de l'échantillon soient croisées avec un individu unique (le taureau dans le cas présent) ou avec des individus différents, mais tous hétérozygotes.

ANNEXE

Cas d'une population panmictique

Si on peut admettre, a priori, que la population échantillonnée est panmictique, l'application de la méthode du maximum de vraisemblance donne les résultats suivants :

Soient $(1-q)^2$, $2q(1-q)$, q^2 les probabilités respectives dans la population des 3 génotypes A^A/A^A, A^A/A^a, A^a/A^a. Les 3 catégories observées ont alors les probabilités d'apparition :

q^2 pour les individus A^a/A^a

$\frac{q(1-q)}{2}$ pour les individus [A] ayant donné un produit A^a/A^a

$(1-q)^2 + \frac{3q(1-q)}{2}$ pour les individus [A] ayant donné un produit [A]

$$\log L = (N-n-r)\left[\log(1-q) + \log\left(1+\frac{q}{2}\right)\right] + r\left[\log q + \log(1-q)\right] + 2n\log q + C^{te}$$

$$\frac{\partial \log L}{\partial q} = -\frac{N-n}{1-q} + \frac{2n+r}{q} + \frac{N-n-r}{2+q}$$

L'équation $\frac{\partial \log L}{\partial q} = 0$ est du second degré. La racine positive constitue la solution cherchée :

$$\hat{q} = \frac{-(N+n+2r) + \sqrt{(N+n+2r)^2 + 16N(r+2n)}}{4N}$$

La variance de $\hat{q}$ est donnée par $\frac{\partial^2 \log L}{\partial q^2}$ en fonction de q.

Le calcul numérique donne $\hat{q} = 0{,}5560$; $\sqrt{\mathrm{Var}(\hat{q})} = 0{,}020$. Il est intéressant de comparer l'efficacité de cette estimation à celle de $q' = \sqrt{\frac{n}{N}}$

L'estimation que donne cette dernière méthode est très voisine ($\hat{q}' = 0{,}557$), mais son écart-type est double ($\sqrt{\text{Var}\,(\hat{q}')} = 0{,}042$).

Reçu pour publication en octobre 1963.

REMERCIEMENTS

Nous tenons à remercier le Pr G. MALÉCOT, de l'aide qu'il a bien voulu nous apporter dans l'élaboration de cette note.

SUMMARY

NOTE ON AN ESTIMATION OF GENE FREQUENCY IN A STUDY OF BLOOD GROUPS

In the course of their study of the blood-groups in the *Montbéliard* cattle breed, GROSCLAUDE and MILLOT (1962) encountered among others, the following problem : take, in a population represented by a sample of N females, a locus A, with two alleles A^A and A^a, the former being dominant, so that the phenotype [A] represents the two genotypes A^A/A^A and A^A/A^a ; it is proposed to estimate the frequency q_a of the gene A^a in the population, and the variance of this estimation without assuming the hypothesis of random mating, but knowing that each female crossed with a heterozygous male gave an offspring of a known phenotype.

This note is to justify the solution, previously offered without a demonstration by GROSCLAUDE and MILLOT, of this problem.

The principle of the method of resolution is to use both the information provided by the number of homozygous recessive mothers and that furnished by the number of mothers of phenotype [A] having given a homozygous recessive offspring.

If n and r are the respective numbers of mothers forming these two categories, the application of the maximum likelihood method gives the following results :

$$\hat{q}_a = \frac{n + 2r}{N} \qquad Var\,(\hat{q}_a) = \frac{1}{N^3}\left[n\,(N\text{-}n) + 4r\,(N\text{-}r\text{-}n)\right]$$

The hypothesis of random mating is then tested by comparing the observed numbers of mothers of the three categories, n, r, $N\text{-}r\text{-}n$, with the corresponding theoretical values :

$$N\hat{q}^2 \qquad \frac{N}{2}\,\hat{q}\,(1\text{-}\hat{q}) \qquad N\left[(1\text{-}\hat{q})^2 + \frac{3}{2}\,q\,(1\text{-}\hat{q})\right]$$

RÉFÉRENCES BIBLIOGRAPHIQUES

COTTERMAN C. W., 1954. Estimation of gene frequencies in nonexperimental populations. In KEMPTHORNE O., BANCROFT T. A., GOWEN J. W., LUSH J. L., *Statistics and Mathematics in Biology*, 449-465, Iowa State College Press, Ames.

FISHER R. A., 1958. *Statistical Methods for Research Workers*, 299-335, 13e édition, Oliver and Boyd, Edinburgh, London.

GROSCLAUDE F., MILLOT P., 1962. Contribution à l'étude des groupes sanguins de la race bovine *Montbéliarde*. *Ann. Biol. anim. Bioch. Biophys.*, 2, 185-208.

KEMPTHORNE O., 1957. *An Introduction to Genetic Statistics*, 21-22, 164-181, John Wiley and Sons, New York.