Ann. Génét. Sél. anim., 1971, **3** (4), 463-477.

CALCUL *A PRIORI*, INTRA- ET INTER-POPULATIONS, DES VARIANCES ET COVARIANCES GÉNOTYPIQUES ENTRE APPARENTÉS QUELCONQUES

C. CHEVALET

Laboratoire de Génétique cellulaire,
Centre de Recherches de Toulouse, I. N. R. A.,
31 - Castanet-Tolosan

RÉSUMÉ

La distinction, dans une population d'effectif limité, entre les espérances mathématiques prises par rapport à la distribution des fréquences alléliques, et par rapport à la distribution des structures d'identité et de non-identité possibles ; l'interprétation des coefficients d'identité au niveau d'une population unique ; l'hypothèse qu'un caractère quantitatif est gouverné par un grand nombre de gènes indépendants ayant des contributions homogènes, aboutissent à des formules nouvelles de variances et de covariances génotypiques entre apparentés quelconques.

Alors que les formules de GILLOIS ne s'appliquent que si l'on dispose d'un grand nombre de réplications du même pedigree, ce qui est rare dans la pratique, les nouvelles expressions sont valables pour tout couple d'individus consanguins et apparentés au sein d'une population unique, et peuvent donc déboucher sur des applications en amélioration des plantes et en génétique animale.

INTRODUCTION

La forme théorique des variances et covariances entre des individus apparentés a évolué lentement depuis les premiers travaux de FISHER et de WRIGHT. FISHER (1918) a calculé les corrélations pour des parentés simples ; dans le cas d'additivité, WRIGHT (1921-1922) les a étendues avec l'usage du « coefficient of inbreeding » et de la théorie des « path-coefficients ». MALÉCOT (1946-1948) a introduit les notions probabilistes dans ce domaine, en définissant les coefficients de consanguinité et de parenté. De nombreux auteurs ont peu à peu généralisé les résultats de MALÉCOT aux cas de caractères gouvernés par plusieurs loci, avec epistasis et linkage, mais toujours en excluant la consanguinité. GILLOIS (1964, 1965, 1966 *a*, *b*) a introduit la relation d'identité entre les gènes, les situations et les coefficients d'identité et

résolu le problème dans le cas d'un caractère avec dominance, gouverné par un ou plusieurs loci indépendants dans une population diploïde et pour des degrés quelconques de parenté et de consanguinité. HARRIS (1964) a obtenu le même résultat par une méthode qui s'apparente plus aux travaux de KEMPTHORNE (1954) et de TRUSTRUM (1960). A la suite de GILLOIS, BOUFFETTE (1966) a étendu le résultat aux espèces tétraploïdes ; GALLAIS (1970) au cas d'un caractère gouverné par plusieurs loci épistatiques et soumis au linkage, dans une population diploïde. Ces derniers travaux traitent toujours d'un grand nombre de populations analogues, ils ne permettent pas l'étude d'une population unique et réelle.

En reprenant les calculs de GILLOIS, et en exploitant les propriétés des coefficients d'identité nous obtenons de nouvelles formules qui, sous des hypothèses qui seront précisées, concernent une population particulière.

I. — LE MODÈLE ET LES HYPOTHÈSES GÉNÉTIQUES DU CALCUL

A. — *Le modèle mathématique*

Le caractère quantitatif est gouverné par un ensemble de loci indépendants. Chaque locus (noté α) a une contribution Z^α, qui est décomposée en :

$$Z^\alpha = X^\alpha + X'^\alpha + D^\alpha$$

où X^α et X'^α sont les contributions additives des gènes, D^α le résidu de dominance. Ces quantités sont définies comme chez FISCHER dans une population panmictique où tous les individus sont indépendants, sans parenté ni consanguinité.

Comme chez GILLOIS, nous noterons avec l'indice p (E_p) les espérances mathématiques prises dans ces conditions d'indépendance. L'indice supérieur (q) précisera qu'il s'agit d'une espérance prise par rapport à la distribution des fréquences alléliques dans une population donnée. Par convention les variables aléatoires sont centrées dans une population panmictique infinie en équilibre.

$$E_p^q(X^\alpha) = E_p^q(D^\alpha) = E_p^q(Z^\alpha) = 0 \; ; \; E_p^q(X^\alpha D^\alpha) = 0$$

B. — *Les lois conjointes des variables aléatoires génotypiques*

Dans une population infinie, en équilibre panmictique, la loi d'une variable aléatoire génotypique Z^α est définie par la connaissance des probabilités (q_i^α) des classes d'isoaction (A_i) (allèles) au locus (α), et la loi de Hardy-Weinberg : chacun des gènes présents au locus (α) a, indépendamment des autres, la probabilité q_i^α d'être l'allèle A_i^α.

Dans une population où deux gènes distincts peuvent être identiques par descendance mendélienne (sans mutation), la loi de Hardy-Weinberg ne permet plus de déterminer les fréquences des génotypes d'après les fréquences alléliques. L'hypothèse faite par GILLOIS, généralisant MALÉCOT, est que la distribution conjointe des génotypes de plusieurs individus est définie si l'on connaît d'une part la distribution des fréquences alléliques dans la population, et d'autre part la situation d'identité réalisée entre les gènes de ces individus.

Alors chaque groupe de gènes identiques a indépendamment des autres groupes, la probabilité q_i^α d'être formé d'allèles A_i^α.

Ainsi par exemple, si quatre gènes G_1, G_2, G_3 et G_4 d'un locus où il y a deux classes d'isoaction A_1 et A_2, de fréquences respectives q_1 et $q_2 = 1 - q_1$ dans la population, sont dans la situation d'identité S_2 :

$$(G_1 \equiv G_2 \equiv G_3) \not\equiv (G_4),$$

les génotypes possibles sont :

$(G_1 \equiv G_2 \equiv G_3)\,\varepsilon A_1$	et $(G_4)\,\varepsilon A_1$	avec la probabilité	$(q_1)^2$
$(G_1 \equiv G_2 \equiv G_3)\,\varepsilon A_1$	et $(G_4)\,\varepsilon A_2$	— —	$q_1 q_2$
$(G_1 \equiv G_2 \equiv G_3)\,\varepsilon A_2$	et $(G_4)\,\varepsilon A_1$	— —	$q_2 q_1$
$(G_1 \equiv G_2 \equiv G_3)\,\varepsilon A_2$	et $(G_4)\,\varepsilon A_2$	— —	$(q_2)^2$

Cela suppose que l'on fait l'hypothèse, vérifiée dans une population « isogamique » (MALÉCOT, 1939), de l'équivalence entre l'identité et la dépendance stricte des gènes isoactifs. Cette restriction a été levée par GILLOIS (1966 *c*, 1967 *a*, *b*) avec les notions de contrainte, de lois jointes, et de génotypes généralisés. Les situations d'identité et l'hypothèse d'équivalence entre non-identité et indépendance définissent une classe particulière de lois jointes, que nous avons retenue dans le modèle.

C. — *Les variables aléatoires indicatrices de situation d'identité et les coefficients d'identité*

Un couple d'individus diploïdes, I et J, étant donné, dans une population particulière, et un locus (α) étant désigné, une seule des quinze situations d'identité possibles entre les quatre gènes est réalisée, mais est inconnue. Connaissant la généalogie de la population, on peut définir la probabilité qu'une situation S_k soit réalisée : c'est le coefficient d'identité δ_k associé à la situation S_k dans une généalogie. En ce sens un coefficient d'identité est un attribut de la généalogie et non d'une population particulière ; il peut s'interpréter comme la limite de la fréquence avec laquelle une situation d'identité apparaît dans l'ensemble de N populations, $\omega_1, \omega_2, \ldots, \omega_n$, possédant la même généalogie, quand N croît indéfiniment.

Pour caractériser l'état d'une population particulière (ω) décrite par la relation d'identité nous introduisons des indicateurs de situation d'identité : un couple (I, J), et un locus (α) étant désignés, on définit les indicateurs $\Delta_k^\alpha(\omega)$ par :

$\Delta_k^\alpha(\omega) = 1$ si les gènes (α) du couple (I, J) sont, dans la population (ω), en situation S_k.

Et

$$\Delta_k^\alpha(\omega) = 0 \text{ sinon } (k = 1, 2, \ldots, 15).$$

Si la généalogie est fixée, et si l'on considère toutes les populations (ω) vérifiant cette généalogie, les quantités $\Delta_k^\alpha(\omega)$ sont les réalisations d'une variable aléatoire, définie sur l'espace de toutes les structures d'identité compatibles avec la généalogie. La loi de cette variable aléatoire est définie par les coefficients d'identité δ_k :

$$P_r(\Delta_k^\alpha = 1) = \delta_k$$
$$P_r(\Delta_k^\alpha = 0) = 1 - \delta_k, \text{ pour tout locus } \alpha.$$

On peut écrire que le coefficient δ_k est l'espérance, prise par rapport à l'espace

des réalisations (ω) possibles de toute variable Δ_k^α : $\delta_k = E^\omega(\Delta_k^\alpha(\omega))$ quel que soit α ; et que δ_k est une limite de fréquence :

$$\delta_k = \lim_{N \to \infty} \frac{1}{N} (\Delta_k^\alpha(\omega_1) + \Delta_k^\alpha(\omega_2) + \ldots + \Delta_k^\alpha(\omega_N))$$

pour N réalisations $\omega_1, \omega_2, \ldots, \omega_N$ de la généalogie.

Mais une autre interprétation des coefficients d'identité peut être donnée, au niveau d'une population particulière (ω). Si $\alpha_1, \alpha_2, \ldots, \alpha_L$ sont L loci ségrégeant indépendamment les uns des autres, les variables indicatrices $\Delta_k^{\alpha_1}, \Delta_k^{\alpha_2}, \ldots, \Delta_k^{\alpha_L}$ sont indépendantes et équidistribuées.

Les réalisations $\Delta_k^{\alpha_1}(\omega), \Delta_k^{\alpha_2}(\omega), \ldots, \Delta_k^{\alpha_L}(\omega)$ dans une population particulière (ω) représentent alors L réalisations d'une même variable aléatoire, $\Delta_k(\omega)$, définie sur l'espace de tous les loci indépendants imaginables, et de loi :

$$P_r(\Delta_k(\omega) = 1) = \delta_k$$
$$P_r(\Delta_k(\omega) = 0) = 1 - \delta_k, \text{ quel que soit } (\omega)$$

et δ_k peut s'interpréter comme l'espérance des $\Delta_k(\omega)$ par rapport à l'espace des loci indépendants (α) :

$$\delta_k = E^\alpha(\Delta_k^\alpha(\omega)), \text{ quel que soit } (\omega),$$

et comme limite de fréquence :

$$\delta_k = \lim_{L \to \infty} \frac{1}{L} (\Delta_k^{\alpha_1(\omega)} + \Delta_k^{\alpha_2(\omega)} + \ldots + \Delta_k^{\alpha_L}(\omega))$$

pour L loci $\alpha_1, \alpha_2, \ldots, \alpha_L$ indépendants.

L'usage des coefficients d'identité dans leur première interprétation conduit à des expressions qui sont caractéristiques de la généalogie, et s'interprètent statistiquement comme des moyennes sur un grand nombre de populations vérifiant cette généalogie : telles sont les formules de GILLOIS. La seconde interprétation nous conduit, dans certaines conditions, à des expressions caractéristiques d'une population particulière (ω).

D. — *Propriétés des variables aléatoires indicatrices de situation d'identité*

Il résulte immédiatement des définitions que :

$$\Delta_k^\alpha(\omega) \cdot \Delta_k^\alpha(\omega) = \Delta_k^\alpha(\omega), \text{ quel que soit } (\omega, \alpha, k),$$

et que :

$$\Delta_k^\alpha(\omega) \cdot \Delta_l^\alpha(\omega) = 0 \text{ si } k \neq l$$

On définit de même les variables $F_I^\alpha(\omega)$, indicatrices de l'identité des gènes du locus α, dans l'individu I, pour la population (ω).

Comme entre probabilités on a les relations :

$$F_I^\alpha(\omega) = \Delta_1^\alpha(\omega) + \Delta_2^\alpha(\omega) + \Delta_3^\alpha(\omega) + \Delta_6^\alpha(\omega) + \Delta_7^\alpha(\omega).$$
$$F_J^\alpha(\omega) = \Delta_1^\alpha(\omega) + \Delta_4^\alpha(\omega) + \Delta_5^\alpha(\omega) + \Delta_8^\alpha(\omega) + \Delta_6^\alpha(\omega).$$

D'où il résulte la formule du produit :

$$F_I^\alpha(\omega) \cdot F_J^\alpha(\omega) = \Delta_1^\alpha(\omega) + \Delta_6^\alpha(\omega). \quad (1)$$

On posera aussi, par analogie avec la relation entre le coefficient de parenté φ_{IJ} du couple (I, J) et les coefficients d'identité :

$$\varphi_{IJ}^\alpha(\omega) = \Delta_1^\alpha(\omega) + \frac{1}{2}(\Delta_2^\alpha(\omega) + \Delta_3^\alpha(\omega) + \Delta_4^\alpha(\omega) + \Delta_5^\alpha(\omega) + \Delta^\alpha(\omega) + \Delta_{12}^\alpha(\omega)) + 1/4\ (\Delta_{10}^\alpha(\omega) + \Delta_{11}^\alpha(\omega) + \Delta_{13}^\alpha(\omega) + \Delta_{14}^\alpha(\omega)).$$

II. — L'EXPRESSION DES VARIANCES ET COVARIANCES GÉNOTYPIQUES THÉORIQUES DANS UNE POPULATION PARTICULIÈRE

Soit (I, J) un couple d'individus dans une population particulière (ω). En chacun des loci (α) qui contribuent au caractère étudié, les gènes du couple sont dans une situation d'identité $S^\alpha(\omega)$. Pour homogénéiser les notations chaque locus (α) est décrit par l'ensemble des quinze indicatrices de situation d'identité

$$\Delta_k^\alpha(\omega)(k = 1,2, \ldots, 15).$$

Les espérances mathématiques sont prises par rapport à la distribution des fréquences alléliques (q), et sont conditionnées par la réalisation particulière (ω) de la généalogie.

A. — *L'expression générale de la covariance théorique entre apparentés*

Soit Z la variable aléatoire génotypique totale :

$$Z = \sum_{\alpha=1}^{\alpha=L} Z^\alpha$$

Les loci étant indépendants, la covariance entre Z_I et Z_J est la somme des covariances entre les variables Z_I^α et Z_J^α. Il suffit de calculer cette covariance, conditionnée par la réalisation (ω), au locus (α) :

$$\mathrm{Cov}^{q/\omega}(Z_I^\alpha Z_J^\alpha) = E^{q/\omega}(Z_I^\alpha Z_J^\alpha) - E^{q/\omega}(Z_I^\alpha) \cdot E^{q/\omega}(Z_J^\alpha).$$

Or on a :

$$E^{q/\omega}(Z_I^\alpha) = F_I^\alpha(\omega)\ E_c^q(D^\alpha).$$

$$E^{q/\omega}(Z_I^\alpha Z_J^\alpha) = \sum_{k=1}^{k=15} \Delta_k^\alpha(\omega)\ E_k^q(Z_I^\alpha Z_J^\alpha)$$

où l'indice (c) exprime que l'espérance est prise sous la condition d'identité des gènes ; E_k^q est l'espérance conditionnée par la réalisation de la situation S_k entre I et J au locus considéré ; une seule des quinze quantités $\Delta_k^\alpha(\omega)$ est non nulle et égale à 1.

Les expressions des E_k^q sont données par GILLOIS (1964. p. 72 ou 1965, p. 61 corrigé par 1966-*b*).

Donc :

$$\mathrm{Cov}^{q/\omega}\,(Z_I^\alpha Z_J^\alpha) = \sum_{k=1}^{k=15} \Delta_k^\alpha(\omega)\ E_k^q(Z_I^\alpha Z_J^\alpha) - F^\alpha(\omega)\ F_J^\gamma(\omega)\ (E_c^q(D^\alpha))^2$$

Compte tenu de la relation (1) et des expressions des E_k^q, cette covariance s'écrit :

$$\begin{aligned}\mathrm{Cov}^{q/\omega}\,(Z_I^\alpha Z_J^\alpha) = {} & 2\varphi_{IJ}^\alpha(\omega)\ (2E^q((X^\alpha)^2)) \\ & + (\Delta_9^\alpha(\omega) + \Delta_{12}^\alpha(\omega)) \cdot E_p^q((D^\alpha)^2) \\ & + \Delta_1^\alpha(\omega) \cdot (E_c^q((D^\alpha)^2) - (E_c^q(D^\alpha))^2) \\ & + (4\Delta_1^\alpha(\omega) + \Delta_2^\alpha(\omega) + \Delta_3^\alpha(\omega) + \Delta_4^\alpha(\omega) + \Delta_5^\alpha(\omega)) \cdot E_c^q(X^\alpha D^\alpha)\end{aligned}$$

expression qui fait apparaître pour le locus (α) :

— la variance génétique additive $2E^q(X^\alpha)^2 = \mathrm{Var}\,(Y^\alpha)$ en notant $Y^\alpha = 2X^\alpha$;

— la variance génétique de dominance classique $E_p^q((D^\alpha)^2) = \mathrm{Var}_p(D^\alpha)$;

— la variance génétique de dominance dans la condition d'identité :

$$E_c^q((D^\alpha)^2) - (E_c^q(D^\alpha))^2 = \mathrm{Var}_c(D^\alpha)\ ;$$

— l'interaction entre les effets additifs et de dominance, dans la condition d'identité :

$$E^q(X^\alpha D^\alpha) = \mathrm{Cov}_c(X^\alpha D^\alpha).$$

La covariance entre Z_I et Z_J est alors simplement :

$$\mathrm{Cov}^{q/\omega}\,(Z_I Z_J) = \sum_\alpha \mathrm{Cov}^{q/\omega}\,(Z_I^\alpha Z_J^\alpha).$$

B. — *Valeurs approchées des espérances, variances, et covariances génotypiques dans une population*

Les expressions précédentes dépendent de la réalisation (ω), de la structure d'identité (ω) qui est inconnue. Mais, dans les conditions qui sont précisées au paragraphe suivant, chacune des sommes telles que :

$$\sum_\alpha \Delta_1^\alpha(\omega)\ \mathrm{Var}_c(D^\alpha)$$

peut être approchée par son espérance E^ω :

$$\delta_1 \sum_\alpha \mathrm{Var}_c\ (D^\alpha)$$

qui est une quantité indépendante de (ω).

On peut alors écrire, quelle que soit la réalisation (ω) :

$$\begin{aligned}\mathrm{cov}^{q/\omega}\,(Z_I Z_J) \simeq {} & 2\,\varphi_{IJ}\ \mathrm{var}\ (Y) \\ & + (\delta_9 + \delta_{12})\ \mathrm{var}_p\ (D) \\ & + \delta_1\ \mathrm{var}_c\ (D) \\ & + (4\,\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5)\ \mathrm{cov}_c\ (XD).\end{aligned} \qquad (2)$$

et de même pour la variance :

$$\begin{aligned}\mathrm{var}^{q/\omega}\,(Z_I) \simeq {} & (1 + f_I)\ \mathrm{var}\ (Y) \\ & + (1 - f_I)\ \mathrm{var}_p\ (D) \\ & + f_I\ \mathrm{var}_c\ (D) \\ & + 4\,f_I\ \mathrm{cov}_c\ (XD).\end{aligned} \qquad (3)$$

où Var (Y), Var_p (D), Var_c (D) et Cov_c (XD) sont les sommes des quantités analogues relatives aux différents loci (α).

On aura aussi, dans les mêmes conditions :

$$\mathrm{E}^{q/\omega}(\mathrm{Z_I}) \simeq f_{\mathrm{I}} \sum_{\alpha} \mathrm{E}_c^q(\mathrm{D}^\alpha). \tag{4}$$

C. — *Les conditions et la valeur de l'approximation*

Chacune des quantités $\mathrm{Cov}^{q/\omega}(\mathrm{Z_I^\alpha Z_J^\alpha})$ s'écrit aussi :

$$\begin{aligned}\mathrm{cov}^{q/\omega}(\mathrm{Z_I^\alpha Z_J^\alpha}) = {} & (\Delta_{10}^\alpha(\omega) + \Delta_{11}^\alpha(\omega) + \Delta_{13}^\alpha(\omega) + \Delta_{14}^\alpha(\omega))\,1/2\ \mathrm{var}(\mathrm{Y}^\alpha) \\ & + (\Delta_9^\alpha(\omega) + \Delta_{12}^\alpha(\omega)) \cdot (\mathrm{var}(\mathrm{Y}^\alpha) + \mathrm{var}_p(\mathrm{D}^\alpha)) \\ & + (\Delta_2^\alpha(\omega) + \Delta_3^\alpha(\omega) + \Delta_4^\alpha(\omega) + \Delta_5^\alpha(\omega)) \cdot (\mathrm{var}(\mathrm{Y}^\alpha) + \mathrm{cov}_c(\mathrm{X}^\alpha\mathrm{D}^\alpha)) \\ & + \Delta_1^\alpha(\omega) \cdot (2 \cdot \mathrm{var}(\mathrm{Y}^\alpha) + \mathrm{var}_c(\mathrm{D}^\alpha) + 4 \cdot \mathrm{cov}_c(\mathrm{X}^\alpha\mathrm{D}^\alpha)).\end{aligned}$$

et est la somme de quatre termes dont chacun est le produit d'une variable indicatrice d'une réunion de situations d'identité, et d'une composante de la variabilité due au locus (α). Un terme tel peut s'écrire :

$$\Gamma_i^\alpha(\omega) \cdot \mathrm{C}_i^\alpha$$

(par exemple : $\Gamma_2^\alpha(\omega) = \Delta_9^\alpha(\omega) + \Delta_{12}^\alpha(\omega)$ et $\mathrm{C}_2^\alpha = \mathrm{var}(\mathrm{Y}^\alpha) + \mathrm{var}_{\mathrm{D}}(\mathrm{D}^\alpha)$)

en fonction de cette écriture la covariance totale devient :

$$\mathrm{cov}^{q/\omega}(\mathrm{Z_I Z_J}) = \sum_{i=1}^{i=4} \left(\sum_{\alpha=1}^{\alpha=\mathrm{L}} \Gamma_i^\alpha(\omega)\ \mathrm{C}_i^\alpha \right).$$

Chaque somme en (α) est une somme de L, variables aléatoires indépendantes. Nous faisons les hypothèses génétiques suivantes.

H-1. *Le caractère étudié, Z, est gouverné par un grand nombre* (L) *de loci indépendants.*

H-2. *Chacune des quantités* C_i^α *est bornée par une quantité* μ_i, *indépendante de* α. Cette hypothèse consiste seulement à dire que les contributions des loci sont bornées, elle n'exclut pas le cas des gènes majeurs.

H-3. *Chacune des quantités* C_i^α *est supérieure en valeur absolue, à un nombre fixe* λ_i. Cela signifie qu'on exclut, dans l'analyse de la variabilité de Z, tous les gènes quasiment fixés, c'est-à-dire tous les allèles dont la fréquence dans la population est très proche de 0 ou de 1.

On notera $\rho_i = \mu_i/\lambda_i$, et $\rho = \sup_i \rho_i$

La présence de gènes majeurs se traduit par une valeur importante de ρ, qui affecte la précision des approximations.

Dans ces conditions le théorème de Liapounov s'applique aux sommes d'aléatoires indépendantes, qui s'écrivent :

$$\sum_{\alpha=1}^{\alpha=\mathrm{L}} \Gamma_i^\alpha(\omega)\ (\mathrm{C}_i^\alpha) = \gamma_i \sum_{\alpha=1}^{\alpha=\mathrm{L}} \mathrm{C}_i^\alpha\ (1 + e_i(\omega))$$

où γ_i est la somme des probabilités des situations d'identité indiquées par $\Gamma_i^\alpha(\omega)$ (par exemple $\gamma_2 = \delta_9 + \delta_{12}$) ; et où e_i est une variable aléatoire dont la loi est voisine d'une loi normale, qui est centrée, et dont l'écart-type est :

$$\left(\frac{1}{L}\,\frac{1-\gamma_i}{\gamma_i}\,(1+r_i^2)\right)^{1/2}$$

où r_i^2 est un « coefficient de dispersion » des C_i^α vérifiant :

$$r_i^2 < 1/4\,(\rho_i + \frac{1}{\rho_i} - 2) \leqslant 1/4\,(\rho + \frac{1}{\rho} - 2)$$

La vérification des hypothèses du théorème de Liapounov et la démonstration de cette dernière inégalité sont présentées en annexe.

Moment du premier ordre.

Dans la somme de la formule (4), il faut distinguer trois sommes partielles : $L_\prime = L_{\prime 1} + L_{\prime 2} + L_{\prime 3}$. En $L_{\prime 1}$ loci les contributions sont nulles ; en $L_{\prime 2}$ loci les contributions sont positives ; en $L_{\prime 3}$ loci elles sont négatives. Notons r_p^2 et r_n^2 les coefficients de dispersion des sommes positives et négatives, respectivement. Alors,

$$E_\prime^{q/\omega}(Z_1) = f_i \sum_{\alpha=1}^{\alpha=L} E_c^q(D^\alpha)$$
$$+ \left(\frac{f_i(1-f_i)(1+r_p^2)}{L_{\prime 2}}\right)^{1/2} \left(\sum_{\alpha\in L_2} E_c^q(D^\alpha)\right) \varepsilon_2(\omega)$$
$$+ \left(\frac{f_i(1-f_1)(1+r_n^2)}{L_{\prime 3}}\right)^{1/2} \left(\sum_{\alpha\in L_3} E_c^q(D^\alpha)\right) \cdot \varepsilon_3(\omega).$$

où ε_2 et ε_3 sont deux variables aléatoires indépendantes proches de la loi normale N(0, 1).

Variance.

$$\text{Soit } V^\alpha = \text{var}(Y^\alpha) - \text{var}_p(D^\alpha) + \text{var}_c(D^\alpha) + 4\,\text{cov}_c(X^\alpha D^\alpha),$$

$V = \sum_\alpha V^\alpha$ et r_v le cœfficient de dispersion correspondant :

$$\text{var}_\prime^{q/\omega}(Z_1) = (1+f_i)\,\text{var}(Y) + (1-f_i)\,\text{var}_p(D)$$
$$+ f_i\,\text{var}_c(D) + 4\,f_i\,\text{cov}_c(XD)$$
$$+ \left(\frac{f_1(1-f_1)(1+r_v^2)}{L_\prime}\right)^{1/2} V \cdot \varepsilon(\omega)$$

où ε est une variable aléatoire dont la loi s'approche, quand $L_\prime$ est grand, de la loi normale réduite N(0,1).

Covariance.

Dans l'expression de la covariance interviennent quatre ou cinq (1) aléatoires ε_i corrélées du fait que dans l'expression de

$$\text{cov}_\prime^{q/\omega}(Z_I^\alpha Z_J^\alpha)$$

un seul des quatre termes peut être non nul.

(1) Les composantes C_3^α peuvent être positives ou négatives, alors que les autres ($i \neq 3$) sont positives : on distingue deux classes de loci ($L = L_4 + L_5$) où les composantes sont respectivement toutes positives ou toutes négatives. Le terme aléatoire comprend donc deux termes, l'un en $\sqrt{\frac{1}{L_4}}$, l'autre en $\sqrt{\frac{1}{L_5}}$.

Notant $C_i = \sum_\alpha C_i^\alpha$ il vient, pour le i^{ieme} terme [1] :

$$\sum_\alpha \Gamma_i^\alpha(\omega)\, C_i^\alpha = \gamma_i\, C_i + \left(\frac{\gamma_i(1-\gamma_i)(1+r_i^2)}{L}\right)^{1/2} C_i\, \varepsilon_i(\omega).$$

ε_i est proche d'une variable normale N(0, 1).

Soient :

$$C = \sup_i |C_i|$$

$$r^2 = \sup_i r_i^2 < \frac{1}{4}\left(\rho + \frac{1}{\rho} - 2\right)$$

$$\gamma(1-\gamma) = \sup_i \gamma_i(1-\gamma_i) \leqslant \frac{1}{4} \text{ puisque } \gamma_i \leqslant 1.$$

$$L' = \inf(L_4, L_5) \quad \text{si} \quad L_4 L_5 \neq 0 \text{ [1]}.$$

La somme des termes aléatoires, qui est une variable aléatoire, asymptotiquement normale, centrée, est majorée en valeur absolue :

$$|e(\omega)| < \left(\frac{\gamma(1-\gamma)(1+r^2)}{L'}\right)^{1/2} C \sum_i |\varepsilon_i(\omega)|.$$

En élevant les deux membres au carré et en prenant les espérances E^ω, l'inégalité demeure. En majorant le second membre par application de l'inégalité de Schwarz, il vient, quelles que soient les corrélations entre les ε_i :

$$E(e^2) < \frac{1}{L'}\gamma(1-\gamma)(1+r^2)\, C^2 \cdot \left(\sum_i (E(\varepsilon_i^2))^{1/2}\right)^2$$

et comme les ε_i sont proches de la loi N(0, 1) et que e est centrée :

$$\text{var}(e) = \sigma^2 < \frac{25}{L'}\gamma(1-\gamma)(1+r^2)\, C^2$$

ε désignant une variable proche de N(0,1) et σ admettant la majoration précédente, la formule de covariance s'écrit alors :

$$\begin{aligned}\text{cov}^{q/\omega}(Z_I Z_J) = {} & 2\,\varphi_{IJ}\,\text{var}(Y) + (\delta_9 + \delta_{12})\,\text{var}_p(D) \\ & + \delta_1\,\text{var}_c(D) + (4\,\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5)\,\text{cov}_c(XD) \\ & + \sigma \cdot \varepsilon(\omega).\end{aligned}$$

III. — ESPÉRANCES, VARIANCES ET COVARIANCES GÉNOTYPIQUES « INTER-POPULATIONS » : FORMULES DE GILLOIS

Dans cette troisième partie nous retrouverons, avec les notations introduites, les formules originales de GILLOIS, pour les comparer aux expressions approchées des covariances « intra-population » que nous avons établies dans la seconde partie et préciser la signification de chacun des deux groupes de formules.

Les formules de GILLOIS sont obtenues quand on prend simultanément les espérances mathématiques par rapport à la distribution (q) des allèles et par rapport à

la distribution de toutes les structures d'identité (ω) compatibles avec une généalogie. Cela peut s'écrire, pour la covariance :

$$\mathrm{cov}^{\omega q}(Z_I Z_J) = E^{\omega q}(Z_I Z_J) - E^{\omega q}(Z_I) \cdot E^{\omega q}(Z_J)$$

et chaque espérance $E^{\omega q}$ peut se décomposer par rapport aux différentes conditions (ω).

$$E^{\omega q}(Z_I Z_J) = E^{\omega}(E^{q/\omega}(Z_I Z_J)).$$

Il vient ainsi :

$$\begin{aligned}\mathrm{cov}^{\omega q}(Z_I Z_J) = {} & E^{\omega}(E^{q/\omega}(Z_I Z_J)) - E^{\omega}(E^{q/\omega}(Z_I)\, E^{q/\omega}(Z_J)) \\ & + E^{\omega}(E^{q/\omega}(Z_I)\, E^{q/\omega}(Z_J)) - E^{\omega}(E^{q/\omega}(Z_I)) \cdot E^{\omega}(E^{q/\omega}(Z_J)).\end{aligned}$$

La première ligne fait apparaître l'espérance (E^{ω}) de la covariance conditionnée par (ω), qui n'est autre que le second membre de l'équation (2).
La deuxième ligne fait apparaître la covariance, par rapport à la distribution des (ω), des espérances conditionnées des aléatoires génotypiques Z_I et Z_J.

On a d'une part ; d'après les expressions établies en II-A :

$$\begin{aligned}E^{q/\omega}(Z_I)\, E^{q/\omega}(Z_J) &= \Big(\sum_{\alpha} F_I^{\alpha}(\omega)\; E_c^q\,(D^{\alpha})\Big)\Big(\sum_{\beta} F_J^{\beta}(\omega)\; E_c^q\,(D^{\beta})\Big) \\ &= \sum_{\alpha} F_I^{\alpha}(\omega)\, F_J^{\alpha}(\omega) \cdot (E_c^q\,(D^{\alpha}))^2 \\ &\quad + \sum \sum_{\alpha \neq \beta} F_I^{\alpha}(\omega)\, F_J^{\beta}(\omega)\, E_c^q\,(D^{\alpha})\, E_c^q\,(D^{\beta})\end{aligned}$$

dont l'espérance E^{ω}, d'après la formule (1) et l'indépendance des loci, est :

$$(\delta_1 + \delta_6) \sum_{\alpha} (E_c^q\,(D^{\alpha}))^2 + f_I f_J \sum \sum_{\alpha \neq \beta} E_c^q\,(D^{\alpha})\, E_c^q\,(D^{\beta})$$

On a d'autre part :

$$E^{\omega}(E^{q/\omega}(Z_I)) = f_I \sum_{\alpha} E_c^q\,(D^{\alpha})$$

Il en résulte que la deuxième ligne a pour valeur :

$$(\delta_1 + \delta_6 - f_I f_J) \sum_{\alpha} (E_c^q\,(D^{\alpha}))^2.$$

Et les formules « inter-populations » de Gillois s'écrivent :

(5)
$$E^{\omega q}(Z_I) = f_I \sum_{\alpha} E_c^q\,(D^{\alpha}).$$

(6)
$$\begin{aligned}\mathrm{var}^{\omega q}(Z_I) = {} & (1 + f_I)\, \mathrm{var}(Y) + (1 - f_I)\, \mathrm{var}_p\,(D) \\ & + f_I\, \mathrm{var}_c\,(D) + 4 f_I\, \mathrm{cov}_c\,(XD) \\ & + f_I(1 - f_I) \sum_{\alpha} (E_c^q\,(D^{\alpha}))^2\end{aligned}$$

(7)
$$\begin{aligned}\mathrm{cov}^{\omega q}(Z_I Z_J) = {} & 2\, \varphi_{IJ}\, \mathrm{var}\,(Y) + (\delta_9 + \delta_{12})\, \mathrm{var}_p\,(D) \\ & + \delta_1\, \mathrm{var}_c\,(D) + (4\,\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5)\, \mathrm{cov}_c\,(XD) \\ & + (\delta_1 + \delta_6 - f_I f_J) \sum_{\alpha} (E_c^q\,(D^{\alpha}))^2.\end{aligned}$$

L'identité avec les formules données par GILLOIS (1966 *a* ou 1966 *b*) est immédiate : il suffit dans les formules (6) et (7) de décomposer Var_c (D) en

$$\mathrm{Var}_c\,(D) = E_c(D^2) - (E_c(D))^2$$

pour faire apparaître les termes $\delta_1\, E_c(D^2)$, et : $(\delta_6 - f_1 f_J)(E_c\,(D))^2$. Les autres termes sont identiques.

CONCLUSION

La distinction entre espérances mathématiques prises par rapport à la distribution des fréquences alléliques dans une population et par rapport à la distribution des structures génétiques compatibles avec la ségrégation mendélienne dans une population d'effectif limité ouvre de nouvelles possibilités de calcul des corrélations génétiques entre apparentés. Elle nous a permis d'exploiter les propriétés des coefficients d'identité au niveau d'une population, et pas seulement au niveau d'un grand nombre de populations présentant le même pedigree comme l'avait fait GILLOIS.

Nous avons conservé les hypothèses fondamentales du calcul de GILLOIS, c'est-à-dire principalement celle de l'équivalence entre la non identité des gènes et leur indépendance stochastique. L'introduction de lois jointes plus générales, et compatibles avec des phénomènes de sélection serait possible mais il ne semble pas que leur théorie soit encore assez développée à ce jour.

Ne faisant aucune hypothèse sur le nombre de loci impliqués dans l'expression d'un caractère, ni sur la grandeur relative de leurs contributions, GILLOIS (1964) et BOUFFETTE (1966) dans le cas d'indépendance, GALLAIS (1970) dans le cas de linkage ne peuvent introduire les coefficients d'identité qu'en prenant des espérances par rapport à toutes les structures d'identité compatibles avec la généalogie. Les formules qui en résultent sont caractéristiques de la généalogie, elles ne peuvent s'interpréter statistiquement qu'au niveau d'un certain nombre de populations issues d'une même population d'origine, respectant la même généalogie et ayant connu des évolutions indépendantes. En revanche, elles sont valables pour un caractère déterminé par un nombre quelconque de gènes, dont certains peuvent avoir des effets très grands. Mais elles ne sont pas applicables à la plupart des situations zootechniques : la construction de populations analogues et indépendantes n'est pas économiquement possible, d'une part. La structure des populations très consanguines comme celles des élevages avicoles ne fournit pas, d'autre part, de familles indépendantes ni même non corrélées comme le supposent les modèles classiques de la génétique quantitative : après quinze années de sélection en effectif limité, le coefficient de parenté entre individus « non-apparentés » (de pères différents) y atteint 20 p. 100 à 30 p. 100.

Les nouvelles formules approchées que nous avons établies sont au contraire applicables à une population particulière, unique, qui vérifie une généalogie. Elles expriment en effet que, dans certaines conditions, des espérances conditionnées par une structure inconnue sont quasiment indépendantes de cette structure. Ces conditions se ramènent à deux points, qui se dégagent des formules du paragraphe II-*c* :

1° le nombre des loci est grand ;

2° le rapport de la plus grande contribution à la variabilité, à la plus petite contribution non nulle est petit devant le nombre des loci ; ce qui exprime qu'aucun gène majeur n'intervient dans la variabilité du caractère. Ce sont en réalité des hypothèses classiques en Génétique quantitative, qui expliquent notamment les distributions sensiblement gaussiennes de caractères quantitatifs héritables.

Conjointement avec une méthode de calcul automatique des coefficients d'identité (CHEVALET, 1971), ces formules ouvrent la possibilité d'applications à des populations végétales et surtout animales où les réplications de populations identiques sont ou bien fastidieuses (drosophiles) ou simplement impossibles (animaux domestiques) pour des raisons économiques. Dans le cas des formules de GILLOIS, l'estimation des paramètres génétiques était rendue impossible pour cette raison ; avec les nouvelles formules la difficulté est d'établir des estimateurs à partir d'un ensemble d'observations qui sont liées et dont les corrélations s'expriment avec les paramètres à estimer. Ce n'est pas une situation classique en statistique, mais il est possible de définir des conditions sur les pedigrees qui permettent ces estimations.

Reçu pour publication en août 1971.

SUMMARY

A PRIORI INTRA- AND INTERPOPULATIONS CALCULUS OF GENOTYPIC VARIANCES AND COVARIANCES BETWEEN INBRED RELATIVES

Through the distinction, in a limited population, between the expectations taken with respect to the distribution of allelic frequencies, and with respect to the distribution of all possible identity situations ; through the meaning of identity-coefficients in a single population ; and through classical hypotheses in population genetics, new formulae of genotypic variance and covariance between inbred relatives are derived. When GILLOIS's formulae only apply if many replications of the same pedigree are given, and hence very rarely in actual situations, these new expressions hold between two inbred relatives within a single population, and then allow applications in plant or animal breeding.

RÉFÉRENCES BIBLIOGRAPHIQUES

BOUFFETTE J., 1966. *Expression de la covariance génotypique chez les tétraploïdes.* Thèse 3e cycle, Fac. Sciences, Lyon.

CHEVALET C., 1971. Calcul automatique des coefficients d'identité. *Ann. Génét. Sél. anim.* **3**, 449-462.

FISHER R. A., 1918. The correlations between relatives on the supposition of mendelian inheritance. *Trans. Roy. Soc. Edinburgh*, **52**, 399-433.

GALLAIS A., 1970. Covariances entre apparentés quelconques avec linkage et épistasie. I. Expression générale. *Ann. Génét. Sél. anim.*, **2**, 281-310.

GILLOIS M., 1964. *La relation d'identité en génétique.* Thèse, Fac. Sciences, Paris, 294 p.

GILLOIS M., 1965. Relation d'identité en génétique. *Ann. Inst. Henri-Poincaré*, B, **2**, 1-94.

GILLOIS M., 1966 (*a*). Le concept d'identité et son importance en génétique. *Ann. Gén.*, **9**, 58-65.

GILLOIS M., 1966 (*b*). Note sur la variance et la covariance génotypiques entre apparentés. *Ann. Inst. Henri-Poincaré, B*, **2**, 349-532.

GILLOIS M., 1966 (*c*). La relation de dépendance en génétique. *Ann. Inst. Henri-Poincaré, B*, **2**, 261-278.

GILLOIS M., 1967 (*a*). La notion de génotype. *Ann. Gén.*, **10**, 201-202.

GILLOIS M., 1967 (*b*). Les lois conjointes des variables aléatoires génétiques. *Ann. Gén.*, **10**, 203-206.

HARRIS D. L., 1964. Genotypic covariances between inbred relatives. *Genetics*, **50**, 1320-1348.

KEMPTHORNE O., 1954. The correlations between relatives in a random mating population. *Proc. Roy. Soc., B*, **143**, 103-113.

MALECOT G., 1939. *Théorie mathématique de l'hérédité mendélienne généralisée.* Thèse, Fac. Sciences, Paris. — Republiée dans : MALECOT G., 1966. *Probabilité et hérédité.* I. N. E. D., cahier n° 47, P. U. F.

MALECOT G., 1946. La consanguinité dans une population limitée. *C. R. Acad. Sci.* Paris, **222**, 841-843.

MALÉCOT G., 1948. *Les mathématiques de l'hérédité.* Masson et Cie, Paris, 1 vol., 60 p.

TRUSTRUM G. B., 1960. The correlations between relatives in a random mating diploïd population. *Proc. Cambo. Phil. Soc.*, **57**, 315-320.

WRIGHT S., 1921. Systems of mating. *Genetics*, **6**, 111-178.

WRIGHT S., 1922. Coefficients of inbreeding and relationships. *Amer. Nat.*, **56**, 330-338.

ANNEXE

I. — *Les hypothèses du théorème de Liapounov*

Chacune des sommes de la forme :

$$\sum_{\alpha} \Gamma_i^{\alpha}(\omega)\, C_i^{\alpha}$$

est du type

$$W_L = \sum_{\alpha=1}^{\alpha=L} U^{\alpha}(\omega)$$

où les U^{α} sont des aléatoires indépendantes (hypothèse H-1) de lois :

$$Pr\,(U^{\alpha} = C^{\alpha}) = \gamma$$
$$Pr\,(U^{\alpha} = 0) = 1 - \gamma$$

D'après l'hypothèse H-2, les U^{α} sont bornées dans leur ensemble par un nombre μ. D'après H-3, la variance de U^{α} est uniformément minorée :

$$\text{var}\,(U^{\alpha}) = \gamma(1-\gamma)(C^{\alpha})^2 > \gamma(1-\gamma)\,\lambda^2.$$

et par conséquent :

$$b_L^2 = \text{var}\,(W_L) \quad \text{tend vers l'infini avec L.}$$

Il en résulte, d'après un énoncé élémentaire du théorème de Liapounov, que la loi de :

$$\frac{W_L - E\,(W_L)}{b_L}$$

tend vers la loi normale réduite N(0, 1) quand L tend vers l'infini.
La somme W_L peut alors s'écrire :

$$W_L = E(W_L) \cdot \left(1 + \frac{b_L}{E\,(W_L)} \frac{W_L - E\,(W_L)}{b_L}\right)$$

Le terme d'erreur relative est une aléatoire ayant une loi proche de la loi normale si L est grand (H-1), et d'écart-type :

$$\frac{b_L}{E(W_L)} = \frac{(\gamma(1-\gamma)\sum_\alpha (C^\alpha)^2)^{1/2}}{\gamma \sum_\alpha C^\alpha} = \left(\frac{1-\gamma}{\gamma}\frac{1}{L}(1+r^2)\right)^{1/2}$$

II. — *La majoration du coefficient de dispersion* r^2 *(principe de la démonstration).*

r^2 est une fonction des quantités C^α, que l'on supposera être toutes du même signe, positif par exemple.

En posant :

$$C = \frac{1}{L}\sum_\alpha C^\alpha$$

On a :

$$r^2 = \frac{1}{L}\sum_\alpha \left(\frac{C^\alpha - C}{C}\right)^2$$

On cherche les extrema de r^2 quand les C^α peuvent varier dans un intervalle (λ, μ) :

$$0 < \lambda \leqslant C^\alpha \leqslant \mu \qquad \text{quel que soit } \alpha,$$

et on suppose que chacune des valeurs extrêmes λ et μ est atteinte au moins une fois.

On démontre alors que si q $(2 \leqslant q \leqslant L)$ quantités C^α sont fixées aux bornes de l'intervalle, la fonction r^2 de $L - q$ variables admet un seul extremum dans l'hypercube ouvert $]\lambda, \mu[^{L-q}$ et c'est un minimum. Le maximum de cette fonction r^2, définie sur l'hypercube fermé, est donc atteint à la surface de cet hypercube, c'est-à-dire quand une variable supplémentaire C^α est fixée en λ ou en μ. Cette surface est elle-même un hypercube $]\lambda, \mu[^{L-q-1}$.

En raisonnant par récurrence, il résulte que le maximum de r^2 est atteint quand toutes les quantités C^α sont fixées aux bornes. Par symétrie, la valeur de r^2 est alors définie par le nombre M tel que M quantités C^α sont fixées en λ et $L - M$ en μ $(1 \leqslant M \leqslant L - 1)$. Le maximum est obtenu quand M prend pour valeur la partie entière N de :

$$L\frac{\mu}{\mu + \lambda}$$

ou bien la valeur extrême (1 ou $L - 1$) la plus proche de N.

Posant $\rho = \mu/\lambda$, ce maximun admet la majoration :

$$\max(r^2) \leqslant \frac{1}{4}\left(\rho + \frac{1}{\rho} - 2\right).$$

Par ailleurs le minimum est :

$$\min (r^2) = \frac{(\rho - 1)^2}{L(\rho^2 + 1) - (\rho - 1)^2} .$$

il est atteint si, à l'exception de deux quantités C^1 et C^2 fixées respectivement en λ et μ, toutes les autres quantités C^α sont égales à : $\frac{\lambda^2 + \mu^2}{\lambda + \mu}$.

Ces deux cas extrêmes où toutes les contributions ne prennent qu'une ou deux valeurs distinctes, n'ont guère de signification biologique. Il est remarquable que des variations importantes des contributions, telles que $\rho = 10$ par exemple (alors $r^2 < 2$) affectent peu la précision des formules approchées (2), (3), (4), l'écart-type de l'erreur étant alors multiplié par $\sqrt{3}$ par rapport au cas idéal où toutes les contributions des différents loci seraient égales.