Génét. Sél. Evol., 1988, **20** (2), 211-226

A statistical model for genotype determination at a major locus in a progeny test design

J.M. ELSEN, Jacqueline VU TIEN KHANG
and Pascale LE ROY

Institut National de la Recherche Agronomique,
Station d'Amélioration Génétique des Animaux,
Centre de Recherches de Toulouse, B.P. 27, 31326 Castanet-Tolosan Cedex, France

Summary

Considering a normally distributed quantitative trait whose genetic variation is controlled by both an autosomal major locus and a polygenic component, and whose expression is influenced by environmental factors, a mixed model was developed to classify sires and daughters for their genotypes at the major locus in a progeny test design. Repeatability and genetic parameters reflecting the polygenic variation were assumed to be known. Posterior distribution of the sire genotypes and that of the daughters given the sire genotypes were derived. A method was proposed to estimate these posterior probabilities as well as the unknown parameters, and a method using the likelihood ratios to test specific genetic hypotheses was suggested. An iterative two-step procedure similar to the EM (expectation-maximization) algorithm was used to estimate the posterior probabilities and the unknown parameters. The operational value of this approach was tested with simulated data.

Key words : major locus, progeny test, genotypic classification, maximum likelihood.

Résumé

Un modèle statistique pour la détermination du génotype à un locus majeur dans un test sur descendance

S'appliquant à un caractère quantitatif à distribution normale, dont la variabilité génétique est contrôlée à la fois par un locus majeur autosomal et par une composante polygénique et dont l'expression est influencée par des facteurs de milieu, un modèle mixte est développé afin de déterminer le génotype (au locus majeur) des pères et de leurs filles dans un test sur descendance. La répétabilité et les paramètres génétiques relatifs à la composante polygénique sont supposés connus. La loi *a posteriori* des génotypes des pères et celles des génotypes de leurs filles, conditionnellement aux génotypes des pères, sont établies. Une méthode est proposée pour estimer ces probabilités *a posteriori*, ainsi que les paramètres inconnus, et une méthode utilisant les rapports de vraisemblance est suggérée afin de tester des hypothèses génétiques spécifiques. Une procédure itérative en deux étapes, similaire à l'algorithme EM *(expectation-maximization)*, est présentée afin d'estimer les probabilités *a posteriori* et les paramètres inconnus. L'intérêt opérationnel de cette approche est éprouvé sur des données simulées.

Mots clés : gène majeur, test sur descendance, détermination du génotype, maximum de vraisemblance.

I. Introduction

PIPER & BINDON discovered, in 1982, a major gene, named *Booroola,* affecting ovulation rate and litter size of ewes. Many data have confirmed this discovery since (DAVIS *et al.*, 1982 *a, b ;* DAVIS & KELLY, 1983). The favourable allele and the wild-type allele are symbolized by *F* and + respectively.

Some differences have been found between the reproductive biology of carrier and non-carrier ewes (*see* the review of BINDON (1984)). However, up till now the only measurements actually used to classify females according to their genotype (*FF, F+* or ++) are ovulation rate and litter size. The most used criterion is that proposed by DAVIS *et al.* (1982 *b*) : a ewe is classified *FF* when, in a series of measurements, it has at least one ovulation rate of 5 or more ; a ewe is said to be *F+* when its maximum ovulation rate recorded is 3 or 4 ; a ewe is identified as ++ when its ovulation rate never exceeds 2.

As far as the choice of males is concerned, the only possibility at the moment is the progeny test : a ram is mated to a large enough number of ++ ewes, for its genotype to be assessed from the observation of its progeny (100, 50, or 0 % of *F+* daughters).

However, even if they are sufficient at the moment, these criteria may be criticized (ELSEN & ORTAVANT, 1984 ; PIPER *et al.*, 1985 ; OWENS *et al.*, 1985) :

1) the threshold values (3 and 5) were derived from observations on *Merino* ewes whose basal level of prolificacy is low. Their mean ovulation rate is about 1.5 for ++ females, 3 for *F+* and 4.5 for *FF.* Obviously, such thresholds could not be used in the case of prolific breeds. Moreover, many sources of variation (age, season, body weight, feeding) influence the ovulation rate, within the breed. Such factors must be considered when choosing a threshold ;

2) the polygenic variability of the ovulation rate is a bias source already shown by DAVIS *et al.* (1982 *a*). For example, an *FF* ram may have a very low breeding value for ovulation rate (compared to the mean of the *FF*) which will lower the percentage of its *F+* daughters and rank him as a heterozygote ;

3) since the penetrance is incomplete, it is necessary to repeat ovulation rate measurements. Unfortunately, the probability of a ++ female with an ovulation rate of 3 or more is not null (even more so when the prolificacy of the breed is higher) and the risk of classifying some ++ ewes as *F+* (or some *F+* as *FF*) increases with the number of measurements. It is generally considered that 3 measurements are necessary for the *Merinos,* but this is not a rule.

Considering these difficulties, OWENS *et al.* (1985) proposed the use of cluster analysis to classify females according to their genotypes : the candidate population is subdivided into three groups by minimizing the sum of squared deviations from the within group means. This solution has the advantage of avoiding the choice of a threshold and of a number of observations per female, but it does not take into account the error sources stated above.

Because of the problems caused by the identification of genotypes in the case of the *Booroola* major gene, we suggest a general approach for determining the genotype at a major locus in a progeny test design, in the case of a quantitative trait with a normal distribution ; the case of a discrete trait is studied in the same way by FOULLEY

& ELSEN (1988). The proposed method, based on maximum likelihood methods, is derived from works concerning mixtures of distributions (DAY, 1969 ; AITKIN & WILSON, 1980 ; EVERITT, 1984) and segregation analysis (ELSTON & STEWART, 1971 ; MORTON & MC LEAN, 1974 ; LALOUEL *et al.*, 1983).

II. Definitions and hypotheses

A. *Genetic model and progeny test design*

1) The genetic variation of the quantitative considered trait has two sources : a polygenic and a monogenic component depending on an autosomal major locus with two alleles F and $+$.

2) In the parental population of the progeny tested sires, there is genetic independence or linkage equilibrium between the major gene and the genes controlling the polygenic variability.

3) The progeny test is made by mating 9 with $++$ dams the sires whose prior distribution of the genotypes at the major locus is assumed to be known. The choice of mates is at random. These matings give birth to daughters ($F+$ or $++$) measured, once or more, for the quantitative trait involved. Several sources of variation can modify the expression of the trait.

4) The measured daughters are not inbred. This means that the sires are not related to their mates.

5) The only relationship between two measured daughters can be due to a possible common father. This means that :

— there are no full sibs in the population of measured daughters,
— the sires are not related,
— their mates are not related.

B. *Notation for genotypes, performances, and probabilities*

1. *Notation for genotypes*

Genotypes of sires and their daughters are considered as random variables with the following notation :

G_t refers to the genotype of the t^{th} sire, t being between 1 and T, the total number of sires

G_{ti}, the genotype of the i^{th} daughter of the t^{th} sire, i being between 1 and n_t, the number of the t^{th} sire's daughters

$\mathbf{\Gamma} = \{G_1, G_2 \ldots, G_T\}$ the vector of the sires' genotypes

$\mathbf{\Gamma}_t = \{G_{t1}, G_{t2}, \ldots, G_{tn_t}\}$ the vector of the genotypes of the t^{th} sire's daughters.

The realizations of these random variables are denoted g_t, g_{ti}, $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_t$, respectively.

2. *Notation for performances*

The random variable Y_{tij} denotes the j^{th} observation of the i^{th} daughter of sire t ($j = 1$ to n_{ti}).

$\mathbf{Y}_{ti}$ is the vector of Y_{tij} variables concerning the i^{th} daughter of sire t.

$\mathbf{Y}_t$ is the vector of all the variables concerning sire t.

$\mathbf{Y}$ is the vector of all the variables.

The realizations of these random variables are denoted y_{tij}, $\mathbf{y}_{ti}$, $\mathbf{y}_t$ and $\mathbf{y}$ respectively.

3. *Notation for probabilities*

For ease of presentation, we shall use the same notation the denote an event as well as the value taken by a random variable when this event is realized : the event « random variable $\mathbf{Y}$ is equal to $\mathbf{y}$ » will be noted « $\mathbf{y}$ » instead of « $\mathbf{Y} = \mathbf{y}$ ». For example, the symbol prob($\boldsymbol{\gamma}/\mathbf{y}$) means prob($\boldsymbol{\Gamma} = \boldsymbol{\gamma}/\mathbf{Y} = \mathrm{y}$), i.e., the probability that the realization of $\boldsymbol{\Gamma}$ is $\boldsymbol{\gamma}$, given that the random variable $\mathbf{Y}$ is $\mathbf{y}$.

C. *Modelling of performances*

1. *Effects considered in the model*

Daughters' performances are described through a linear model with the following effects :

— fixed effects independent of the daughter's major genotype ($\mathbf{b}$ vector),
— fixed effects dependent on the daughter's major genotype ($\boldsymbol{\beta}$ vector),
— a random sire effect accounting for the polygenic part of the variation, and whose distribution depends on the daughter's major genotype ($\mathbf{U}$ vector),
— a residual whose distribution depends on the daughter's major genotype ($\mathbf{E}$ vector).

The $\boldsymbol{\beta}$ vector may be split into two parts ($\boldsymbol{\beta}_{l++}$ and $\boldsymbol{\beta}_{lF+}$) only one of which is applicable depending on the daughter's genotype (++ of $F+$). Similarly, the $\mathbf{U}$ vector may be split into two parts, $\mathbf{U}_{t++}$ and $\mathbf{U}_{tF+}$.

2. *Distribution of random variables*

The vector $\mathbf{U}_t = \begin{pmatrix} \mathbf{U}_{t++} \\ \mathbf{U}_{tF+} \end{pmatrix}$ of sire t effects, depending on daughters' genotypes, follows a binormal distribution :

$$f(u_t) = \frac{1}{2\pi\ |\mathbf{D}|^{1/2}} \exp\left(-\frac{1}{2} \cdot u'_t \cdot \mathbf{D}^{-1} \cdot u_t\right)$$

$$\text{with } \mathbf{D} = \begin{pmatrix} \sigma^2_{u/++} & \rho \cdot \sigma_{u/++} \cdot \sigma_{u/F+} \\ \rho \cdot \sigma_{u/++} \cdot \sigma_{u/F+} & \sigma^2_{u/F+} \end{pmatrix}$$

The vector of residuals $\mathbf{E}_{ti/g_{ti}}$ conditional on genotype g_{ti} of daughter *ti* is supposed to be multinormally distributed with zero mean and a $n_{ti} \times n_{ti}$ variance-covariance matrix :

$$\mathbf{R}_{ti/g_{ti}} = \sigma^2_{E/g_{ti}} \begin{pmatrix} 1 & r & r & \dots & r \\ r & 1 & r & \dots & r \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ r & r & r & \dots & 1 \end{pmatrix}$$

where r is the repeatability of the trait, supposed independent of the genotype.

There is independence between :

— the different random sire effects,
— the residuals of the performances of different daughters,
— the sire effects and the residuals.

With this model, two heritabilities have to be defined, reflecting the polygenic relationship between a sire and its daughters, depending on whether they are ++ or *F*+ :

$$h^2_{++} = \frac{4 \cdot \sigma^2_{u/++}}{\sigma^2_{u/++} + \sigma^2_{E/++}}$$

$$h^2_{F+} = \frac{4 \cdot \sigma^2_{u/F+}}{\sigma^2_{u/F+} + \sigma^2_{E/F+}}$$

In this context, the ρ parameter can be defined as a genetic correlation.

3. *Notation for incidence matrices*

The random vector $\mathbf{Y}_{ti/g_{ti}}$ of the performances of the t^{th} sire's i^{th} daughter conditional on its genotypes g_{ti} can be written :

$$\mathbf{Y}_{ti/g_{ti}} = \mathbf{X}_{ti} \cdot \mathbf{b} + \mathbf{W}_{ti/g_{ti}} \cdot \boldsymbol{\beta} + \mathbf{Z}_{ti/g_{ti}} \cdot \mathbf{U} + \mathbf{E}_{ti/g_{ti}}$$

where $\mathbf{X}_{ti}$, $\mathbf{W}_{ti/g_{ti}}$ and $\mathbf{Z}_{ti/g_{ti}}$ are the incidence matrices corresponding to vectors **b**, $\boldsymbol{\beta}$ and **U** respectively.

The common part of $\mathbf{W}_{ti/++}$ and $\mathbf{W}_{ti/F+}$ is noted $\mathbf{W}_{ti}$.

We shall have :

$$\mathbf{W}_{ti/++} \cdot \boldsymbol{\beta} = (\mathbf{W}_{ti},\ 0) \cdot \begin{pmatrix} \boldsymbol{\beta}_{/++} \\ \boldsymbol{\beta}_{/F+} \end{pmatrix} \quad \text{and} \quad \mathbf{W}_{ti/F+} \cdot \boldsymbol{\beta} = (0,\ \mathbf{W}_{ti}) \cdot \begin{pmatrix} \boldsymbol{\beta}_{/++} \\ \boldsymbol{\beta}_{/F+} \end{pmatrix}$$

Thus $\mathbf{W}_{ti/++} \cdot \boldsymbol{\beta} = \mathbf{W}_{ti} \cdot \boldsymbol{\beta}_{/++}$ and $\mathbf{W}_{ti/F+} \cdot \boldsymbol{\beta} = \mathbf{W}_{ti} \cdot \boldsymbol{\beta}_{/F+}$ ·

Similary, we have $\mathbf{Z}_{ti/g_{ti}} \cdot \mathbf{U} = \mathbf{Z}_{ti} \cdot \mathbf{U}_{/g_{ti}}$.

Finally, the preceding incidence matrices will be generalized in $\mathbf{X}_t$, $\mathbf{W}_t$, $\mathbf{Z}_t$ and **X**, **W**, **Z** when considering random vectors $\mathbf{Y}_t$ and **Y**, respectively.

4. *Expression of performance distribution conditionally on the genotype*

According to the assumptions and notations presented above, the joint density of the random vector of the t^{th} sire's daughters' performances $\mathbf{Y}_{t/\gamma_t}$, conditional on their genotypes γ_t, is multinormal with

— a mean $\boldsymbol{\mu}_{ti/\gamma_t} = \mathbf{X}_t \cdot \mathbf{b} + \mathbf{W}_{ti/\gamma_t} \cdot \boldsymbol{\beta}$

— a variance-covariance matrix $\mathbf{V}_{ti/\gamma_t} = \mathbf{Z'}_{ti/\gamma_t} \cdot \mathbf{D} \cdot \mathbf{Z}_{ti/\gamma_t} + \mathbf{R}_{ti/\gamma_t}$

where

$$\mathbf{R}_{ti/\gamma_t} = \begin{pmatrix} \mathbf{R}_{t1/g_{t1}} & 0 & \dots & 0 \\ 0 & \mathbf{R}_{t2/g_{t2}} & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & \mathbf{R}_{tn_t/g_{tn_t}} \end{pmatrix}$$

Similarly, the mean vector and variance-covariance matrix of the random vector $\mathbf{Y}_{ti/g_{ti}}$ of the ti^{th} daughter performances, conditional on its genotype g_{ti}, are denoted $\boldsymbol{\mu}_{ti/g_{ti}}$ and $\mathbf{V}_{ti/g_{ti}}$, respectively.

III. Objectives

The prior distribution of sire genotypes is assumed to be known. These sires being unrelated, we obtain

$$\text{prob}(\boldsymbol{\gamma}) = \prod_t \text{prob}\ (g_t).$$

With the method described here, the genotypic classification of sires and their daughters is given by estimating the posterior distribution of sire genotypes prob$(g_t/\mathbf{y}_t)$, and, conditional on these genotypes, the posterior distribution of their daughters' genotypes prob$(g_{ti}/\mathbf{y}_t$ and $g_t)$.

IV. Methods

A. *Expression of the posterior probabilities of sire and daughter genotypes, conditionally on the sire random effect* $\mathbf{U}_t$, *the parameters of the model being assumed to be know*

1. *Posterior distribution of sire genotypes*

The aim is to calculate prob$(\boldsymbol{\gamma}/\mathbf{y})$. Under our assumptions, we can write :

$$\text{prob}(\boldsymbol{\gamma}/\mathbf{y}) = \prod_t \text{prob}(g_t/\mathbf{y}_t).$$

We are looking for the T probabilities prob$(g_t/\mathbf{y}_t)$. Bayes theorem gives :

$$\text{prob}\ (g_t/\mathbf{y}_t) = \frac{\text{prob}(g_t) \cdot f\ (\mathbf{y}_t/g_t)}{\sum_{g'_t} \text{prob}(g'_t) \cdot f\ (\mathbf{y}_t/g'_t)}$$

The quantity prob(g_t) is the prior probability that the genotype of sire t is g_t.

The density $f(\mathbf{y}_t/g_t)$ can be described by the sum :

$$f(\mathbf{y}_t/g_t) = \sum_{\gamma_t} \text{prob}(\boldsymbol{\gamma}_t/g_t) \cdot f(\mathbf{y}_t/g_t \text{ and } \boldsymbol{\gamma}_t)$$

where the summation of the 2^{n_t} possible vectors $\boldsymbol{\gamma}_t$ forms a complete sum of events.

Practically the sum over the 2^{n_t} possible vectors $\boldsymbol{\gamma}_t$ is impossible as soon as the number of daughters exceeds 10. In order to avoid this difficulty, we shall work conditionally on the random sire effect $\mathbf{U}_t$:

$$f(\mathbf{y}_t/g_t) = \int\int f(\mathbf{u}_t) \cdot f(\mathbf{y}_t/g_t \text{ and } \mathbf{u}_t)\, d\mathbf{u}_t.$$

But, conditionally on genotype G_t and polygenic effect $\mathbf{U}_t$ of their sire t, the performances $\mathbf{Y}_{ti}$ and $\mathbf{Y}_{ti'}$ of two distinct daughters are independent :

$$f(\mathbf{y}_t/g_t \text{ and } \mathbf{u}_t) = \prod_i f(\mathbf{y}_{ti}/g_t \text{ and } \mathbf{u}_t)$$

and $$f(\mathbf{y}_{ti}/g_t \text{ and } \mathbf{u}_t) = \sum_{g_{ti}} \text{prob}(g_{ti}/g_t) \cdot f(\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)$$

where $f(\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)$ is the density function of a normal distribution with a mean $\boldsymbol{\mu}_{ti/g_{ti}} + \mathbf{u}_{t/g_{ti}}$ and a variance-covariance matrix $\mathbf{R}_{ti/g_{ti}}$.

Consequently the desired density-function can be written

$$f(\mathbf{y}_t/g_t) = \int\int f(\mathbf{u}_t) \cdot \{ \prod_i [\sum_{g_{ti}} \text{prob}(g_{ti}/g_t) \cdot f(\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)]\} \cdot d\mathbf{u}_t.$$

2. *Posterior distribution of daughter genotypes conditional on their sires' genotypes*

The aim is to calculate $\text{prob}(g_{ti}/\mathbf{y}_t \text{ and } g_t)$. As before we shall work conditionally on the random sire effect $\mathbf{U}_t$:

$$\text{prob}(g_{ti}/\mathbf{y}_t \text{ and } g_t) = \int\int f(\mathbf{u}_t) \cdot \text{prob}(g_{ti}/\mathbf{y}_t,\ g_t,\ \mathbf{u}_t) \cdot d\mathbf{u}_t.$$

But, taking into account the assumptions adopted,

$$\text{prob}(g_{ti}/\mathbf{y}_t,\ g_t,\ \mathbf{u}_t) = \text{prob}(g_{ti}/\mathbf{y}_{ti},\ g_t,\ \mathbf{u}_t).$$

Using Bayes theorem and substituting $f(\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)$ to $f(\mathbf{y}_{ti}/g_{ti},\ g_t,\ \mathbf{u}_t)$ as well as $\text{prob}(g_{ti}/g_t)$ to $\text{prob}(g_{ti}/g_t \text{ and } \mathbf{u}_t)$ — because of our assumptions —, we can write :

$$\text{prob}(g_{ti}/\mathbf{y}_{ti},\ g_t,\ \mathbf{u}_t) = \frac{\text{prob}(g_{ti}/g_t) \cdot f\ (\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)}{\sum_{g'_{ti}} \text{prob}(g'_{ti}/g_t) \cdot f\ (\mathbf{y}_{ti}/g'_{ti} \text{ and } \mathbf{u}_t)}$$

Our assumptions enable us to write :

$$\text{prob}(g_{ti}/\mathbf{u}_{ti},\ g_t,\ \mathbf{u}_t) = \int\int f(\mathbf{u}_t) \times \frac{\text{prob}(g_{ti}/g_t) \cdot f(\mathbf{y}_{ti}/g_{ti} \text{ and } \mathbf{u}_t)}{\text{prob}(g'_{ti}/g_t) \cdot f\ (\mathbf{y}_{ti}/g'_{ti} \text{ and } \mathbf{u}_t)}\, d\mathbf{u}_t$$

B. *Estimation of the unknown parameters and of the posterior probabililites of the genotypes*

Heritabilities h^2_{++}, and h^2_{F+}, genetic correlation ρ, and repeatability r are assumed to be known. The unknown parameters to be estimated ($\boldsymbol{\theta}$ vector) are the location parameters ($\mathbf{b}$ and $\boldsymbol{\beta}$) and some of the dispersion parameters (sires and residual

variances). These parameters could be estimated by the maximum likelihood method, i.e. by maximizing the probability of observing the measures :

$$f(\mathbf{y}) = \prod_t f(\mathbf{y}_t) = \prod_t \left[\sum_{g_t} f(\mathbf{y}_t/g_t)\ \text{prob}(g_t)\right].$$

Expression of $f(\mathbf{y}_t/g_t)$ is given in section IV.A.1.

Then we shall use the subscripts θ or $\hat{\theta}$ in denoting the probabilities of the different events and their estimates.

Although it is numerically possible to integrate $f(\mathbf{y}_t/g_t)$ with respect to $\mathbf{u}_t$ when $\boldsymbol{\theta}$ parameters are known, we did not find any practical solution when $\boldsymbol{\theta}$ parameters are to be estimated. Our proposition, therefore, is to estimate $f(\mathbf{y}_t/g_t)$ by $f_{\hat{\boldsymbol{\theta}}}(\mathbf{y}_t/g_t$ and $\hat{\mathbf{u}}_t)$ where $\hat{\mathbf{u}}_t$ is the mode of the distribution of $\mathbf{U}_t$ conditional on $\mathbf{Y}_t$, noting that $\hat{\mathbf{u}}_t$ maximizes the joint density of the $\mathbf{Y}_t$ and $\mathbf{U}_t$, f_θ ($\mathbf{u}_t$ and $\mathbf{y}_t$).

This approach will be discussed later. We use it according to GIANOLA & FOULLEY (1983) who clearly showed its limits and its value in the context of Bayesian theory of selection indices.

Looking simultaneously for the estimates of $\boldsymbol{\theta}$ parameters and the modal value of the distribution of $\mathbf{U}$, conditional on $\mathbf{Y}$, drives us to maximize, with respect to $\mathbf{u}_t$ values and $\boldsymbol{\theta}$ parameters, the quantity $\prod_t f_\theta(\mathbf{y}_t$ and $\mathbf{u}_t)$.

Then, $\text{prob}_{\hat{\boldsymbol{\theta}}}(g_{ti}/g_t,\ \mathbf{y}_t$ and $\hat{\mathbf{u}}_t)$ can be deduced firstly, $\text{prob}_{\hat{\boldsymbol{\theta}}}(g_t/\mathbf{y}_t$ and $\hat{\mathbf{u}}_t)$ secondly.

V. Solutions

To avoid burdening this paper with unnecessary algebra, it can be simply stated that the solutions were obtained by equating to zero the first derivatives of the logarithm of the density $\prod_t f_\theta(\mathbf{y}_t$ and $\mathbf{u}_t)$.

The proposed solution is an iterative two step procedure :

— the first step is to estimate $\boldsymbol{\theta}$ and $\mathbf{u}$, given the probability P_{ti} that each female *ti* would be *F+* ;

— the second step is to estimate, given the $\hat{\boldsymbol{\theta}}$ parameters and $\hat{\mathbf{u}}$ values, the posterior probabilities :

$$\text{prob}_{\hat{\boldsymbol{\theta}}}(g_{ti}/g_t,\ \mathbf{y}_{ti}\ \text{and}\ \hat{\mathbf{u}}_t)$$
$$\text{prob}_{\hat{\boldsymbol{\theta}}}(g_t/\mathbf{y}_t\ \text{and}\ \hat{\mathbf{u}}_t)$$

$$\text{thus } P_{ti} = \text{prob}_{\hat{\boldsymbol{\theta}}}(G_t = FF/\mathbf{y}_t,\ \hat{\mathbf{u}}_t) + \text{prob}_{\hat{\boldsymbol{\theta}}}(G_t = F+/\mathbf{y}_t,\ \hat{\mathbf{u}}_t) \cdot \text{prob}_{\hat{\boldsymbol{\theta}}}(G_{ti} = F+/G_t = F+,\ \mathbf{y}_{ti},\ \hat{\mathbf{u}}_t).$$

At this point, we can return to the parameters estimation step and continue until the results converge. To that end, the successive values of the estimated parameters or of the density $\prod_t f_{\hat{\boldsymbol{\theta}}}(\mathbf{y}_t$ and $\hat{\mathbf{u}}_t)$ must be compared.

A. *Estimation of the* **b**, **β** *and* **u** *vectors*

Estimates of the **b**, **β** and **u** vectors are obtained by simultaneously solving the system :

MAT · PARA = RHS, with :

$$\mathbf{MAT} = \begin{pmatrix} \mathbf{X}' \cdot (\mathbf{R}_{++}^{-1} + \mathbf{R}_{F+}^{-1}) \cdot \mathbf{X} & \mathbf{X}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{W} & \mathbf{X}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{W} & \mathbf{X}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{Z} & \mathbf{X}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{Z} \\ \mathbf{W}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{X} & \mathbf{W}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{W} & 0 & \mathbf{W}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{Z} & 0 \\ \mathbf{W}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{X} & 0 & \mathbf{W}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{W} & 0 & \mathbf{W}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{Z} \\ \mathbf{Z}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{X} & \mathbf{Z}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{W} & 0 & \mathbf{Z}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{Z} + \mathbf{\Delta}_{00}^{-1} & \mathbf{\Delta}_{01}^{-1} \\ \mathbf{Z}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{X} & 0 & \mathbf{Z}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{W} & \mathbf{\Delta}_{10}^{-1} & \mathbf{Z}' \cdot \mathbf{R}^{-1} \cdot \mathbf{Z} + \mathbf{\Delta}_{11}^{-1} \end{pmatrix}$$

$$\mathbf{PARA} = \begin{pmatrix} \mathbf{b} \\ \boldsymbol{\beta}_{/++} \\ \boldsymbol{\beta}_{/F+} \\ \mathbf{u}_{/++} \\ \mathbf{u}_{/F+} \end{pmatrix} \qquad \mathbf{RHS} = \begin{pmatrix} \mathbf{X}' \cdot (\mathbf{R}_{++}^{-1} + \mathbf{R}_{F+}^{-1}) \cdot \mathbf{y} \\ \mathbf{W}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{y} \\ \mathbf{W}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{y} \\ \mathbf{Z}' \cdot \mathbf{R}_{++}^{-1} \cdot \mathbf{y} \\ \mathbf{Z}' \cdot \mathbf{R}_{F+}^{-1} \cdot \mathbf{y} \end{pmatrix}$$

The $\mathbf{R}_{++}^{-1}$ matrix is a block diagonal one, the block *ti* being given by $\mathbf{R}_{ii/++}^{-1}(1 - P_{ti})$. In the same way, the matrix $\mathbf{R}_{F+}^{-1}$ is made of blocks $\mathbf{R}_{F+}^{-1} \cdot P_{ti}$.

With $\mathbf{I}_T$ being the $T \times T$ identity matrix, we get :

$$\mathbf{\Delta}_{00}^{-1} \times \frac{\sigma^2_{u/F+}}{\sigma^2_{u/++}\,\sigma^2_{u/F+}(1 - \rho^2)} \times \mathbf{I}_T$$

$$\mathbf{\Delta}_{11}^{-1} = \mathbf{\Delta}_{00}^{-1} \times \frac{\sigma^2_{u/++}}{\sigma^2_{u/F+}}$$

$$\text{and} \quad \mathbf{\Delta}_{10}^{-1} = \mathbf{\Delta}_{01}^{-1} = \frac{-\rho \times \sigma_{u/++} \times \sigma_{u/F+}}{\sigma^2_{u/F+}} \times \mathbf{I}_T$$

Thus, estimates of the **b** and **β** parameters and of the **u** modal values are obtained, after each iteration, by solving a linear system of equations quite similar to the BLUP (HENDERSON, 1973).

B. *Variance estimation*

Estimates of the variances of sire effects are given by solving the following system :

$$\frac{a_{11}}{\sigma^2_{u/++}} + \frac{a_{12}}{\sigma_{u/F+} \cdot \sigma_{u/++}} = b_1 \qquad \text{and} \qquad \frac{a_{21}}{\sigma_{u/F+} \cdot \sigma_{u/++}} + \frac{a_{22}}{\sigma^2_{u/F+}} = b_2$$

$$\text{with } a_{12} = a_{21} = -\frac{\rho}{1 - \rho^2} \sum_t (\mathbf{u}_{t/F+} \cdot \mathbf{u}_{t/++})$$

$$a_{11} = \frac{(\Sigma \mathbf{u}^2_{ti/++})}{1 - \rho^2} + \{\sum_t \sum_i (1 - P_{ti}) \cdot \mathbf{z}'_{ti/++} \cdot \mathbf{R}^{-1}_{++} \cdot \mathbf{z}_{ti/++}\} \cdot k^2_{++}$$

$$a_{22} = \frac{(\Sigma \mathbf{u}^2_{ti/F+})}{1 - \rho^2} + \{\sum_t \sum_i P_{ti} \cdot \mathbf{z}'_{ti/F+} \cdot \mathbf{R}^{-1}_{F+} \cdot \mathbf{z}_{ti/F+}\} \cdot k^2_{F+}$$

where k^2_{F+} and k^2_{++} are the ratios of sire/residual variances ($k^2_g = h^2_g/(4 - h^2_g)$)

and where $\mathbf{z}_{ti/g_{ti}}$ is the vector of the deviations : $\mathbf{y}_{ti/g_{ti}} - \boldsymbol{\mu}_{ti/g_{ti}} - \mathbf{u}_{t/g_{ti}}$.
Finally, b_1 and b_2 are given by :

$$b_1 = \mathbf{T} + \sum_t \sum_i n_{ti} (1 - P_{ti})$$

$$b_2 = \mathbf{T} + \sum_t \sum_i n_{ti} P_{ti}$$

The sire variances are found simply by solving a second degree equation. The residual variances follow.

C. *Estimates of the posterior probabilities of genotypes*

Given the values of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{u}}$, we estimate the genotypic probabilities and suggest the following steps :

— the corrected records are given $\hat{\mathbf{z}}_{ti/g_{ti}}$ (*see* before)

— the probabilities of the records of each daughter may be calculated :

$$f_{\hat{\theta}}(\mathbf{y}_{ti}/g_{ti} \text{ and } \hat{\mathbf{u}}_t) = \frac{1}{(2\,\Pi^{n_{ti}} \mid \hat{\mathbf{R}}_{ti/g_{ti}} \mid)^{1/2}} \cdot \exp(-\,1/2\,\hat{\mathbf{z}}'_{ti/g_{ti}} \cdot \hat{\mathbf{R}}^{-1}_{g_{ti}} \cdot \hat{\mathbf{z}}_{ti/g_{ti}})$$

— for each daughter, we estimate the quantities :

$\hat{q}(g_{ti}/g_t) = \text{prob}(g_{ti}/g_t) \cdot f_{\hat{\theta}}(\mathbf{y}_{ti}/\hat{\mathbf{u}}_t \text{ et } g_{ti})$

— and for each sire, the quantities $\hat{q}(g_t) = \text{prob}(g_t) \cdot \prod_i \sum_{g_{ti}} \hat{q}(g_{ti}/g_t)$

— then we obtain $\text{prob}_{\hat{\theta}}(g_{ti}/g_{ti} \text{ and } \mathbf{y}_t) = \dfrac{\hat{q}(g_{ti}/g_t)}{\sum_{g_{ti}} \hat{q}(g_{ti}/g_t)}$

and $\text{prob}_{\hat{\theta}}(g_t/\mathbf{y}_t) = \dfrac{\hat{q}(g_t)}{\sum \hat{q}(g_t)}$

— finally $P_{ti} = \text{prob}_{\hat{\theta}}(G_{ti} = F+/G_t = F+ \text{ and } \mathbf{y}_t) \cdot \text{prob}_{\hat{\theta}}(G_t = F+/\mathbf{y}_t)$
$+ \text{prob}_{\hat{\theta}}(G_t = FF/\mathbf{y}_t)$.

At this moment, we can return to the parameters estimation step and continue until the results converge. To that end, the successive values of the estimated parameters or of the density $\prod_t f_{\hat{\theta}}(\mathbf{y}_t \text{ and } \hat{\mathbf{u}}_t)$ must be compared.

VI. Illustration

As the computations corresponding to the proposed method are long, the results given here concern only a limited number of simulations (10 per case). Thus, they must be considered just as indicative tendencies. In order to show the properties and limits of the method, we studied different situations for the number of sires (5, 10 and 20), daughters per sire (10, 20, 30, 50, 100, 150), mean value μ_{F+} of the $F+$ daughters' measurements (from 0.5 to 3.5), variances σ^2_{F+} of $F+$ daughter's measurements (1, 2, 3 and 4) and heritabilities (0.1 to 0.6). In all cases, the two previously defined heritabilities, h^2_{++} and h^2_{F+}, are assumed to be equal (they will be denoted h^2), and the following parameters are given the values :

— prior probabilities of the genotypes : 0.5 for the $F+$ and 0.5 for the $++$, corresponding to the general situation during the fixation of a major gene into a new breed,
— mean values μ_{++} of the $++$: 0,
— variance σ^2_{++} of the $++$: 1,
— genetic correlation ρ : 0.8,
— number of measurements per daughter : 1.

TABLE 1

Percentage P'_α of errors among the classified sire, percentage P''_α of sires whose genotypes remain undetermined, and estimates (with standard deviations) of means and variances for various values of assumed μ_{F+}, number of sires, and number of daughters per sire (with $h^2 = 0.2$ and $\sigma^2_{++} = \sigma^2_{F+} = 1$)

a) with $\mu_{F+} = 1$

Number of		$P'_{0.5}$	$P'_{0.9}$	$P''_{0.9}$	$\hat{\mu}_{++}$	$\hat{\mu}_{F+}$	$\hat{\sigma}^2_{++}$	$\hat{\sigma}^2_{F+}$
sires	daughters							
5	10	34	29	24	0.05 ± .19	1.15 ± .75	0.83 ± .36	0.19 ± .23
5	20	40	40	3	0.11 ± .27	0.92 ± .70	0.85 ± .26	0.14 ± .19
5	30	18	18	0	− .10 ± .24	1.37 ± .36	0.86 ± .14	0.54 ± .35
10	10	35	33	20	− .05 ± .24	1.31 ± .48	0.68 ± .18	0.27 ± .24
10	20	27	27	7	0.02 ± .16	1.14 ± .37	0.89 ± .18	0.46 ± .34
10	30	16	15	22	− .08 ± .14	1.26 ± .34	0.89 ± .12	0.64 ± .19

b) with $\mu_{F+} = 2$

Number of		$P'_{0.5}$	$P'_{0.9}$	$P''_{0.9}$	$\hat{\mu}_{++}$	$\hat{\mu}_{F+}$	$\hat{\sigma}^2_{++}$	$\hat{\sigma}^2_{F+}$
sires	daughters							
5	10	14	12	2	− .08 ± .28	2.05 ± .54	0.79 ± .26	0.36 ± .26
5	20	14	14	0	− .08 ± .19	1.76 ± .55	0.75 ± .15	0.64 ± .25
5	30	2	2	0	0.02 ± .16	2.05 ± .27	0.88 ± .11	0.76 ± .23
10	10	11	10	9	− .01 ± .15	2.05 ± .41	0.76 ± .19	0.34 ± .17
10	20	3	3	0	− .03 ± .20	2.03 ± .06	0.80 ± .06	0.66 ± .23
10	30	3	3	0	− .03 ± .11	2.05 ± .24	0.91 ± .11	0.82 ± .17
20	50	2	2	0	− .01 ± .07	1.97 ± .07	0.90 ± .04	0.85 ± .10
20	100	1	1	0	0.01 ± .05	1.98 ± .07	0.95 ± .02	0.91 ± .06
20	150	0	0	0	0.0 ± .05	2.00 ± .02	0.91 ± .03	0.90 ± .08

A sire was classified in a genotypic class ($F+$ or $++$) if the estimate of the posterior probability of this genotype was more than a threshold α.

Each simulation gives the estimated posterior probabilities of the genotypes and the estimates of the parameters. Deprived of any objective measurement of the quality of the probability estimation, we chose to give the percentage P'_α of errors among the sires classified by using the following criterion : a sire is classified in a genotypic class ($F+$ or $++$) if the estimate of posterior probability of its genotype is more than a threshold α (0.5 or 0.9). When the threshold is 0.9, some sires cannot be classified and we give also the percentage of sires whose genotype remains undetermined. Concerning the parameters, we give the averaged values and standard deviation of the means (μ_{++}, μ_{F+}) and of the variances (σ^2_{++}, σ^2_{F+}).

Results are given in tables 1 and 2. As expected, the quality of the classification and of the parameter estimation increased with the number of sires and more drastically with the number of their daughters. A minimum of 20 daughters per sire seems necessary for a sufficient accuracy. Differences between the two probability criteria P'_α ($P'_{0.5}$ and $P'_{0.9}$) are notable : the percentages of misclassified sires are quite similar when the mean value μ_{F+} is high (excluding the extreme situation where sires are tested on 10 daughters) but rather different when this mean value is only 1 standard deviation. In fact, the second criterion P'_α shows that the general situation for $\mu_{F+} = 2$ is that the posterior probabilities are near 0 or 1 but that, for $\mu_{F+} = 1$, the prior information is dominant (unless the number of daughters is high) leading to probabilities near 0.5.

TABLE 2

Percentage P'_α of errors among the classified sires, percentage P''_α of sires whose genotypes remain undetermined, and estimates (with standard deviations) of means and variances for various values of parameters μ_{F+}, h^2, and σ^2_{F+} (with 10 sires and 20 daughters per sire)

Parameter	Value	$P'_{0.5}$	$P'_{0.9}$	$P''_{0.9}$	$\hat{\mu}_{++}$	$\hat{\mu}_{F+}$	$\hat{\sigma}^2_{++}$	$\hat{\sigma}^2_{F+}$
	0.5	41	37	9	0.03 ± .17	0.69 ± .58	0.89 ± .08	0.18 ± .20
	1.0	28	28	12	− .04 ± .20	1.13 ± .44	0.78 ± .09	0.57 ± .48
	1.5	20	20	2	0.11 ± .33	1.81 ± .26	0.88 ± .18	0.45 ± .27
μ_{F+} (*)	2.0	2	1	0	0.03 ± .07	2.31 ± .24	0.88 ± .08	0.62 ± .20
	2.5	3	2	0	0.01 ± .14	2.49 ± .20	0.88 ± .08	0.74 ± .18
	3.0	3	3	0	0.07 ± .19	2.91 ± .38	0.85 ± .09	0.84 ± .29
	3.5	0	0	0	0.04 ± .11	3.49 ± .15	0.86 ± .10	0.65 ± .17
	0.1	2	1	2	0.01 ± .11	2.13 ± .13	0.87 ± .10	0.75 ± .20
	0.3	6	6	0	0.02 ± .18	1.99 ± .23	0.76 ± .14	0.67 ± .13
h^2 (+)	0.4	10	10	0	− 0.3± .16	1.94 ± .31	0.76 ± .11	0.56 ± .17
	0.5	10	10	0	0.01 ± .14	2.16 ± .24	0.80 ± .10	0.70 ± .23
	0.6	8	8	0	− 0.6± .18	2.04 ± .35	0.77 ± .08	0.51 ± .17
	2	4	4	1	− .08 ± .12	2.12 ± .37	0.82 ± .10	1.18 ± .47
σ^2_{F+} (•)	3	7	7	0	− .04 ± .08	2.31 ± .52	0.87 ± .13	1.51 ± .64
	4	4	3	1	− .08 ± .11	2.51 ± .80	0.92 ± .14	1.93 ± 1.0

A sire was classified in a genotypic class ($F+$ or $++$) if the estimate of the posterior probability of this genotype was more than an threshold α.

(*) $h^2 = 0.2$ and $\sigma^2_{F+} = 1$ (+) $\mu_{F+} = 2$ and $\sigma^2_{F+} = 1$ (•) $h^2 = 0.2$ and $\mu_{F+} = 2$

Table 2 gives some more information for the case where 10 sires are tested on 20 daughters. The first part concerns the magnitude of the differences between means μ_{F+}-μ_{++}. A threshold appears around a deviation of 2 units and the power seems poor for differences of 1 standard deviation or less. The heritability is not a very important parameter even if, as expected, the accuracy of the method decreases when this parameter increases, the separation between major gene and polygenic variation being more and more difficult. The difference between the variances of the two genotypes σ^2_{++} and σ^2_{F+} does not play a great role in the discrimination.

VII. Discussion and conclusion

A. *Discussion concerning the proposed method*

Solutions obtained depend on a number of assumptions and simplifications which have to be emphasized.

1. *Assumptions*

Only the case where dams are known to be homozygous ++ was considered. As mentioned above, this is the general situation when progeny testing sires in a structured design for fixation of a new major gene in a breed (*see* for instance ELSEN *et al.*, 1985). Nevertheless, when intercrossings are made, at the end of such a process, in order to create *FF* animals, the assumption falls down. Then daughter genotypes will have to be determined simultaneously. Approaches similar to that described here could probably be followed.

We assumed here that the progeny tested sires were unrelated. In the opposite case, two levels of complications would occur : the prior probabilities of genotypes cannot be written as the product of separated terms and off diagonal non zero terms appear in the variance-covariance matrix of the polygenic random sire effect. The second point could probably be neglected when the heritability and genetic relationships are low, whereas the first one seems very crucial since all the daughters of sires related to a particular sire will inform on its own genotype. The computations will be simplified if the group of sires can be partitioned into independent families.

We studied a gene with only two alleles (*F* and +). Generalization to a larger number of alleles does not cause any difficulties and is given in FOULLEY & ELSEN (1988).

Finally, we assumed that the sire effect was a bivariate phenomenon, defining two heritabilities and a genetic correlation. Other assumptions could be made. The first one is a unique random sire effect leading to the definition of a unique error variance if the heritability is still given and assumed to be the same for both genotypes, or to the estimation of different heritabilities if the total calculated variances may be different. A second approach would be to define a proportionality coefficient c and to describe the sire random effect as $\mathbf{U}_i$ or $c.\mathbf{U}_i$ depending on the genotype of the daughter. Whatever the hypothesis, the problem of prior information on these parameters appears and requires preliminary investigations.

2. *Simplifications*

A major point in the proposed method is the replacement in the likelihood function of the integration over $\mathbf{u}$ by searching for the modal value of the posterior

random sire effect **U**. As suggested by GIANOLA & FOULLEY (1983), the validity of these methods depends on the form of the posterior distribution of **U**, the hypothesis being that it is symmetric and sharp enough. This must be checked relative to current parameters. Using rapid computers, the possibility of integration over **u** cannot be neglected, at least when the numbers of animals are not too high.

B. *Discussion concerning the classification criteria*

The posterior probabilities described here are useful when describing a population. Nevertheless, they cannot be directly used for decisions when carriers are to be kept and non-carriers to be eliminated. In the illustration, we suggested a decision criterion based on the comparison between the probability value and a threshold. Other methods could be adopted considering for instance the costs of the errors.

We suggest a test for a hypothesis H_0 concerning sire genotypes. This hypothesis is that the realization of the genotypes vector Γ is $\gamma_0 = (g_1, g_2, \ldots g_T)$. Strictly speaking, there is no general hypothesis for sire genotypes and this causes two difficulties : firstly, the hypothesis to be tested being not nested in a general one, the classical asymptotic properties of the maximum likelihood ratio test can no longer be used, resulting in more complicated methods (Cox, 1961). Secondly, there is no absolute reference to compare a particular hypothesis and H_0 has to be tested against a^{T-1} other hypotheses concerning vector Γ (a being the number of possible genotypes per sire).

To prevent this difficulty, we suggest use of a process similar to segregation analysis, introducting the probability p_t that a sire t gives the F allele to one daughter. Biologically, this probability can only take the values 0, 1/2 or 1. But we suppose here that p_t can take any value in the interval [0, 1]. We shall denote as $\mathbf{p}(\gamma)$ the vector of probabilities $(p_1, p_2, \ldots, p_T)$; $\mathbf{p}(\gamma_0)$ will be this vector under the hypothesis H_0 : $\mathbf{p}(\gamma_0) = \{p_{1/0}, p_{2/0}, \ldots p_{T/0}\}$.

The proposed test is done as follows (*see* the appendix for details) :

- H_1 hypothesis : $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{u}}$ are determined by maximizing the density $M_1(\boldsymbol{\theta}, \mathbf{u}, \mathbf{p}(\gamma)/\mathbf{y})$:

$$\prod_t \{f_{\boldsymbol{\theta}}(\mathbf{u}_t) \cdot \prod_i [p_t \cdot f_{\boldsymbol{\theta}}(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = F+) + (1 - p_t) \cdot f_{\boldsymbol{\theta}}(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = ++)]\}$$

- H_0 hypothesis : $\hat{\hat{\boldsymbol{\theta}}}$, $\hat{\hat{\mathbf{u}}}$ are determined by maximizing the density $M_0(\boldsymbol{\theta}, \mathbf{u}, \mathbf{p}(\gamma_0)/\mathbf{y})$:

$$\prod_t \{f_{\boldsymbol{\theta}}(\mathbf{u}_t) \cdot \prod_i [p_{t/0} \cdot f_{\boldsymbol{\theta}}(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = F+) + (1 - p_{t/0}) \cdot f_{\boldsymbol{\theta}}(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = ++)]\}$$

- the ratio $l(\gamma_0) = -2.\log \dfrac{M_0(\hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\mathbf{u}}}, \mathbf{p}(\gamma_0)/\mathbf{y})}{M_1(\hat{\boldsymbol{\theta}}, \hat{\mathbf{u}}, \mathbf{p}(\gamma)/\mathbf{y})}$ is calculated

- this ratio $l(\gamma_0)$ has to be compared to a threshold $t(\alpha)$. If $l(\gamma_0) > t(\alpha)$, H_0 hypothesis is rejected at the α level.

Unfortunately, M_0 and M_1 not being real likelihood functions, $l(\gamma_0)$ does not seem to converge to the classical χ^2 with T degrees of freedom as would make a true likelihood ratio. Thus, this point needs further research, involving for instance integration over **u**.

Received June 4, 1987.

Accepted November 15, 1987.

References

AITKIN M., WILSON G.T., 1980. Mixture models, outliers and the EM algorithm. *Technometrics*, **22,** 325-331.

BINDON B.M., 1984. Reproductive biology of the *Booroola Merino* sheep. *Aust. J. Biol. Sci.*, **37,** 163-189.

COX D.R., 1961. Tests of separate families of hypotheses (Proc. 4th Berkeley Symp.). *Math. Statist. Prob.*, **1,** 105-123.

DAVIS G.M., KELLY R.W., 1983. Segregation of a major gene influencing ovulation rate in progeny of *Booroola* sheep in commercial and research flocks. *Proc. N.Z. Soc. Anim. Prod.*, **43,** 197-199.

DAVIS G.M., MONTGOMERY G.W., ALLISON A.J., KELLY R.W., BRAY A.R., 1982 *a*. Fecundity in *Booroola Merino* sheep. Further evidence of major gene. *Proc. Aust. Soc. Reprod. Biol.*, **13,** 5-6.

DAVIS G.M., MONGOMERY G.W., KELLY R.W., 1982 *b*. Estimates of the repeatability of ovulation rate in *Booroola* cross ewes. *In : Proceedings of the 2nd World Congress of Genetics Applied to Livestock Production, Madrid, October 4-8, 1982, vol. 8,* 674-679, Editorial Garsi, Madrid.

DAY N.E., 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56,** 463-474.

ELSEN J.M., ORTAVANT R., 1984. Le gène *Booroola*. Mise en évidence. Fonctionnement. Perspectives d'utilisation. *In : 9es Journées de la Recherche Ovine et Caprine, Paris, 5-6 décembre 1984,* 415-451, INRA-ITOVIC, Paris.

ELSEN J.M., VU TIEN J., BOUIX J., RICORDEAU G., 1985. Linear programming model for incorporating the *Booroola* gene into another breed. *In :* LAND R.B., ROBINSON D.W. (ed.), *Genetics of reproduction in sheep,* 175-181, Butterworths, London.

ELSTON N.E., STEWART J., 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.*, **21,** 523-542.

EVERITT B.S., 1984. Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions ; a comparison of different algorithms. *The Statistician*, **33,** 205-215.

FOULLEY J.L., ELSEN J.M., 1988. Posterior probability of the sire's genotype at a major locus based on progeny-test results for discrete characters. *Génét. Sél. Evol.*, **20,** 227-238.

GIANOLA D., FOULLEY J.L., 1983. Sire evaluation for ordered categorical data with a threshold model. *Génét. Sél. Evol.*, **15,** 201-224.

HENDERSON C.R., 1973. Sire evaluation and genetic trends. *In : Proc. Anim. Breed. Genet. Symp. in honor of Dr. J.L. Lush,* 10-41, American Society of Animal Science and American Dairy Science Associations, Champaign, Illinois.

LALOUEL J.M., RAO D.C., MORTON N.E., ELSTON R.C., 1983. A unified model for complex segregation analysis. *Am. J. Hum. Genet.*, **35,** 816-826.

MORTON N.E., MC LEAN C.J., 1974. Analysis of family resemblance. 3. Complex segregation analysis of quantitative traits. *Am. J. Hum. Genet.*, **26,** 489-503.

OWENS J.L., JOHNSTONE P.D., DAVIS G.M., 1985. An independent statistical analysis of ovulation rate data used to segregate *Booroola-Merino* genotypes. *N.Z. J. Agric. Res.*, **28,** 361-363.

PIPER L.R., BINDON B.M., 1982. The *Booroola Merino* and the performance of *medium non-Peppin* crosses at Armidale. *In :* PIPER L.R., BINDON B.M., NETHERY R.D. (ed.), *The Booroola Merino,* 9-20, CSIRO, Melbourne.

PIPER L.R., BINDON B.M., DAVIS G.H., 1985. The single gene inheritance of the high litter size of the *Booroola Merino*. *In :* LAND R.B., ROBINSON D.W., (ed.), *Genetics of reproduction in sheep,* 115-125, Butterworths, London.

Appendix

Proposition of a test for the determination of genotypes

Hypothesis H_0 will be tested by comparing the estimated probabilities of recorded data $f_\theta(\mathbf{y})$ under H_0 and under H_1. These probabilities may be written :

$$\prod_t \{\int\int f_\theta(\mathbf{u}_t) \cdot \prod_i [p_t \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = F+) + (1 - p_t) \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = ++)] \cdot d\mathbf{u}_t\}$$

The likelihood will be obtained through the maximization of these probabilities with respect to $\boldsymbol{\theta}$ (and also to $\mathbf{p}$ under H_1).

As before, we do not integrate with respect to $\mathbf{u}$ but approach $f_\theta(\mathbf{y})$ by $f_\theta(\mathbf{y} \text{ and } \hat{\mathbf{u}})$ where $\hat{\mathbf{u}}$ is the modal value of the distribution of $\mathbf{U}$ conditional to $\mathbf{Y}$.

Then $M_1(\boldsymbol{\theta}, \mathbf{u}, \mathbf{p}(\gamma)/\mathbf{y})$ is

$$\prod_t \{f_\theta(\mathbf{u}_t) \cdot \prod_i [p_t \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = F+) + (1 - p_t) \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = ++)]\}$$

and $M_0(\boldsymbol{\theta}, \mathbf{u}, \mathbf{p}(\gamma_0)/\mathbf{y})$ is

$$\prod_t \{f_\theta(\mathbf{u}_t) \cdot \prod_i [p_{t/0} \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = F+) + (1 - p_{t/0}) \cdot f_\theta(\mathbf{y}_{ti}/\mathbf{u}_t \text{ and } G_{ti} = ++)]\}$$

The algorithm presented for the estimation of the genotypes probabilities can be transposed for this test. Only two points are to be modified : the probability p_{ti} used in the successive estimations of the parameters is defined in another way and we have to calculate at each step the probability p_t.

We, now, have :

$$\hat{P}_{ti} = \frac{\hat{p}_t \cdot f_\theta(\mathbf{y}_{ti}/\hat{\mathbf{u}}_t \text{ and } g_{ti} = F+)}{\hat{p}_t \cdot f_\theta(\mathbf{y}_{ti}/\hat{\mathbf{u}}_t \text{ and } g_{ti} = F+) + (1 - \hat{p}_t) \cdot f_\theta(\mathbf{y}_{ti}/\hat{\mathbf{u}}_t \text{ et } g_{ti} = ++)}$$

The probabilities $\hat{p}_t$ are given by :

$$\hat{p}_t = [\sum_i \text{prob}_\theta(g_{ti} = F+/\mathbf{y}_{ti} \text{ and } \hat{\mathbf{u}}_t)]/n_t$$

We shall have a two steps procedure :

— estimation of the p_t, **PARA,** and variances,
— estimation of the p_{ti}.

Finally, it has to be noted that the results (estimated of **PARA,** of the variances and of the posterior probabilities) are the same as the estimates obtained with the first method when the genotypes of the T sires are fixed. In this case, we have (for the distribution estimation and for the genotypes test, respectively) :

— either : $\text{prob}(G_t = FF)$ and $p_t = 1$
— either : $\text{prob}(G_t = F+)$ and $p_t = 1/2$
— or : $\text{prob}(G_t = ++)$ and $p_t = 0$.