Theor Appl Genet (1988) 76:311–319

Statistical techniques for detection of major genes in animal breeding data*

I. Hoeschele **

Department of Animal Science, Iowa State University, Ames, IA 50011, USA

Received July 6, 1987; Accepted November 11, 1987
Communicated by L. D. Van Vleck

Summary. Statistical techniques for detection of major loci and for making inferences about major locus parameters such as genotypic frequencies, effects and gene action from field-collected data are presented. In field data, major genotypic effects are likely to be masked by a large number of environmental differences in addition to additive and nonadditive polygenic effects. A graphical technique and a procedure for discriminating among genetic hypotheses based on a mixed model accounting for all these factors are proposed. The methods are illustrated by using simulated data.

Key words: Major genes – Mixed inheritance – Mixture data – Normal probability plot – Genetic hypotheses

Introduction

Animal breeding theory for quantitative traits is based on the polygenic model of inheritance. It assumes the breeding value of an individual to be the sum of small and additive effects of many genes. Polygenic traits comprise continuous measurements on production traits such as milk yield and discontinuous traits with threshold character (Falconer 1965). Many secondary traits (reproduction and health) are considered as threshold characters. However, Hanset (1982) and Roberts and Smith (1982) have reviewed examples of single loci accounting for an appreciable amount of the genetic variance in quantitative traits (major loci). Hanset (1982) conjectured that twinning, calving ease, size and resistance to metabolic, infectious and parasitic diseases are "candidates" for mixed major gene and polygenic inheritance and required statistical analysis to detect major genes. Famula (1986) suggested an application of major gene indices (Karlin et al. 1979) to animal breeding data. In human genetics, likelihood-based tests of genetic hypotheses (Elton 1987) are employed to discriminate among modes of inheritance. These approaches are computationally demanding, and it is not obvious how to proceed when major gene effects are masked by a large number of environmental differences in addition to additive and nonadditive polygenic variation.

In this paper, alternative techniques are presented for detection of major loci and for making inferences about the number of major genotypes, major genotypic frequencies, effects and major gene action from field-collected data.

Graphical method for detecting a major locus

Data

Consider the model

$$y_{ik} = g_k + \Delta_i'\theta + e_{ik} \tag{1}$$

where y_{ik} is an observation on the i^{th} individual, g_k is the k^{th} major genotypic value, $\theta' = [\beta', u']$ is a vector of environmental effects (β) and polygenic effects (u), u may be partitioned into additive ($\mathbf{u}_1$) and nonadditive ($\mathbf{u}_2$) effects according to Henderson (1985) with $\mathbf{u} \sim N(\mathbf{O}, \mathbf{G})$, Δ_i is the i^{th} row of the incidence matrix $\Delta = [\mathbf{X}, \mathbf{Z}]$, and e_{ik} is a residual with $\mathrm{var}(e_{ik}) = \sigma_e^2$. If $\mathbf{u}$ is a vector of additive polygenic effects, $\mathbf{G} = \mathbf{A}\,\sigma_u^2$ with $\mathbf{A}$ being a matrix of additive genetic relationships. Denote known major genotype membership by $I_i = k$ with $k \in (1, \ldots, m)$. Then, assuming

[*] Journal Paper No. J-12733 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 1901

[**] Present address: Department of Dairy Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

normality

$$y_i \,|\, I_i = k, g, \theta, \sigma_e^2 \sim N(g_k + \Delta_i'\theta, \sigma_e^2). \qquad (2)$$

With unknown major genotypic membership I_i, y_i has an m-component mixture distribution with mean

$$E_h(y_i \,|\, \mathbf{p}_i, g, \theta, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)\, g_k + \Delta_i'\theta \qquad (3)$$

and variance

$$\mathrm{var}_h(y_i \,|\, \mathbf{p}_i, g, \theta, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)(g_k - \mu_i)^2 + \sigma_e^2, \qquad (4)$$

where $\mathbf{p}_i$ is an $m \times 1$ vector with elements $p(I_i = k)$, the probability that the ith individual has major genotype memberships k ($k = 1, \ldots, m$) and $\mu_i = \sum_{k=1}^{m} p(I_i = k)\, g_k$. Also, h denotes expectation and variance with respect to the density

$$h(y_i \,|\, \mathbf{p}_i, g, \theta, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)\, f(y_i \,|\, I_i = k, g, \theta, \sigma_e^2) \qquad (5)$$

where h(.) is an m-component mixture density and f(.) is the kth component density of the distribution in [2].

Graphical technique

A variety of graphical techniques dealing with mixture data have been developed (Titterington et al. 1985). The data situation in (5) suggests using such techniques to obtain evidence for the presence of several major genotypic means (g_k). For mixtures, the normal quantile-quantile (Q-Q) plot has a typical configuration and is very competitive with respect to sensitivity to departures from normality. A numerical procedure based on the plot can be devised to estimate component means (major genotypic means) and mixing weights (major genotypic frequencies). The Q-Q plot and comparable methods are discussed in the context of detecting mixtures by Harding (1949), Fowlkes (1979) and Titterington et al. (1985).

The theoretical Q-Q plot

Let x and y be random variables with $x \sim N(0, 1)$ and $y \sim N(\mu_y, \sigma_y^2)$. Then, $y = \mu_y + \sigma_y x$ and $\delta y / \delta x = \sigma_y$. Also, let $F^{-1}(P)$ and $\Phi^{-1}(P)$ denote the inverse cumulative distribution function (cdf) of y and the standard normal deviate x, and P be the cumulative probability ($0 \leqq P \leqq 1$). Plotting $F^{-1}(P) = y$ against $\Phi^{-1}(P) = x$ gives a straight line with intercept μ_y and slope σ_y. Now, if y has a mixture distribution with m normal components, its cdf is

$$F(y) = \sum_{k=1}^{m} p_k \,\Phi\left\{\frac{y - \mu_k}{\sigma_k}\right\} \qquad (6)$$

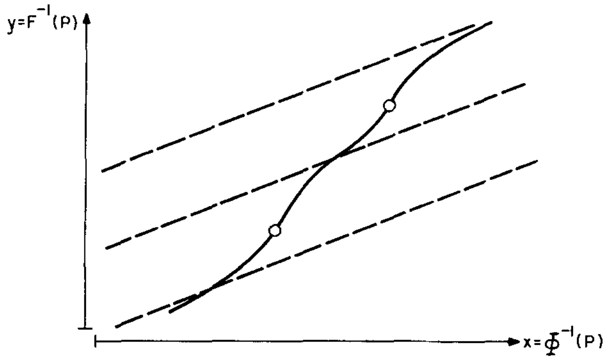


Fig. 1. The theoretical quantile-quantile plot for a three-normal-component mixture

where p_k is the mixing weight for the kth component and Φ(.) denotes the standard normal integral. It follows that

$$\frac{\partial y}{\partial x} = \frac{e^{-x^2/2}}{\sum_{k=1}^{m} p_k\, e^{-(y-\mu_k)^2/(2\sigma_k^2)}} \qquad (7)$$

which is clearly a nonlinear relationship. In this situation, the Q-Q plot has a characteristic s-shaped configuration (Fowlkes 1979) illustrated for m = 3 in Fig. 1.

This plot has two points of inflection and three asymptotes indicated by straight lines corresponding to the normal components. Harding (1949) uses the points of inflection, the slopes and the intercepts of the asymptotes to obtain estimates of the mixing weights, component means and standard deviations of the component distribution directly from the plot.

The practical Q-Q plot

In practice, the Q-Q plot is obtained by using packages 'such as Univariate in SAS or probability graph paper. In the empirical version of the theoretical $F^{-1}(P)$ versus $\Phi^{-1}(P)$ plot, the ordered data points y_i ($y_1 \leqq y_2 \leqq \ldots \leqq y_N$) referred to as mixture quantiles are plotted against the standard normal quantiles $\Phi^{-1}(q_i)$ with $q_i = i \,|\, (N + 1)$, $i = 1, \ldots, N$, and N being the sample size. Different plotting positions (q_i values) have been suggested (Kimball 1960), but no important differences were found (Looney and Gulledge 1985). If N is large, the data are partitioned into histogram intervals and the means of the intervals y_i are plotted against $\Phi^{-1}(q_i)$, with $q_i = \sum_{j=1}^{i} n_j \,|\, N$ where n_j is the number of observations in the jth interval.

Application to animal breeding data

Because of the superposition of systematic environmental and genetic factors, the elements of θ in model (1), the observed phenotypes cannot be used as mixture quantities to construct the Q-Q plot. Hence, we fit the linear model ignoring major genotypes:

$$\mathrm{y} = \Delta\boldsymbol{\theta} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{u} + \mathbf{e} \tag{8}$$

where $\boldsymbol{\beta}$ is a vector of environmental effects and $\mathbf{u}$ is a vector of additive genetic effects. Based on model (8), Best Linear Unbiased Predictions (Henderson 1973) of $\mathbf{u}$, ($\hat{\mathbf{u}}$), are obtained for continuous data. Estimates of $\mathbf{u}$ on an underlying scale can be obtained from categorical data by using the method of Gianola and Foulley (1983). The u's or $\mathrm{u_i}$'s represent additive sire effects (half of the sire's breeding value transmitted to its progeny) or breeding values of individuals. Consider, for example, a major locus with two alleles A and A. It can be quickly verified that the genetic difference between the progeny of all sires having major genotypes $\mathrm{g_{AA}}$ and $\mathrm{g_{aa}}$ is $\mathrm{t+d\,[p(a)-p(A)]}$, with d being the degree of dominance and $\mathrm{t = g_{AA} - g_{aa}}$ the displacement effect. Hence, for a not to small t, examining the u's should reveal the existence of a major locus. With unbalanced data the variance of the ith element in $\hat{\mathbf{u}}$ is $\mathrm{var(\hat{u}_i) = \sigma_u^2 - var\,(u_i\,|\,y)}$, where $\mathrm{var(u_i|\,y)}$ is the ith diagonal element of $\left[\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}\,\frac{-1\sigma_e^2}{\sigma_u^2}\right]^{-1}$ with $\mathbf{M} = \mathbf{R}^{-1} - \mathbf{R}^{-1}$ $\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}$, $\mathbf{A}$ is the matrix of additive genetic relationships among the elements in $\mathbf{u}$ (Henderson 1973) and $\mathbf{R} = \mathrm{var}(\mathbf{e})$. Then, the quantities $\mathrm{\hat{u}_i\,|\sqrt{\sigma_u^2 - var(u_i\,|\,y)}}$ are used as mixture quantiles to construct the Q-Q plot. It should be noted that these quantities are dependent, i.e., the variance-covariance matrix of $\hat{\mathbf{u}}$ has nonzero off-diagonal elements. Transformation to uncorrelated variables requires taking linear combinations of the û's, which would affect the configuration of the Q-Q plot. This problem also arises in testing normality with residuals estimated from regression analysis, and the possibility of ignoring the dependence has been supported by Pierce and Gray (1982) and Pierce (1985).

Estimation of major genotypic means and frequencies

Harding (1949) suggested using the points of inflection of the Q-Q plot configuration to obtain a crude estimation of the mixing weights of the components. Because the differential equation in (7) has no explicit solution, the points of inflection are determined by first finding an approximate fit of the Q-Q plot configuration and then setting its second derivatives with respect to x equal to zero. After some experimentation, the following function was used to fit the Q-Q plot configuration of a three-component mixture:

$$\mathrm{y} \cong \mathrm{b_1 + b_2\,x} + \frac{1}{\mathrm{b_3 - e^{-b_4(b_6-x)}}} + \frac{1}{\mathrm{b_8 + e^{-b_5(x-b_7)}}}\,. \tag{9}$$

In [9], the y's (x's) are the mixture (standard normal) quantiles and the $\mathrm{b_i}$'s are unknown parameters that were estimated by using a nonlinear least-squares algorithm. An improved fit and accommodation of unequal variances could be achieved by replacing the numerator in (9) by $\mathrm{b_9 + b_{10}\,x}$ and $\mathrm{b_{11} + b_{12}\,x}$, respectively. This would increase the number of unknown parameters from 8 to 12 and would only be suitable for large sample sizes ($\mathrm{N} \gg 100$).

Second derivatives of [9] with respect to x are:

$$\frac{\partial^2 \mathrm{y}}{(\partial \mathrm{x})^2} = \frac{\mathrm{b_4^2\,e^{-b_4(b_6-x)}}}{\mathrm{[b_3 - e^{b_4(b_6-x)}]^2}} + 2\,\frac{\mathrm{b_4^2\,[e^{-b_4(b_6-x)}]^2}}{\mathrm{[b_3 - e^{b_4(b_6-x)}]^3}}$$

$$- \frac{\mathrm{b_5^2\,e^{-b_5(x-b_7)}}}{\mathrm{[b_8 + e^{b_5(x-b_7)}]^2}} + 2\,\frac{\mathrm{b_5^2\,[e^{-b_5(x-b_7)}]^2}}{\mathrm{[b_8 + e^{b_5(x-b_7)}]^3}}\,. \tag{10}$$

Setting (10) equal to zero gives a polynomial in x with roots $\mathrm{x_1^*}$ and $\mathrm{x_2^*}$. Using Harding (1949), estimates of the mixing weights (in this case, genotype frequencies) are

$$\mathrm{\hat{p}_1} = \Phi(\mathrm{x_1^*}),\quad \mathrm{\hat{p}_2} = \Phi(x_2^*) - \mathrm{\hat{p}_1} \quad \text{and} \quad \mathrm{\hat{p}_3} = 1 - \mathrm{\hat{p}_2} - \mathrm{\hat{p}_1}\,. \tag{11}$$

Fowlkes (1979) used a similar procedure for a two-component mixture and calculated the bias in estimates from (11). Next, the ordered sample of size N is partitioned into three subsets $\mathrm{y_1}$, $\mathrm{y_2}$, and $\mathrm{y_3}$ of sizes $\mathrm{[N\,p_1]}$, $\mathrm{[N\,p_2]}$, and $\mathrm{[N\,p_3]}$, respectively, where $\mathrm{[N\,p_i]}$ stands for integer $\mathrm{(N\,p_i + 0.5)}$. The empirical standard normal quantiles are rescaled by

$$\mathrm{x_{j1}} = \Phi^{-1}(\mathrm{q_{1j}}) = \Phi^{-1}(\mathrm{q_j/\hat{p}_1}) \quad \text{and}$$

$$\mathrm{x_{jk}} = \Phi^{-1}(\mathrm{q_{kj}}) = \Phi^{-1}\left[\left(\mathrm{q_j} - \sum_{l=1}^{k-1} \mathrm{p_l}\right)\Big/\mathrm{\hat{p}_k}\right], \quad \mathrm{k} = 2, 3\,.$$

Based on the model $\mathrm{y_{ik}} = \mu_\mathrm{k} + \sigma\,\mathrm{x_{ik}} + \mathrm{e_{ik}}$, μ_k ($\mathrm{k} = 1, 2, 3$) and σ are estimated by ordinary least-squares (OLS) separately from the three partitions of the sample. Given the estimates $\hat{\mu}_\mathrm{k}$ and $\hat{\sigma}$, with ϕ denoting the standard normal density, the quantities

$$\mathrm{p(I_i = k\,|\,y_i)} = \frac{\mathrm{\hat{p}_k}\,\phi[(y_i - \hat{\mu}_\mathrm{k})/\hat{\sigma}]}{\sum_{l=1}^{3} \mathrm{\hat{p}}_l\,\phi[(\mathrm{y_i} - \hat{\mu}_l)/\hat{\sigma}]}, \qquad k = 1, 2, 3 \tag{12}$$

are computed for all observations. If $\mathrm{max\;p(I_i = k\,|\,y_i)} < \varepsilon$ (e.g., $\varepsilon = 0.8$), the ith observation was discarded. This occurs for $\mathrm{y_i}$'s located in the overlapping regions of neighboring components. With the remaining observations, the major genotypic means are estimated as

$$\mathrm{g_k} = \hat{\mu} + \frac{1}{\mathrm{[N\,\hat{p}_k]}} \sum_{\mathrm{i}=1}^{\mathrm{[N\hat{p}_k]}} \mathrm{y_{ki}}\,\sigma_\mathrm{u} \tag{13}$$

with $\mathrm{y_{ki} = \hat{u}_{ki}/\sqrt{var(\hat{u}_{ki})}}$, the $\mathrm{u_{ki}}$'s being breeding values (2 × sire effects,), μ is an overall mean, and when assuming mainly additive gene action at the major locus.

Discrimination between genetic hypotheses

In finite mixture models such as (5), hypothesis testing based on maximum likelihood inference has been employed when the number of components is uncertain (Aitkin and Rubin 1985; Titterington et al. 1985). Likelihood ratio tests are used in segregation analysis (Elston and Stewart 1971) and complex segregation analysis (Morton and McLean 1974; Bonney 1986; Elston 1987) to discriminate between different modes of inheritance. The main interest is (i) to test whether the data suggest absence or presence of major locus, (ii) to find the "most likely" number of major genotypes, and (iii) to make inference about the gene action (additive, dominant) at the major locus. These testing problems involve the following hypotheses:

(i) H_0: $g_1 = g_2 = \ldots = g_m$ versus

H_A: $g_k \neq g_l$ for all $l \neq k \in \{1, \ldots, m\}$,

(ii) H_0: $g_1 \neq g_2 \neq \ldots \neq g_m$ versus

H_A: $g_1 \neq g_2 \neq \ldots \neq g'_m = g_{m+1} \cdots = g'_{m'}$, $m' > m$,

(iii) H_0: $g_{Aa} = g_{aa} + 0.5\,t$ versus

H_A: $g_{Aa} = g_{aa} + dt$, $d \neq 0.5$, $d \geqq 0$

assuming a major locus with genotypes g_{AA}, g_{Aa},. Additive gene action implies that $g_{Aa} = g_{aa} + 0.5\,t$.

In complex segregation analysis (Morton and McLean 1974; Bonney 1986; Elston 1987), choosing between H_0 and H_A is based on comparing likelihoods maximized numerically over the parameter spaces under H_0 and H_A. Typically, the likelihoods are specified for pedigrees of limited size and a small number of unknown parameters (e.g., 10). Hoeschele (1988) considered analyzing field-collected data based on model (1) with the vector of explanatory factors $\boldsymbol{\theta}$ of large order and interest in obtaining estimates for polygenic effects (**u**).

A method of simultaneously estimating major genotypic values (**g**), frequencies (**p**), polygenic effects (**u**), environmental effects ($\boldsymbol{\beta}$), and variance components ($\boldsymbol{\sigma}'$, e.g., $\boldsymbol{\sigma}' = [\sigma_u^2, \sigma_e^2]$) has been proposed (Hoeschele 1988). Using Bayes' theorem (Box and Tiao 1973) and assuming $\boldsymbol{\sigma}$ known, the parameters are estimated by maximizing the posterior density

$$h(\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u} \mid \boldsymbol{\sigma}, \mathbf{y}) \propto g(\mathbf{y} \mid \mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma})\, f(\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}), \tag{14}$$

which is proportional to the product of likelihood g(.) and prior density f(.). It is assumed that **p**, **g**, $\boldsymbol{\beta}$, and **u** are independent a priori and that prior knowledge on **p**, **g**, and $\boldsymbol{\beta}$ is vague, meaning that each value is equally likely a priori and implying $f(\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}) \propto f(\mathbf{u})$. If **u** is a vector of additive polygenic effects, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$ by the central limit theorem.

Now consider testing problem (i). Under H_0, the vector of unknown parameters is $\gamma'_0 = [\mu, \boldsymbol{\beta}', \mathbf{u}']$, where μ is the overall mean. Under H_A, the vector of unknown parameters is $\gamma'_A = [\mathbf{p}', \mathbf{g}', \boldsymbol{\beta}', \mathbf{u}']$. The ratio of "averaged" likelihoods, i.e., likelihoods unconditional with respect to **u**, is used for discriminating between H_0 and H_A. Using (14), this is

$$\Lambda_{1a} = \frac{\sup\limits_{\mu, \boldsymbol{\beta}} \int\limits_{R_{\mathbf{u}}} g(\mathbf{y} \mid \mu, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)\, f(\mathbf{u})\, \sigma_u^2)\, d\mathbf{u}}{\sup\limits_{\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}} \int\limits_{R_{\mathbf{u}}} g(\mathbf{y} \mid \mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)\, f(\mathbf{u} \mid \sigma_u^2)\, d\mathbf{u}} \tag{15}$$

If the likelihood ratio statistic $-2 \log \Lambda_{1a}$ were asymptotically chi-square distributed under H_0, a test for H_0 versus H_A would consists of rejecting H_0 if

$$-2 \log \Lambda_{1a} = 2[\log L(H_A) - \log L(H_0)] > \chi^2_{\alpha, (\dim \gamma_A - \dim \gamma_0)},$$

where α is the size of the test (Lindgren 1976), and $L(H_0)$ and $L(H_A)$ denote likelihoods maximized under H_0 and H_A, respectively. Application of this test involves problems discussed below.

Computation of Λ_{1a} for continuous data assuming normality is illustrated in the Appendix. In segregation analysis using regressive models (Bonney 1986), the denominator of (15) is maximized numerically over the range of **g** and $\boldsymbol{\beta}$. However, in analyses of field data, the order of **y** and $\boldsymbol{\beta}$ can be large. Then, computation of (A.5) and, hence, of (15) becomes difficult, and we might consider the following large sample ($N \to \infty$) approximation:

$$\Lambda^*_{1a} = \frac{\sup\limits_{\mu, \boldsymbol{\beta}, \mathbf{u}} g(\mathbf{y} \mid \mu, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)\, f(\mathbf{u} \mid \sigma_u^2)}{\sup\limits_{\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}} g(\mathbf{y} \mid \mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)\, f(\mathbf{u} \mid \sigma_u^2)}. \tag{16}$$

The approximation (16) will be close to Λ_{1a} only if the likelihood is very peaked ($N \to \infty$) with most of its probability mass concentrated over a small region about the ML estimates.

Also, exact alternatives to (15) need to be considered for two reasons. First, in testing problems (i) and (ii), under H_0 some elements of **p** are zero. Because zero is the boundary of the parameter space of **p**, $-2 \ln \Lambda_{1a}$ will not be asymptotically χ^2-distributed under H_0. This problem can be overcome by replacing the denominator of (15) by

$$\sup_{\mathbf{g}, \boldsymbol{\beta}} \int\limits_{R_{\mathbf{u}}} \left[\int\limits_{R_{\mathbf{p}}} g(\mathbf{y} \mid \mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2)\, f(\mathbf{p})\, d\mathbf{p} \right] f(\mathbf{u} \mid \sigma_u^2)\, d\mathbf{u}.$$

However, this requires specifying the form of a proper prior distribution of **p**, which is the Dirichlet distribution (Hoeschele 1988) and its parameters. The additional integration also increases the computational difficulty. Alternatively, consider a transformation of **p**, $t^{-1}(\mathbf{p}) = \boldsymbol{\varrho}$, with $\varrho_k \in [a, b]$, $k \in \{1, \ldots, m\}$, and $t^{-1}(0) \neq a$, $t^{-1}(l) \neq b$, and replace the denominator of (15) by

$$\sup_{\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}} \int\limits_{R_{\mathbf{u}}} \{ g(\mathbf{y} \mid \mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)|_{\mathbf{p} = t^{-1}(\boldsymbol{\varrho})}\, f(\mathbf{u} \mid \sigma_u^2) \} \left| \frac{\partial \mathbf{p}}{\partial \boldsymbol{\varrho}'} \right|.$$

Akaike (1977) suggested computing the information criteria $\mathrm{AIC}(\mathrm{H}_0)$ and $\mathrm{ASIC}(\mathrm{H}_A)$ with

$$0.5[\mathrm{AIC}(\mathrm{H}_0)-\mathrm{AIC}(\mathrm{H}_A)] = \log \mathrm{L}(\mathrm{H}_A) - \log \mathrm{L}(\mathrm{H}_0) + \dim \gamma_0 - \dim \gamma_A .$$

A positive difference, i.e., $\mathrm{AIC}(\mathrm{H}_A) < \mathrm{AIC}(\mathrm{H}_0)$, would suggest choosing H_A in the light of the data. Titterington et al. (1985) and Akaike (1977) report this criterion to be useful in many practical situations.

Application to simulated data

A small simulation study was conducted to illustrate and examine the proposed techniques.

Data

Phenotypes were generated using a mixed model including herd-year-season effect (hys_i), major genotype (g_j), polygenic effect (u_k) and residual, so that

$$\mathrm{y}_{ijkl} = \mathrm{hys}_i + \mathrm{g}_j + \mathrm{u}_k + \mathrm{e}_{ijkl} . \tag{17}$$

Dispersion assumptions were:

$$\mathrm{var}(\mathrm{y}_{ijkl}) = \sigma^2 = \sigma^2_{\mathrm{hys}} + \sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{u}} + \sigma^2_{\mathrm{e}} ,$$

$$\sigma^2_{\mathrm{g}} = \sum_{j=1}^{3} \mathrm{p}_j (\mathrm{g}_j - \mu_{\mathrm{g}})^2 ,$$

$$\{\mathrm{u}_k\} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\frac{\sigma^2_{\mathrm{u}}}{4}\right) \text{ if the } \mathrm{u}_k\text{'s were sire effects (sire model),}$$

$$\{\mathrm{u}_k\} \sim \mathrm{N}(\mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{u}}) \text{ if the } \mathrm{u}_k\text{'s were breeding values (animal model),}$$

$$\{\mathrm{hys}_i\} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{h}}), \quad \text{and} \quad \{\mathrm{e}_{ijkl}\} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{e}}) .$$

Discontinuous phenotypes were obtained by using [17] and

$$\mathrm{Y}_{ijkl} = \begin{cases} 0 & \text{if } \mathrm{y}_{ijkl} < \Phi^{-1}(0.7) \\ 1 & \text{otherwise} \end{cases}$$

where 0.7 is the frequency in the category coded by 1.

The following parameter sets were used:

Parameter set	p(A)	σ^2	σ^2_{hys}	σ^2_{g}	σ^2_{u}	σ^2_{e}	h^2 %	t/2
(1)	0.3	50^2	25^2	22.4^2 $=0.2\,\sigma^2$	11.2^2	35.4^2	25	35.4
(2)	0.3	50^2	25^2	18.0^2 $=0.13\,\sigma^2$	17.3^2	35.4^2	25	27.8
(3)	0.3	46.7^2	25^2	0.0	17.3^2	35.4^2	14	0.0

where p(A) is the frequency of allele A at a major locus with alleles A and a, h^2 is heritability with $\mathrm{h}^2 = (\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{u}})/(\sigma^2_{\mathrm{hys}} + \sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{u}} + \sigma^2_{\mathrm{e}})$, and the displacement effect $\mathrm{t} = \mathrm{g}_{AA} - \mathrm{g}_{aa}$ was computed assuming additive gene action. Six data sets with unbalanced design were generated:

Data set	Phenotypes	Sample size	Model	Parameter set
I	continuous	5,000	sire model	(1)
II	continuous	5,000	sire model	(2)
III	continuous	5,000	sire model	(3)
IV	continuous	150	animal model	(1)
V	continuous	150	animal model	(2)
VI	discontinuous	5,000	sire model	(1)

Data sets generated by the sire model included 100 sires.

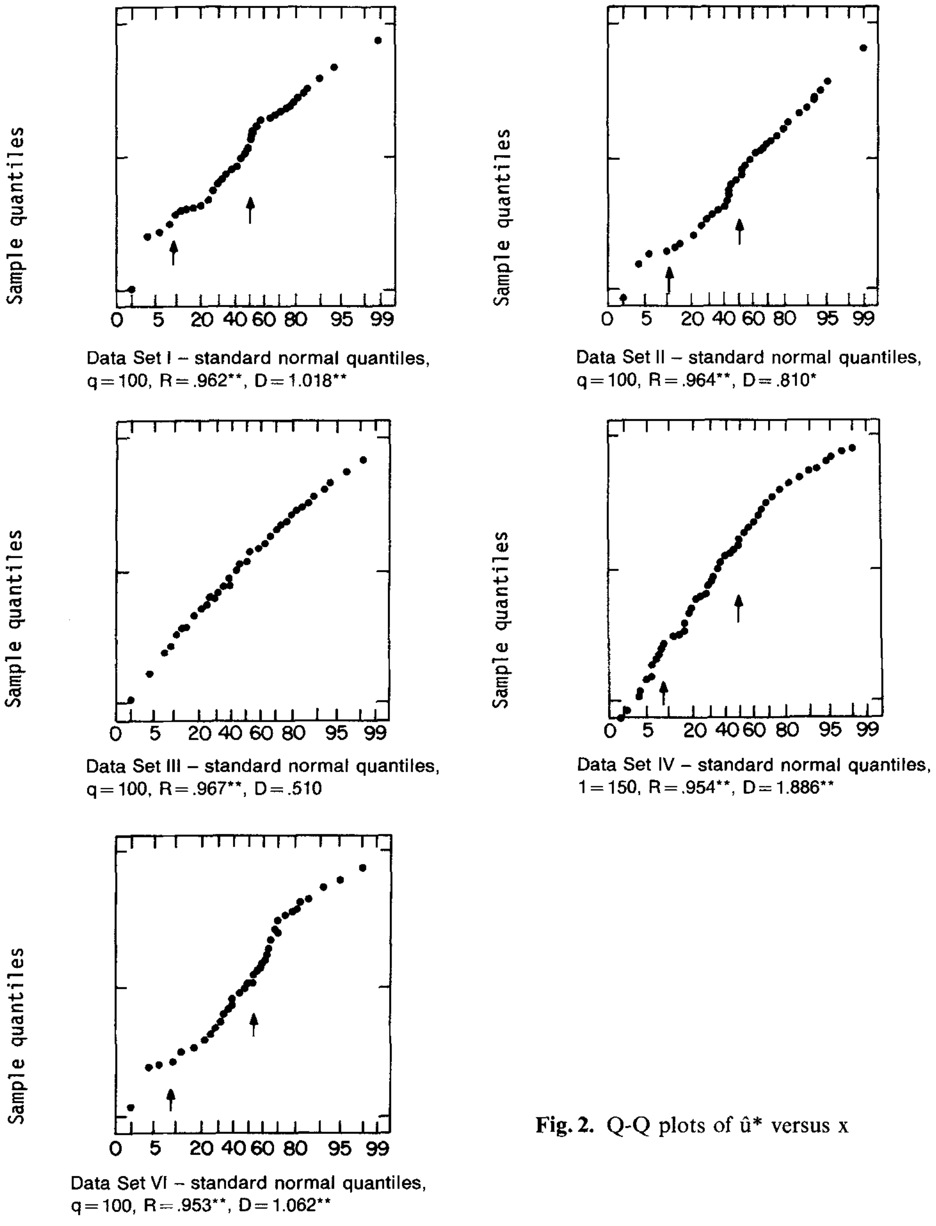
Graphical technique

For data sets I, II, III, IV and VI, the polygenic effects (u's) were estimated by fitting a mixed model (Henderson 1973; Gianola and Foulley 1983) ignoring the major genotypes (g) and standardized as described earlier. Let $\hat{\mathbf{u}}^*$ be the ordered vector of standardized estimates of u with $\hat{\mathrm{u}}^*_1 < \hat{\mathrm{u}}^*_2 < \ldots < \hat{\mathrm{u}}^*_q$ (sample quantiles) and $\mathbf{x}$ the vector of standard normal quantiles with $\mathrm{x}_i = \Phi^{-1}(\mathrm{q}_i)$, $\mathrm{q}_i = \frac{i}{q+1}$, $i = 1, \ldots, q$. The Q-Q plots are shown in Fig. 2. Test statistics for assessing departures from normality such as the Q-Q plot correlation coefficient R (Johnson and Wichern 1982) and the Anderson-Darling statistic D (Lindgren 1976) were also computed and found significant except for data set III, where D was not significant. These tests have to be interpreted with caution because of the lack of independence of the sample quantiles.

The plots for data sets I, II, IV, and VI show evidence for the presence of a mixture due to major genotypes when compared with the plot for data set III. However, they demonstrate that it is hardly possible to obtain reasonable estimates of the means ($\mu_k \equiv \mathrm{g}_k$) and frequencies ($\mathrm{p}_k$) directly from the plot. Also, it becomes apparent that the plot helps determine the number of components m only when m is small, in particular, for m = 3.

Estimates from the Q-Q plot sample quantiles

When inspecting the graphs in Fig. 2, one may conclude that three components (major genotypes) exist with the approximate location of the two points of inflection indicated by arrows. Let x_1 and x_2 be the approximate x-values at the points of inflection. Next, the vector of ordered sample quantiles $\hat{\mathbf{u}}^*$ was partitioned in three sub-



Data Set I – standard normal quantiles, q=100, R=.962**, D=1.018**

Data Set II – standard normal quantiles, q=100, R=.964**, D=.810*

Data Set III – standard normal quantiles, q=100, R=.967**, D=.510

Data Set IV – standard normal quantiles, 1=150, R=.954**, D=1.886**

Data Set VI – standard normal quantiles, q=100, R=.953**, D=1.062**

Fig. 2. Q-Q plots of û* versus x

vectors according to the frequencies estimated from x_1 and x_2 by using (11). Let the mean of the elements of the subvectors be $\tilde{\mu}_1$, $\tilde{\mu}_2$, and $\tilde{\mu}_3$. Then, by using the asymptotes ($x \to -\infty$, $x \to \infty$), starting values for the parameters in (9) were computed as

$$\begin{array}{ll}
b_1^{[0]} = \tilde{\mu}_2, & b_6^{[0]} = \Phi^{-1}(p_1^*) = x_1^*, \\
b_2^{[0]} = \tilde{\sigma}, & b_7^{[0]} = \Phi^{-1}(p_1^* + p_2^*) = x_2^*, \\
b_3^{[0]} = 1/(\tilde{\mu}_1 - \tilde{\mu}_2), & b_4^{[0]}\text{: from (9) and the other starting values or from a theoretical plot,} \\
b_8^{[0]} = 1/(\tilde{\mu}_3 - \tilde{\mu}_2), & b_5\text{: as } b_4.
\end{array}$$

Starting values for b_4 and b_5 cannot be obtained as easily as for the other parameters because one can show, using (7), that b_4 and b_5 are rather complicated expressions, depending on differences in component means, variances and mixing weights. Deriving initial guesses on b_4 and b_5 from a theoretical Q-Q plot would involve using μ_1, μ_2, μ_3, σ and an initial pair (x, y) in (7) to generate a sequence of (x, y) pairs by solving this ordinary differential equation. Next, y's would be computed by using the approximation (9) and compared with the exact solutions to (7) for different b_4 and b_5 values. Given the starting values, the parameters were estimated by nonlinear least-squares fitting and used in (10) to determine x_1 and x_2 more

Table 1. Estimates of major genotypic means and frequencies from the Q-Q plot sample quantiles

Parameters	Data set I		Data set II		Data set VI	
	True values	Estimates	True values	Estimates	True values	Estimates
p_1	0.09	0.058	0.09	0.021	0.09	0.045
p_2	0.42	0.406	0.42	0.411	0.42	0.395
p_3	0.49	0.549	0.49	0.568	0.49	0.560
g_1	414.0	411.9	422.0	411.0	−0.868	−0.687
g_2	430.0	447.0	450.0	447.0	0.0	0.310
g_3	486.0	478.8	478.0	466.6	0.868	0.836

Table 2. Approximate criterion for discriminating between polygenic and mixed inheritance

Criterion	Data set				
	I	II	III	IV	V
AIC(Λ^*_{1a})	59.3	10.1	1.1	76.3	15.1

precisely. Next, estimates of p_k and g_k ($k = 1, 2, 3$) were obtained from (11) and (13). Of course, this technique can only provide very crude estimates of p_k and g_k being considerably biased, in particular for very different p_k values, with the smallest frequency being underestimated (Fowlkes 1979). For some data sets, estimates are listed in Table 1.

Discrimination between genetic hypotheses

The ability to discriminate between polygenic and mixed inheritance in testing problem (i) was examined by computing $0.5[\mathrm{AIC}(H_0) - \mathrm{AIC}(H_A)]$ using the approximation (16) which is denoted by AIC(Λ^*_{1a}).

When the major locus accounted for 20% of σ^2_p, the criterion clearly suggested a better explanation of the data by mixed inheritance. Evidence for the presence of major genotypes was considerably reduced when the major locus accounted for only 13% of σ^2_p; for data set III, there was no evidence for a major locus.

Conclusion

Elston (1988) reviewed several approaches for discrimination between alternative modes of inheritance with pedigree data, mainly includind complex segregation analysis (e.g., Morton and McLean 1974) and regressive models (Bonney 1986). In animal breeding, these methods seem very useful for analyzing experimental data but not (large) sets of field-collected data (Hoeschele 1988). For field data, Famula (1986) suggested using a major gene index (Karlin et al. 1979). However, the major gene index was able to detect a major locus only if it contributed more than 20% of the phenotypic variation and could not provide information about the number of major genotypes, their effects and frequencies. When extended to several generations, the major gene index may become more sensitive and might be used to check if the nature of the factor causing a mixture is truly genetic and not environmental. Crude estimates of the major genotypic values and freuqencies obtained by the technique proposed in this paper may serve as starting values for the iterative method of estimating major genotypic values, frequencies, polygenic effects and variance components as proposed by Hoeschele (1988).

The methods assume normal distribution of the polygenic effects conditional on major genotype in the observed continuous scale or in the unobserved underlying scale (of discontinuous data). Departures from normality would affect both the Q-Q plot configuration (nonlinearities in the components) and the discrimination between genetic hypotheses. Therefore, for continuous data, Box-Cox transformations or estimation of the power transformation (Gianola et al. 1987) would need to be considered. It is also crucial to correctly specify the mixed model in (1), e.g., to account for nonadditive genetic effects if necessary.

One might also consider techniques for assessing departure from multivariate normality such as chi-square probability and scatter plots (Andrews et al. 1973) and chi-square tests (Moore and Stubblebine 1976). However, interpretation of a systematic curvature in a graph indicating a mixture will be less clear in the multivariate than in the univariate case, where crude parameter estimates can be obtained from the Q-Q plot.

In conclusion, the simulation results indicate that the proposed techniques may be potentially useful for suggesting existence and providing information about major genotypes in field data.

Acknowledgements. This research was conducted while I. Hoeschele was a visiting scientist at Iowa State University on leave from the Institute of Animal Breeding and Animal Husbandry, Hohenheim University, Stuttgart, West-Germany. Financial support from the Deutsche Forschungsgemeinschaft and from the Department of Animal Science, Iowa State University, is gratefully acknowledged. Dr. W. H. Kliemann is thanked for helpful suggestions.

Appendix

Under normality, the density functions in (15) are:

$$g(\mathbf{y} \mid \mu, \boldsymbol{\beta}, \mathbf{u}; \sigma^2_e)$$

$$= \frac{1}{(2\pi)^{N/2} \, |\mathbf{R}|^{1/2}} \exp\{-\tfrac{1}{2}\sigma_e^{-2} \, [(\mathbf{y} - \mathbf{1}\mu - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\hat{\mathbf{u}})'$$

$$\cdot (\mathbf{y} - \mathbf{1}\mu - \mathbf{X}\beta - \mathbf{Z}\hat{\mathbf{u}}) + (\mathbf{u} - \hat{\mathbf{u}})\mathbf{Z}'\mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}})]\}, \quad (A.1)$$

$$g(\mathbf{y}\,|\,\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}; \sigma_e^2)$$

$$= \frac{1}{(2\pi)^{N/2}|\mathbf{R}|^{1/2}} \sum_{\mathbf{K}} p(\mathbf{I}=\mathbf{K}) \exp\{-\tfrac{1}{2}\sigma_e^{-2}[(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}})'$$

$$\cdot(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}})+(\mathbf{u}-\hat{\mathbf{u}})\mathbf{Z}'\mathbf{Z}(\mathbf{u}-\hat{\mathbf{u}})]\}, \quad \text{(A.2)}$$

and

$$f(\mathbf{u}\,|\,\sigma_u^2) = \frac{1}{(2\pi)^{q/2}|\mathbf{G}|^{1/2}} \exp\{-\tfrac{1}{2}\sigma_u^{-2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\}. \quad \text{(A.3)}$$

Above, $N = \dim(\mathbf{y})$, $q = \dim(\mathbf{u})$, $\mathbf{R} = \mathbf{I}\sigma_e^2$, $\mathbf{G} = \mathbf{A}\sigma_u^2$ and

$$\sum_{\mathbf{K}} p(\mathbf{I}=\mathbf{K}) = \sum_{k_1=1}^{m} \cdots \sum_{k_N=1}^{m} p(I_1 = k_1, \ldots, I_N = k_N)$$

is a nested sum.

Using results of the multivariate normal theory (Zellner 1971), we obtain the numerator of (15):

$$\int_{R_{\mathbf{u}}} g(\mathbf{y}\,|\,\mu, \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2)\, f(\mathbf{u}\,|\,\sigma_u^2)\, d\mathbf{u}$$

$$= \frac{1}{(2\pi)^{N/2}|\mathbf{G}|^{1/2}(\sigma_e^2)^{N/2}|\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda|^{-1/2}}$$

$$= \exp\left\{-\frac{1}{2\sigma_e^2}[(\mathbf{y}-\mathbf{1}\mu-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y}-\mathbf{1}\mu-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}}) + \tilde{\mathbf{u}}^1(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)\tilde{\mathbf{u}}+\hat{\mathbf{u}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{u}}']\right\} \quad \text{(A.4)}$$

where

$$\tilde{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{1}\mu-\mathbf{X}\boldsymbol{\beta}),$$

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{1}\mu-\mathbf{X}\boldsymbol{\beta}), \quad \text{and} \quad \lambda = \sigma_e^2/\sigma_u^2.$$

It can be shown that the supremum of (A.4) is achieved at

$$\hat{\boldsymbol{\beta}}^* = [\mathbf{X}^{*\prime}\mathbf{X}^* - X^{*\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{X}^*]^{-1}$$

$$\cdot[\mathbf{X}^*\mathbf{y}' - \mathbf{X}^{*\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}']\,\mathbf{y}$$

$$= [\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{X}^*]^{-1}\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{y}$$

with

$$\boldsymbol{\beta}^{*\prime} = [\mu, \boldsymbol{\beta}'], \quad \mathbf{X}^* = [\mathbf{1}, \mathbf{X}] \quad \text{and} \quad \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{I}\sigma_e^2.$$

Similarly, the denominator of (15) is

$$\int_{R_{\mathbf{u}}} g(\mathbf{y}\,|\,\mathbf{p}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2)\, f(\mathbf{u}\,|\,\sigma_u^2)\, d\mathbf{u}$$

$$= \frac{1}{(2\pi)^{N/2}|\mathbf{G}|^{1/2}(\sigma_e^2)^{N/2}|\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda|^{-1/2}} \sum_{\mathbf{K}} p(\mathbf{I}=\mathbf{K})$$

$$\times \exp\{-\tfrac{1}{2}\sigma_e^{-2}[(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\boldsymbol{\lambda}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{u}})$$

$$+ \tilde{\mathbf{u}}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)\tilde{\mathbf{u}}+\hat{\mathbf{u}}\mathbf{Z}'\mathbf{Z}\hat{\mathbf{u}}]\}, \quad \text{(A.5)}$$

where

$$\tilde{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\mathbf{X}\beta) \quad \text{and}$$

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{W}_k\mathbf{g}-\mathbf{X}\boldsymbol{\beta}).$$

References

Aitkin M, Rubin DB (1985) Estimation and hypothesis testing in finite mixture models. JR Stat Soc B 47:67–75

Akaike H (1977) On entropy maximation principle. In: Krishnaiah PR (ed) Application of statistics. North-Holland, Amsterdam, pp 27–41

Andrews DF, Gnanadesikan R, Warner JL (1973) Methods for assessing multivariate normality. In: Krishnaiah PR (ed) Multivariate analysis, vol III. Proc. 3rd Int Symp Multivariate Analysis. Wright State University, Dayton, Ohio, June 1972. Academic Press, New York London

Bonney GE (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42–611–625

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading, Mass., 588 pp

Elston TC (1988) Models for discrimination between alternative modes of inheritance. In: Gianola D, Hammond K (ed) Advances in statistical methods for genetic improvement of lievestock. Springer, New York Berlin (in press) ■

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523

Falconer DS (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann Hum Genet 29:51–76

Famula TR (1986) Identifying single genes of large effect in quantitative traits using best linear unbiased prediction. J Anim Sci 63:68–76

Fowlkes EB (1979) Some methods for studying the mixture of two normal (lognormal) distributions. J Am Stat Assoc 74:561–575

Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. Genet Sel Evol 15:201–223

Gianola D, Im S, Fernando RL, Foulley JL (1988) Mixed model methodology and the Box-Cox theory of transformations: A Bayesian approach. In: Gianola D, Hammond K (eds) Advances in statistical methods for genetic improvement of livestock. Springer, New York Berlin (in press) ■

Hanset R (1982) Major genes in animal production, examples and perspectives: cattle and pigs. In: 2nd World Congr Genet Appl Livst Prod, vol VI. Publicaciones Agrarias, Madrid, Spain, p 439–453

Harding JP (1949) The use of probability paper for the graphical analysis of polymodal frequency distributions. J Mar Biol Assoc 28:141–153

Henderson CR (1973) Sire evaluations and genetic trends. Proc Anim Breed Genet Symp in Honor of Dr JL Lush, Blacksburg, Virginia, July 29, 1982. American Society of Animal Science and American Dairy Science Association, Champaign, Ill, pp 10–41

Henderson CR (1985) Best linear unbiased prediction of nonadditive genetic merit in noninbred populations. J Anim Sci 60:111–117

Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. Theor Appl Genet 76:81–92

Johnson RA, Wichern DW (1982) Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey, 564 pp

Karlin S, Carmelli D, Williams R (1979) Index measures for assessing the model of inheritance of continuously distributed traits. I. Theory and justification. Theor Popul Biol 16:81–106

Kimball BF (1960) On the choice of plotting positions on probability paper. J Am Stat Assoc 55:546–560

Lindgren BW (1976) Statistical theory. Macmillan, New York: Collier Macmillan Publishers, London, 614 pp

Looney SW, Gulledge TR (1985) Use of the correlation coefficient with normal probability plots. Am Stat 39:75–79

Moore DS, Stubblebine JB (1976) Chi-square tests for multivariate normality with application to common stock prices. Department of Statistics, Purdue University, Mimeo Ser No 78-17

Morton NE, McLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–502

Pierce DA, Gray RJ (1982) Testing normality of errors in regression models. Biometrika 69:233–236

Pierce DA (1985) Testing normality in autoregressive models. Biometrika 72:293–297

Roberts RC, Smith C (1982) Genes with large effects – theoretical aspects in livestock breeding. In: 2nd World Congr Genet Appl Livst Prod, vol VI. Publicaciones Agrarias, Madrid, Spain, p 420–437

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley and Sons, Chichester New York Brisbane Toronto Singapore, 243 p

Zellner A (1971) An introduction to Bayesian inference in econometrics. Wiley and Sons, New York London Sydney Toronto, 431 p