Identification of a major gene in F1 and F2 data when alleles are assumed fixed in the parental lines

Llg Janss, Jhj van Der Werf

© Elsevier/INRA

Original article

Identification of a major gene in F_1 and F_2 data when alleles are assumed fixed in the parental lines

LLG Janss, JHJ Van Der Werf

Wageningen Agricultural University, Department of Animal Breeding
PO Box 338 6700 AH Wageningen, The Netherlands

(Received 9 August 1991; accepted 27 August 1992)

Summary – A maximum likelihood method is described to identify a major gene using F_2, and optionally F_1, data of an experimental cross. A model which assumed fixation at the major locus in parental lines was investigated by simulation. For large data sets (1 000 observations) the likelihood ratio test was conservative and yielded a type I error of 3%, at a nominal level of 5%. The power of the test reached > 95% for additive and completely dominant effects of 4 and 2 residual SDs respectively. For smaller data sets, power decreased. In this model assuming fixation, polygenic effects may be ignored, but on various other points the model is poorly robust. When F_1 data were included any increase in variance from F_1 to F_2 biased parameter estimates and led to putative detection of a major gene. When alleles segregated in parental lines, parameter estimates were also biased, unless the average allele frequency was exactly 0.5. The model uses only the non-normality of the distribution due to the major gene and corrections for non-normality due to other sources cannot be made. Use of data and models in which alleles segregate in parents, *eg* F_3 data, will give better robustness and power.

cross / major gene / maximum likelihood / hypothesis testing

Résumé – **Identification d'un gène majeur en F_1 et F_2 quand les allèles sont supposés fixés dans les lignées parentales.** *Cet article décrit une méthode de maximum de vraisemblance pour identifier un gène majeur à partir de données F_2, et éventuellement F_1, d'un croisement expérimental. Un modèle supposant un locus majeur avec des allèles fixés dans les lignées parentales est étudié à l'aide de simulations. Pour des fichiers de grande taille (1 000 observations), le test du rapport de vraisemblance est conservateur, avec une erreur de première espèce de 3%, à un niveau nominal de 5%. La puissance du test d'identification d'un gène majeur atteint plus de 95% pour des effets additifs et de dominance de 4 et 2 écarts-types respectivement. Pour des fichiers de taille plus petite, la puissance baisse rapidement. Dans le modèle utilisé la variance polygénique peut être négligée mais sur d'autres points le modèle est peu robuste. Si des données F_1 sont incluses, toute augmentation de la variance entre F_1 et F_2 introduit un biais sur les paramètres estimés et peut mener à la détection d'un faux gène majeur. Quand les allèles ségrègent* *dans les lignées parentales, les paramètres estimés sont également biaisés si la fréquence allélique moyenne n'est pas exactement de 0,5. Finalement, le modèle n'utilise que la non normalité de la distribution due au gène majeur, et ne peut pas corriger pour une non normalité due à d'autres raisons. L'utilisation d'un modèle ou les allèles ségrègent chez les parents, par exemple sur des données F_3, doit améliorer la robustesse et la puissance du test.*

croisement / gène majeur / maximum de vraisemblance / test d'hypothèse

INTRODUCTION

In animal breeding, crosses are used to combine favourable characteristics into one synthetic line. It is useful to detect a major gene as soon as possible in such a line, because selection could be carried out more efficiently, or repeated backcrosses be made. Once a major gene has been identified it can also be used for introgression in other lines.

Major genes can be identified using maximum likelihood methods, such as segregation analysis (Elston and Stewart, 1971; Morton and MacLean, 1974). Segregation analysis is a universal method and can be applied in populations where alleles segregate in parents. However, when applied to F_1, F_2 or backcross data assuming fixation of alleles in parental lines, genotypes of parents are assumed known and all equal and this analysis leads to the fitting of a mixture distribution without accounting for family structure.

Fitting of mixture distributions has been proposed when pure line and backcross data as well as F_1 and F_2 data are available, and when parental lines are homozygous for all loci (Elston and Stewart, 1973; Elston, 1984). Statistical properties of this method, however, were not described, and several assumptions may not hold. For example, not much is known concerning the power of this method when only F_2 data are available, which is often the case when developing a synthetic line. Furthermore, homozygosity at all loci in parental lines is not tenable in practical animal breeding. Here it is assumed that many alleles of small effect, so-called polygenes, are segregating in the parental lines. Alleles at the major locus are assumed fixed. F_1 data could possibly be included, but this is not necessarily more informative because F_1 and F_2 generations may have different means and variances due to segregating polygenes.

The aim of this paper is to investigate by simulation some of the statistical properties of fitting mixture distributions, such as Type I error, power of the likelihood ratio test and bias of parameter estimates when using only F_2 data. To study the properties of the major gene model, polygenic variance is not estimated. The robustness of this model will be checked when polygenic variance is present in the data, and when the major gene is not fixed in the parental lines. The question of whether F_1 data can and should be included will be addressed.

MODELS USED FOR SIMULATION

A base-population of F_1 individuals was simulated, although the F_1 generation may not have had observed records. Consider a single locus A with alleles A_1 and A_2, where A_1 has frequencies f_p and f_m in the paternal and maternal line. Genotype frequencies, values and numeration are given for F_1 individuals as:

Genotype	A_1A_1	A_1A_2	A_2A_2
Number	1	2	3
Frequency	f_pf_m	$f_p(1-f_m)+f_m(1-f_p)$	$(1-f_p)(1-f_m)$
Value	μ_1	μ_2	μ_3

Genotypes of F_1 animals were allocated according to the frequencies given above using uniform random numbers. For the F_2 generation, genotype probabilities were calculated given the parents' genotypes using Mendelian transmission probabilities and assuming random mating and no selection. A random environmental component e_i was simulated and added to the genotype. The observation en individual $i(F_1$ or $F_2)$ with genotype $r(y_i^r)$ is:

$$y_i^r = \mu_r + e_i, \qquad [1]$$

with e_i distributed $N(0,\sigma^2)$. Polygenic effects are assumed to be normally distributed. For base individuals polygenic values were sampled from $N(0,\sigma_g^2)$, where σ_g^2 is the polygenic variance. No records were simulated for F_1 individuals when polygenic effects were included. For F_2 offspring, phenotypic observations y_{ij}^r were simulated as:

$$y_{ij}^r = \mu_r + 1/2a_p + 1/2a_m + \phi_i + e_{ij}, \qquad [2]$$

where ϕ_i is the Mendelian sampling term, sampled from $N(0,\sigma_g^2/2)$, a_p and a_m are paternal and maternal polygenic values and e_{ij} is distributed $N(0,\sigma^2)$. Additionally, data were simulated with no major gene or polygenic effect:

$$y_i = e_i, \qquad [3]$$

where e_i is distributed $N(0,\sigma^2)$. A balanced family structure was simulated, with an equal number of dams, nested within sire, and an equal number of offspring for each dam. Random variables were generated by the IMSL routines GGUBFS for uniform variables and GGNQF for normal variables (Imsl, 1984).

MODELS USED FOR ANALYSIS

The test for the presence of a major gene is based on comparing the likelihood of a model with and without a major gene. Polygenic effects are not included in the model, and the model without a major gene therefore contains random environment only. Apart from major gene or no major gene, models can account for only F_2 data, or for both F_1 and F_2 data. This results in a total of 4 models to be described.

Model for F_2 data with environment only

For F_2 data, with n observations, the model can be written:

$$y_i = \beta + e_i;\ \mathrm{E}(y_i) = \beta;\ \mathrm{var}(y_i) = \mathrm{var}(e_i) = \sigma^2. \qquad [4]$$

The logarithm of the joint likelihood for all observations, assuming normality and uncorrelated errors, is:

$$L_1 = -\frac{n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{(y_i - \beta)^2}{-2\sigma^2} \qquad [5]$$

Maximising [5] with respect to β and σ^2 yields as the maximum likelihood (ML) estimate for the mean, $\widehat{\beta} = \Sigma_i y_i/n$, and the ML estimate for the variance is $\widehat{\sigma}^2 = \Sigma_i(y_i - \widehat{\beta})^2/n$.

Model for F_1 and F_2 data with environment only

Data on F_1 and F_2 are combined, with $n_1 + n_2 = N$ observations. The observation on animal j from generation $i (i = 1, 2)$ is:

$$y_{ij} = \beta_i + e_{ij};\ \mathrm{E}(y_{ij}) = \beta_i;\ \mathrm{var}(y_{ij}) = \mathrm{var}(e_{ij}) = \sigma^2 \qquad [6]$$

where β_i is the mean for generation i. Observations for F_1 and F_2 are assumed to have equal environmental variance. The joint log-likelihood is given as:

$$L_1^* = -\frac{N}{2}\ \ln(2\pi\sigma^2) + \sum_{i=1}^{2}\sum_{j=1}^{n_i} \frac{(y_{ij} - \beta_i)^2}{-2\sigma^2} \qquad [7]$$

The ML estimates for β_i are simply the observed means for each generation, *ie* $\widehat{\beta}_1 = \Sigma_j y_{1j}/n_1$, and $\widehat{\beta}_2 = \Sigma_j y_{2j}/n_2$. The ML estimate for the variance is $\widehat{\sigma}^2 = \Sigma_i\Sigma_j(y_{ij} - \widehat{\beta}_i)^2/N$.

Model with major gene and environment for F_2 data

When alleles are assumed fixed in parental lines, all F_1 individuals are known to be heterozygous. If no polygenic effects are considered, this means that all F_2 individuals have the same expectation, and conditioning on parents is redundant. In the likelihood for such data, summuations over the parents' possible genotypes can be omitted and families can be pooled. The model is given as:

$$y_i^r = \mu_r + e_i;\ e_i \sim N(0, \sigma^2) \qquad [8]$$

and the log-likelihood equals:

$$L_2 = \sum_{i=1}^{n} \ln\left(\sum_{r=1}^{3} P_r f(y_i | G_i = r)\right) \qquad [9]$$

In [9] G_i is the genotype of individual i, P_r denotes the prior probability that $G_i = r$, which equals 1/4, 1/2 and 1/4 for $r = 1$, 2 and 3 (or A_1A_1, A_1A_2 and A_2A_2). The total number of F_2 individuals is given as n, and the function f is given as:

$$f(y_i|G_i = r) = (2\pi\sigma^2)^{-1/2}\exp\left\{-1/2\sigma^{-2}(y_i - \mu_r)^2\right\} \qquad [10]$$

Model with major gene and environment for F_1 and F_2 data

In the F_1 generation only one genotype occurs; hence F_1 data are distributed around a single mean, with a variance equal to the residual variance in the F_2 generation. Due to possible heterosis shown by the polygenes a separate mean is modelled, but the possible heterogeneity in variance caused by polygenes is not accounted for. The model for individual j from generation i for genotype r is:

$$y_{ij}^r = \mu_r + \beta_i + e_{ij};\ e_{ij} \sim N(0, \sigma^2). \qquad [11]$$

where β_i is a fixed effect for generation i. Model [11] is overparameterised because genotype means and 2 general means are modelled. We chose to put $\beta_2 = 0$. In that case the mean of F_1 individuals, which all have known genotype $r = 2$, can be written as $\mu_{F_1} = \mu_2 + \beta_1$. The joint log-likelihood for F_1 and F_2 data, using μ_{F_1} is:

$$L_2^* = \sum_{j=1}^{n_1}\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_{1j} - \mu_{F_1})^2}{2\sigma^2}\right\} + \sum_{j=1}^{n_2}\ln\left\{\sum_{r=1}^{3} P_r f(y_{2j}|G_j = r)\right\} \qquad [12]$$

where n_1 and n_2 are number of observations in the F_1 and F_2 generation. The ML estimate for μ_{F_1} is equal to $\widehat{\beta}_1$ in [6].

ML estimates for $\mu_r (r = 1, 2, 3)$ and σ^2 in models [8] and [11] cannot be given explicitly. These parameters were estimated by minimising minus log-likelihood L_2 in [9] and L_2^* in [12], using a quasi-Newton minimisation routine. A reparameterisation was made using the difference between homozygotes $t = \mu_3 - \mu_1$, and a relative dominance coefficient $d = (\mu_2 - \mu_1)/t$, as in Morton and MacLean (1974). By experience, this parameterisation was found more appropriate than the parameterisation using 3 means μ_1, μ_2 and μ_3, because convergence is generally reached faster due to smaller sampling covariances between the estimates. The mean was chosen as the midhomozygote value: $\mu = 1/2\mu_1 + 1/2\mu_3$.

Parameters t and d are easier to interpret than 3 means, and therefore results are also presented using these parameters. Parameter t indicates the magnitude of the major gene effect and can be expressed either absolutely or in units of the residual standard deviation. Parameter t was constrained to be positive, which is arbitrary because the likelihood for the parameters μ, t and d is equal to the likelihood for the parameters μ, $-t$ and (1–d). Parameter d was estimated in the interval [0,1]. Problems were detected when this constraint was not used, because t could become zero, leading to infinitely large estimates for d. This occurred frequently when the effects where small and dominant. Minimisation by IMSL routine ZXMIN (Imsl, 1984) specified 3 significant digits in the estimated parameters as the convergence criterion.

HYPOTHESIS TESTING

The null hypothesis (H_0) is "no major gene effect", whereas the alternative hypothesis (H_1) is "a major gene effect is present". The log-likelihoods L_1 in [5] and L_2 in [9] are the likelihoods for each hypothesis when only F_2 data are present. When F_1 data are included the likelihoods L_1^* in [7] and L_2^* in [12] apply. A likelihood ratio test is used to accept or reject H_0. Twice the logarithm of the likelihood ratio is given as:

$$\tau = 2(L_2 - L_1), \quad \text{for } F_2 \text{ data only}$$

or

$$\tau = 2(L_2^* - L_1^*), \quad \text{for } F_1 \text{ and } F_2 \text{ data.}$$

Two important aspects of any test are the type I and type II errors. The type I error is the percentage of cases in which H_0 is rejected, although it is true. The H_0 model is simulated by [3]. The type II error is the percentage of cases in which H_1 is rejected, although it was true. Here, the type II error is not used, but its complement, the power, which is the percentage of cases in which H_1 is accepted, when H_1 is true. The H_1 model is simulated by model [1]. Fixation of alleles in parental lines is simulated by taking $f_p = 1$ and $f_m = 0$.

Type I error

The distribution of τ when H_0 is true is expected asymptotically to be χ^2 with 2 degrees of freedom, because the H_1 model has 2 parameters more than the H_0 model (Wilks, 1938). Since in practice data sets are always of finite size, it is interesting to know whether and when the distribution of τ is close enough to the expected asymptotic distribution, so that quantiles from a χ^2 distribution can be used as critical values. Type I errors were estimated for data sets of 100 up to 2 000 observations, simulating 1 000 replicates for each size of data set. Three critical values were used, corresponding to nominal levels of 10, 5 and 1%. The nominal level is defined as the expected error rate, based on the asymptotic distribution. Exact binomial probabilities were used to test whether the estimates differed significantly from the nominal level. When the observed number of significant replicates does not differ significantly, a χ^2 distribution is considered suitable to provide critical values. Also, when the observed number is lower than expected the asymptotic distribution might remain useful. The nominal tye I error is in that case an upper bound for the real type I error.

Power of the test and estimated parameters

The power is investigated for additive ($d = 0.5$) and completely dominant ($d = 1$) effects, with a residual variance of 100, and t varying from 10–40, *ie* from 1 to 4 SDs. The additive genetic variance caused by this locus equals $t^2/8$, when t is absolute. Heritability in the narrow sense therefore varies from 0.11–0.67. Each data set contained 1 000 observations, and each situation was repeated 100 times. The power of the test for smaller data sets was investigated for one relatively small effect and one relatively large effect.

Robustness

Investigation of the type I error and the power considered situations where either H_0 or H_1 was true, satisfying all assumptions in the models. The robustness of this test and usefulness of the assumption of fixation in parents for parameter estimation was investigated for situations which violate 2 assumptions:

– when there is a covariance between error terms. This was induced by simulation of polygenic variance by model [2]. The total variance was held constant at 100, so that the power of the test could not change due to a change in total variance;

– when fixation of alleles is not the case. The data were simulated by model [1], in which f_p and f_m were not equal to 0 and 1, resulting in segregation of alleles in the F_1 parents. Firstly, 3 situations were simulated where the average allele frequency remains 0.5. In that case only the assumption that all F_1 parents are heterozygous was violated. Secondly, 3 situations were simulated where the average allele frequency was not 0.5. In that case, the assumption that genotype frequencies in F_2 are 1/4, 1/2, and 1/4 was also violated.

Inclusion of F_1 data

A major gene which starts segregating in the F_2 not only renders the distribution non-normal, but also increases the phenotypic variance in the F_2 relative to the F_1. When F_1 data are included, this increase in variance may be taken as supplementary evidence, apart from any non-normality, for the existence of a major gene. Assessing the relative importance of the 2 sources of information is useful so as to judge the robustness of the model including F_1 data. The effects on non-normality and increased F_2 variance due to the major gene should therefore be distinguished. This was accomplished by simulating different residual variances in F_1 and F_2. Four situations were investigated, combining all combinations of non-normality in F_2 and increased variance in F_2 (table I). In general, 500 F_1 and 1 000 F_2 observations were simulated. For situation 3, data sets with 1 000 F_1 and 1 000 F_2 observations were also investigated. Data for situations 1 and 3 were simulated by model [3], whereas data for situations 2 and 4 were simulated by model [1].

Table I. The effect on variance and non-normality in the F_2, when F_1 and F_2 data are combined, for various situations investigated.

Situation	*Description*	*F_2 distribution normal*	*Larger variance in F_2*
1	H_0 (no major gene)	Yes	No
2	H_1 (major gene)	No	Yes
3	H_0 with increased F_2 variance	Yes	Yes
4	H_1 with decreased F_2 variance	No	No

RESULTS

Type I error and parameter estimates under the null hypothesis

Estimated type I errors, based on 1 000 replicates, have been given in table II for different sizes of the data set. Estimates decreased, and more or less stabilised when the size of the data set exceeded 1 000 observations, especially for a nominal level of 10%, which were most accurate. For these large data sets, however, the type I errors were too low ($P < 0.01$), which means that critical values obtained from a χ^2_2 distribution would provide a too conservative test. For example, application of the χ^2_2 95-percentile to data sets with 1 000 observations will not result in the expected type I error of 5%, but rather in a type I error of ≈ 3%.

Table II. Estimated Type I errors (%) at 3 nominal levels for different size of the data set.

	Nominal level					
	10%		*5%*		*1%*	
N	*Estimate*	P	*Estimate*	P	*Estimate*	P
100	9.5	0.3216	5.0	0.5375	0.8	0.3317
250	7.8	0.0099	3.3	0.0059	0.9	0.4573
500	6.9	0.0004	2.9	0.0007	0.4	0.0287
1 000	6.1	0.0000	3.1	0.0022	0.5	0.0661
2 000	6.0	0.0000	2.5	0.0001	0.6	0.1289

N: Number of observations in the data set; *P*: critical level for test whether estimate is equal to the nominal level, based on exact binomial probabilities.

When no major gene effect was present, stil on average a considerable effect could be found. Parameter estimates for the major gene model have been given in table III, simulating just a normally distributed error effect with variance 100. The empirical standard deviation for estimated t-values ranged between $7(N = 100)$ and $5(N = 2000)$ (not in table). The average estimate for t is therefore biased, and many of the individual estimates were significantly different from zero if a t-test was applied. The average estimated d is 0.5, which is expected because the simulated distribution was symmetrical.

Parameter estimates and power of the test

Results for the different situations studied under a major gene model are in table IV. The χ^2_2 95-percentile was used as critical value for the test. The power reached over 95% for additive effects ($d = 0.5$) with a t-value of 40, which is 4 σ (residual standard deviations). For completely dominant effects ($d = 1$), 100% power was reached for an effect of $t = 20$ (2σ). Phenotypic distributions for these 2 cases are unimodal, although not normal (fig 1).

For small genetic effects ($t \leqslant 10$, *ie* 1σ) t was overestimated, in particular when $t = 0$, as was already mentioned. For larger genetic effects, t was overestimated for

Table III. Average major gene parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) under the null-hypothesis for varying size of the data set.

N	t	d	σ^2
100	15.90	0.50	57.1
250	13.72	0.50	67.0
500	12.54	0.49	73.2
1 000	11.35	0.51	77.2
2 000	10.51	0.50	81.3

Simulated: $\sigma^2 = 100$; N: number of observations in the data set.

Table IV. Power of the test and average parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) in different situations (data sets with 1 000 observations, 100 resplicates).

Simulated parameters			*Power*	*Estimated parameters*		
σ^2	d	t		t	d	σ^2
100	–	0	3.1	11.4	0.51	77.2
100	0.50	10	3	12.6	0.44	84.7
		15	7	14.0	0.47	95.4
		20	12	18.2	0.47	100.2
		25	29	23.4	0.48	104.4
		30	38	28.1	0.50	108.6
		35	82	34.9	0.50	99.2
		40	96	39.8	0.50	103.3
100	1.00	10	1	14.1	0.93	61.6
		15	70	18.2	0.83	90.9
		20	100	22.4	0.87	94.0
		25	100	27.2	0.89	95.6
		30	100	32.7	0.90	94.0
		35	100	37.6	0.90	94.8
		40	100	40.9	0.96	97.4

Power: Number significant at nominal 0.05 level (total = 100); first line: based on 1 000 simulations under H_0 (tables I and II).

$d = 1$ and was underestimated for $d = 0.5$. For $d = 0.5$, average estimates for t and d differed from the simulated values by $< 1\%$ when the power reached near 100%. For $d = 1$, however, the bias in t was still 10% when the power had reached 100%. This bias reduced gradually, and was $< 1\%$ for a genetic effect of $t = 40$.

In figure 2 power of the test is depicted for varying sizes of the data set. Two additive effects were chosen, with $t = 25$ and $t = 35$. Each point in the figure is on average of 100 replicates. The power increased with increasing number of observations. Increasing the number of observations $> 1\,000$ gave relatively less improvement in power, especially for the smaller effect (= 25). For a small number

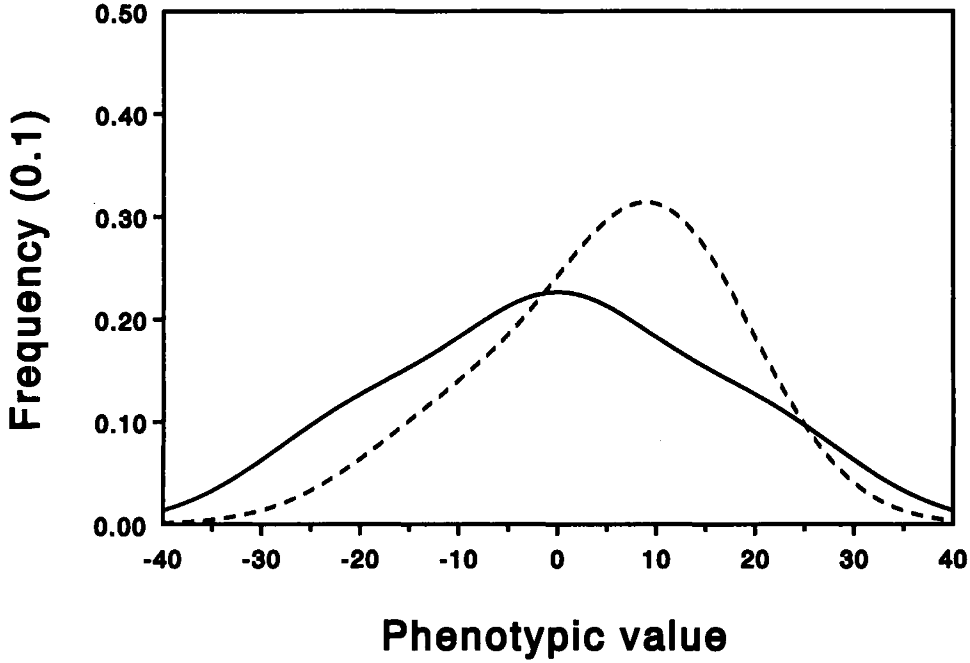

Fig 1. Phenotypic distributions on which over 95% power was reached for the identification of a major gene: $t = 40$, $d = 0.5$ (solid line) and $t = 20$, $d = 1$ (dashed line); $\sigma = 10$.

of observations this graph is expected to level off at the type I error (nominally 5%), but sampling makes results somewhat erratic.

Robustness when ignoring polygenic variance

Data following model [2] were simulated with $d = 0.5$ and $t = 35$ and different proportions of polygenic and residual variance. The data set contained 20 sires with 5 dams each and 10 offspring per dam; each situation was repeated 100 times. Estimated parameters and resulting power are in table V. Parameter estimates for t and d, and the power of the test were not affected when a part of the variance was polygenic. The total estimated variance was equal to the sum of simulated variances.

Robustness when ignoring segregation in the parental lines

Data following model [1] were simulated with $d = 0.5, t = 35, \sigma^2 = 100$ and various values for f_p and f_m. The genotype probabilities in parents (F_1) and offspring (F_2) have been given in table VI. For the first 3 situations, genotype probabilities in the F_2 were 1/4, 1/2 and 1/4, as assumed under the fixation assumption. For the last 3 situations, however, genotype probabilities were different, because the allele

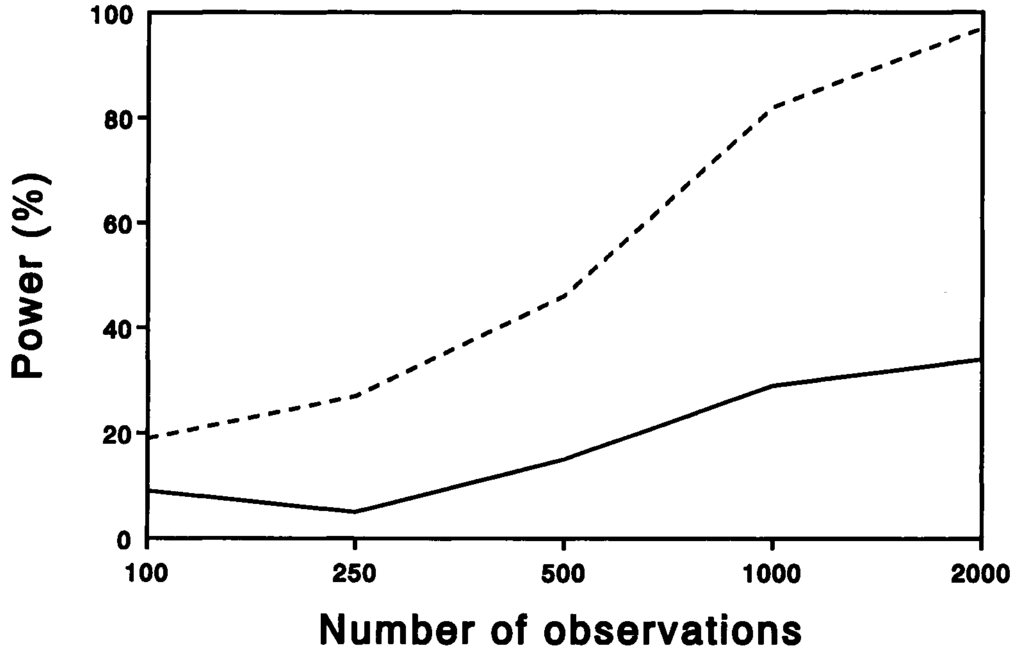

Fig 2. The power for detection of a major gene in relation to the size of the data set shown for 2 situations: $t = 25$ (solid line) and $t = 35$ (dashed line); $d = 0.5$ and $\sigma = 10$.

Table V. Power of the test and average parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) when polygenic variance is present (data sets with 1 000 observations, 100 replicates).

Simulated parameters		*Power*	*Estimated parameters*		
σ_g^2	σ_e^2		t	d	σ^2
0	100	82	34.9	0.50	99.2
20	80	87	35.0	0.50	99.6
40	60	80	34.4	0.51	102.5
60	40	78	34.5	0.50	101.4
80	20	90	35.3	0.50	96.7
100	0	80	34.5	0.50	100.0

σ_g^2, σ_e^2: simulated polygenic and residual variance; other parameters simulated: $t = 35$, $d = 0.5$; power: number significant at nominal 0.05 level (total = 100).

frequency was not 0.5 on average. High average allele frequencies were simulated, but because only additive effects are considered, results are equally valid for low allele frequencies. The power remained equal, as long as genotype probabilities in F_2 remained 1/4, 1/2 and 1/4, and parameter estimates are unbiased (table VII). In case the allele frequency did not average 0.5, however, parameter estimates were biased. The power of the test increased, because in this situation the distribution became skewed. The situation with $d = 0.5$ and $t = 35$ for data where the gene is fixed in parental lines (table IV), with a power of 82%, may serve as a reference.

Table VI. Genotype probabilities in F_1 and F_2 for different allele frequencies in the parental lines.

			F_1 probabilities			*F_2 probabilities*	
f_p	f_m	A_1A_1	A_1A_2	A_2A_2	A_1A_1	A_1A_2	A_2A_2
0.9	0.1	0.09	0.82	0.09	0.25	0.50	0.25
0.8	0.2	0.16	0.68	0.16	0.25	0.50	0.25
0.6	0.4	0.24	0.52	0.24	0.25	0.50	0.25
0.9	0.3	0.07	0.66	0.27	0.16	0.48	0.36
0.9	0.5	0.05	0.50	0.45	0.09	0.42	0.49
0.9	0.7	0.03	0.34	0.63	0.04	0.32	0.64

f_p, f_m: frequency of A_1 allele in paternal and maternal line.

Table VII. Power of the test and parameter estimates for genetic effect (t), dominance coefficient (d) and variance (σ^2) when alleles are segregating by various frequencies in the parental lines (data sets with 1 000 observations, 100 replicates).

$\mathrm{f_p}$	$\mathrm{f_m}$	*Power*	t	d	σ^2
0.9	0.1	76	34.37	0.50	103.9
0.8	0.2	83	34.66	0.51	101.5
0.6	0.4	76	34.14	0.50	105.6
0.9	0.3	81	31.99	0.58	113.4
0.9	0.5	92	26.02	0.77	127.2
0.9	0.7	99	21.17	0.96	115.9

Simulated: $t = 35, d = 0.5, \sigma^2 = 100$; f_p, f_m: allele frequency in paternal an maternal line; power: number significant at nominal 0.05 level (total = 100).

Inclusion of F_1 data

Five hundred, or 1 000, F_1 observations were also simulated, with additive major gene effects (table VIII). With no major gene effect ($t = 0$ and hence $\sigma^2_{mg} = 0$), and with equal variances in F_1 and F_2 (situation 1) the average estimated t was much smaller than in the model using only F_2 data (table III). In the second situation (table VIII) a major gene effect of $t = 20$ was simulated which corresponds to the given major gene variance of 50. When using only F_2 data, the test had a power of only 12% for detection of an additive effect of $t = 20$ (table IV). When including F_1 data, however, the power was 100% (table VIII). From the situations 3 and 4 considered in table VIII, however, it becomes apparent that when F_1 data were included, the major gene was detected only by its effect on variance, considering a power near the type I error rate as irrelevant. When the variance in F_2 increased by 50%, but when in fact no major gene was present, a major gene was found in 100% of the cases. For smaller increases of the variance (10%) major genes were still detected, and the probability of detection increased with the size of the data set (alternative 3* with more F_1 observations). A major gene was totally undetectable, on the other hand, when the total variance in F_1 was equal to the total variance in F_2 (situation 4). This shows that the ability to detect a major gene can even be worsened when F_1 data are included. If only F_2 data were used a major gene with similar effect was detected in 12% of the cases (table IV).

Table VIII. Power of the test and parameter estimates for genetic effect (t) and variance (σ^2) in different situations when 500 F_1 and 1 000 F_2 observations are combined.

	F_1	F_2			*Estimated parameters*	
Situation	σ_e^2	σ_e^2	σ_{mg}^2	*Power*	t	σ^2
1	100	100	0	1	3.03	97.9
2	100	100	50	100	19.43	100.8
3	100	150	0	100	19.62	99.3
3	100	110	0	15	7.72	99.1
3*	100	110	0	25	8.11	99.3
4	150	100	50	2	5.05	145.3

Situation: refers to table II; 3*: alternative with 1 000 F_1 observations instead of 500; σ_e^2, σ_{mg}^2: simulated residual and major gene variance; power: number significant at nominal 0.05 level (total = 100).

DISCUSSION AND CONCLUSIONS

Type I error

Nominal levels for type I errors were based on Wilks (1938) who proved asymptotic convergence of the likelihood ratio test statistic to a χ^2 distribution. Type I errors decreased and stabilised for larger data sets, as expected. The estimated type I errors, however, were significantly too low. It is unlikely that the type I error, after having first decreased, would increase for even larger data sets as studied here. It can be concluded therefore, that type I errors are significantly lower than expected in the asymptotic case, and that for large data sets the likelihood ratio test is conservative. It has been investigated whether the constraint used on the dominance coefficient could have caused the too low type I errors. However, this was not the case, because even with no constraint, too low type I errors were found of 7.5% and 3.9% at nominal levels of 10 and 5%.

For the investigation of power we have chosen to use the theoretical asymptotic quantiles, although they were shown to give a conservative test. The nominal level for the type I error is then an upper bound, and the experimenter still has a reasonably good idea of the risk of making a type I error. When the actual type I error would be above the expected level, however, the test would become of less use.

A second reason for still using theoretical asymptotic quantiles is that adapting the test is difficult and of little practical use. A difficulty is, for instance, that estimated quantiles would be subject to sampling and the obtained point estimate is therefore only expected to give the correct test. Therefore, 2 experimenters investigating the same test, will find different critical values and the test applied will depend on the experimenter. Also in practice such a procedure would be difficult to apply since the calculated quantile would only hold for the same model and data sets of similar size and structure.

Power of the test

Using only F_2 data, the power of this test was poor for additive effects (dominance coefficient = 0.5). This can be explained by the resulting symmetrical distribution which is similar to the distribution under H_0. In this case, the genetic effect has to be about 4σ to be detectable, which corresponds to a heritability of 0.67 in the F_2 generation. When the dominance coefficient is 1, an effect of 2σ was detectable. These results are based on data sets with 1 000 observations, but it was shown that the power decreased dramatically for smaller data sets.

Power increased when F_1 data was included in the analysis, and additive effects of 2σ could be detected. In that case the increase in variance in F_2, caused by the major gene, was taken as an important indication for the presence of a major gene. The power to detect a major gene in F_2 data may also increase if alleles were not fixed in the parental lines, or alternatively F_3, instead of F_2, data were used. This corresponds more to the situation in a usual population, where between-family variation will arise. For F_3 data, for example, when pure lines were homozygous, the allele frequency will be 0.5, and parents will be in Hardy-Weinberg equilibrium. For such a situation, Le Roy (1989) found a power of 25% for an additive effect of 2σ in a data set of 400 observations (20 sires with 20 half-sib offspring each). In figure 2, the power for a data set of similar size can be seen to be only $\approx$ 10% for an even larger effect of $2.5\sigma (t = 25)$. This indicates that an increase in power may be expected when the F_3 generation is observed, despite the facts that more parameters have to be estimated, and that parents' genotypes are no longer known.

The power for detection of a major gene is related to the unexplained variance in the model of analysis. The inclusion of fixed and polygenic effects will therefore make the major gene easier to detect, provided that all these effects can be accurately estimated.

Parameter estimates

For additive effects simulated ($d = 0.5$), bias for the average estimated genetic effect t and dominance coefficient d was less than 1% when the power approached 100%.

For dominant effects ($d = 1$), however, t was overestimated by 10% when the power for detection of a major gene reached 100%. This overestimate is probably related to the underestimate for d, which resulted from the applied constraint. As mentioned, this constraint was applied to prevent t from going to zero, at which point d tended to go to infinity. When such a constraint was not applied with, for instance, an effect of $t = 10$ and $d = 1$, analyses gave in 100 replicates an average estimated d of 2.93. This is an average overestimate of ≈ 200%. The average estimate using the constraint was 0.93, showing that indeed better estimates were obtained under the restriction, even when the true value was on the border of the allowed parameter space. In practice, of course, overdominance cannot be excluded and parameter estimates could be compared with and without this constraint. A small, near zero, estimate for t and a large estimate for d would suggest a possible overestimation of d.

For very small or absent effects, the ML estimates were considerably biased. In this situation, the asymptotic properties of ML estimates, *ie* consistency, are far from being attained. In the absence of a major gene, average estimates were presented for increasing size of the data set. This showed that the average estimate decreased, and will probably reach the true value when the number of observations is very much larger. Bias of ML estimates in finite samples also resulted in significant t-values when no effect was present. This indicates that the presence of a major gene should not be judged by the estimates and their standard errors. The standard errors discussed here were empirical standard errors. In practice such standard errors will have to be obtained using the inverse of an estimated Hessian matrix, or some other quadratic approximation of the likelihood curve near the optimum. Using the estimated Hessian matrices, we found roughly the same standard errors, although they were not very accurate. In our study, the quasi-Newton algorithm was started close to the optimum and not enough iterations are then carried out to estimate the Hessian matrix accurately.

Robustness of model and test

Inclusion of F_1 data results in a poorly robust test when differences in variances would arise between the F_1 and F_2 due to other causes than a major gene. An increase in variance from F_1 to F_2 can result in a putative major gene being detected. An increase in variance of 10%, for instance, gave 25% false detections when 1 000 F_1 and 1 000 F_2 observations were combined. Such increases are not unlikely, due to, for instance, polygenes. The major gene test is then merely a test for homogeneous variance in F_1 and F_2. The inclusion of F_1 data could also worsen the detection of a major gene, when the environmental variance in F_2 was less. Therefore any differences in variance, due to other causes than the major gene effect, will bias the parameter estimates. Also in a model that allows for segregation, such biases will remain.

It was shown that the model is robust when polygenic effects were ignored. This can be explained by the fact that the test uses only the non-normality of the distribution as a criterion. It must be noted however that, when polygenic effects can be accurately estimated, including a polygenic effect in the model will increase power because it reduces the residual variance.

Another aspect of robustness concerns the assumption of fixed alleles in parental lines. It was shown that parameter estimates were not biased when alleles segregated, as long as the average frequency in the 2 lines was 0.5. In that case the assumed fitting proportions 1/4, 1/2 and 1/4 are still correct. If the average frequency in parental lines differed from 0.5, t was underestimated and, because skewness was introduced, estimates for d deviated from 0.5. This second situation is more likely to occur than the situation where the average frequency is exactly 0.5. Because it could be difficult to justify the fixation assumption *a priori*, application of a more general model that allows for segregation in parental lines might have to be considered.

A final aspect of robustness concerns non-normality of the distribution not due to a major gene. As stated earlier a mixture distribution is fitted and the detection of a major gene in F_2 data, assuming fixation, relies solely on the non-normality caused by the major gene. This means that in fact only a significant non-normality is proven. The method would therefore be poorly robust against any non-normality due to another cause. The robustness might be improved using data in which alleles segregate in parents. This is guaranteed in F_3 data, but may also arise in F_2 data, when alleles were not fixed in parental lines. If segregation in parents is the case, evidence for a major gene is no longer only in the non-normality of the overall distribution, but also for instance in heterogeneous within family variances. Therefore a model that allows for segregation is not only preferred to increase power, but is also preferred to improve robustness.

ACKNOWLEDGMENTS

This research was supported financially by the Dutch Product Board for Livestock and Meat, the Dutch Pig Herdbook Society, Bovar BV, Euribrid BV, Fomeva BV and the VOC Nieuw-Dalland BV. Profs Brascamp and Grossman and Drs Van Arendonk and Van Putten are acknowledged for helpful comments and editorial suggestions. Comments of 2 referees have helped to shorten the manuscript and to improve the discussion section.

REFERENCES

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. *Hum Hered* 21, 523-542

Elston RC, Stewart J (1973) The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. *Genetics* 73, 695-711

Elston RC (1984) The genetic analysis of quantitative trait differences between two homozygous lines. *Genetics* 108, 733-744

IMSL (1984) *Library Reference Manual Edition 9.2.* International and Statistical Libraries, Houston, TX

Le Roy P (1989) Méthodes de détection de gènes majeurs; application aux animaux domestiques. Doctoral Thesis, Université de Paris-Sud, Centre d'Orsay

Morton NE, MacLean CJ (1974) Analysis of family resemblance III. Complex segregation of quantitative traits. *Am J Hum Genet* 26, 489-503

Wilks SS (1938) The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat* 9, 60-62