STATISTICAL METHODS FOR THE DETECTION OF MAJOR GENES IN FARM ANIMAL POPULATIONS

Sara Anne Knott

Ph.D.
University of Edinburgh
1990

CONTENTS

ACKNOWLEDGEMENTS

I would like to thank Professor W.G. Hill, Dr R. Thompson and Dr C.S. Haley for supervising my Ph.D. studies. In particular, I appreciate the struggles of Robin in attempting to explain his ideas to me, and although I have often complained, progress has been made! Also, special thanks to Chris who has always been interested in my work and has spent much of his time listening to my problems.

Whilst doing this work, I was fortunate in having contact with people outside Scotland who were interested in my work. I would like to thank Jean-Michel Elsen and Pascale Le Roy for some interesting discussions. Thanks also to Johan van Arendonk for organising my stay in Wageningen, The Netherlands, and to the faithful attenders of our discussion meetings there.

The computing involved in this work has been excessive. I would like to thank the operators/managers of the various computers I have used, both within the AFRC and at Edinburgh University.

On a social note, thanks to the animal breeding students and staff in Edinburgh, and, of course, to Chris for his support and understanding.

Chapter 1 is based on a paper written with Bill Hill for the conference 'Advances in Statistical Methods for Genetic Improvement of Livestock' which took place in February, 1987 in Armidale, Australia. The proceedings have been published subsequently (Eds. D. Gianola and K. Hammond, 1990, Springer-Verlag, Berlin). Robin Thompson drew my attention to the papers by I. Hoeschele which suggested the line of study described in chapter 5. Robin has been responsible for some of the ideas which initiated areas of my research, in particular the use of Hermite integration, section 5.5 and section 6.6 and he also helped with the work involved in the derivation of the REML variance equations (Appendix 4).

This work was carried out at the AFRC Institute of Animal Breeding and Genetics Research whilst in receipt of an AFRC studentship.

ABSTRACT

Animal breeding theory is based on the assumption that traits are controlled by many genes each having small effect, however, genes with a large effect have been identified in favourable circumstances. Where major genes can be identified, and individual animals genotyped, exploitation of the genetic variation can be optimised. Segregation analysis has been proposed as a suitable method for detecting major genes. It involves maximising and comparing the likelihood of the data under different genetic models to ascertain the most likely genetic structure. To identify a major gene the likelihood of the data under a polygenic model is maximised and compared with the maximum likelihood under the mixed model (i.e. containing a major gene and polygenic component). A significant improvement in the likelihood obtained by incorporating the major gene gives evidence for its existence.

Equations for the exact mixed model and polygenic likelihoods can be obtained, however the mixed model likelihood involves the integration of a complex function. Several approximations to this likelihood have been investigated. The first effectively retains the integration and approximates crossproduct terms involving the major gene and the polygenic component. The second (Herm) approximates the integration with a summation using the Hermite polynomial. The likelihood has been maximised using a quasi-Newton algorithm. The third and fourth methods are extensions of mixed model methods (in the statistical sense, i.e. including fixed and random effects), which are already familiar to animal breeders. One replaces the integration with a single estimate of the mode of each sire's transmitting ability distribution (ME1), the other estimates three modes one for each possible major genotype of the sire (ME3). These have been implemented using an expectation-maximisation algorithm.

The first approximation was thought too complex to extend to include, for example, fixed effects. The operational characteristics of the other three methods have been investigated using simulated data. The Monte-Carlo simulation program uses Boolean algebra to describe the genotype of individuals at each locus and the inheritance of the alleles. Different genetic models have been considered and the data was analysed twice, firstly assuming that the polygenic heritability was known and fixing it at the expected value, secondly estimating the heritability from the data. For all the analyses the simulated data contained 50 sires each with 20 half-sib offspring.

The test statistic obtained using Herm when analysing polygenic data was not significantly different from the expected distribution: a χ^2 distribution with three degrees of freedom. However, the other two approximations gave lower test statistics for the same data and a high percentage resulted in a polygenic model with zero test statistic.

When analysing data containing a major gene, Herm was the most powerful. The frequency of detection of a major gene depended on the proportion of the genetic variance explained by the major gene and whether the major gene caused the distribution to be skewed. A dominant major gene explained the highest proportion of the genetic variance (79%) and caused skewness of the distribution and was detected in virtually all the analyses with all the methods. In terms of power, the approximations were most different when the simulated major gene was additive with equal allele frequencies and explained 61% of the genetic variance and the heritability was fixed in the analyses, Herm detected a major gene in 75% of analyses, ME3 in 33% and ME1 in only 4%. With the heritability fixed the Herm and ME3 analyses were more powerful, however, the analyses are not robust to incorrect values for the polygenic heritability and if under estimated a major gene can be inferred to explain the additional genetic variance not explained by the heritability. Using Herm in the majority of analyses the mixed model was most likely if not significantly so. Using ME1 and, to a lesser extent, ME3, the analyses often resulted in a polygenic or major gene model. The parameter estimates, when a mixed model resulted, were, on average, in reasonable agreement with the expected value for all three approximations. When a model containing a major gene resulted the three methods were similar in their ability to genotype individuals at the major locus.

Segregation analysis is capable of detecting a major gene segregating in a population and can accurately estimate its effect and frequency. Approximations to the mixed model likelihood make the method feasible for large data sets.

CHAPTER 1

DETECTION OF GENES WITH LARGE EFFECT

1.1 INTRODUCTION

Animal breeding theory is based on the assumption that traits are controlled by many genes each having small effect. The action of individual genes cannot be observed directly and traits are generally described in terms of summary statistics such as the heritability. However, genes with large effect on commercial traits have been identified in favourable circumstances. Notable examples are the dwarfing gene in poultry (Merat and Ricard, 1974), the Booroola gene affecting ovulation rate in sheep (Piper and Bindon, 1982; Piper *et al.*, 1985), the double muscling gene in cattle (Rollins *et al.*, 1972; Hanset and Michaux, 1985a, b), and the gene determining halothane sensitivity in pigs (Smith and Bampton, 1977). Where genes can be identified and individual animals genotyped exploitation of the genetic variance can be optimised. Major genes could also provide raw material for genetic engineering programmes.

Despite large phenotypic effects, major genes are often not immediately apparent due to the obscuring effects of polygenic and environmental variation. Hence, it is likely that genes of lesser effect than those already known but still of major phenotypic and potential economic importance remain to be detected.

The aim of this work is to develop statistical methods for the detection of major genes in farm animal populations. First a review of existing methods will be given.

1.2 REVIEW

1.2.1 Prior Information - Number of Genes

In farm animals there is little information about the numbers and effects of genes influencing quantitative traits, apart from the few major genes which have been identified. In laboratory animals, and particularly *Drosophila*, data are much more extensive. There are several questions that can be posed; for example: How many genes influence the trait? How much selection response can be attributed to one or a few loci? What proportion of the variation in a population is accounted for by segregation at the most important locus, the second most important locus, etc? Most completely: what is the distribution of gene effects on the trait in the population, supplemented, if possible, by the distribution of gene frequencies? As background, the methods and results of estimates of gene number are briefly reviewed.

Effective Number of Genes. The classical estimate of effective number of genes $\frac{\text{range}^2}{\text{variance x 8}}$ devised by Wright (Castle, 1921; Wright, 1952) has been very widely used and does, in principle, demonstrate the presence of genes of large effect if the estimate of number is very low. Various modifications have recently been suggested (Lande, 1981; Comstock and Enfield, 1981; Cockerham, 1986), but the estimate is likely to be biassed downwards by linkage of genes in coupling. Estimates from different species and experiments vary greatly (see, for example, Falconer, 1981), and it is difficult to draw general conclusions from them.

Genotype Assay. The genotype assay method proposed by Jinks and Towey (1976) can be used in plants were rapid inbreeding can be practised, and genes are identified by segregation within sublines drawn from lines having already had two or more generations of selfing following a cross.

Chromosomal and Intra-Chromosomal Analysis. The identification of effects of individual chromosomes has been carried out in analyses of selected lines of *Drosophila* using marker and cross-over suppressor techniques (e.g. Mather and Jinks, 1982). These analyses can be extended to analyse effects within chromosomes by forming recombinants against multiple marked chromosome stocks and, in principle, if continued with enough effort, leads to the mapping of all genes differentiating a pair of extreme lines (Thoday, 1961).

The most complete data are from analyses of lines of *Drosophila* selected for bristle number, where chromosome manipulation can be used to attribute effects to chromosomes and to sites within chromosomes. Analyses made by Mather and others (Mather and Jinks, 1982) have shown that selected lines usually differ at all chromosomes. Analyses of recombinants within chromosomes have suggested, at least in some cases, that most selection response is due to very few (two or three) genes; Thompson and Thoday (1979) give a review. A frequency distribution of effects has been compiled by Shrimpton (1981). It shows gene effects on bristles of up to more than two standard deviations, but with an increasingly higher frequency of genes of smaller effect, down to about one-half of a standard deviation. Below that size of effect, the method used does not enable the genes to be ascertained separately. A model with an exponential form seems appropriate to describe the distribution of effects, but hard evidence for it is lacking. It seems likely from such detailed analysis, that there are some genes of large effect segregating for most traits.

1.2.2 Methods

Methods have been proposed for the detection of genes of large effect both from the analysis of differences between populations and where the gene is segregating within a population. In animals, the use of population differences has mainly been in species which can be experimentally manipulated, especially where inbred lines can be maintained. For genes segregating within a population there has been substantial work in recent years by human geneticists, developing methods for identifying genes of large effect on continuous and discrete traits such as disease incidence, motivated by attempts to understand the basis of their inheritance for use in counselling. Most of the tests are based on finding departures from normal distributions; some involve simple computations but others require heavy computation of likelihoods. Their utility depends on their ease of use, power in detecting segregation of a major gene, and sensitivity to breakdown of assumptions, notably of normality of environmental and background genetic distributions.

Segregation In Crosses and Backcrosses

Perhaps the standard method for identifying genes of large effect is from the analysis of segregation in crosses and backcrosses among homozygous lines, particularly those which differ substantially for the metric trait of interest. However, if the lines are inbred, then nothing can be done until the F2 and backcrosses are obtained, when it is possible to search for non-normality possibly using likelihood fitting routines. With outbreeding lines, a significant increase in the variance of the F2 and backcrosses may be an indication that a gene of large effect is segregating. In any event, the analysis has to be continued for at least one more generation to test putative genotypes. When family information on two or more generations is combined, then maximum likelihood techniques such as segregation analysis can be used. It is clear, however, that two or more closely linked genes in coupling cannot be distinguished from a single major gene in the early generations after line crossing.

Repeated Backcrossing and Selection

Wright (1952) suggested that genes of large effect could be identified by repeated backcrossing of, e.g., crosses of a high scoring line to a low scoring line, and selecting for high score. This method leads to a halving of allele frequency in the absence of selection, so only genes having large effect in the heterozygote, such that fitness of animals carrying them is effectively doubled, can be maintained against the backcrossing force. Thus, selection for prolificacy by the Seears brothers maintained the Booroola gene despite their use of only bought-in rams (Piper and Bindon, 1982).

Departures from Normality

If no major genes are segregating, many genes with small effect are acting additively on the trait, and environmental deviations are continuously distributed additive to genetic effects and normally distributed, the central limit theorem implies that, after appropriate scaling, observations of traits on individuals and their relatives follow a multivariate normal distribution. Various tests have been suggested that are based on departures from multivariate normality: skewness or kurtosis of distributions of observations and of family means, non-linearity of regression of performance of progeny on parent, asymmetry of responses to high and low selection, heterogeneity of variance within families, and association between variance and level of performance. These will be reviewed only briefly, for all such methods use only part of the information contained in the data and, therefore, as computing power increases, are likely to be superseded by methods making fuller use of the data, notably maximum likelihood.

Heterogeneity of Variance. If a gene of large effect is segregating in the population, heterogeneity of variance within families is expected. Further, as Pearson (1904) was the first to point out, the variability of progeny about the parental value will be greater for intermediate scoring parents than for extremes (see also Felsenstein, 1973; Smith *et al.*, 1978), and the variability will be greater within families of intermediate than of extreme mean performance (Fain, 1978). Similarly, Penrose (1969) showed how the differences between the full-sib and offspring-parent correlations depended on the number of genes and mean score. Matthysse *et al.* (1979) suggested that the correlation between the within-family variance and the family mean be used as a statistic. The efficiency and sensitivity of these methods have been analyzed by their proposers and by others (Mayo *et al.*, 1983), who considered their power, or lack thereof, and sensitivity to assumptions. These include genes having a geometric distribution of effects (Matthysse *et al.*, 1979), differences in environmental variance between homozygotes and heterozygotes (Mayo *et al.*, 1980), dominance (Felsenstein, 1973), and so on.

Skewness and Kurtosis. Segregation of genes of large effect leads to both skewness, except at specific gene frequencies (Fisher *et al.*, 1932), and kurtosis, and the degree of kurtosis can be used as an estimator of gene number (O'Donald, 1971). The method rests very strongly on the basic normal assumptions of, for example, the environmental error distribution, and is sensitive to heterogeneity of variance. Hammond and James (1972) demonstrate the lack of power using data from *Drosophila*.

These methods were extended by Merat (1968), who suggested that deviations from family mean in families of high and low variance should be pooled, and skewness and kurtosis checked on the deviations. The method was discussed and used on *Drosophila* data by Hammond and James (1970).

Non-Linearity of Regression of Progeny on Parent. Robertson (1977) analysed the influence of genes of large effect and other factors such as skewed environmental distributions on departures from linearity of the regression of progeny on parent and, consequently, on asymmetry of response to high and low selection for the trait. Departures are largest for recessive genes at low frequency. Maki-Tanila (1982) discussed this further, and considered the use of sib on sib regression in addition to the offspring on parent regression. The power of tests for single genes using non-linearity has not be investigated.

Q-Q plots. The use of quantile-quantile plots has been suggested for obtaining evidence for the presence of mixtures (Titterington *et al.*, 1985; Everitt and Hand, 1981). Hoeschele (1988a) suggests its use for the detection of a segregating major gene in animal breeding data. For a sire model she proposes predicting the additive genetic effects for each sire assuming a model without major genotypes, i.e. a polygenic model, and using these as the mixture quantiles. There is a problem with the fact that covariances exist between the estimates, but Hoeschele suggests ignoring them. The points of inflection, slopes and intercepts can be used to obtain estimates of the mixing proportions, the component means and the standard deviation. Hoeschele (1988a) suggests fitting a function to this curve, the second derivatives of which give the mixing weights. Using these the other estimates can be obtained by least squares. Everitt and Hand (1981) indicate that unless the difference between the means is large, the method has very little power.

Structured Exploratory Data Analysis

A simple and quite different method, subsequently called structured exploratory data analysis (SEDA), was proposed by Karlin *et al.* (1979), and has been extended to include a group of tests, which, it has been suggested, should be applied together to indicate the presence of a major gene. The three tests that have been most used are the major gene index (MGI), offspring between-parents function (OBP), and mid-parental correlation coefficient (MPCC).

The MGI has received most attention and is based on the argument that if a major gene is segregating, the deviation of the observation on an individual from the parent mean would tend to be larger than the geometric mean of deviations from the individual parents. Subsequently Famula (1986) suggested an improvement to the method and also showed how mixed model methods could be employed to estimate fixed effects.

OBP defines the proportion of offspring within an interval of defined length around the mid-parental value. A larger proportion should be nearer the mid-parental value with multifactorial compared with monogenic inheritance. MPCC is the correlation of the offspring on the mid-parental value. This should be greater than zero if variation in the trait is genetically controlled.

Other tests and graphical methods have also been suggested. Analyses of properties of the methods have been undertaken by Karlin *et al.* (1979, 1981), Karlin and Williams (1981), Mayo *et al.* (1983), Morton *et al.* (1982), Famula (1986), and Kammerer *et al.* (1984). SEDA is not based on normal assumptions, which should make it more robust. However, although quick and simple to apply, the method seems rather *ad hoc* and, as pointed out by Mayo *et al.* (1983), for models of unequal effects gives values similar to those for multifactorial models. Kammerer *et al.* (1984) found that although reasonably sensitive in detecting the presence of a major gene, SEDA lacked specificity and consistently classified polygenic traits as due to a major gene. It seems likely that SEDA will be superseded by more formal methods.

Segregation Analysis

Segregation analysis combines information on distributions and genetic relationships using maximum likelihood techniques. Originally, formally described for the situation of a major gene segregating within a population (Elston and Stewart, 1971) and later considered for the analysis of crosses originating from homozygous lines (Elston and Stewart, 1973). The method involves maximising and comparing the likelihood of the data under different genetic models to ascertain the most likely genetic structure. Possible genetic models could be the major gene model, the polygenic model or the mixed model (a combination of the former two models; Morton and MacLean, 1974). It is computationally demanding, especially for the mixed model, and hence only small pedigrees can be analysed. Therefore methods need to be simplified in order to be suitable for large animal breeding data sets. The use of maximum likelihood means that parameter estimates to describe the genetic model being considered are obtained.

Although analyses in human data structures, generally nuclear families consisting of parents and their full-sib offspring, suggest that the method is fairly powerful (MacLean *et al.*, 1975) data structures relevant to animal breeding have not been considered. However, the method is sensitive to non-normality of the data, and any non-normality might be misinterpreted as a major gene (MacLean *et al.*,1975).

Bonney (1984, 1986) has proposed the use of regressive models in which the phenotypes of relatives are fitted as covariates in computing the likelihood under different genetic models. These methods may lead to substantial improvements in computing efficiency, and recently their equivalence with the mixed model has been shown under certain conditions (Demenais and Bonney, 1989).

Use of Physiological Markers

On the assumption that more basic traits, such as levels of a hormone or metabolite, are influenced by fewer genes than traits of commercial importance such as growth rate or egg production, then it may be useful in analyses of crosses to monitor such traits. Any evidence of discontinuity or bimodality may point to genes affecting the physiological trait, and their effect on the economic trait can then be estimated. Segregation analysis can be employed to increase efficiency.

Application of this method within a segregating population is less likely to be useful, when there is no prior information. If, however, such data were collected for other purposes there might be benefit in analysing them. Also the use of lines selected high and low for a trait of interest offers an efficient method for the detection of major genes. If there is little or no difference in the physiological trait between the selected lines, there is no point continuing. Further, if there is a difference between the lines but there is no increased variance or bimodality in the F2, there is also no point in continuing the analysis of correlated variables as potential indicators of commercial traits. Only genes producing very large differences in the indicator trait are likely to be of interest because the consequent difference in the commercial trait is likely to be smaller.

Use of Linked Markers

By use of marker genes, which are now becoming available in large numbers as restriction fragment length polymorphisms (RFLPs) and minisatellites (Jeffreys *et al.*, 1985), the effects on quantitative traits of regions of chromosomes can be estimated (e.g. Soller and Beckmann, 1982, 1985; Elston, 1990b). For markers associated with regions of large effect, further generations of crossing can be performed to detect whether the effect is due mostly or entirely to a single gene and is close to the marker. Of course, as more markers become available and associations with more traits can be tested, the greater the chance that spurious effects will be detected unless the power of individual tests is reduced. Markers such as the RFLPs and minisatellite probes to particular sites are likely to be more useful than the available minisatellites which hybridize to many sites. This is so because although the latter are hypervariable, their allelism is usually difficult to ascertain and bands of the same mobility may not correspond to the same locus.

This method has been discussed for farm animals by Soller and Beckmann (1982, 1985) and reviewed further by Elston (1990). Major genes detected by association with a marker can be easily manipulated both for selection within a population and introgression to another population. Crosses from inbred lines will be easiest to utilise as in outbreeding populations the phase of the linkage will have to be assessed for each family.

Miscellanea

Non-Normally Distributed Traits. Segregation analysis is appropriate for traits which, in the absence of segregation of a major gene, are normally distributed or can be transformed to normality, and for all-or-none characters in which a normal-threshold model can be assumed. Some traits in farm animals, however, such as litter size in pigs or in prolific breeds of sheep, have discrete distributions with many classes; these traits are likely to be close enough to normal for crude but not for fine-scale analysis. Other traits have notably non-normal distributions: particularly egg number in poultry and body size in fish after rearing under competitive conditions. These distributions are skewed and not normalised by any standard transformation. It is clear that developments in the formal methods are needed, but it is not clear how these should proceed. Nevertheless, *ad hoc* methods can be used successfully. For example, Piper and Bindon (1982) adopted an arbitrary cut-off at three lambs born, and were thereby able to demonstrate major gene inheritance of litter size in the Booroola strain of sheep. Also, Hanrahan and Owen (1985) used the fact that the repeatability of litter size is likely to be much higher in sheep populations in which a major gene is segregating. Neither of these authors tested alternative hypotheses, however, and the high repeatability was associated with a high mean and heritability.

Breeding from Extreme Animals. In view of the low power of detection of single genes by assessing departures from normality, it may be worthwhile to breed from extreme animals on a regular basis, as suggested by Roberts and Smith (1982). In this way more data is accumulated for testing, and if indeed a major gene is present, progress would be made towards its identification.

Use of Selected Populations. If widely divergent populations for analysis through crossing and backcrossing methods as discussed previously are not available, it may be possible to create these by selecting high and low individuals from some base. Then, any genes of large effect which are segregating in the population will contribute a large part of the high-low difference. Such a scheme can only utilise variants segregating initially and so it would seem to have little to offer for gene identification over an analysis using maximum likelihood directly in the base population. There is the potential benefit that in the cross between high and low lines the genes may be at intermediate frequency, but the disadvantage that in the crosses the effects of different genes are correlated by linkage disequilibrium.

Selection has more promise as a technique for identifying genes through their effects on other physiological or structural variables. If the trait is expensive to measure then use of divergent lines is more efficient than analysis of the base population. For example, assume selection is for high and low growth rate. As Bulfield (1985) has suggested, two-dimensional (2-D) gel electrophoresis can be used to identify differences in amount and structure of proteins in several tissues. If such differences are found, the protein can, in principle, be identified and ultimately cloned. Such analyses can give us information about the nature of genes which are associated with genetic changes in the quantitative trait, and have the potential of enabling subsequent manipulation using the cloned gene. Although 2-D gel analyses offer the possibility of screening very large numbers of loci at one time, unless gels are of very high quality differences in protein positions and intensity are hard to detect.

Molecular Manipulation. Transposable elements and retroviruses might be a useful way of detecting major genes if the potential gene has been created by using these as mutagenic agents, as the mutant genes can be identified by tagging with the element. (Mackay, 1985). Alternatively the transposon might provide a probe for an RFLP enabling analysis as if they were linked markers, although the mutant gene itself is being located. (Soller and Beckmann, 1985). However, the chance of advantageous mutants occurring is low.

1.2.3 Use of the major gene

Selection. There have been a few studies considering the benefit of including a major gene in the breeding programme. Mainly they have concentrated on its effect when incorporated in a selection scheme (Smith, 1967; Roberts and Smith, 1982). The genotypes of all individuals are required before the information can be utilised and the benefit of the major gene depends on the proportion of the genetic variance explained by the major gene. If selection using usual methods is effective, the additional information contributes little to the rate of improvement. However, if the heritability is low or where indirect selection has to be practised the rate can be increased substantially. The gain will be maximised at intermediate allele frequencies for the major gene (Smith, 1967). As emphasised by Roberts and Smith (1982) and Smith and Webb (1981) reliable information is required on the effects of the genotypes on all the traits of economic importance. Otherwise the selection programme could be misdirected. Recently, considering the benefits of selecting for certain milk proteins, Gibson *et al.* (1990) have confirmed the necessity of accurate estimates of the effects of the genotypes, demonstrating that incorrect estimates possibly lead to losses in improvement relative to ignoring the major gene information.

Fixation. An alternative would be to make the population homozygous for the advantageous allele, possibly using a linked marker. However, Smith (1967) states that the response from fixing the better allele is normally less than the response from selecting on all available information. Genes closely linked to the major gene will also be selected, and these might have a deleterious effect, and genes that are advantageous might be lost.

Heterozygote advantage. Smith and Webb (1981) consider benefits from designing a breeding strategy to exploit the halothane gene, where, it was suggested that the heterozygote would be the most advantageous genotype. However, this would mean maintaining the less advantageous homozygotes, which in this case would mean keeping sires which had a risk of suffering from porcine stress syndrome and might reduce the rate of future genetic response and be costly or difficult to maintain.

Introgression. This method to exploit the major gene has had little attention. However, transfer of the major gene into another line that does not have the gene might be a good way to improve the other breed. Obviously, the gene might have other deleterious effects in the new breed that had already been eliminated from the old breed and these would have to be selected against.

1.3 DISCUSSION

A large number of methods for identifying genes of large effect have been reviewed, but a more quantitative analysis and discussion is still needed. Several workers have investigated the efficiency of individual methods, but it is necessary to bring these together and compare their powers and limitations for a range of models. For example, how powerful are they at detecting a single gene when all the rest are of infinitesimal effect, or when the rest have a distribution of effects? How sensitive are the methods to heteroscedasticity of environmental variance, or to non-normality of environmental deviations? How readily can they cope with sex-limited traits and nearly continuous traits such as litter size?

The formal methods, notably segregation analysis, for detection of genes within populations have been developed for data on man. Their extension to farm animal populations is clearly necessary and the data available are generally more suitable. For livestock there are repeated records, large family sizes, data on individuals from previous generations recorded at the same age, and large contemporary environmental groups. If human geneticists can get anything from segregation analysis, then surely so can the animal breeders. There are problems, for example, of selection; in principle, however, maximum likelihood techniques can handle these, so there is obviously room for a lot of work and for a very large computer. It is also important to consider the power of the methods and the size and design of experiments to provide data for such analyses. It seems likely to be of little benefit, except in an exploratory.sense, to pursue simple methods based on departures from normality. Karlin's SEDA procedure would seem to have the benefit of being less dependent on normality of underlying variation, but has only an *ad hoc.* foundation. There is clearly a need to adapt maximum likelihood methods to non-normal data, for example, on egg production of poultry.

There are obvious benefits in being able to attribute variation between and within populations to single genes. This is illustrated by the use in breeding practice of the dwarf, halothane, double muscling, and Booroola genes in poultry, pigs, cattle, and sheep, respectively. Ironically, the Booroola gene has too large an effect on litter size for many management systems and even its inheritance was not clarified until many years after the Booroola flock was known.

CHAPTER 2

SEGREGATION ANALYSIS

2.1 INTRODUCTION

Segregation analysis has been developed and widely used by human geneticists in order to ascertain the mode of inheritance of disease traits. Although segregation analysis is likely to be the most generally appropriate method when determining mechanisms of genetic control in animal populations, it has had little attention from animal breeders. The extension to livestock is clearly necessary, where the pedigrees can be much larger than in human data sets and with larger family sizes, but often only simple relationships are considered, for example paternal half-sibs. With farm animals there is the advantage of being able to manipulate them experimentally and, for example, have planned mating schemes, repeat records or records on different individuals at the same age. The use of a maximum likelihood (ML) method means that alternative models of genetic determination can be compared and estimates of the parameters involved in the model can be obtained within a sound statistical framework.

2.2 REVIEW

Segregation analysis was formally described by Elston and Stewart (1971) to encourage the use of the whole pedigree in human genetics when trying to ascertain the mode of inheritance. Prior to this nuclear families (parents and their full-sib offspring) had generally been analysed, assuming independence between families. Although this ignores many of the potentially useful relationships, the aim of most of the analyses had been to establish whether a dichotomous trait was inherited in a dominant or recessive fashion, and hence the loss of information was not important. However, the wish to test more complex models of inheritance including polygenic or environmental components, or several loci required a more powerful method.

The aims of the method were: to establish whether there is evidence for a genetic component in the control of variation of a trait, and if there is to elucidate the mechanism of control, and to calculate the 'risk' that an individual, either in the population or yet to be born, has a particular genotype. The method suggested by Elston and Stewart (1971) obtains this information by maximising the likelihood of the data under different genetic models and comparing these MLs to find the mode of inheritance that best explains the data.

Elston and Stewart (1971) describe many genetic models, all in terms of the transmission of genetic effects from parent to offspring. To test for the presence of a major gene, a model with a major gene with Mendelian transmission probabilities of 0,1/2 and 1 (for the probability that parents of genotype aa, Aa and AA, respectively, transmit allele A to their offspring) and an environmental component is fitted and the ML compared with that from a non-genetic model, with equal transmission probabilities. A significant increase in the likelihood when the major gene is included, and no further significant increase when the transmission frequencies are estimated, relaxing the assumption of Mendelian transmission, is evidence for a major gene. Later Lalouel *et al.* (1983) called this the general transmission single locus model. This single locus model can be easily extended to include sex-linkage or many loci. As the number of loci increases, considering each one separately quickly becomes cumbersome. In the limiting case as the number of loci tends to infinity, i.e. polygenic inheritance, a more simple expression is derived. However, the direct comparison of the likelihoods of a single gene and a polygenic model can not be made as the parameters to be estimated to describe the two models are different, i.e. the two hypotheses are not nested with one being a more general form of the other. A model with the phenotype resulting from the joint effects of a major locus, a polygenic component and random environment, which would allow for the test of a major locus and polygenic component, was suggested but thought to be computationally infeasible.

An alternative method was developed by Morton and MacLean (1974). Again based on a full pedigree (and now called complex segregation analysis) they suggested the 'mixed model' including a major locus, a polygenic component, random and common environmental effects. (Throughout this thesis, the term 'mixed model', following the terminology of human geneticists, will be used to describe a model containing both a major gene and polygenic component, rather than the mixed model in the animal breeding sense, containing random and fixed effects). The common family environmental effects were incorporated to prevent any such effect from being interpreted as dominance at a major locus (Morton and MacLean, 1974).

To test for a major gene Morton and MacLean (1974) fit a normal distribution in which genetic and environmental variances are estimated, and then the same model plus parameters for the frequency, effect and degree of dominance of a single gene, i.e. the mixed model. A significant increase in the ML indicates that a major gene is segregating.

Complex segregation analysis is more powerful than other methods based on normality because all the information contained in the data is used. In considering its efficiency, account has to be taken both of the possibility of detecting individual genes which are not actually present (false positives) and of missing genes of large effect that are segregating (false negatives). False positives could be caused by non-normality of the data, for example by skewness of the distribution, particularly when continuous genetic, individual and family environmental distributions, and a segregating locus are fitted (Elston, 1979; Eaves, 1983). Power transformations of the data (MacLean *et al.*, 1976) can be used to transform to normality allowing for one, two, or three underlying distributions, and to reduce the chances of false detection. In any case, if two distributions fit better than one, this is suggestive of a major gene. There have been no analyses of the power of complex segregation analysis in data structures relevant to animal breeding, the usual structure being that of nuclear families of man, comprising parents and full-sibs. For such a structure, results have been given by Go *et al.* (1978) and MacLean *et al.* (1975) and these are not presented in a form readily transferable to the animal breeding context. The power of the model is, of course, dependent on the effect and frequency of the major gene and on family size, and is greater for continuously distributed than for all-or-none traits because more information is present in the data. Simulation results of MacLean *et al.* (1975) serve as an illustration of the power. They assumed 500 families with records on each of the father, mother, and four sibs on a trait with normally distributed genetic and environmental distributions. For an additive gene with frequency 0.1 and effect, expressed as the difference between homozygotes, of 0.5, 1.0, and 1.5 residual phenotypic standard deviations, the power of detecting its presence was, respectively, negligible, about one-half, and close to one.

More recently, the approaches of Elston and Stewart and of Morton and MacLean have been unified (Lalouel *et al.*, 1983) by including terms for both continuous genetic variation and non-Mendelian transmission probabilities. The mixed model was found to be sensitive to the distribution of the data (MacLean *et al.*,1975) and if skewed a spurious major gene might be suggested. To overcome this, Lalouel *et al.* (1983) suggest that a polygenic model and a major gene model should be first tested against a non-genetic model to confirm that there is a genetic component in the control of variation in the trait. Both of these models can then be compared with a mixed model. If there is evidence for both components, i.e. that the mixed model is a significant improvement over both the major gene and polygenic models, then the mixed model should be tested against a 'mixed model' in which the transmission probabilities are estimated. If skewness of the distribution of phenotypes was not caused by a major gene, this general model, although difficult to interpret genetically, should be a significant improvement over the mixed model and hence prevent the spurious detection of a major gene. The major gene model should also be tested against the 'major gene' model in which the transmission frequencies are not restricted.

A major problem is the very laborious computation of the likelihood for large pedigrees and considerable ingenuity has gone into designing efficient algorithms (for example, Cannings *et al.*, 1978; Lalouel and Morton, 1981). MacCluer *et al.* (1983) have compared the ability of different computer packages to determine the mode of inheritance. Although tests were based on only 40 nuclear families, the results obtained were in fairly good agreement, with some programs giving parameter estimates very similar to those simulated. As computing costs fall and methods of programming become more sophisticated, it is likely that the use of ML will become more widespread.

2.3 LIKELIHOODS

To illustrate the principles of segregation analysis a simple sire model, with balanced structure, will be used. All parents are assumed to be unrelated and randomly mated. A single trait is considered with one observation on each of the offspring. For the development of the likelihoods, fixed effects, such as herd or year, will be ignored, although the extensions to include these or more complex relationships are possible in theory.

The trait is assumed to be controlled by many unlinked loci, each with 2 alleles segregating at equal frequency and with equal effect, and an individual random environmental component. The aim is to see whether there is also an allele of large effect segregating against this background. Hence only two likelihoods will be considered, the polygenic and the mixed model.

The model to describe the data under a particular genotype for offspring j of sire i can be represented as:

$$y_{ij} = \mu + \mu_d + u_i + e_{ij}$$

Where: y_{ij} is the performance of the jth offspring of the ith sire.

μ is the overall population mean of the polygenic and environmental components.

d is the offspring major genotype, set to zero for the polygenic model.

μ_d is the effect of major genotype d (for polygenic model μ_0 equals zero).

u_i is the random effect for sire i (i.e. polygenic component) independent of μ_d; $u \sim N(0,\sigma_u^2)$

e_{ij} is the residual random effect for each individual, independent of u_i and μ_d; $e \sim N(0,\sigma_w^2)$

2.3.1 Polygenic likelihood

Under the polygenic model the likelihood for each individual is composed of two likelihoods:

i) The conditional likelihood of the offspring's phenotype given the sire's transmitting ability (u_i), denoted $k_0(y_{ij} \mid \mu, u_i, \sigma^2_w)$.

$$k_0(y_{ij} \mid \mu, u_i, \sigma^2_w) = \frac{1}{\sqrt{2\pi\sigma^2_w}} \exp\left[-\frac{1}{2\sigma^2_w}(y_{ij}-\mu-u_i)^2 \right]$$

Where exp[...] means e raised to the power of the function in brackets.

ii) The likelihood of the sire's transmitting ability (u_i), denoted $h(u_i)$.

$$h(u_i) = \frac{1}{\sqrt{2\pi\sigma^2_u}} \exp\left[-\frac{1}{2\sigma^2_u}(u_i)^2 \right]$$

Given the sire's genotype, the genotypes of the offspring are independent of each other and the conditional likelihood of the phenotypes of the n sibs is the product of the likelihood for each one. This likelihood is conditional on the sire's genotype (which is unknown and can take any value between minus infinity and plus infinity). Hence, the likelihood of the sibship is obtained by weighting the conditional likelihood of the sibship by the likelihood of the sire's transmitting ability, and integrating this function over all possible values of the transmitting ability. Sires were assumed to be unrelated, therefore the likelihood of the complete data is the product of the likelihood for each sibship. The following likelihood is obtained:

$$L(\text{poly}) = \prod_{i=1}^{s} \int_{-\infty}^{+\infty} h(u_i) \prod_{j=1}^{n} k_0(y_{ij} \mid \mu, u_i, \sigma^2_w) . du_i$$

Where: s is the number of sires.

n is the number of offspring per sire.

This likelihood can be integrated to give the following expression:

$$L(\text{poly}) = \prod_{i=1}^{s} \frac{1}{\sqrt{(2\pi)^n\left(\sigma_u^2\sigma_w^{2n}\left(\frac{n}{\sigma_w^2}+\frac{1}{\sigma_u^2}\right)\right)}}\exp\left[-\frac{1}{2}\left(\sum_{j=1}^{n}\frac{(y_{ij}-\mu)^2}{\sigma_w^2}-\left(\frac{\sum_{j=1}^{n}(y_{ij}-\mu)}{\sigma_w^2}\right)^2\left(\frac{n}{\sigma_w^2}+\frac{1}{\sigma_u^2}\right)^{-1}\right)\right]$$

In matrix notation, using the equivalent model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$$

Where: **y** is the vector of phenotypes for the offspring.

μ is the overall population mean of the polygenic and environmental components.

u is the vector of sires' transmitting ability.

e is the vector of individual random effects.

1 is a vector of 1s.

Z is a design matrix, linking offspring to their sires.

the likelihood is:

$$L(\text{poly}) = \frac{1}{(2\pi)^{sn/2}|\mathbf{V}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{1}\mu)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{1}\mu)\right] \qquad [2.1]$$

Where: **V** - variance (**y**) = **Z Z'** σ_u^2 + **I** σ_w^2

2.3.2 Mixed model likelihood.

The likelihood for the mixed model can be obtained in a similar way, including the probability of the major gene. For each offspring it is composed of the following:

i) The conditional likelihood of the offspring's phenotype given the sire's transmitting ability and the offspring's major genotype, denoted $k_d(y_{ij} \mid \mu,\mu_d,u_i,\sigma_w^2)$.

$$k_d(y_{ij} \mid \mu,\mu_d,u_i,\sigma_w^2) = \frac{1}{\sqrt{2\pi\sigma_w^2}}\exp\left[-\frac{1}{2\sigma_w^2}(y_{ij}-\mu-\mu_d-u_i)^2\right]$$

ii) The probability of the offspring's major genotype given the sire's major genotype, i.e. based on Mendelian transmission probabilities and the allele frequency in the dam, denoted trans(d|c), where c is the major genotype of the sire. These probabilities are shown in figure 2.1 and, for example, the transmission probability of a sire with genotype AA having an offspring with genotype Aa can be written trans(2 |1) and is equal to (1-p).

Figure 2.1 *Probability of the genotype of the offspring given the genotype of the sire, for a single locus.*

		Genotype of offspring		
		AA	Aa	aa
	AA	p	(1-p)	0
Genotype of sire	Aa	$\frac{p}{2}$	$\frac{1}{2}$	$\frac{(1-p)}{2}$
	aa	0	p	(1-p)

Assuming random mating and that the frequency of allele A in dams is p.

iii) The likelihood of the sire's genotype, which includes the transmitting ability denoted $h(u_i)$, as before, and now also the probability of his major genotype (p(c)), ignoring family information. That is, p(c) is the population genotype frequency for sires.

These likelihoods can be combined as for the polygenic model to obtain the mixed model likelihood, now, however the possible major genotype combinations for the offspring and sire need to be considered. The likelihood that the jth offspring has phenotype y_{ij} given the sire's major genotype is c and his polygenic contribution is u_i is:

$$\sum_{d=1}^{m} \text{trans}(d\,|\,c)\; k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)$$

In the same way as for the polygenic component, given the major genotype of the sire the offspring genotypes are independent and, hence, the conditional likelihood for the whole sibship is the product of this likelihood for each offspring. This likelihood is calculated for each major genotype of the sire, and for all possible values of the sire's transmitting ability, and weighted by the probabilities of these genotypes. Hence the following likelihood is obtained for the mixed model:

$$L(MM)=\prod_{i=1}^{s}\int_{-\infty}^{+\infty}\sum_{c=1}^{m}p(c)\,h(u_i)\prod_{j=1}^{n}\sum_{d=1}^{m}\mathrm{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)\,.\,du_i \qquad [2.2]$$

Where: m is the number of major genotypes, assumed to be equal to 3.

After rearrangement, this can be integrated and using matrix notation, gives the following likelihood for the mixed model:

$$L(MM)=\prod_{i=1}^{s}\frac{1}{(2\pi)^{n/2}|\mathbf{V}_i|^{1/2}}\sum_{c=1}^{m}\sum_{D=1}^{m^n}p(c)\,\mathrm{trans}(D\,|\,c)\exp\left[-\frac{1}{2}(\mathbf{y}_i-\mathbf{1}\mu-\mathbf{W}_D\mu_d)'\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{1}\mu-\mathbf{W}_D\mu_d)\right]$$

[2.3]

Where: $\mathbf{y}_i$ is the vector of offspring phenotypes for sire i.

$\mathbf{V}_i$ is the variance covariance matrix for the offspring of sire i.

D is one of the m^n major genotype combinations for the offspring of a sire.

trans(D|c) is the transmission probability for the sibship when their major genotypes are D, which equals the product of the transmission probability for each offspring.

$\mathbf{W}_D$ is an n x m matrix containing 1 or 0, depending on the genotype being considered for each offspring.

μ_d is the vector containing the mean effect of the major genotypes.

2.3.3 Restricted Maximum Likelihood

The parameter estimates obtained at the maximum of the likelihoods given above will be ML estimates. The variance estimates obtained in this way can be shown to be biassed (Shaw, 1987) as they do not take account of the degrees of freedom lost due to the use of the same data to estimate fixed effects. An alternative procedure, restricted maximum likelihood (REML), has been suggested (Patterson and Thompson, 1971) to correct for this by calculating the likelihood of a series of error contrasts (Harville, 1977). In this case the likelihood contains additional determinants (Searle, 1979). One based on the fixed effects structure and the other on the variance-covariance matrix for fixed effects. For a polygenic model the likelihood can be written as:

$$L(\mathrm{poly})=\frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{(2\pi)^{(sn-t)/2}|\mathbf{V}|^{1/2}|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta)\right]$$

Where: $\mathbf{X}$ is the design matrix for fixed effects

β is the vector of fixed effects

t is the rank of $\mathbf{X'X}$

$\mathbf{X'X}$ is constant for a given fixed effect design and does not alter the parameter estimates. $\mathbf{X'V^{-1}X}$ effects the equation for the variance estimates, reducing the degrees of freedom associated with the sums of squares and hence will give different parameter estimates compared with ML. The variance estimates are now unbiassed. However, differences between the REML likelihoods cannot be used as a measure of goodness of fit when comparing different fixed effect models, only when comparing different random effect models with constant fixed effect structure (R. Thompson, personal communication). In the comparison of the mixed model and polygenic model, although the major genotypes are not fixed effects in the usual sense they are effectively treated as such and, hence, a model with one fixed effect, the mean, is being compared with a model with three fixed effects, the three genotype means, hence the REML likelihoods obtained will not be comparable. Further investigation is required to overcome this problem.

2.4 TEST STATISTIC

A test statistic is provided by twice the difference between the natural logarithms of the MLs under the mixed model and the polygenic model. This test statistic is expected asymptotically to follow a χ^2 distribution with degrees of freedom (d.f.) equal to the number of parameters fixed under the polygenic model (poly) but estimated under the mixed model (MM) (Wilks, 1938).

$$\text{Test statistic} = 2\ (\ \ln L\ (MM)\ -\ \ln L\ (poly))\ \sim\ \chi^2 \qquad [2.4]$$

Wilks obtained this distribution for the comparison of a general likelihood with one where some of the parameters were fixed, and showed it was approximate, ignoring terms of the order $\frac{1}{\sqrt{N}}$, where N was the number of observations. Several assumptions were made in deriving this distribution for the test statistic and there has been little work to test whether these are valid in the comparison of the mixed and polygenic models. However, in the related area of mixture models, where the aim is generally to identify the number of component distributions in a mixture of such distributions, there is evidence that this sampling distribution for the test statistic is inappropriate (McLachlan and

Basford, 1987; Titterington *et al.*, 1985). Under the null hypothesis the mixing proportions lie on the boundary of the parameter space. For example, when testing whether a distribution is composed of two normal distributions, one with mean μ_1 and variance σ_1^2 and the other with μ_2 and σ_2^2 and the proportion in the first distribution p, or composed of one normal distribution, the null hypothesis of one distribution can be described using the full model and fixing p at 1, a boundary. With this restriction μ_2 and σ_2^2 cannot be estimated. Alternatively if the means and variances of the two distributions are constrained to be equal then it is impossible to estimate the mixing proportion. Hence, although the hypothesis of two normal distributions estimates three parameters more than the null hypothesis, the comparison can be made enforcing only one restriction, that p is equal to zero or 1.

On the basis of a simulation study Wolfe (1971) suggested the following modified test statistic to be used as a guideline when testing the number of component distributions:

$$-\frac{2}{N}\left(N-1-v-\frac{r'}{2}\right)\ln\left(\frac{Lr}{Lr'}\right)\sim\chi^2_{2v(r'-r)\,df}$$

Where: N is the sample size.

v is the number of variables.

r and r' are the number of component distributions under the two hypotheses; r' > r.

Lr is the likelihood under model r.

Further investigations of this criterion (e.g. Everitt, 1981) confirm that although this distribution may be appropriate in certain circumstances, it is not generally applicable and further work is required to ascertain the true distribution of the test statistic.

The interpretation of this work in the context of segregation analysis is not clear, although it is obvious that care should be taken before assuming that the distribution of the test statistic will be χ^2. When comparing the likelihoods from the mixed and polygenic models an hypothesis with the distribution composed of three component distributions (r'=3) is being compared with an hypothesis of just one distribution (r=1). Following Wolfe's modification the relevant χ^2 distribution would have four d.f. (2(r'-r)). However, unlike the mixture model situation, when considering the mixed model there is information about the relative proportions of the three component distributions. For example, if Hardy-Weinberg equilibrium is assumed, then the frequencies of the distributions can be explained by a single parameter and the proportions will be p^2, 2p(1-p) and $(1-p)^2$.

2.5 GENOTYPING AT THE MAJOR LOCUS

In order to utilise the knowledge that an allele with large effect on the trait of interest is segregating in the population, identification of the genotype of each individual at that locus is required. Mating schemes could then be designed to optimise improvement.

The genotypes of animals can not be determined with complete certainty but the most probable genotype can be obtained assuming that the mode of inheritance is known and that all parameters in the likelihood are known. General expressions for genotyping individuals have been considered by Elston and Stewart (1971) and an iterative scheme by Arendonk *et al.* (1988). However the possibility of a trait being controlled by both a major gene and polygenic component was ignored. Essentially, to obtain the probability of an individual being a particular genotype the ratio of the likelihood of the pedigree assuming that the individual is that genotype to the total likelihood for the pedigree is obtained.

For the sire model being considered, extension to include a polygenic component is straightforward (Elsen *et al.*, 1988). Sires are assumed to be independent of each other and, hence, the conditional probability of each genotype for each sire is dependent on the phenotypes of his half-sib offspring only. Using the notation described already (section 2.2) and assuming that the trait is controlled both by a major gene and a polygenic component, the conditional probability for sire i having genotype c is :

$$q_i(c) = \frac{p(c) \int_{-\infty}^{+\infty} h(u_i) \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d \mid c)\, k_d(y_{ij} \mid \mu, \mu_d, u_i, \sigma_w^2)\, .\, du_i}{\sum_{c'=1}^{m} p(c') \int_{-\infty}^{+\infty} h(u_i) \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d \mid c')\, k_d(y_{ij} \mid \mu, \mu_d, u_i, \sigma_w^2)\, .\, du_i} \qquad [2.5]$$

In practice the parameters required to describe the effect and frequency of the alleles at the major locus and the polygenic and residual variances will be replaced by their ML estimates.

2.6 DISCUSSION

The likelihoods have been derived for the polygenic and mixed models assuming a simple pedigree structure. With this structure, it is possible to write these in a form enabling exact calculation of the likelihood. However, the exact mixed model likelihood involves a summation for each combination of major genotypes for the pedigree, hence as the number of offspring increases calculation of the likelihood soon becomes infeasible, i.e. with 5 offspring, 3^6 (729) summations are required and with 10 offspring 3^{11} (177147) summations.

Although segregation analysis is appealing, having a general application and enabling both the testing of hypotheses and obtaining estimates of the parameters involved, the mixed model likelihood would be impossible to calculate in most animal breeding situations. Hence in order to be able to use this method approximations to the mixed model will be required.

CHAPTER 3

APPROXIMATION 1 - PEDIGREE ANALYSIS PACKAGE

3.1 INTRODUCTION

The Pedigree Analysis Package (PAP) (version 2.0) (Hasstedt and Cartwright, 1981) is a Fortran computer package which can calculate and maximise the likelihood of the data under many different genetic models for pedigrees of any complexity. It is based on segregation analysis described by Elston and Stewart (1971) and has been written primarily for human geneticists.

The package is composed of a 'driver' and many subroutines. The user incorporates subroutines relevant to their requirements and hence a large choice of possible genetic models exists. For example, to calculate the likelihood under a major gene model subroutines need to be combined to describe the frequency of the major genotypes in the founder population, the transmission of alleles from parent to offspring and the penetrance (or the probability of the phenotype given the genotype). In the same way more complicated models can be obtained, including the possibility of incorporating a polygenic component, having linkage to a marker gene, sex linked traits or age dependent traits. Once the subroutines have been assembled there is still some flexibility as to the model, for example enforcing dominance or estimating the transmission probabilities.

Three sources of information are required to run PAP: a file containing the phenotypes, another file containing the pedigree information and the last the model. The phenotype file contains a list of identification numbers, each specific to an individual within each pedigree, and associated information on the sex of the individual, the phenotype for the traits to be analysed and if relevant the genotype at any marker loci of interest. The pedigree file gives the pedigree in the form of nuclear families indicating which members of the family also appear in subsequent families. The final information specifies the genetic model, giving initial parameter estimates for the maximisation process, indicating which parameters are to be estimated, and giving bounds within which the parameter estimates have to lie. It is also possible to specify, for example, dominance, by enforcing two parameters to take the same value.

When running the program several options are available: to calculate the likelihood for a set of parameters at fixed values, to maximise the likelihood from a given starting values, to grid the likelihood with respect to one or two parameters or to calculate the risk that an individual has a particular genotype given the model. The last option is of importance in genetic counselling especially as risks can be obtained for individuals which are not yet born. A recent revision of the package (version 3.0) (Hasstedt, 1989) is more flexible, allowing the analysis of dichotomous traits with a continuous underlying genetic component, and additional run time options, for example the calculation of standard errors for given maximum likelihood parameter estimates.

The program has been written so that any complexity of pedigree can be analysed, including pedigrees containing loops, for example when the two parents of an individual are related within the pedigree, i.e. the individual is inbred (other types of loops are described by Cannings *et al.*, 1978). The likelihood is calculated using a method called peeling. The speed of calculation of the likelihood depends on the complexity of the pedigree and on the order in which the individuals within a pedigree are considered. The onus is on the user to supply an efficient ordering, and algorithms have been designed to obtain optimum orders (for example, Thomas, 1986b).

PAP contains the minimisation routine GEMINI (Lalouel, 1979). This is a quasi-Newton routine that retains generality by estimating the 1st and 2nd derivatives by finite difference, hence requiring only the calculation of the likelihood value for specified parameter combinations. It is possible to constrain the parameters using bounds. Maximisation can be slow because the parameters being estimated are on different scales; for example the frequencies are constrained between zero and one while the means can take any value. For efficient maximisation the parameters need to be scaled so that one unit change in the parameter value causes one unit change in the log likelihood. The new version of PAP (3.0) incorporates the possibility of scaling the parameters by adding and multiplying by constants, however more sophisticated reparameterisations cannot be included in order to retain flexibility, for example to enable the option of any number of major genotypes, the frequency of each genotype cannot easily be constrained to be between zero and one.

The new version (3.0) also enables preliminary transformation of the data (MacLean *et al.*, 1976) to one or more normal distributions.

Fixed effects cannot be incorporated in the model but have to be estimated and the data adjusted prior to the segregation analysis, which is a disadvantage for animal breeders.

A brief description of the algorithm used to calculate the polygenic and mixed model likelihoods in the PAP package is given below and the package has been used to analyse some data of α-glucosidase activity in cattle to look for evidence for a major gene.

3.2 DESCRIPTION OF POLYGENIC AND MIXED MODEL LIKELIHOOD CALCULATIONS

Two of the models under which the likelihood of the data can be calculated using PAP are the polygenic and the mixed model. These two models require the same subroutines and the polygenic model is obtained by setting the number of major genotypes to one. In version 2.0 of the PAP package (Hasstedt and Cartwright, 1981) two mixed models are offered, an exact calculation suitable for pedigrees with up to 10 members, and an approximation (Hasstedt, 1982). In both cases the corresponding polygenic likelihood is exact and the algorithm for its calculation similar.

3.2.1 Polygenic model

Classically, in the field of human genetics, pedigrees have been considered in the form of nuclear families, that is, parents and their full-sib offspring. However, this ignores many useful relationships. PAP makes use of this additional information, including relationships between the nuclear families. Any relationship can be considered, leading to pedigrees of any complexity. In animal breeding terminology this would be an animal model.

$$\mathbf{y} = \mathbf{X}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}$$

Where: $\mathbf{X}$ is a design matrix for fixed effects

μ is a vector of fixed effect class means

$\mathbf{Z}$ is a design matrix for random genetic effects

$\mathbf{g}$ is a vector of additive polygenic breeding values; $\mathbf{g}\sim N(0,\sigma_g^2)$

$\mathbf{e}$ is an individual random effect; $\mathbf{e}\sim N(0,\sigma_e^2)$

In PAP $\mathbf{X}$ is treated as a vector of 1s ($\mathbf{1}$) and μ a scalar equal to the population mean (μ)

Following equation [2.1], and assuming that all N individuals have phenotypic information, the likelihood can be written as follows:

$$\ln L(\text{poly}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{1}\mu)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{1}\mu) \qquad [3.1]$$

Where: $\mathbf{Z}$ is now an identity matrix ($\mathbf{I}$)

$\mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$

σ_g^2 is the genetic variance

σ_e^2 is the environmental variance

A is the additive genetic relationship matrix

The inverse of the variance matrix is:

$$\mathbf{V}^{-1} = \frac{\mathbf{I}}{\sigma_e^2} - \frac{\mathbf{I}}{\sigma_e^2}\left(\frac{\mathbf{A}^{-1}}{\sigma_g^2} + \frac{\mathbf{I}}{\sigma_e^2}\right)^{-1}\frac{\mathbf{I}}{\sigma_e^2}$$

Hence the likelihood can be written containing a part that is dependent only on the individual being considered: $-\frac{1}{2}(\mathbf{y}-\mathbf{1}\mu)'\,\frac{\mathbf{I}}{\sigma_e^2}\,(\mathbf{y}-\mathbf{1}\mu)$ and a part that takes into account the relationships with other individuals in the pedigree: $\frac{1}{2}\,(\mathbf{y}-\mathbf{1}\mu)'\,\frac{\mathbf{I}}{\sigma_e^2}\left(\frac{\mathbf{A}^{-1}}{\sigma_g^2} + \frac{\mathbf{I}}{\sigma_e^2}\right)^{-1}\frac{\mathbf{I}}{\sigma_e^2}\,(\mathbf{y}-\mathbf{1}\mu)$.

The calculations involving all related individuals are obviously computationally demanding requiring the inverse of two matrices of the order of the number of individuals. Fortunately, Henderson (1976) showed that the inverse of the relationship matrix can be obtained directly without the need to invert **A**. However the inversion of $\left(\frac{\mathbf{A}^{-1}}{\sigma_g^2} + \frac{\mathbf{I}}{\sigma_e^2}\right)$ is still required. PAP overcomes the need to invert the whole matrix using a method called peeling.

Peeling

Elston and Stewart (1971) suggested a recursive method for the calculation of probabilities on pedigrees that successively reduces the size of the pedigree by collapsing information from offspring onto their parents. Cannings *et al.* (1976 and 1978) extended this idea and described a general method which allowed collapsing of information both 'upwards' onto parents and 'downwards' onto offspring. This process of collapsing information was called 'peeling' by Cannings *et al.* (1976). In essence, the method involves defining individuals or sets of individuals, called the cutset, that split the pedigree into at least two parts, so that the parts are connected only through the cutset. A sequence of such sets is required so that one of the parts always contains all the peeled animals and, eventually, the whole pedigree is within this set. Given the genotype of the cutset members, the two parts of the pedigree, peeled and unpeeled, are independent. Hence the probability of all genetic and genealogical information in the peeled set conditional on (if parents have not yet been peeled) or joint with (if parents are peeled) genotypes of the individuals in the cutset can be accumulated, by considering the recently peeled individuals. The possibility of peeling information onto a set of individuals jointly allows for pedigrees of arbitrary complexity to be considered.

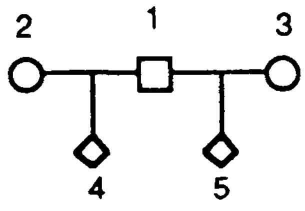
The most common means of implementing this method is to consider one complete marriage in one operation, i.e. considering nuclear families in turn. Then the order of marriages specifies the cutset sequence. However, this may not be the optimum peeling sequence and Thomas (1986 a and b) considers methods of obtaining this optimum sequence for complex pedigrees by including the possibility of having several independent components to the pedigree with the cutset removed.

For pedigrees without loops the peeling process is relatively simple. Under these conditions it is possible to consider the nuclear families in an order so that there is only one member of each nuclear family that will also be present in a family not yet considered, i.e there is only ever one cutset member to consider and, hence, only one probability function. With loops in the pedigree the whole process becomes much more complex and time consuming, with several cutset members required.

Example

To illustrate the process a simple pedigree will be considered with two nuclear families connected by a common sire.

Pedigree:



This pedigree can be decomposed into the following nuclear families:

Sire	Dam	Offspring
1	2	4
1	3	5

With this example it is obvious that, by 'removing' the sire to the cutset, two discrete pedigrees are produced and hence, given the genotype of the sire the two nuclear families are independent. Information from the first nuclear family can be peeled onto the sire, at the same time calculating the contribution to the likelihood of this family independent of the sire. Then the second family can be considered, in the same way as for the first family, with the sire containing additional information from the first family.

Multiplying $\left(\frac{\mathbf{A}^{-1}}{\sigma_g^2} + \frac{\mathbf{I}}{\sigma_e^2}\right)^{-1}$ by σ_g^{-2} gives $(\mathbf{A}^{-1} + k)^{-1}$ where k is $\frac{\sigma_g^2}{\sigma_e^2}$, which is a more convenient structure to consider and will be denoted, **c**.

When constructing **c** all individuals are considered either to be a founder or an offspring, and if an offspring they are assumed to have two known parents. This can easily be obtained by incorporating a parent with unknown phenotype if the parent is missing. By considering nuclear families the genetic relationship matrix is always relatively simple to create. The matrix for the first nuclear family is:

$$\mathbf{c} = \begin{pmatrix} 1.5+k & 0.5 & -1 \\ 0.5 & 1.5+k & -1 \\ -1 & -1 & 2+k \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{14} \\ c_{21} & c_{22} & c_{24} \\ c_{41} & c_{42} & c_{44} \end{pmatrix}$$

The vector containing the phenotypes for this nuclear family can be written as follows:

$$\mathbf{b} = \begin{pmatrix} y_1 - \mu \\ y_2 - \mu \\ y_4 - \mu \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_4 \end{pmatrix}$$

Hence the data part of the likelihood for the first nuclear family can be written as:

$$-\frac{1}{2\sigma_e^2}\mathbf{b'b} + \frac{k}{2\sigma_e^2}\mathbf{b'c^{-1}b}$$

First, the pedigree can be reduced by peeling individual 4. The contribution to the likelihood from his own phenotype, ignoring all relatives, is calculated as:

$$-\frac{1}{2\sigma_e^2}(b_4)^2 + \frac{k}{2\sigma_e^2}(b_4)^2(c_{44})^{-1}$$

However, there are also terms which are functions of both individual 4 and his parents. This information can be transferred to the parents, reducing the size of the matrix. Giving the following matrices:

$$\begin{pmatrix} c_{11} - \frac{c_{14}\,c_{41}}{c_{44}} & c_{12} - \frac{c_{14}\,c_{42}}{c_{44}} \\ c_{21} - \frac{c_{24}\,c_{41}}{c_{44}} & c_{22} - \frac{c_{24}\,c_{42}}{c_{44}} \end{pmatrix} = \begin{pmatrix} c_{11}^{*} & c_{12}^{*} \\ c_{21}^{*} & c_{22}^{*} \end{pmatrix} = \mathbf{c}^{*}$$

$$\begin{pmatrix} b_1 - b_4\dfrac{c_{14}}{c_{44}} \\ \\ b_2 - b_4\dfrac{c_{24}}{c_{44}} \end{pmatrix} \quad = \quad \begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix} \quad = \quad \mathbf{b}^*$$

i.e. the original likelihood can be rewritten as

$$-\frac{1}{2\sigma_e^2}\mathbf{b'b} + \frac{k}{2\sigma_e^2}\mathbf{b'c^{-1}b} \quad = \quad -\frac{1}{2\sigma_e^2}(b_4)^2 + \frac{k}{2\sigma_e^2}(b_4)^2(c_{44})^{-1} - \frac{1}{2\sigma_e^2}\mathbf{b}^{*\prime}\mathbf{b}^* + \frac{k}{2\sigma_e^2}\mathbf{b}^{*\prime}\mathbf{c}^{*-1}\mathbf{b}^*$$

After any further offspring have been peeled the contribution from the dam can be calculated. In the same way as for the offspring, there is a contribution that is independent of the remaining members of the pedigree (the sire) which can be added to the likelihood already calculated from the offspring:

$$-\frac{1}{2\sigma_e^2}(b_2^*)^2 + \frac{k}{2\sigma_e^2}(b_2^*)^2(c_{22}^*)^{-1}$$

and a joint contribution from the sire and dam which can be collapsed onto the sire as follows:

$$\left(c_{11}^* - \frac{c_{12}^* c_{21}^*}{c_{22}^*}\right) \quad = \quad c_{11}^{**}$$

$$\left(b_1^* - b_2^*\frac{c_{12}^*}{c_{22}^*}\right) \quad = \quad b_1^{**}$$

Hence all the information on the first nuclear family has either been incorporated into the likelihood function (from dam and offspring) or anything relating to the sire is stored in c_{11}^{**} and b_1^{**}. Now the second nuclear family can be considered, the matrices will look similar to those set up for the first family except that the sire will already have information relating to his first nuclear family.

$$\mathbf{c} = \begin{pmatrix} c_{11}^{**}+0.5 & 0.5 & -1 \\ 0.5 & 1.5+k & -1 \\ -1 & -1 & 2+k \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} b_1^{**} \\ y_3 - \mu \\ y_5 - \mu \end{pmatrix}$$

As for the first family, the information from the offspring and dam can be peeled onto the sire. Hence, after both families have been considered all relevant information on the sire has been obtained and the contribution of the likelihood due to him (and related terms) can be calculated:

$$-\frac{1}{2\sigma_e^2}(b_1^\circ)^2 + \frac{k}{2\sigma_e^2}(b_1^\circ)^2(c_{11}^\circ)^{-1}$$

where $^\circ$ indicates that all information from both families has been peeled.

For a more complex pedigree the procedure is the same although information will have to be stored on additional individuals.

The determinant required is a function of **c** and can be calculated at the same time as peeling information to calculate the data part of the likelihood. Partitioning the matrix into four sub-matrices gives the following expression for the determinant:

$$|\mathbf{c}| = \begin{vmatrix} \mathbf{D}\mathbf{E} \\ \mathbf{E}'\mathbf{F} \end{vmatrix} = |\mathbf{F}|\,|\mathbf{D} - \mathbf{E}\,\mathbf{F}^{-1}\mathbf{E}'|$$

If **F** is a single element, | **F** | is equal to the value of the element **F** and | $\mathbf{D} - \mathbf{E}\,\mathbf{F}^{-1}\mathbf{E}'$ | is the matrix $\mathbf{c}^*$ after **F** has been peeled. Hence, repeatedly using this equality, with **F** corresponding to each individual in the pedigree in turn, gives an expression for the determinant in terms of the product of the diagonals of matrix **c** after information has been peeled.

3.2.2 Mixed models

Exact likelihood

The form of the mixed model likelihood calculated is given in equation [2.3] with **V** as given in [3.1]. In effect the procedure described above is repeated for each possible major genotype combination for the pedigree. For each individual μ is the effect of the relevant major genotype for the combination being considered. There are many combinations and hence the likelihood calculation takes a long time.

Approximate likelihood

In the approximation (Hasstedt, 1982), rather than setting up the matrices for each combination of major genotypes as described above, the matrices are used just once with a weighted mean for each individual instead of the major genotype mean. A correction is made for the use of this weighted mean for each combination of major genotypes.

The mixed model for a particular genotype, c, for individual i can be written as:

$$y_i = \mu_c + \frac{g_s + g_d}{2} + z_i + e_i$$

Where: μ_c is the effect of major genotype c.

g_s is the polygenic contribution to the genotype of the sire of individual i (i.e. the sire's polygenic breeding value).

g_d is the polygenic contribution to the genotype of the dam of individual i.

z_i is the Mendelian sampling component for i; $z \sim N\left(0, \frac{\sigma_g^2}{2}\right)$

e_i is the individual random effect; $e \sim N\left(0, \sigma_e^2\right)$

Note that $g_i \;=\; \frac{g_s + g_d}{2} + z_i$

With the model written in this form the likelihood for individual i with sire s and dam d can be written as follows:

$$L(MM) = p(c_s)\,p(c_d)\,\mathrm{trans}(c_i \,|\, c_s, c_d)\,\frac{1}{\sqrt{2\pi\sigma_e^2}}\,\frac{1}{\sqrt{\pi\sigma_g^2}}$$

$$\int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left((y_i - \mu_c - g_i)^2\frac{1}{\sigma_e^2} + \left(g_i - \left(\frac{g_s + g_d}{2}\right)\right)^2 \frac{1}{\sigma_g^2/2}\right)\right].dg_i$$

Integrating over g_i and substituting

$$(y_i - \mu_c)(g_s + g_d) \qquad \text{with} \qquad (y_i - \hat{\mu}_i)(g_s + g_d) + (\hat{\mu}_i - \mu_c)(\hat{g}_s + \hat{g}_d)$$

gives the following expression for the mixed model likelihood:

$$L(MM) = p(c_s)\,p(c_d)\,\mathrm{trans}(c_i \mid c_s, c_d)\,\frac{1}{\sqrt{(2\pi)\left(\sigma_e^2 \frac{\sigma_g^2}{2}\,\frac{(k+2)}{\sigma_g^2}\right)}} \qquad [A]$$

$$\exp\left[-\frac{1}{2}\left(-(y_i-\hat{\mu}_i)^2 \frac{\sigma_g^2}{(k+2)\,\sigma_e^4}\right.\right. \qquad [A]$$

$$-(g_s+g_d)^2 \frac{1}{(k+2)\,\sigma_g^2} - 2(y_i-\hat{\mu}_i)(g_s+g_d)\,\frac{1}{(k+2)\,\sigma_e^2} + (g_s+g_d)^2 \frac{1}{2\sigma_g^2} \qquad [B]$$

$$+(y_i-\mu_c)^2 \frac{1}{\sigma_e^2} - 2(y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_c)\frac{\sigma_g^2}{(k+2)\,\sigma_e^4} + (\hat{\mu}_i-\mu_c)^2 \frac{\sigma_g^2}{(k+2)\,\sigma_e^4} \qquad [C]$$

$$\left.\left.-2(\hat{\mu}_i-\mu_c)(\hat{g}_s+\hat{g}_d)\,\frac{1}{(k+2)\,\sigma_e^2}\right)\right] \qquad [C]$$

Where: A is the individual contribution as described for the polygenic model calculated using the weighted mean, $\hat{\mu}_i$.

B is the contribution from the sire and dam calculated implicitly in the polygenic model.

C is the adjustment for the use of the approximate mean in A and the individual contribution using the correct major genotype.

Whereas A and B are the same for all major genotypes and hence need only be calculated once, in the same way as for the polygenic model described above, C needs to be calculated for every genotype combination of parents and their full-sib offspring. However while working through these combinations the genotype of the parents is specified and hence the offspring can be treated independently. This reduces the number of calculations required, especially for large families. For example, for a nuclear family there are 9 possible parental genotype combinations, which, if they have one offspring result in a total of 15 different combinations of major genotype for the family. With two offspring, if treated independently, there would be 15 combinations involving each offspring. The combinations for the family treated as a whole are given below.

Treating offspring independently gives 15n combinations each with 3 individuals, if they are not treated independently $4+4(2^n)+3^n$ combinations of n are required. The second option very soon becomes infeasible with increasing n, even if, as in this example an efficient algorithm excluding all impossible combinations is used.

Parental genotypes		Number of possible combinations 1 offspring	 2 offspring	 n offspring
1	1	1	1	1
1	2	2	2^2	2^n
1	3	1	1	1
2	1	2	2^2	2^n
2	2	3	3^2	3^n
2	3	2	2^2	2^n
3	1	1	1	1
3	2	2	2^2	2^n
3	3	1	1	1
		15	29	$4+4(2^n) + 3^n$

The accuracy of the approximation depends on the values for $\hat{g}_s$, $\hat{g}_d$ and $\hat{\mu}_i$ used. Obviously if $\hat{g}_s = g_s$ and $\hat{g}_d = g_d$ the approximation would be exact. However the values of g_s and g_d are not known and have to be estimated.

The best estimates for $\hat{g}_s$ and $\hat{g}_d$ are BLUP estimates obtained using the relevant major genotype mean to adjust the data. The BLUP equations can be written as follows:

$$(\mathbf{A}^{-1}k^{-1} + \mathbf{I})\ (\mathbf{g}) = (\mathbf{y} - \boldsymbol{\mu})$$

where: $\mathbf{A}$, $\mathbf{I}$, $\mathbf{g}$, $\mathbf{y}$, $\boldsymbol{\mu}$ contain all relatives of the sire and dam.
$\boldsymbol{\mu}$ is a vector of major genotype means

However, estimates are only required for the sire and dam of the nuclear family being considered ($\hat{g}_s$ and $\hat{g}_d$), hence the information from relatives can be absorbed into the equations for these two animals, so that only two equations need to be solved simultaneously. The peeling process described above has already accumulated terms similar to these, containing information on all related individuals already peeled, and stored them in matrices **c** and **b**. The problem is that rather than the data being adjusted for the major genotype being considered for that individual a weighted mean has been used ($\hat{\mu}_i$). A correction for the use of the incorrect mean is required.

At any one time, a major genotype is only being considered for three individuals, the sire, dam and one offspring, and the equations for $\hat{g}_s$ and $\hat{g}_d$ can be corrected for the major genotype of these individuals. However, the expected frequency of the major genotype for any other full-sibs will be altered depending on the genotypes being considered for parents. Therefore, rather than using the weighted mean calculated previously for these full-sibs, a new weighted mean can be calculated which weights the genotype means by their expected frequency given the parental genotypes:

$$\sum_{c=1}^{m} \mu_c \, \text{trans}(c \mid c_s, c_d)$$

For all other individuals no correction is made, which is equivalent to using the original weighted mean calculated for these individuals.

Considering the pedigree described before, with two nuclear families with a common father, when calculating the likelihood contribution from the first nuclear family it is only known that the sire has one offspring, hence only these three individuals are involved when calculating $\hat{g}_s$ and $\hat{g}_d$. For the sire the information from the offspring and dam have already been absorbed into the sire and hence $\hat{g}_s$ can be calculated as:

$$\hat{g}_s = b_1^{**} (c_{11}^{**})^{-1}$$

However b_1^{**} contains $\hat{\mu}_i$ for the three members of the family :

$$b_1^{**} = (y_1 - \hat{\mu}_1) - \left((y_2 - \hat{\mu}_2) - (y_4 - \hat{\mu}_4) \frac{c_{24}}{c_{44}}\right) \frac{c_{12}}{c_{22}} - (y_4 - \hat{\mu}_4) \frac{c_{14}}{c_{44}}$$

this can be corrected for a particular major genotype combination as follows when considering offspring 4:

$$b_1 = b_1^{**} + (\hat{\mu}_1 - \mu_s) + \left((\hat{\mu}_2 - \mu_d) - (\hat{\mu}_4 - \mu_i) \frac{c_{24}}{c_{44}}\right) \frac{c_{12}}{c_{22}} + (\hat{\mu}_4 - \mu_i) \frac{c_{14}}{c_{44}}$$

Where: μ_s, μ_d, μ_i are the relevant major genotype means.

In the approximation by Hasstedt (1982), instead of completely replacing the weighted mean with the correct value, half the difference between the two is used:

$$\left(y - \frac{\hat{\mu}_i - \mu_c}{2}\right)$$

A similar equation can be derived for the dam using the relevant parts of **c** after the information from the offspring has been absorbed:

$$\hat{g}_d = (b_2^{\bullet} - c_{21}^{\bullet}\, \hat{g}_s)\, c_{22}^{\bullet\,-1}$$

As before, $b_2^{\bullet}$ can be corrected for the use of the weighted mean for the dam and offspring.

In this way estimates for the polygenic component of the sire and dam are estimated under the hypothesis of different major genotype combinations and C calculated.

When the next nuclear family is considered the sire already has some information from the first nuclear family, which is included in the calculations of his breeding value but not adjusted for the major genotype.

The value of the approximate mean also affects the likelihood value. The equation used to calculate this mean in PAP is as follows:

$$\hat{\mu}_i = \sum_{c=1}^{m} \mu_c P_i(c)$$

Where: $P_i(c)$ is the probability of genotype c for individual i, based on pedigree information already accumulated, which can be written as:

$$P_i(c) = \frac{p(c)\, \exp\left[-\frac{(y_i-\mu_c)^2}{2\sigma_e^2}\right] p(c|I_i)}{\sum_{c'=1}^{m} p(c')\, \exp\left[-\frac{(y_i-\mu_{c'})^2}{2\sigma_e^2}\right] p(c'|I_i)}$$

Where: $p(c|I_i)$ is the probability of individual i having genotype c given all the information from individuals already peeled. For example, when individuals are first encountered there is no information on their major genotype and hence for parents the population frequencies are used for these probabilities and for offspring the probabilities are set to one. For an individual who has already been considered in another nuclear family, the probability of being each major genotype given the information from all relatives already peeled will have been stored and this is used.

The determinant can be calculated in the same way as for the polygenic model, as this is not dependent on the data and hence not on the major genotypes.

3.2.3 Discussion

The likelihood of the data under a polygenic model is calculated as written in [3.1] and hence the inverse of the variance matrix, **V**, is required. The procedure described allows any complexity of pedigree to be analysed without the need to set up this complete matrix. The mixed model approximation follows the same procedure and hence, effectively retains the integration but approximates the crossproducts between the major genotype and polygenic components.

Algorithms for the calculation of the likelihood under a polygenic model have been considered in animal breeding situations (for example, Graser *et al.*, 1987). The pedigree is not usually partitioned into nuclear families but considered as a whole. However, when computing the likelihood use of the knowledge of the structure of the matrix and sparse matrix techniques are used to reduce the computation and storage requirements. The method of Graser *et al.* (1987), which has been implemented (and extended) in programmes such as DFREML (Meyer, 1988), involves rewriting the likelihood, giving several determinants and a quadratic in terms of the phenotypes, and then making use of extended mixed model (random and fixed effects) matrices including the data (Meyer, 1989). The use of Gaussian elimination on this extended matrix, which is the procedure described previously to remove rows and columns of a matrix, results in the required quadratic of the data and the possible accumulation of the determinants (Meyer, 1989). Unlike the approximation in PAP, this method is easy to extend to incorporate fixed effects or additional random components such as common family environment. However, it cannot be used for the mixed model as this would involve the calculation required for the polygenic model to be repeated for each major genotype combination for the pedigree.

3.3 ANALYSIS OF α-GLUCOSIDASE ACTIVITY IN CATTLE

3.3.1 Introduction

A deficiency of the lysozomal α-glucosidase causes excessive tissue accumulation of glycogen and a lack of glucose. The disease associated with this trait is generalised glycogenosis type II or Pompe's disease and is known to affect humans as well as cattle (Howell *et al.*, 1981). Affected animals are clinically normal at birth, although have decreased α-glucosidase activity and excessive glycogen deposition, but fail to grow as rapidly as their contemporaries and eventually die. Two clinical forms of the disease have been observed in cattle (Howell *et al.*, 1981) One is similar to the infantile onset seen in children, where the animals die within the first months of life usually due to cardiac failure. In the second form, more closely related to the childhood form observed in humans, the affected animals remain clinically normal for about 9 months before showing gradual loss of condition and muscular weakness and eventually require to be slaughtered by 16 months of age.

The condition is thought to be controlled by a recessive allele at a single autosomal locus, with additive effect (Howell *et al.*, 1981, Healy *et al.*, 1987). The two forms of the disease have been observed within one family and are thought to be related and possibly caused by the same genetic lesion with variation in clinical expression.

The aim of this study was to determine the mode of inheritance of the disease using segregation analysis on data kindly provided by C.P. McPhee.

3.3.2 Description of the data

The data consisted of mononuclear blood cell enzyme activities recorded for α-glucosidase and two reference enzymes, β-galactosidase and hexosaminidase (methods of enzyme analysis are given in Healy, 1982), on individuals in one herd of Brahman cattle in Australia taking part in a programme for the control of Pompe's disease. The data contained the identities of the sire and dam of each individual, but not necessarily other information on these parents. The enzyme activities had been recorded over several overlapping generations, and contained records of dams being mated more than once and offspring also appearing as parents.

Originally, unlike α-glucosidase, variation in the activity of the reference enzymes was not thought to be controlled genetically within this population but dependent on the environment. Correlations between these three enzymes exist which must, therefore, be caused by the environment. From analyses of 180 clinically normal dairy cattle the following multiple regression equation was obtained to predict the activity of α-glucosidase based on the activities of the reference enzymes:

$$\text{predicted } \alpha\text{-glucosidase} = 0.03496 + 0.09585 \text{ GAL} + 0.003187 \text{ HEX}$$

Where GAL and HEX are the activities of β-galactosidase and hexosaminidase, respectively. (Reichmann *et al.*, 1987).

The standard procedure has been to express observed α-glucosidase activity as a proportion of the predicted activity. Animals with less than 70% observed activity are assumed to be heterozygotes (Reichmann *et al.*, 1987). However, there is no rational basis for this procedure and the use of percentages leads to a trait which is not easily amenable to analysis, especially as the distribution is expected to be non-normal. In this situation it would seem more logical to subtract the predicted value from that observed in order to remove some of the (supposedly) environmentally caused variation. This trait, DEV, would be expected to be distributed more closely to a normal distribution.

3.3.3 Analyses

Summary statistics

The data were analysed assuming that the sires were all randomly mated and all parents unrelated, i.e. assuming that offspring were related either as full-sibs or paternal half-sibs. The data analysed contained :

	No. individuals	No. with observations on enzyme activity
Sires	126	12
Dams	551	167
Offspring	571	571
Total	1248	750

Four traits were considered: the activity of the three enzymes (GLU, GAL and HEX) and the observed minus the predicted activity for α-glucosidase (DEV).

The distributions of the 4 traits for all 750 animals are given in figures 3.1 to 3.4. Summaries of the distributions, in terms of the mean, standard deviation, skewness and kurtosis are given in table 3.1.

Figure 3.1 *Distribution of α-glucosidase enzyme activities for the total data set.*

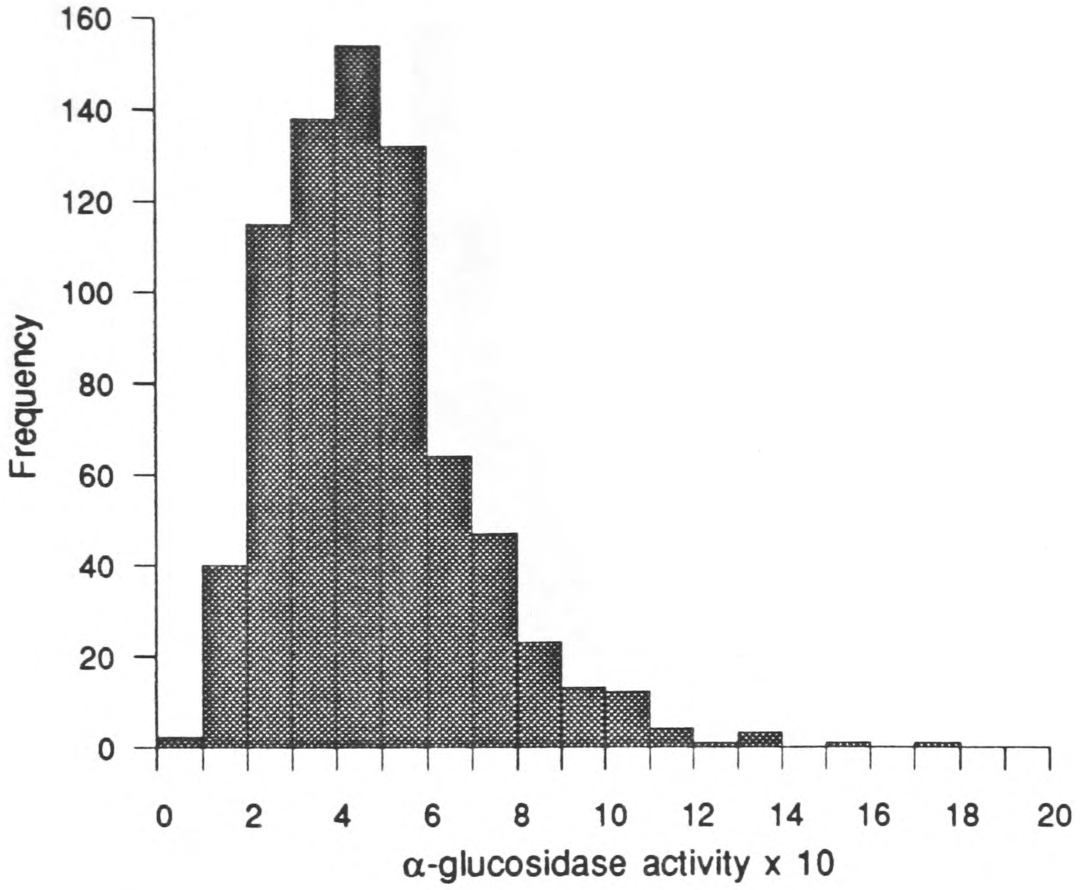


Figure 3.2 *Distribution of β-galactosidase enzyme activities for the total data set.*

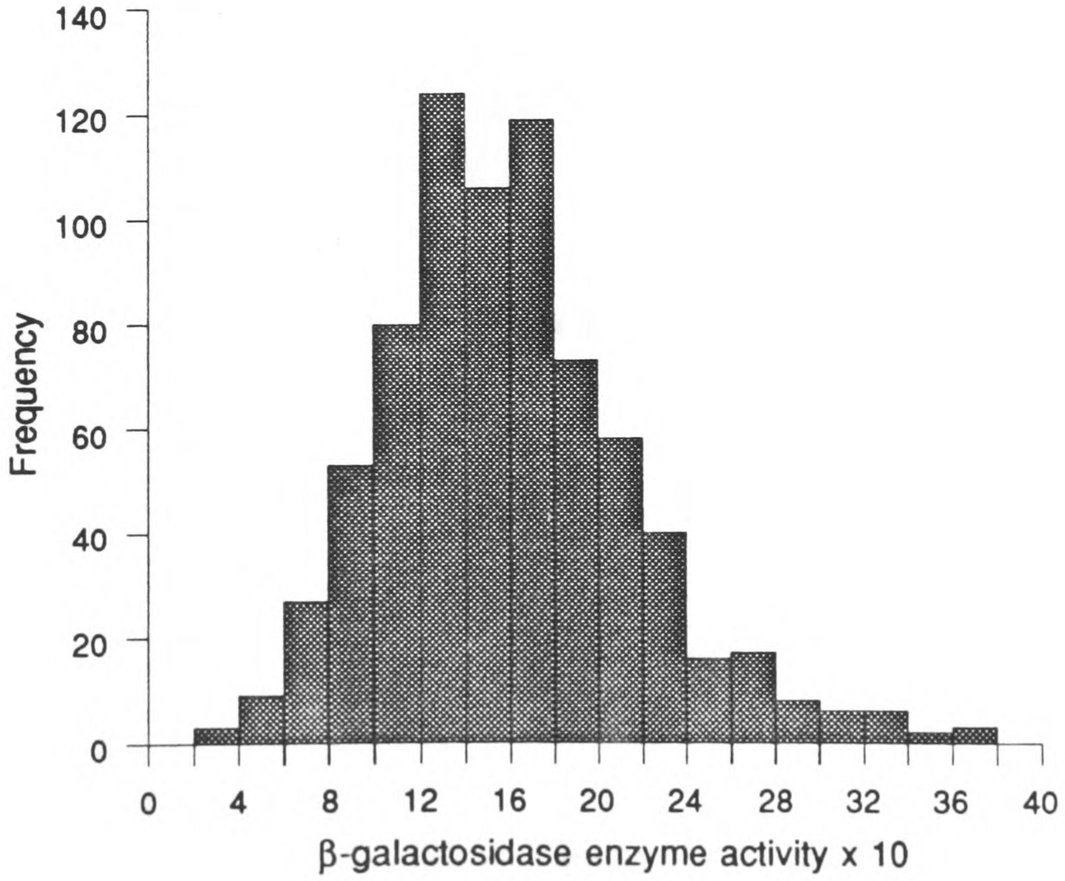


Figure 3.3 *Distribution of hexosaminidase enzyme activities for the total data set.*

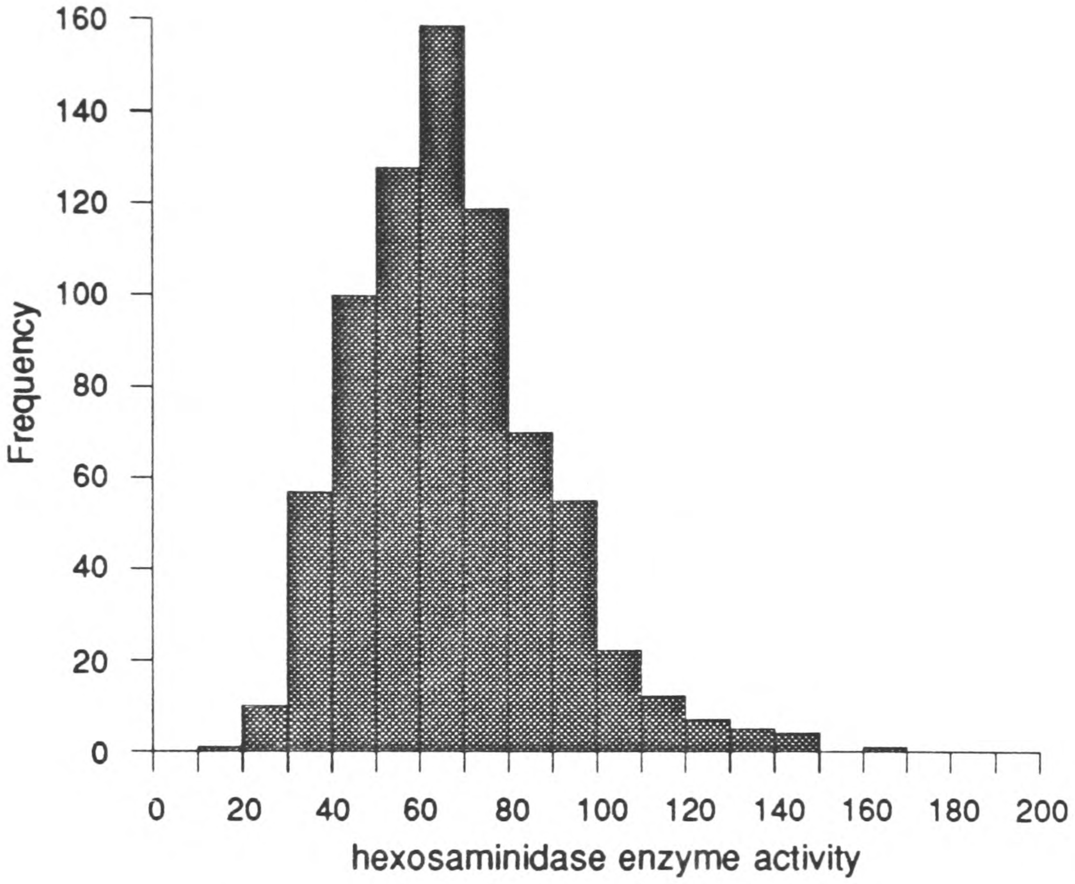


Figure 3.4 *Distribution of observed minus predicted α-glucosidase enzyme activity for the total data set.*

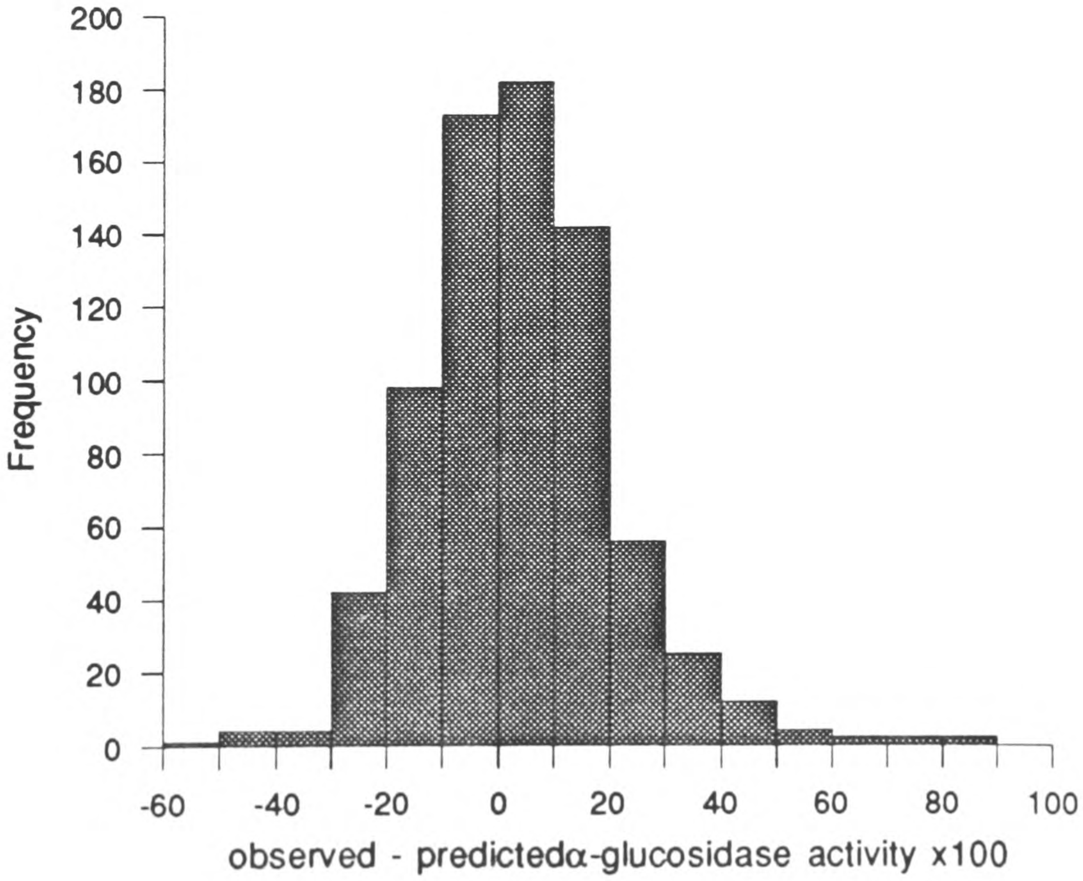


Table 3.1 *Mean, standard deviation, skewness and kurtosis for the total data.*

Trait	mean	standard deviation	skewness	kurtosis
GLU (x10000)	4791.47	2187.25	1.212***	2.917***
GAL (x1000)	1603.26	558.11	0.764***	1.102***
HEX (x100)	6696.83	2182.33	0.791***	1.061***
DEV (x10000)	349.03	1762.95	0.622***	2.218***

Skewness is given as: $\frac{m3}{m2\sqrt{m2}}$ (sd = 0.089) and kurtosis as: $\frac{m4}{(m2)^2} - 3$ (sd = 0.179)

where: $mp = \frac{\sum_i (x_i - \bar{x})^p}{n}$

*** indicates significance at 0.1% level.

Using the 571 offspring records and assuming a sire model, so that all full-sibs were assumed to be half-sibs, Harvey's mixed model least squares program (LSML76, Harvey, 1977) was used to obtain estimates of the heritabilities and correlations, genetic and phenotypic, between the traits. These are given in table 3.2.

Table 3.2 *Heritabilities on the diagonals, phenotypic correlations between traits above and genetic correlations below.*

Trait	GLU	GAL	HEX	DEV
GLU	0.55±0.16	0.57	0.55	0.82
GAL	0.67±0.14	0.49±0.16	0.73	0.08
HEX	0.63±0.21	0.70±0.16	0.24±0.14	0.01
DEV	0.91±0.06	0.36±0.24	0.30±0.33	0.49±0.16

These results are in agreement with the estimates obtained by McPhee and Reichmann (1990) using a larger set of data. There is evidence that variation in the reference enzymes is genetically controlled and positive genetic correlations exist between the traits, however the estimates for the standard errors are large. The genetic correlation between HEX and GLU is much higher here than reported by McPhee and Reichmann (1990) . As expected, the phenotypic correlation between DEV and GAL and DEV and HEX are approximately zero. The genetic correlations between these are positive and although of reasonable size are not significantly different from zero. The estimate of the heritability for DEV is lower than for GLU, which would not be expected if the reference enzymes removed some of the environmentally caused variation.

In PAP the trait is is assumed to be normally distributed within each mode, i.e. that a polygenic model is composed of a single normal distribution and the mixed and major gene models of two or three normal distributions. Non-normal data, caused by non-genetic effects is likely to be interpreted as evidence for a major gene (MacLean *et al.*, 1975). To reduce the possibility of spuriously detecting a major gene the data can be transformed before analysis.

In these analyses SKUMIX (MacLean *et al.*, 1976) was used to find the best transform. SKUMIX is a Fortran program that uses the transform:

$$y_i = \frac{r}{p}\left[\left(\frac{x_i}{r}+1\right)^{p}-1\right]$$

Where: y_i is the transformed variable.

p is the power.

x_i is the original variable, transformed to have a mean of zero and unit variance.

r is a scale parameter, such that $\frac{x_i}{r}+1 > 0$ for all x_i, here set to 5.

All observations are assumed to be independent, i.e. familial relationships are ignored, and using maximum likelihood techniques enables the user to transform the data to a mixture of 1, 2 or 3 normal distributions. Likelihoods can also be calculated under the assumption of different numbers of underlying distributions without transformation. A comparison of the likelihoods will suggest whether the observed distribution is, for example, composed of one skewed distribution or several normal distributions. For these analyses the data were transformed to a single normal distribution, this favouring the polygenic or non-genetic models where a unimodal distribution is expected rather than the mixed or major gene models where a non-normal distribution might be expected and hence reducing the chance of spuriously detecting a major gene. Table 3.3 gives the transforms used and the difference between the likelihoods for a single normal distribution with and without transformation. The distributions of the transformed data sets are given in figures 3.5 to 3.8. Comparing these to the distributions prior to transformation it can be seen that these more closely resemble a normal distribution and, in fact, are no longer significantly skewed or kurtic.

The transformed data were reanalysed using Harvey's program (LSML76), as described above, (with the assumption of normality, based on measures of skewness and kurtosis, now satisfied) and the results are given in table 3.4.

Table 3.3 *Skewness parameters used for complete data, and associated change in likelihood.*

Trait	p	lnL difference
GLU	-0.799	161.07
GAL	-0.138	70.71
HEX	-0.286	82.64
DEV	0.521	21.46

Table 3.4 *Heritabilities on the diagonals, phenotypic correlations between traits above and genetic correlations below for the transformed data*

Trait	GLU	GAL	HEX	DEV
GLU	0.66±0.17	0.58	0.55	0.78
GAL	0.60±0.14	0.49±0.16	0.76	0.05
HEX	0.61±0.19	0.75±0.14	0.26±0.15	-0.02
DEV	0.91±0.06	0.27±0.23	0.25±0.31	0.54±0.16

There is little difference in the correlations between the traits calculated after transformation and those given previously in table 3.2 prior to transformation. The largest change is in the heritability of GLU which has increased.

Figure 3.5 *Distribution of α-glucosidase enzyme activities for the transformed data set.*

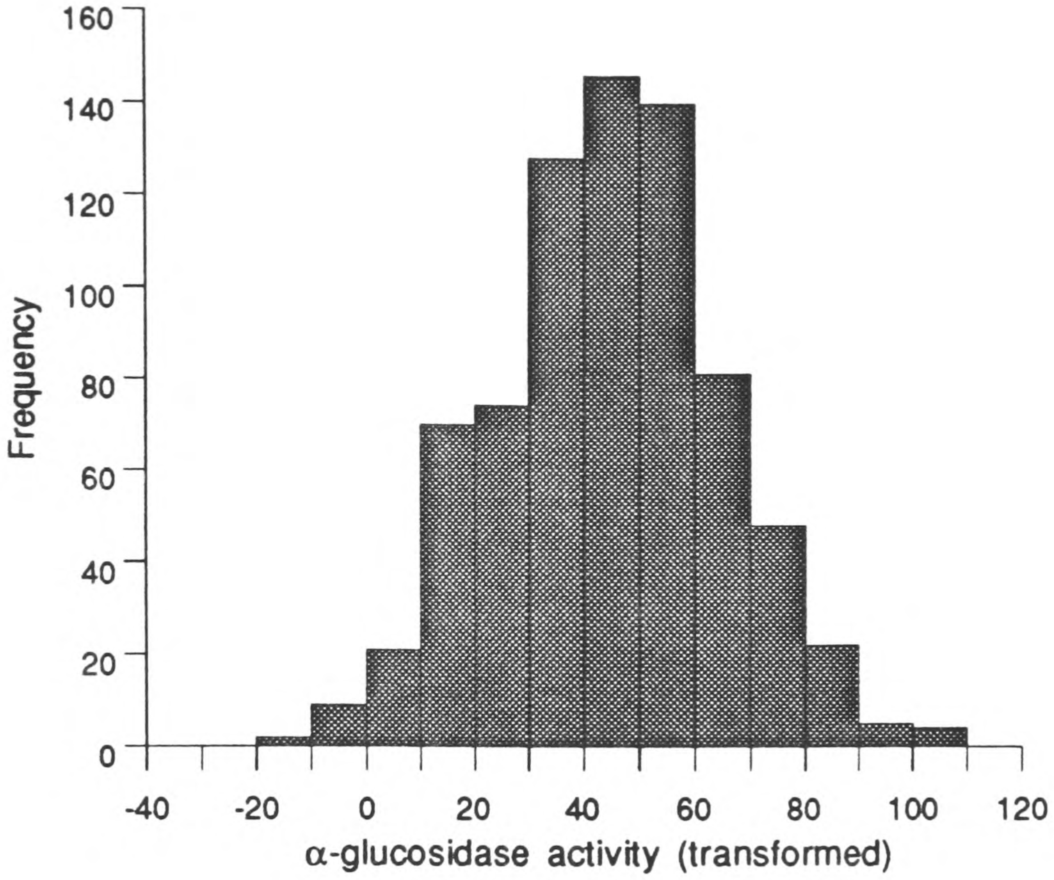


Figure 3.6 *Distribution of β-galactosidase enzyme activities for the transformed data set.*

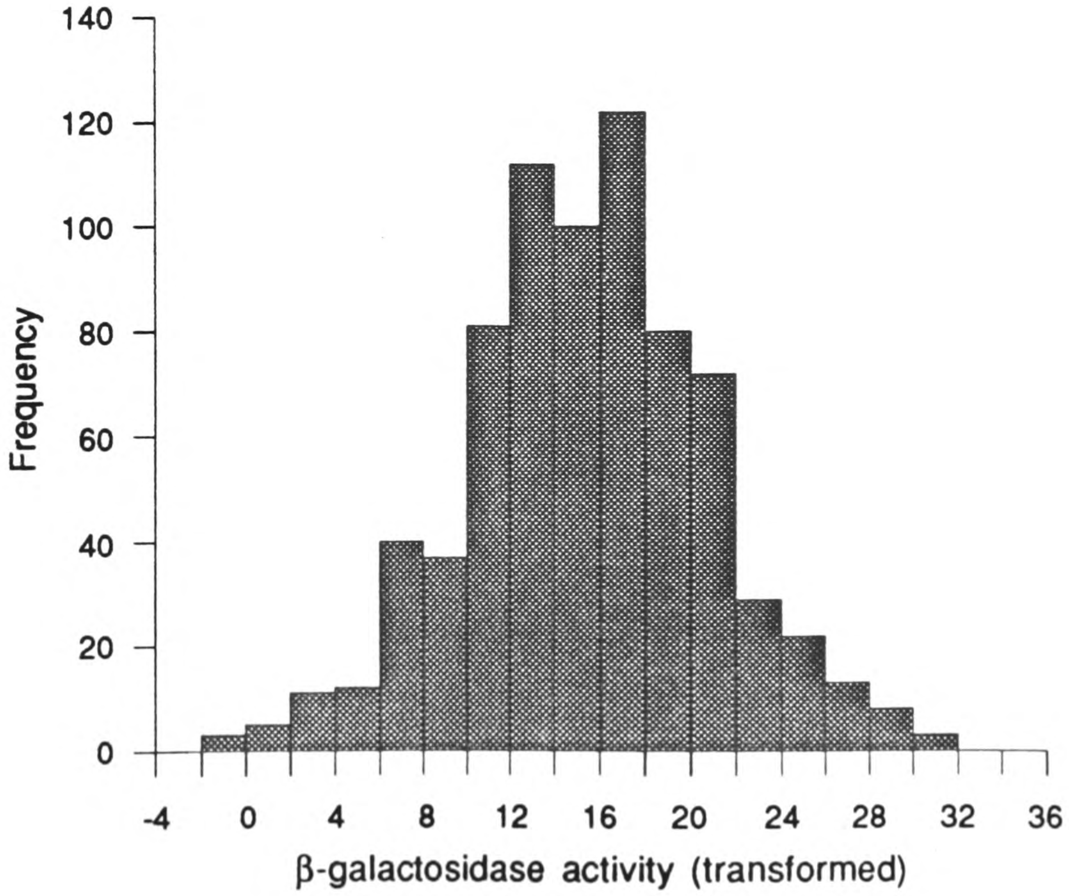


Figure 3.7 *Distribution of hexosaminidase enzyme activities for the transformed data set.*

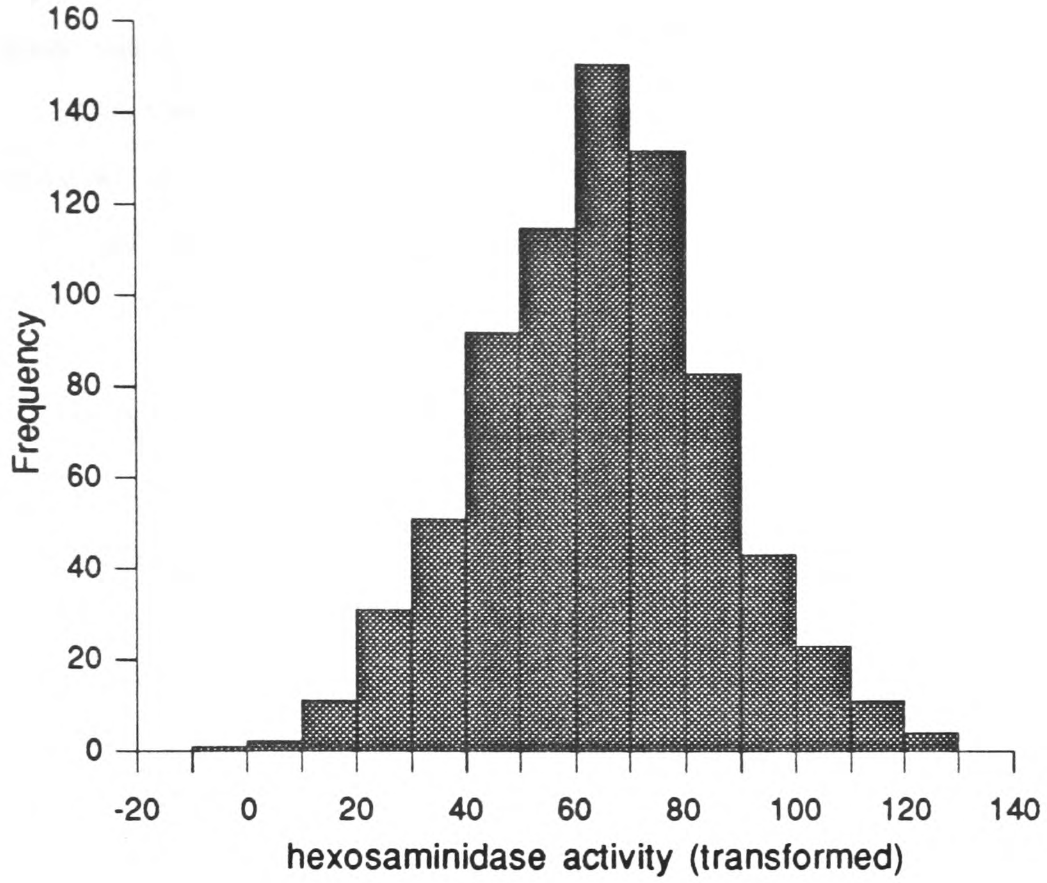
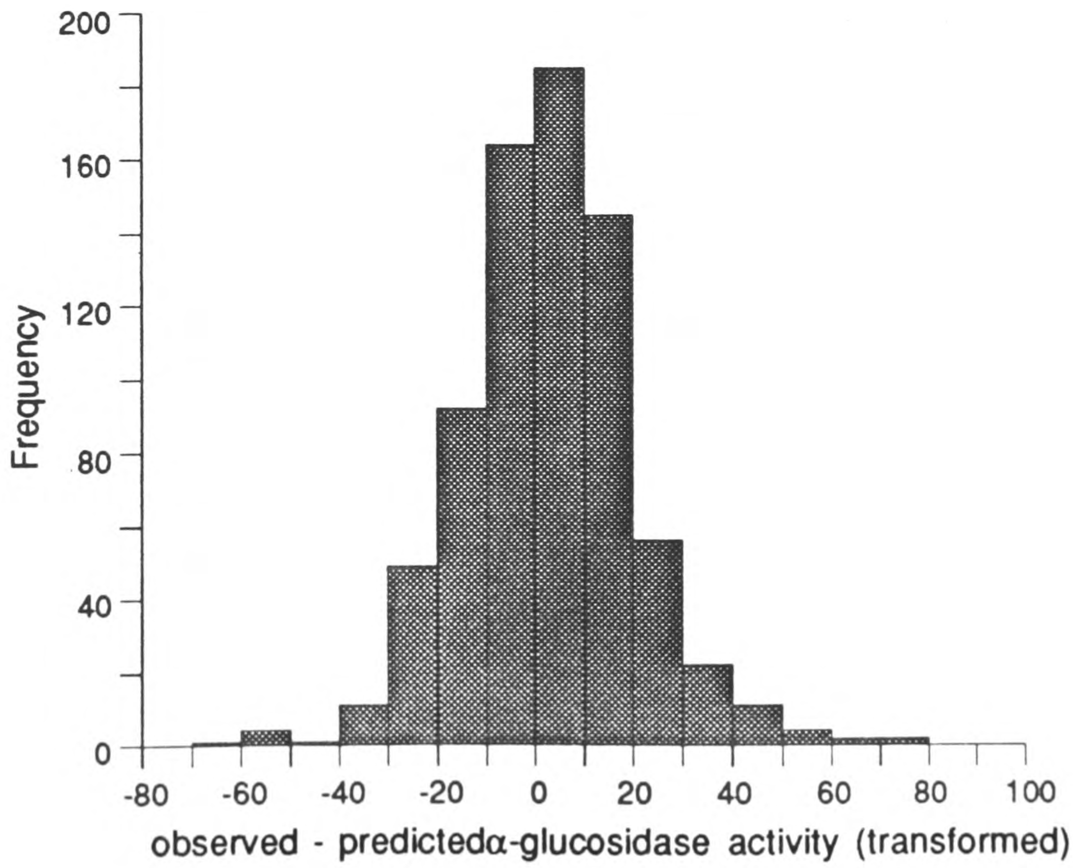


Figure 3.8 *Distribution of observed minus predicted α-glucosidase enzyme activity for the transformed data set.*



Segregation analyses

The maximum likelihoods of the data under the models given in table 3.5 were obtained using the relevant subroutines from PAP. The likelihood under the mixed model was calculated using the approximation described above (Hasstedt, 1982). The population was assumed to be in Hardy-Weinberg equilibrium with respect to the major gene and hence an allele frequency can describe the three genotype frequencies. The parameters required to be estimated for each model are also given in table 3.5. The likelihoods for the models are given in Appendix 1.

Table 3.5 *Models considered and the parameters necessary to describe these models*

Model	Parameters estimated					
Mixed	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	h^2_{poly}	σ^2_e
Polygenic		μ			h^2_{poly}	σ^2_e
Major gene	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}		σ^2_e
Non-genetic		μ				σ^2_e

Hence the tests given in table 3.6 can be made to determine the mode of inheritance, where, as described before, for nested hypotheses, the degrees of freedom are calculated as the difference in the number of parameters estimated under the more general model but fixed under the restricted model. For example, when comparing the mixed model and major gene model the only additional parameter required to explain the mixed model is the polygenic heritability, therefore the test has one d.f.

Table 3.6 *Possible comparison of hypotheses and the relevant degrees of freedom.*

test	ML(general)	ML(restricted)	d.f.
1	Major gene	Non-genetic	3
2	Polygenic	Non-genetic	1
3	Mixed	Major gene	1
4	Mixed	Polygenic	3

Where the test is $2\ln(ML(general) / ML(restricted)) \sim \chi^2$ d.f.

If tests 1 and 2 are significant, that is, if both the major gene model and the polygenic model are significant improvements over the non-genetic model there is evidence that the variation in the trait has a genetic component of control. If only one of them is significant this suggests the mechanism of genetic control. Given that tests 1 and 2 are both significant, if tests 3 and 4 are also significant, so that the mixed model is a significant improvement over both the major gene and polygenic models, there is evidence that the genetic component is composed of both a major gene and a polygenic component. However, if test 3 indicates a significant improvement of the mixed model over the major gene model but test 4 is not significant, this suggests that the trait is controlled by polygenes, and hence the addition of a polygenic component to a major gene model significantly increases the likelihood, but the addition of a major gene to the polygenic model does not. If test 4 gives a significant result but test 3 does not the evidence is for a major gene. A consideration of the parameter estimates obtained is also required because some models may not be sensible in terms of their genetic interpretation.

The inclusion of the small number of parental phenotypes might lead to biassed estimates because some individuals are in the data as different animals. To overcome this possible bias the analyses were repeated using only the offspring phenotypes. The data were transformed using the transforms given in table 3.3. To reduce the possibility that evidence for a major gene was being found because of the presence of a few animals with extreme enzyme activities the data were also reanalysed after removal of the extreme 2% of individuals, followed by transformation to normality. If these outlying observations were in fact caused by the major gene then evidence for this gene might be removed and the parameter estimates might be biassed, hence the results of these analyses have to be considered along with the results from the complete data.

Results of segregation analysis

The test statistics from the different tests are given in tables 3.7, 3.8 and 3.9 for the three data sets. The likelihoods and parameter estimates on the transformed scale for the four traits are given in tables 3.10 to 3.13 for the models assuming Hardy-Weinberg equilibrium for any genetic component. In tables 3.14 to 3.17 the estimates for the means have been transformed back to the original scale and, hence, the within mode standard deviation has been given as a standard deviation on either side of the mean.

Table 3.7 *Test statistics for the four tests described in table 3.6 using the complete data set.*

Test	GLU	GAL	HEX	DEV
1	45.78***	19.76***	12.34**	68.70***
2	55.86***	21.66***	11.68***	66.88***
3	10.88***	4.36*	-	19.78***
4	0.80	2.46	-	21.60***

*** indicates significance at the 0.1% level, ** at the 1% level and * at the 5% level.

Table 3.8 *Test statistics using the data containing offspring phenotypes only.*

Test	GLU	GAL	HEX	DEV
1	32.64***	19.46***	12.26**	50.34***
2	41.46***	25.60***	10.80**	35.16***
3	8.38**	5.72*	0.16	5.78*
4	0.44	0.42	1.62	20.96***

Table 3.9 *Test statistics using the complete data with the extreme scoring 2% of individuals removed.*

Test	GLU	GAL	HEX	DEV
1	27.68***	18.60***	17.92***	38.04***
2	33.56***	24.80***	14.96***	43.50***
3	6.96**	6.64**	0.02	11.46***
4	1.08	0.80	2.98	6.00

Table 3.10 *Parameter estimates for α-glucosidase on the transformed scale with standard errors in parentheses.*

Model	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	h^2_{poly}	σ^2_e
Mixed	0.65 (0.28)	5239 (943)	3851 (678)	3048 (887)	0.32 (0.19)	1854 (194)
Major gene	0.49 (0.08)	6234 (309)	4352 (289)	2615 (415)		1610 (97)
Polygenic		4342 (98)			0.48 (0.08)	2035 (55)
Non-genetic		4436 (75)				2058 (53)

Table 3.14 *Means and variances for α-glucosidase activity (x10000) on the original scale.*

Model	μ_{AA}	μ_{Aa}	μ_{aa}
Mixed	5256	3919	3267
	(7627 - 3531)	(5780 - 2522)	(4903 - 2021)
Major gene	6427	4367	2947
	(8854 - 4626)	(6086 - 3038)	(4250 - 1910)
Polygenic	4358		
	(6614 - 2731)		
Non-genetic	4446		
	(6770 - 2779)		

Values in parentheses are one within genotype standard deviation from the mean.

Table 3.11 *Parameter estimates for β-galactosidase on the transformed scale with standard errors in parentheses.*

Model	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	h^2_{poly}	σ^2_e
Mixed	0.93 (0.06)	1613 (58)	1110 (212)	699 (648)	0.20 (0.12)	513 (26)
Major gene	0.81 (0.12)	1708 (66)	1287 (193)	709 (22)		485 (20)
Polygenic		1542 (25)			0.33 (0.08)	547 (14)
Non-genetic		1542 (20)				547 (14)

Table 3.15 *Means and variances for β-galactosidase activity (x1000) on the original scale.*

Model	μ_{AA}	μ_{Aa}	μ_{aa}
Mixed	1613	1156	845
	(2188 - 1148)	(1624 - 775)	(1242 - 520)
Major gene	1711	1306	852
	(2272 - 1251)	(1778 - 918)	(1227 - 542)
Polygenic	1543		
	(2141 - 1064)		
Non-genetic	1543		
	(2142 - 1064)		

Table 3.12 *Parameter estimates for hexosaminidase on the transformed scale with standard errors in parentheses.*

Model	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	h^2_{poly}	σ^2_e
Mixed	-	-	-	-	-	-
Major gene	0.40 (0.12)	7836 (552)	6932 (480)	5072 (375)		1853 (961)
Polygenic		6394 (94)			0.24 (0.09)	2130 (56)
Non-genetic		6432 (77)				2129 (55)

Table 3.16 *Means and variances for hexosaminidase activity (x100) on the original scale.*

Model	μ_{AA}	μ_{Aa}	μ_{aa}
Mixed	7659	6909	5246
	(10030 - 5772)	(9084 - 5170)	(7004 - 3828)
Major gene	7917	6935	5216
	(10303 - 6012)	(9070 - 5221)	(6929 - 3828)
Polygenic	6400		
	(8740 -45762)		
Non-genetic	6436		
	(8786 - 4605)		

Table 3.13 *Parameter estimates for the observed minus the predicted α-glucosidase activity on the transformed scale with standard errors in parentheses.*

Model	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	h^2_{poly}	σ^2_e
Mixed	0.05 (0.03)	6062 (110)	1884 (457)	-30 (98)	0.47 (0.13)	1555 (57)
Major gene	0.07 (0.04)	5807 (973)	1924 (445)	-123 (107)		1490 (56)
Polygenic		159 (85)			0.56 (0.08)	1738 (48)
Non-genetic		265 (64)				1757 (45)

Table 3.17 *Means and variances for the observed minus the predicted α-glucosidase activity (x10000) on the original scale.*

Model	μ_{AA}	μ_{Aa}	μ_{aa}
Mixed	6941	1948	-26
	(9037 - 4973)	(3697 - 328)	(1562 --1484)
Major gene	6610	1991	-117
	(8596 - 4742.31)	(3668 - 434)	(1394 --1508)
Polygenic	160		
	(1962 --1478)		
Non-genetic	265		
	(2099 --1400)		

GLU (Tables 3.10 and 3.14)

The addition of either a polygenic component or a major gene causes a significant increase in the likelihood compared with the non-genetic model. The loss of the polygenic component from the mixed model causes a significant decrease in the likelihood, however, the loss of the major gene has virtually no effect on the likelihood. Hence there is evidence to suggest that the trait is controlled by many genes of small effect. The estimate for the heritability on the transformed scale is 0.48. Figures 3.9 and 3.10 illustrate the expected distribution of the population under the polygenic model and the mixed model.

Analyses of the other data sets gave essentially the same results, although with the removal of the extreme individuals the parameter estimates were altered slightly. The mixed model resulted in an high scoring allele with dominant effect and low frequency (p=0.22) and the major gene model suggested an allele of high effect with a frequency of 0.32. The polygenic heritability estimate decreased.

GAL (Tables 3.11 and 3.15)

There is strong evidence for a genetic component controlling variation in the activity of GAL, as both the polygenic and major gene models are significant when compared with the non-genetic model. When the mixed model is compared with these two sub-models, the comparison with the major gene model is just significant, whereas the comparison with the polygenic model is not significant. These results support the hypothesis that the trait is polygenically controlled with an heritability of 0.33.

When analysing data containing the phenotypes of the offspring only, the same conclusion was drawn and the parameter estimates were similar. With the extreme 2% of individuals removed, for the mixed model the allele frequency remained extreme but with few high scoring rather than low scoring individuals. The major gene model gave intermediate allele frequencies and an approximately additive model. A polygenic model would still be suggested.

HEX (Tables 3.12 and 3.16)

There is evidence that variation in the activity of HEX has a genetic component of control, with the likelihood under a polygenic model being more significant than under a major gene model although with a slightly lower value for the likelihood. With the extension to a mixed model, convergence to a model containing both components with likelihood greater than the major gene model could not be attained.

Using the other two data sets, a maximum for the mixed model likelihood was attained, but it was not a significant improvement over either the polygenic or the major

Figure 3.9 *Predicted distribution for α-glucosidase activity under the polygenic model.*

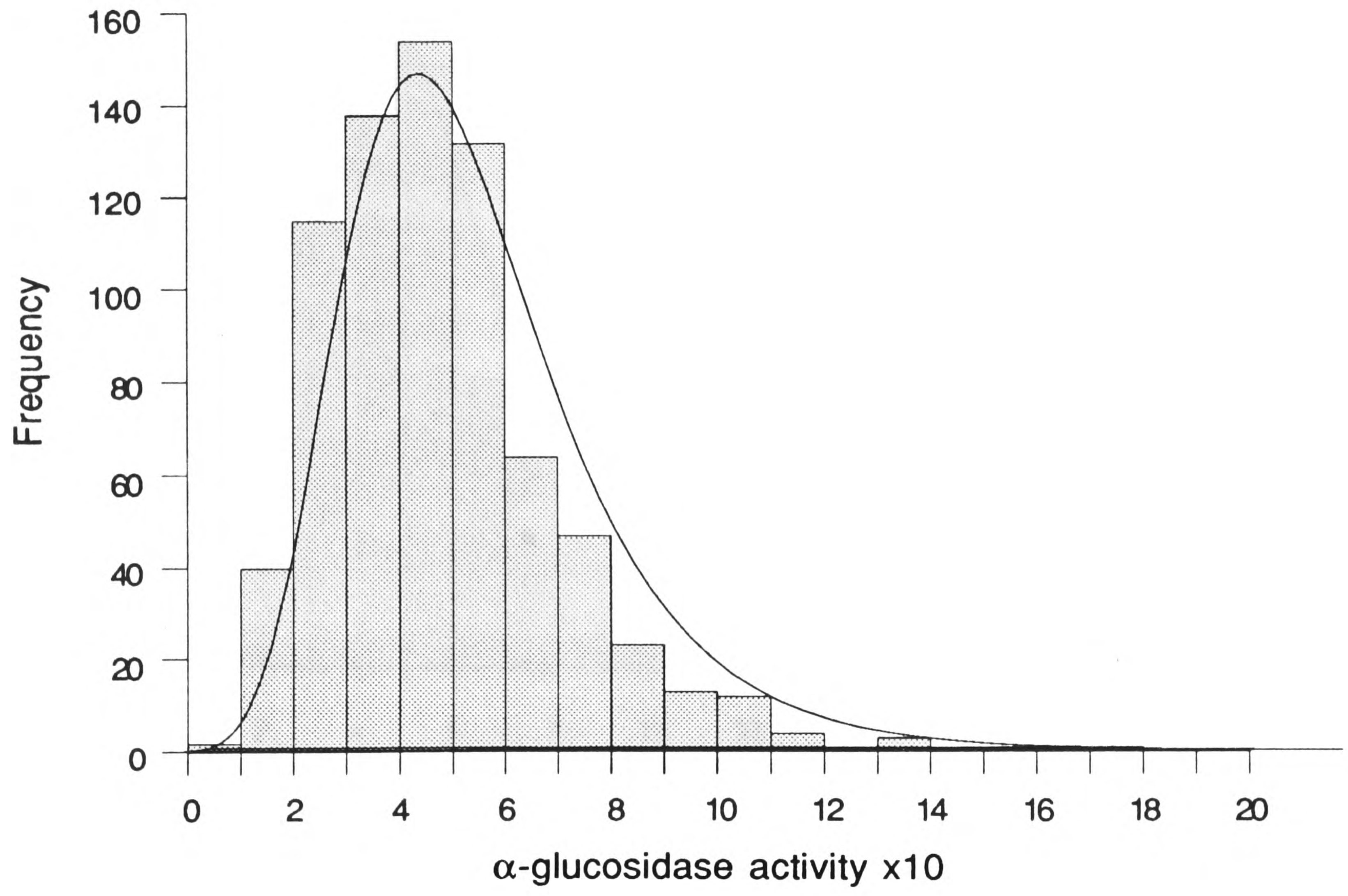
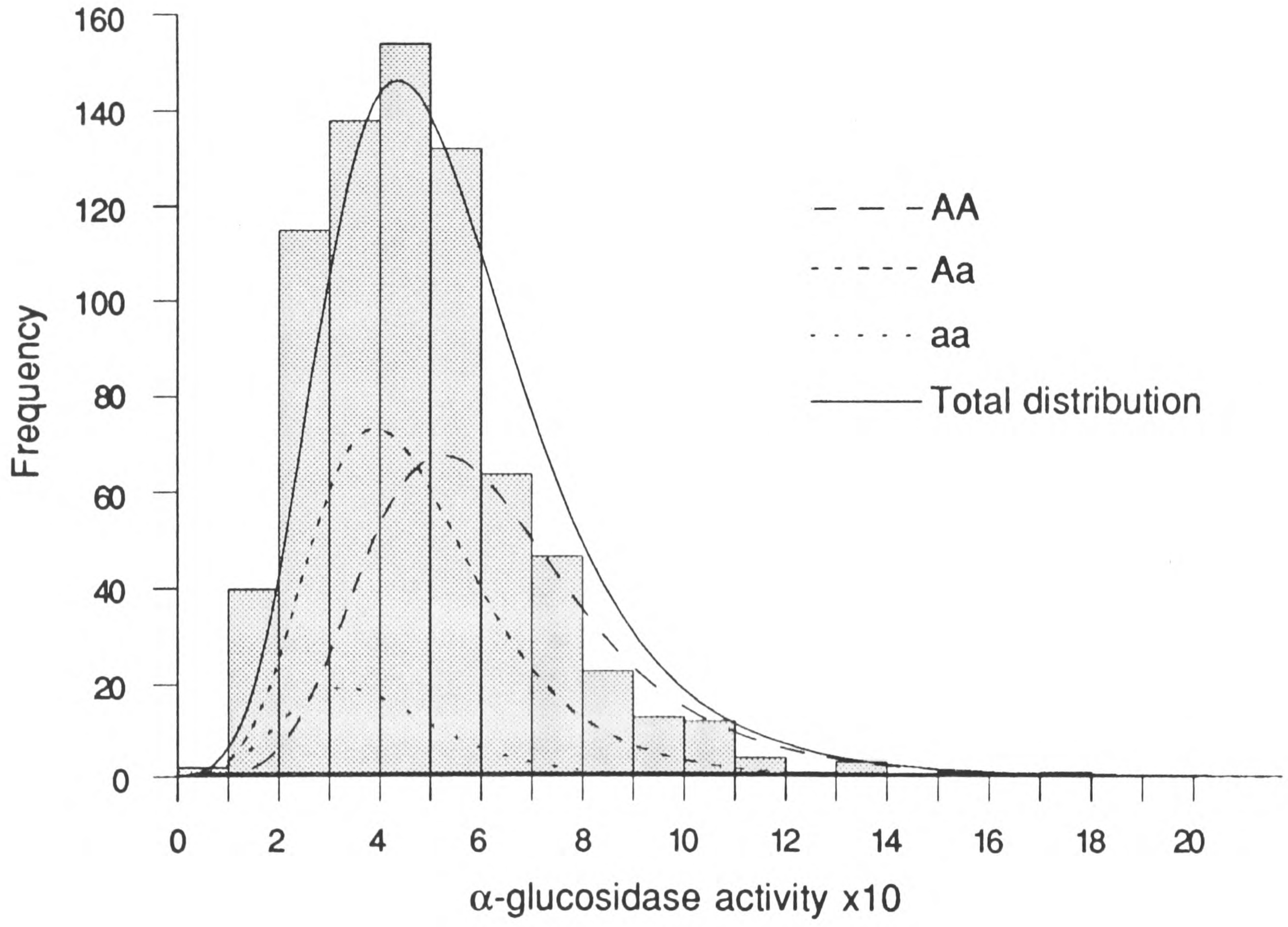


Figure 3.10 *Predicted distribution of α-glucosidase enzyme activity under the mixed model.*



gene models. Both of these models were significantly better than a non-genetic model and it is not possible to suggest the mechanism of inheritance.

When a maximum was attained for the mixed model the parameter estimates for the major gene were similar to the estimates under the major gene model and the heritability estimate was less than 0.1. Parameter estimates from the three data sets were similar.

DEV (Tables 3.13 and 3.17)

Both the likelihood under the major gene model and under the polygenic model are significantly larger than the likelihood of the non-genetic model, suggesting that variation in the trait is genetically controlled. The mixed model is a significant improvement over both of these sub-models. Hence there is evidence for both a major gene and polygenic component. However, the mixed model suggested has a rare allele, with an extremely high mean effect of the homozygous genotype for this allele. The low frequency of this genotype suggests that this model is explaining some remaining non-normality of the data with a slight excess of individuals with high score.

The results using the offspring phenotypes only were essentially the same. However, with the extreme 2% of observations removed there was no longer evidence for a mixed model, and polygenic inheritance would suggested. The major genes suggested have intermediate allele frequencies and the genotype means are less extreme presumably because of the removal of the few individuals causing the results observed with the full data set.

3.3.4 Discussion

The results of the analyses presented above suggest that there is no evidence to support the hypothesis of a major gene controlling α-glucosidase activity. The three enzymes all show evidence of polygenic inheritance for their activity. DEV is a combination of three traits and the variation in each of these traits has been shown to be genetically controlled, hence, the interpretation of the results is difficult.

A major gene model, which associates mortality with the low scoring homozygous genotype, has been suggested for the inheritance of α-glucosidase activity (Howell *et al.* 1981). Hence, except when taken at birth, the offspring are not expected to be in Hardy-Weinberg equilibrium, even if the parents are randomly mated. Also, only two genotypes would be observed in the data. Obviously, the major gene model fitted cannot exactly describe this hypothesised model; if however, the data were derived under this model it would be expected that, when analysed, a model including a major gene would fit better than a polygenic one.

Transforming the data prior to analysis might have removed evidence of a major gene as the original distribution, although non-normal, has only a single mode. Analysis of the data without prior transformation would have given a useful indication of the effect of the transformation, in case the non-normality was a product of a major gene rather than environmental. There is an additional problem that the phenotypes are influenced by fixed effects which could not be accounted for in the analyses presented here. For example, the analyses are sensitive to the age of the animal from which the sample was taken, sex of the animal and length of time from the sample being taken to the assay being carried out (McPhee and Reichmann, 1990). These would be expected to alter the estimates and possibly the suggested mode of inheritance. For example, a sex difference in the trait would mean that there are effectively two normal distributions of enzyme activity within the population and this could be interpreted as a major gene.

Removal of the extreme scoring individuals did not alter the conclusions of the analyses except in the case of DEV which had been identifying a few individuals with extreme scores. For the other traits some of the parameter estimates were altered, in general, these were predictable changes with the major genotype means becoming less extreme.

The assumption that individuals within a generation can only be related as full-sibs or paternal half-sibs and hence that all parents are unrelated, leads to a loss of information. It is not clear how useful this additional information might be; its inclusion would certainly significantly reduce the speed of computation. A small bias might be introduced when individuals with phenotypes appear in the data more than once with a different identification. Analysing the data without the parental phenotypes should, however, remove this bias and the results form these analyses were in agreement with those from the total data set.

The allele frequency estimated is the frequency in the founder or, in this case, the parent generation. The sires in the data leave very unequal numbers of progeny which will effect the frequency of the major genotype in the offspring generation. The distribution of offspring genotype frequencies will depend on the conditional genotype probability of each parent weighted by the number of offspring the matings produce.

α-glucosidase activity is known to be associated with mortality as when activity is too low animals can no longer survive. The analyses above suggest that the level of activity is genetically controlled and hence there must be a relationship between the genotype of an animal and the probability of its death. The effect of this is that selection is being practised, with selection against the low activity genotypes. Hence parameter estimates would be expected to alter over generations. For example, if a major gene was involved the frequency of the allele conveying low activity would be expected to decrease and if polygenically controlled a change in the variance and associated heritability might be observed. The two generation model used cannot take these into account. Information on the individuals which died from Pompe's disease would be useful, although its incorporation into the analyses is not straight forward.

Extensions to the models offered in PAP

The hypothesised mode of inheritance is that α-glucosidase activity is controlled by a single locus with two alleles, the wild type and a mutant allele that decreases the activity to the extent that the homozygous animals for the mutant allele die. If this is correct, the genotype frequencies in the parent population are not expected to be in Hardy-Weinberg equilibrium. In fact the affected homozygote is not expected to reach maturity and hence will be absent from the parent generation. It is possible, however, that the genotype could be present in the offspring population if testing for the enzyme activity was carried out early in development, possibly before clinical signs of the disease were present. Additional parameters to those used in the models above are required to describe this model.

First, parameters are required to describe the fitness of the major genotypes in the parent and offspring generations:

Genotype	AA	Aa	aa
Relative fitness in the parents	1	1	*sp*
Relative fitness to testing in the offspring	1	1	*so*

sp can take a value of zero or one. If set to one, this is the most general model with regards the parental genotype frequencies with the possibility of all three genotypes being present without assumptions about their relative frequency. If *sp* is set to zero the genotype aa is assumed to be absent from the parental generation, which is the hypothesised model and a sub-model of the more general one with *sp* equal to one. The second parameter, *so*, affects the expected frequencies of observed major genotypes in the offspring population given the parental genotypes. Hence, the transmission frequencies are as follows:

Mating type	Offspring genotypes AA	 Aa	 aa
AA x AA	1	0	0
AA x Aa	0.5	0.5	0
Aa x Aa	$\frac{0.25}{0.25+0.5+0.25so}$	$\frac{0.5}{0.25+0.5+0.25so}$	$\frac{0.25so}{0.25+0.5+0.25so}$
AA x aa	0	1	0
Aa x aa	0	$\frac{0.5}{0.5+0.5so}$	$\frac{0.5so}{0.5+0.5so}$
aa x aa	0	0	1

Matings, however, will not be represented in the data if the offspring does not have a record. If the major gene hypothesis is correct and some of the individuals with genotype aa die before being tested there will be fewer matings than expected in the data which give rise to affected progeny. Assuming that each mating produces n offspring and matings are expected to appear in the data if at least one offspring survives, then the expected frequency of each mating type depends on the relative fitness of offspring (*so*) and the number of offspring per mating. In the case of cattle, to a good approximate, each mating produces one calf. Hence the probability of an Aa x Aa mating being present is:

probability that the calf survives

= prob(calf AA) + prob(calf Aa) + prob(calf aa and survives)

$$= \frac{3+so}{4}$$

More generally for n offspring the probability of an Aa x Aa mating being present is:

= prob(at least one offspring survives)

= 1 - prob(no offspring survive)

= 1 - prob(all calves are aa and die)

$= 1 - (\text{prob(offspring dies|aa)prob(aa)})^n$

$$= 1 - \left(\frac{1-so}{4}\right)^n$$

However it would be possible for parents to be remated if the offspring died soon after birth. A correction could be incorporated to allow for the possibility that some of the mating types are present at a higher frequency than would be expected if all the matings which resulted in only dead offspring are omitted. Hence a third parameter, which corrects the expected frequency of the different mating types, is required to describe this model. For one expected progeny per mating the corrections are as follows:

Mating type	Expected frequency present in the data x TOT
AA x AA	$p(AA)\ p(AA)$
AA x Aa	$2\ p(AA)\ p(Aa)$
Aa x Aa	$p(Aa)\ p(Aa)\ \left(\frac{3+so}{4}\right)^{xm}$
AA x aa	$2\ p(AA)\ p(aa)$
Aa x aa	$2\ p(Aa)\ p(aa)\ \left(\frac{1+so}{2}\right)^{xm}$
aa x aa	$p(aa)\ p(aa)\ so^{xm}$

Where: TOT = sum of all the above frequencies, i.e. frequencies are corrected so that they sum to one.

p(c) is the frequency of genotype c in the parental population and $\sum_{c=1}^{m} p(c) = 1$

xm can take any value between zero and one. If set to zero the matings appear in the frequency expected assuming random mating of the parental genotypes, which could be explained by parents being remated to each other if their progeny died. If *xm* is set to one there is a correction for those matings which produced an offspring that died before testing and the expected number of this mating type is reduced. Intermediate values indicate that a proportion of the parents whose offspring die are being remated.

The most general model can be obtained by setting *sp* to one and allowing *so* and *xm* to be estimated. Models with fixed parameter values can then be compared as sub-models of this one. The parameter *xm* could be omitted and possibly with the data structure used there would not be enough information to estimate it.

An alternative model that could be tested assumes that the trait is controlled by many genes of small effect with a threshold, and when the enzyme activity falls below this level the individual dies.

3.4 DISCUSSION

Using PAP (Hasstedt and Cartwright, 1981) the likelihood of a given set of data under many different models can be calculated. Peeling, a recursive procedure, permits the analysis of any complexity of pedigree and is an efficient way of reducing the storage requirements of the program, as at any time matrices are only of the size of the nuclear family being considered and additional information is only being retained on individuals which reappear in the pedigree.

The approximation to the mixed model likelihood is accurate. Although for each likelihood calculation the approximation is much faster than the exact form the approximation is fairly complex involving many calculations. Overall speed could be improved by scaling the parameter estimates in the maximisation routine and, if memory space allows, retaining the phenotypes in memory rather than reading in for each evaluation of the likelihood.

Data from animal populations generally involves many fixed effects, such as herd or year, and these cannot be included in the analysis. Estimation of these effects could take place, and the data adjusted, prior to analysis. Theoretically this would not be the best solution and estimation simultaneously with the genetic parameters would be preferable. Also, unlike in human genetics, the knowledge of breeding values is important and these cannot be obtained immediately from the program as written and additional subroutines would be required.

Although the program is designed to enable the user to alter or add subroutines the complexity of the package means that this cannot always be achieved easily. This is especially true when considering the approximation to the mixed model which is not clearly documented. If extensions were required, for example to include fixed effects or common environment, the use of alternative approximations might prove easier. Increasing the number of parameters to be estimated is difficult.

Many of the models considered are unlikely to be suitable for animal populations, and hence the program could be simplified by reducing the number of possible genetic situations that can be analysed.

Problems were encountered attaining convergence for some models, especially when, rather than assuming that the genotypes were in Hardy-Weinberg equilibrium, they were estimated individually. Although the frequencies of two of the genotypes are constrained between zero and one, the third is obtained by subtracting these two frequencies from one and hence can easily become negative, at which point the likelihood often increases markedly. Although it would be possible to constrain the three frequencies to be between zero and one the retention of a general model enabling any number of alleles makes this difficult.

Appendix 1

A1.1 Models and likelihoods

A1.1.1 Non-genetic model

The non-genetic model assumes that the trait is environmentally controlled. In these analyses the individual random environmental effects (e_i) were assumed to be normally distributed.

Model: $y_i = e_i$

Where: $e \sim N(0,\sigma_e^2)$

Likelihood:

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{1}{2\sigma_e^2} y_i^2\right]$$

or in matrix notation $\mathbf{y} = \mathbf{e}$

and

$$L = \frac{1}{\sqrt{(2\pi\sigma_e^2)^N}} \exp\left[-\frac{1}{2\sigma_e^2} \mathbf{y}'\mathbf{y}\right]$$

Where **y** and **e** are vectors of length N.

A1.1.2 Polygenic model

The polygenic model assumes that the trait is controlled by many genes of small effect, giving a normal distribution of genetic effects (g_i), and an individual random environmental effect (e_i) which is assumed to be normally distributed.

Model: $y_i = \mu + g_i + e_i$

Where: $e \sim N(0,\sigma_e^2)$

$g \sim N(0,\sigma_g^2)$

Likelihood:

$$L = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\prod_{i=1}^{N} \text{ø}_{y_i}(g_i,\sigma_e^2)\right)\left(\prod_{j=1}^{f} \text{ø}_{g_j}(\mu,\sigma_g^2)\right)$$

$$\left(\prod_{k=1}^{n} \text{ø}_{g_k}\left(\frac{(g_{k1}+g_{k2})}{2},\frac{\sigma_g^2}{2}\right)\right) dg_1\, dg_2 \cdots dg_N$$

(Hasstedt, 1982)

The integrations are over the polygenic component for each individual. The first product is over all (N) individuals, the second over all (f) founders and the third all (n) offspring, where f + n = N.

g_{k1} and g_{k2} are the polygenic effects of the parents of g_k.

$ø_y(\mu,\sigma^2)$ is the conditional likelihood that y is from a normal distribution with mean μ and variance σ^2, which is equal to

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

In matrix notation, $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}$

and

$$L = \frac{1}{(2\pi)^{N/2}|\mathbf{V}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{1}\mu)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{1}\mu)\right]$$

Where: $\mathbf{V} = \mathbf{I}\sigma_e^2 + \mathbf{ZAZ}'\sigma_g^2$

$\mathbf{Z}$ is the design matrix for polygenic effects

$\mathbf{1}$ is a vector of ones

μ is the population mean

$\mathbf{A}$ is the additive genetic relationship matrix

A1.1.3 Major gene model

The major gene model describes the situation where the genetic component of the phenotype is controlled by a single locus. The variation within each genotype is assumed to be environmentally controlled and normally distributed. The model for an individual with genotype c is:

Model $y_i = \mu_c + e_i$

Where: $e \sim N(0,\sigma_e^2)$

μ_c is the mean effect of genotype c.

That is, the phenotypic distribution is composed of a mixture of normal distributions, the number of distributions being equal to the number of distinguishable genotype means. For the analyses described here, it was assumed that there were two alleles at this locus and hence three genotypes (m=3). The likelihood is composed of the following:

i) The conditional likelihood of the phenotype given the major genotype - $ø_y(\mu_{ci},\sigma_e^2)$

ii) The probability of the major genotype for a random (founder) individual - $p(c_i)$

iii) The probability of the major gènotype given the major genotype of the parents - $\text{trans}(c_i|c_s,c_d)$

Likelihood

$$L=\sum_{c_1=1}^{m}\sum_{c_2=1}^{m}\cdots\sum_{c_N=1}^{m}\prod_{i=1}^{N}\text{ø}_{y_i}(\mu_{ci},\sigma_e^2)\prod_{j=1}^{f}p(c_j)\prod_{k=1}^{n}\text{trans}(c_k\,|\,c_{k1},c_{k2})$$

[Hasstedt, 1982]

The summations are each over the m genotypes with one for each individual, which is equivalent to having a single summation over the m^N genotype combinations. The first product is over all individuals, the second over all founders and the third all offspring. c_{k1} and c_{k2} are the major genotypes of the parents of k.

In matrix notation the model for one genotype combination is:

$$\mathbf{y} = \mathbf{W}_D\mu_c + \mathbf{e}$$

and

$$L= \frac{1}{(2\pi\sigma_e^2)^{N/2}}\sum_{D=1}^{m^N}p(D)\exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{W}_D\mu_c)'(\mathbf{y} - \mathbf{W}_D\mu_c)\right]$$

Where: D is one of the m^N combinations of major genotype for the pedigree.

p(D) is the probability of genotype combination D, which is equal to the product

of

transmission probabilities for offspring given the parents genotypes and the relevant population frequency for founders.

$\mathbf{W}_D$ is an n x m matrix containing a 1 for the genotype being considered for each individual and a zero otherwise.

μ_c is a vector of major gene effects.

A1.1.4 Mixed model

As described previously (2.3.2) the mixed model assumes that the trait is controlled by both a major gene and polygenes. The variation within each major genotype is assumed to be the same for each genotype and equal to the sum of the polygenic and environmental variances. As before the individual environmental component is assumed to be normally distributed. The model under genotype c is:

Model $$y_i = \mu_c + g_i + e_i$$

Likelihood

$$L = \sum_{c_1=1}^{m} \sum_{c_2=1}^{m} \cdots \sum_{c_N=1}^{m} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{N} ø_{y_i}(g_i, \sigma_e^2) \right) \left(\prod_{j=1}^{f} p(c_j)\, ø_{g_j}(\mu, \sigma_g^2) \right)$$

$$\left(\prod_{k=1}^{n} \mathrm{trans}(c_k \mid c_{k1}, c_{k2})\, ø_{g_k}\left(\frac{(g_{k1}+g_{k2})}{2}, \frac{\sigma_g^2}{2} \right) \right) dg_1\, dg_2 \cdots dg_N$$

[Hasstedt, 1982]

This is a combination of the mixed and polygenic likelihoods, with a summation over all the major genotypes and an integration over the polygenic component for each individual.

In matrix notation the model for a given genotype combination is

$$\mathbf{y} = \mathbf{W}_D \mu_c + \mathbf{Z}\mathbf{g} + \mathbf{e}$$

and

$$L = \frac{1}{(2\pi)^{N/2} |\mathbf{V}|^{1/2}} \sum_{D=1}^{m^N} p(D) \exp\left[-\frac{1}{2} (\mathbf{y} - \mathbf{W}_D \mu_c)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{W}_D \mu_c) \right]$$

CHAPTER 4

APPROXIMATION 2 - USING HERMITE INTEGRATION

4.1 INTRODUCTION

The exact mixed model likelihood [2.2] contains an integration of a complicated function, over all possible values for each sire's transmitting ability. A standard statistical approximation to an integration is to replace it with a weighted summation, so that effectively a continuous density function (c(x)) is replaced by a discrete histogram.

$$\int_a^b c(x)f(x).dx = \sum_{g=1}^{G} w_g f(x_g) \qquad [4.1]$$

Where: G is the number of points in the summation.

x_g are the abscissae within the range a to b.

w_g are the weights.

Suitable weights and abscissae need to be supplied. Obviously as the number of points in the summation increases the approximation improves, an integration being equivalent to an infinite number of points. However, by taking into account the distribution of the function to be integrated, the abscissae and weights can be provided to reduce the number of points in the summation required to provide a reasonable approximation compared with ignoring this information and, for example, using evenly spaced abscissae. In the case of the mixed model likelihood, the variable over which integration takes place appears in the form exp[-x^2] and hence efficient abscissae and weights can be obtained from the Hermite polynomial (Hildebrand, 1974). If G points in the summation are required the abscissae are obtained from the roots of the order G polynomial:

$$H_G(x) = (-1)^G \exp[x^2]\frac{d^G(\exp[-x^2])}{dx^G}$$

Fortunately tables of the weights and abscissae exist for various numbers of points in the summation and are given for a standard curve, symmetrically placed about the origin (e.g. Selby, 1970). A reduction in the number of points required for a reasonable approximation may be obtained by transforming the weights and abscissae, so that they reflect the correct mean and variance of the variable over which the integration is to be made.

4.2 LIKELIHOOD

4.2.1 Derivation of the mixed model likelihood using Hermite integration.

The exact mixed model likelihood can be written as follows, with the integration over each sire's transmitting ability (see chapter 2 for notation and derivation):

$$L(MM)=\prod_{i=1}^{s}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left[\frac{-u_i^2}{2\sigma_u^2}\right]\sum_{c=1}^{m}p(c)\prod_{j=1}^{n}\sum_{d=1}^{m}\text{trans}(d|c)\frac{1}{\sqrt{2\pi\sigma_w^2}}\exp\left[\frac{-(y_{ij}-\mu-\mu_d-u_i)^2}{2\sigma_w^2}\right].du_i$$

However, to allow some flexibility so the summation can be taken around a value other than zero, and the variance of the parameter altered, the transmitting ability (u_i) can be transformed:

$$x_i=\frac{u_i\text{-}uc_i}{V_i}$$

where: uc_i are the location parameters

V_i are the scaling parameters

and the likelihood can be rewritten as follows:

$$L(MM)=\prod_{i=1}^{s}\int_{-\infty}^{+\infty}\frac{\sqrt{2\pi V_i^2}}{\sqrt{2\pi\sigma_u^2}}\exp\left[-\frac{u_i^2}{2\sigma_u^2}+\frac{x_i^2}{2}\right]\frac{1}{\sqrt{2\pi V_i^2}}\exp\left[-\frac{x_i^2}{2}\right]$$

$$\sum_{c=1}^{m}p(c)\prod_{j=1}^{n}\sum_{d=1}^{m}\text{trans}(d|c)\frac{1}{\sqrt{2\pi\sigma_w^2}}\exp\left[\frac{-(y_{ij}-\mu-\mu_d-u_i)^2}{2\sigma_w^2}\right].du_i$$

Rewriting the transform $u_i=V_ix_i+uc_i$

and hence $du_i=V_i.dx_i$

then changing the integration over u_i to an integration over x_i gives the following equation for the mixed model likelihood:

$$L(MM) = \prod_{i=1}^{s} \int_{-\infty}^{+\infty} \frac{\sqrt{2\pi V_i^2}}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{(V_i x_i + u c_i)^2}{2\sigma_u^2} + \frac{x_i^2}{2} \right] \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2} \right]$$

$$\sum_{c=1}^{m} p(c) \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d|c) \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[\frac{-(y_{ij}-\mu-\mu_d-(V_i x_i + u c_i))^2}{2\sigma_w^2} \right] . dx_i \qquad [4.2]$$

Finally replacing the integration over x_i with a summation gives the following mixed model likelihood:

$$MML = \prod_{i=1}^{s} \frac{\sqrt{2\pi V_i^2}}{\sqrt{2\pi\sigma_u^2}} \sum_{g=1}^{G} \left(\sum_{c=1}^{m} p(c) \exp\left[-\frac{(V_i x_g + u c_i)^2}{2\sigma_u^2} + \frac{x_g^2}{2} \right] \right.$$

$$\left. \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d|c) \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[\frac{-(y_{ij}-\mu-\mu_d-(V_i x_g + u c_i))^2}{2\sigma_w^2} \right] \right) W_g \qquad [4.3]$$

The integration in [4.2] is to be taken over x_i. The Hermite polynomial is appropriate when c(x) [4.1] is of the form $\exp[-x^2]$, whereas here the variable to be integrated is $\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2} \right]$ hence the abscissae obtained from standard tables (e.g. Selby, 1970) should be multiplied by $\sqrt{2}$ and the weights divided by $\sqrt{\pi}$. This approximation will be denoted Herm.

4.2.2 Effect of transforming the abscissae.

To investigate the effect of different scaling and location parameters on the number of points in the summation required to obtain a reasonable approximation a simple polygenic model will be considered initially.

Polygenic model

Data were simulated under a polygenic model with balanced structure. An exact expression for the maximum likelihood of the data under this polygenic model can be written in terms of the between and within sire components of variance as follows (Searle, 1971):

$$\ln L(\text{poly}) = \frac{1}{2}\left[sn \ln(2\pi) + s(n-1) \ln\sigma_w^2 + s \ln(\sigma_w^2 + n\sigma_u^2) + (s-1) + s(n-1)\right] \qquad [4.4]$$

An equation for the likelihood under a polygenic model using Hermite integration can be obtained from [4.3] by fixing $\mu_1 = \mu_2 = \mu_3 = 0$. To calculate the likelihood, values for the mean (μ) and the variance components (σ_w^2, σ_u^2) were obtained from the data. The integration is over the sire's transmitting ability and information about this parameter can be obtained by considering the observations on his progeny. Four cases were considered, representing different use of the information on the distribution of transmitting abilities:

Parameter	Case 1	 2	 3	 4
location parameter (uc_i)	0	$\hat{u}_i$	0	$\hat{u}_i$
scaling parameter (V_i)	σ_u	V*	V*	σ_u

Where: $\hat{u}_i = \dfrac{n(\bar{y}_{i.} - \mu)}{n+\lambda}$ $\bar{y}_{i.}$ - progeny mean $\lambda = \dfrac{\sigma_w^2}{\sigma_u^2}$ $V^* = \sqrt{\dfrac{\sigma_w^2}{(n+\lambda)}}$

The first case ignores all the information about the effect of individual sires, and the abscissae are taken around a mean of zero, this being the mean of the transmitting ability distribution assumed in the model. In the second case, the location parameters, around which the abscissae are placed, are the transmitting abilities for each sire ($\hat{u}_i$). With this value for the location parameter it can be shown that the approximation is exact with one point in the summation and $V^2 = \frac{\sigma_w^2}{(n+\lambda)}$. Case 3 uses zeros as the location parameters and the scaling parameter from case 2. Case 4 uses the transmitting abilities for each sire as the location parameter and the square root of the sire variance component as the scaling parameter. For each case a range of the number of points in the summation was tried.

Table 4.1 gives the natural log likelihoods as a deviation from the true log likelihood calculated using [4.4]. As expected, as the number of points in the summation increases the likelihood approximated using Hermite integration approaches the value of the exact likelihood. When the transmitting abilities were used it was possible, with the correct scaling, to obtain the exact likelihood with one point. Even when the scaling parameter was not the most appropriate for the location parameters used (cases 3 and 4) the likelihood calculated with Hermite integration was close to the exact with only 10 or 16 points in the summation although not very good with only a few points. Lack of information about the parameters can be overcome by increasing the number of points in the summation.

Table 4.1 *Polygenic log likelihood using Hermite integration with different location, scaling parameters and number of points in the summation expressed as a deviation from the exact log likelihood.*

Number	Case			
Points	1	2	3	4
32	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	-0.0011
5	0.0375	0.0000	-0.0509	0.2662
3	-0.0997	0.0000	-1.2312	2.3302
1	-7.1383	0.0000	-24.4325	17.2942

Mixed model

The mixed model likelihood cannot be calculated by considering the variance components, however, as shown for the polygenic model, as the number of points in the summation used in Hermite integration increases the approximation to the likelihood should near the exact value. To investigate the number of points in the summation required to obtain a reasonable approximation, data were simulated under a mixed model. The expected values were used for the major genotype means and frequencies and the variance components. Four different combinations of scaling and location parameters were used, equivalent to those tried for the polygenic model. For a mixed model the mean of the progeny cannot be used as a measure of the sire's polygenic transmitting ability, as the mean will also include the major gene component. To see if an improvement in the approximation might be obtained if an estimate of the polygenic transmitting ability was available, the mean of the simulated phenotypic values for the offspring with the effect of their major genotype subtracted, were used where the progeny means had been used in the polygenic model (in cases 2 and 4). The results are given in table 4.2, with the likelihood given relative to that calculated using 32 points in the summation.

Table 4.2 *Mixed model log likelihood using Hermite integration with different location, scaling parameters and number of points in the summation expressed as a deviation from the log likelihood calculated with 32 summations.*

Number	Case			
Points	1	2	3	4
32	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
10	0.0000	-0.0016	-0.0004	0.0030
5	-0.0135	-0.3239	-0.1482	0.0074
3	-0.1518	-2.3570	-1.5178	0.0954
1	-0.6837	-21.0727	-18.8714	-2.8850

There is no further improvement in the approximation by including more than 16 points in the summation, for all four cases considered. With fewer summations, taking account of the polygenic contribution of the sire does not improve the approximation. With a single point in the summation, using the 'true' value for the transmitting ability might make the sire very likely to be one of the homozygotes, whereas if this information was ignored this genotype would not be more likely and information would also come from the other genotypes.

Although for a large pedigree the exact mixed model likelihood is too computationally demanding, for a small pedigree the exact likelihood can be calculated using equation [2.3]. Data were simulated for 10 half-sibs from each of 20 sires. The exact mixed model likelihood was calculated and compared with the approximate likelihood calculated using Hermite integration with 20 points in the summation. Several different parameter values were tried and in all cases the two likelihoods were the same to 8 significant figures. Hence, using Hermite integration with 20 points in the summation is effectively the same as using the exact likelihood.

4.3 MAXIMISATION

4.3.1 Algorithm

The likelihood was maximised explicitly using a quasi-Newton algorithm (E04JAF) from the NAG library (Numerical Algorithms Group, 1988). This algorithm is based on the

Newton-Raphson algorithm which can be derived from a Taylor expansion. Denoting the parameters to be estimated θ, where θ_a is the value of these parameters at iteration a, the following equation can be obtained:

$$\frac{\partial \ln L(\theta_a)}{\partial\theta}=\frac{\partial \ln L(\theta_{a-1})}{\partial\theta}+(\theta_a-\theta_{a-1})\frac{\partial^2 \ln L(\theta_{a-1})}{\partial\theta'\partial\theta}+\cdots=0$$

Ignoring quadratic and higher order terms of $(\theta_a-\theta_{a-1})$ the equation can be rearranged to give the following algorithm for maximisation:

$$\theta_a=-\left(\frac{\partial^2 \ln L(\theta_{a-1})}{\partial\theta'\partial\theta}\right)^{-1}\frac{\partial \ln L(\theta_{a-1})}{\partial\theta}+\theta_{a-1} \qquad [4.5]$$

This requires both the vector of partial 1st derivatives of the log likelihood with respect to the parameters θ (gradient vector) and the Hessian matrix of partial 2nd derivatives of the log likelihood with respect to the parameters. The quasi-Newton algorithm used approximates the gradient vector by finite difference and at each iteration the approximation to the Hessian matrix is updated and improved.

The algorithm minimises a function and so in this case the negative log likelihood for the mixed model is minimised which is equivalent to maximising the likelihood. A routine that calculates the function to be minimised for given parameter values needs to be supplied along with initial parameter estimates from which the minimisation process starts.

The rate of convergence can be improved by scaling the parameters, ideally so that a unit change in the parameter causes a unit change in the function value (Numerical Algorithms Group, 1988). In the mixed model the genotype frequencies are obviously on a different scale to the means and variances. To improve this situation and also to prevent the necessity for constraining the parameters with bounds, the parameters can be transformed. If the population is assumed to be in Hardy-Weinberg equilibrium with respect to the major gene, then only a single allele frequency (p(A)) is required to be estimated in order to describe the frequencies of the major genotypes. Using a logistic transformation of the frequency, $p(A)^* = \ln\left(\frac{p(A)}{1-p(A)}\right)$, enables the parameter to be estimated to take any value between negative and positive infinity, while the allele frequency is effectively constrained to have a value between zero and one. To convert back to the original scale the transform required is $p(A) = \frac{\exp[p(A)^*]}{1+\exp[p(A)^*]}$.

The assumption of the population being in Hardy-Weinberg equilibrium with respect to the major gene can be relaxed and genotype frequencies estimated for the founders. In this case the following reparameterisation can be made to constrain each of the frequencies to be between zero and one, while also constraining the sum of the frequencies to be one.

Frequency	Reparameterisation
freq(AA)	x_1
freq(Aa)	$(1-x_1)x_2$
freq(aa)	1-freq(AA)-freq(Aa)

Where x_1 and x_2 can take values between zero and one. The parameters x_1 and x_2 can be treated in the same way as the allele frequency and transformed using a logistic transformation (giving x_1^* and x_2^*). The transformed parameters are estimated and can be converted back into the genotype frequencies for the likelihood calculation using the following transformations:

Frequency	Transformation
freq(AA)	$\frac{\exp[x_1^*]}{1+\exp[x_1^*]}$
freq(Aa)	$\frac{\exp[x_1^*]}{(1+\exp[x_1^*])(1+\exp[x_2^*])}$
freq(aa)	1-freq(AA)-freq(Aa)

To constrain the variance estimates to be positive the square roots of the variance components (standard deviation) are estimated.

4.3.2 Initial parameter estimates

A small simulation study was undertaken to determine how sensitive the maximisation routine was to the initial parameter estimates used to start the process. Data were simulated under a mixed model containing 20 half-sib progeny from each of 50 sires with all parents unrelated and randomly mated. The additive polygenic variance was equal to one quarter the environmental variance. The major locus had two alleles at equal frequency with effect equal to 2 within major genotype standard deviations between the homozygotes and with additive action. Initially the population was assumed to be in Hardy-Weinberg equilibrium and a single allele frequency was estimated, then the analyses were repeated estimating the genotype frequencies in the sire population and an allele frequency for the dams. The polygenic heritability was assumed to be known and fixed at the expected value (h^2_{poly}). Five different starting places were tried as follows:

Model	p(A)	μ_{AA}	μ_{Aa}	μ	σ^2_w
1	0.5	$2\sqrt{V_w+V_u}$	$\sqrt{V_w+V_u}$	$\bar{y}_{..}-\mu_{Aa}$	V_w^*
2	0.5	$2\sqrt{V_w+V_u}$	$\sqrt{V_w+V_u}$	$\bar{y}_{..}-\mu_{Aa}+5$	V_w^*
3	0.5	$2\sqrt{V_w+V_u}$	$\sqrt{V_w+V_u}$	$\bar{y}_{..}-\mu_{Aa}-5$	V_w^*
4	0.5	$\sqrt{V_w+V_u}$	$0.5\sqrt{V_w+V_u}$	$\bar{y}_{..}-\mu_{Aa}$	V_w^*
5	0.5	0	0	$\bar{y}_{..}$	Vw^*

Where: V_w is the within sire variance component from an analysis of variance

V_u is the between sire variance component from an analysis of variance

$\bar{y}_{..}$ is the mean of the phenotypes

$$V_w^* = (V_w + V_u - V_{mg})\frac{1-h^2_{poly}}{4}$$

$$V_{mg} = 2p(A)\ (1-p(A))\ \mu^2_{Aa}$$

The simulation and analyses were repeated ten times.

Results

If the maximisation process was started from a polygenic model, the end values were also a polygenic model. When assuming that the population was in Hardy-Weinberg equilibrium and estimating an allele frequency, for all but two of the simulated data sets all the analyses converged to the same maximum. For these two sets of data, one model (model 2 in one case and model 3 in the other) resulted in a maximum that was just more likely than the others which all gave the same result. In one of these cases both of the mixed models obtained were a significant improvement over the polygenic model. The parameter estimates and likelihood as a deviation from the polygenic model are given in table 4.3.

When genotype frequencies were estimated, the maximisation process was more sensitive to the initial values of the parameters. Two of the data sets resulted in the same maximum for all four of the initial models tried. For the other simulations, several local maxima were obtained. More than one set of parameter estimates could be obtained with a similar likelihood. Examples of the resulting models and their likelihoods are given in table 4.3.

Table 4.3 *Alternative estimates obtained using different parameter values to start the maximisation process.*

		Sire		Dam/Popn.						
Analysis	p(AA)	p(Aa)	p(aa)	p(A)	μ_{AA}	μ_{Aa}	μ	σ_u^2	σ_w^2	ln L
AF				0.376	17.383	15.264	31.292	4.831	91.788	11.35
				0.703	26.064	11.729	23.042	4.052	76.980	11.74
AF				0.378	3.211	13.243	34.436	5.561	105.663	0.46
				0.309	8.756	7.321	37.188	6.851	130.165	0.29
GF	0.528	0.260	0.212	0.184	1.543	14.499	32.955	5.010	95.197	8.13
	0.681	0.210	0.108	0.378	21.811	9.761	29.784	4.809	91.368	7.70
GF	0.249	0.562	0.189	0.100	20.591	11.627	34.609	5.651	107.370	7.37
	0.184	0.520	0.295	0.053	6.240	-12.772	47.305	5.664	107.607	7.64
GF	0.095	0.015	0.890	0.316	21.481	11.514	36.233	5.326	101.201	7.78
	0.217	0.699	0.085	0.565	19.860	6.892	31.303	4.463	84.804	4.65

ln L - mixed model likelihood as a deviation from the polygenic likelihood.

AF - analysis maximising a population allele frequency.

GF - analysis maximising sire genotype frequencies for and dam allele frequency.

4.4 SIMULATION STUDY

4.4.1 Data

To investigate the approximation, data were simulated using the Fortran program described in Appendix 2. Phenotypes of 20 half-sib progeny from each of 50 sires were simulated, with all parents unrelated and randomly mated. The phenotypes were composed of a polygenic component, an individual environmental component and a major gene. The polygenic component comprised of 24 unlinked loci, with equal effect. At each locus there were two alleles at equal frequency and with an additive effect. Four different models were considered for the major locus, in each there were two alleles (A and a) in the following models:

Model	h^2_{poly}	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}
Add1	0.2	0.5	2	1	0
Add2	0.4	0.5	2	1	0
Dom	0.2	0.5	2	2	0
Rare	0.2	0.2	2	1	0

Where: $h^2_{poly} = \dfrac{4\sigma_u^2}{\sigma_u^2+\sigma_w^2}$

p(A) is the frequency of the high scoring allele in the parent population.

μ_d is the effect of genotype d in units of within major gene standard deviations $\left(\sqrt{\sigma_u^2+\sigma_w^2}\right)$.

4.4.2 Analyses

Hermite integration was used with 20 points in the summation with the abscissae located around zero and using the square root of the sire variance estimate as the scaling parameter (σ_u), i.e. case 1 from table 4.2.

In the analysis the mean effect of the low scoring homozygous genotype (μ) at the major locus and the deviation from this mean of the other two major genotype means (μ_{AA} and μ_{Aa}) were estimated. The population was assumed to be in Hardy-Weinberg equilibrium and hence an allele frequency (p(A)) was estimated. Two analyses of each data set were carried out, the first assuming that the polygenic heritability was known, and estimating just the residual variance, and the second estimating the polygenic heritability as well as the residual variance.

As described above parameter estimates are required from which the maximisation process can start. If these are close to the global maximum, convergence to this maximum is more likely to be obtained. With the assumption of the population being in Hardy-Weinberg equilibrium the method has been shown not to be very sensitive to the initial parameter estimates (see section 4.3.2). In these simulations the expected parameters are known and hence these were used as initial estimates. In practice this would not be the case and several starting points should be tried in order to confirm that the global maximum has been attained.

Genotyping at the major locus

For the knowledge that a major gene exists to be of use, not only are good estimates of the parameters involved required, but also the identification of the genotype of each individual. To obtain an indication of how reliable the method is at genotyping individuals the probability of each genotype for each sire was calculated. Equation [2.5] gave this probability for the exact mixed model, however, in the same way as the likelihood has been rewritten by replacing the integration with a summation, this probability can also be rewritten:

$$q_i(c) = \frac{\sum_{g=1}^{G} p(c)\exp\left[-\frac{(V_i x_g + u c_i)^2}{2\sigma_u^2} + \frac{x_g^2}{2}\right] \prod_{j=1}^{n}\sum_{d=1}^{m} \mathrm{trans}(d|c)\exp\left[\frac{-(y_{ij}-\mu-\mu_d-(V_i x_g + u c_i))^2}{2\sigma_w^2}\right].W_g}{\sum_{g=1}^{G}\sum_{c'=1}^{m} p(c')\exp\left[-\frac{(V_i x_g + u c_i)^2}{2\sigma_u^2} + \frac{x_g^2}{2}\right] \prod_{j=1}^{n}\sum_{d=1}^{m} \mathrm{trans}(d|c')\exp\left[\frac{-(y_{ij}-\mu-\mu_d-(V_i x_g + u c_i))^2}{2\sigma_w^2}\right].W_g}$$

[4.6]

After maximisation of the likelihood, this probability can be calculated using the ML estimates for the parameters. As the sires are assumed to be independent, the genotype that is most likely for each sire will also be the most likely for that sire when considering all the sires together. Although expressions for the genotype probabilities for the offspring can be obtained, these will be less reliable as they are composed of the probabilities of the genotypes of their sire and the deviation of their own phenotype from the major gene means.

4.4.3 Test statistic

A test for the presence of a segregating major gene is the likelihood ratio test [2.4]. Under the null hypothesis of no major gene component, this test statistic is expected asymptotically to follow a χ^2 distribution with three degrees of freedom, as three parameters are estimated in the mixed model but fixed in the polygenic model (p(A), μ_{AA}, μ_{Aa}). To confirm this expected distribution data were simulated under a polygenic model. The simulation program described in Appendix 2 was used with the genetic component being controlled by 25 loci with equal effect and no linkage between them. At each locus there were two alleles, with additive effect and equal frequency. There was also an individual random component giving a heritability of 0.2 or 0.4, these being the polygenic heritabilities in the mixed model simulations. 100 replicates of each model were simulated.

In the same way as for the analyses of mixed model data each data set was analysed assuming that the heritability was known and fixing it at the value used in the simulation, and then repeated estimating the heritability. For the mixed model analyses initial estimates assumed that the major gene explained half of the total variance. The MLs of the data under the polygenic and mixed models were obtained and minus twice the natural log of the likelihood ratio calculated.

Results

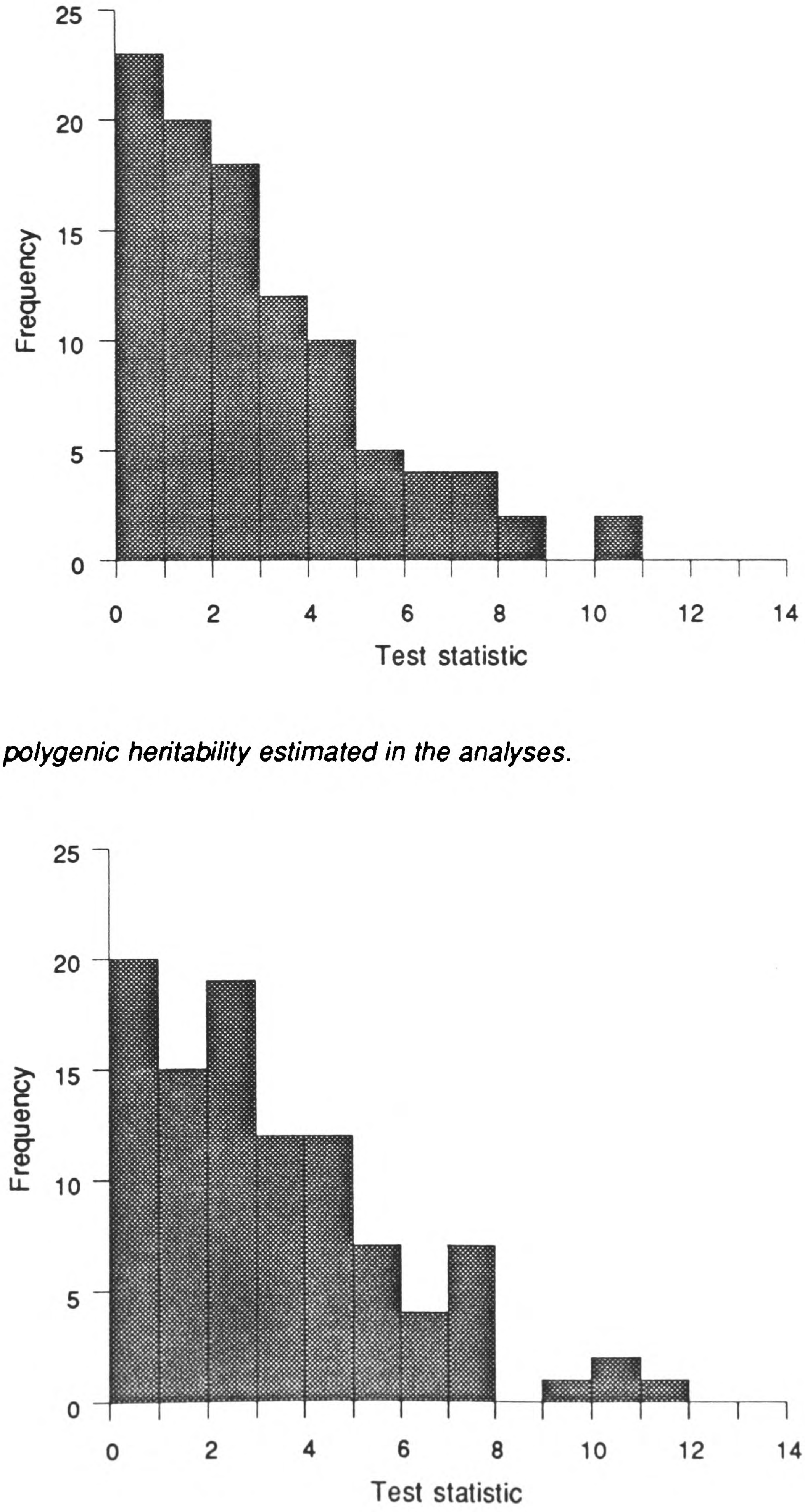
The distribution of test statistics for the data with heritability 0.2 and 0.4, analysed assuming the heritability was known and estimating the heritability are shown in figures 4.1 and 4.2. The expected distribution, a χ^2 distribution with three degrees of freedom, is shown in figure 4.3.

Table 4.4 *Mean and variance of the test statistic and the number of analyses where the test statistic was significant at the 5% and 1% significance levels of a χ^2 distribution with 3 degrees of freedom.*

		Mean	Variance	5%	1%
Expected	(χ^2 3 d.f.)	3		5	1
h^2=0.2	fixed	2.886	5.430	5	0
h^2=0.2	estimated	3.352	6.321	6	0
h^2=0.4	fixed	3.056	8.975	5	3
h^2=0.4	estimated	3.483	10.470	8	4

Based on 100 simulations of each genetic model.

Figure 4.1 *Distribution of the test statistic from analyses of polygenic data with an expected heritability of 0.2.*

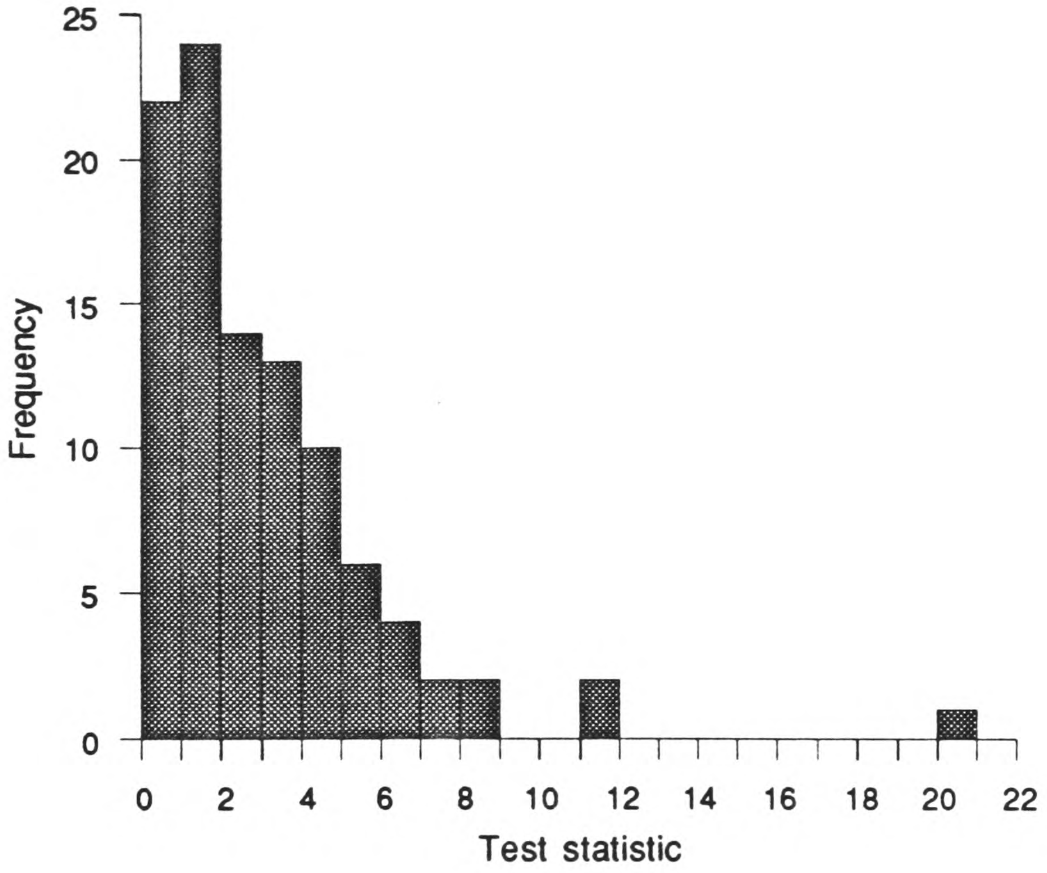


a) With the polygenic heritability assumed to be known in the analyses.

b) With the polygenic heritability estimated in the analyses.

Figure 4.2 *Distribution of the test statistic from analyses of polygenic data with an expected heritability of 0.4.*

a) With the polygenic heritability assumed to be known in the analyses.



b) With the polygenic heritability estimated in the analyses.

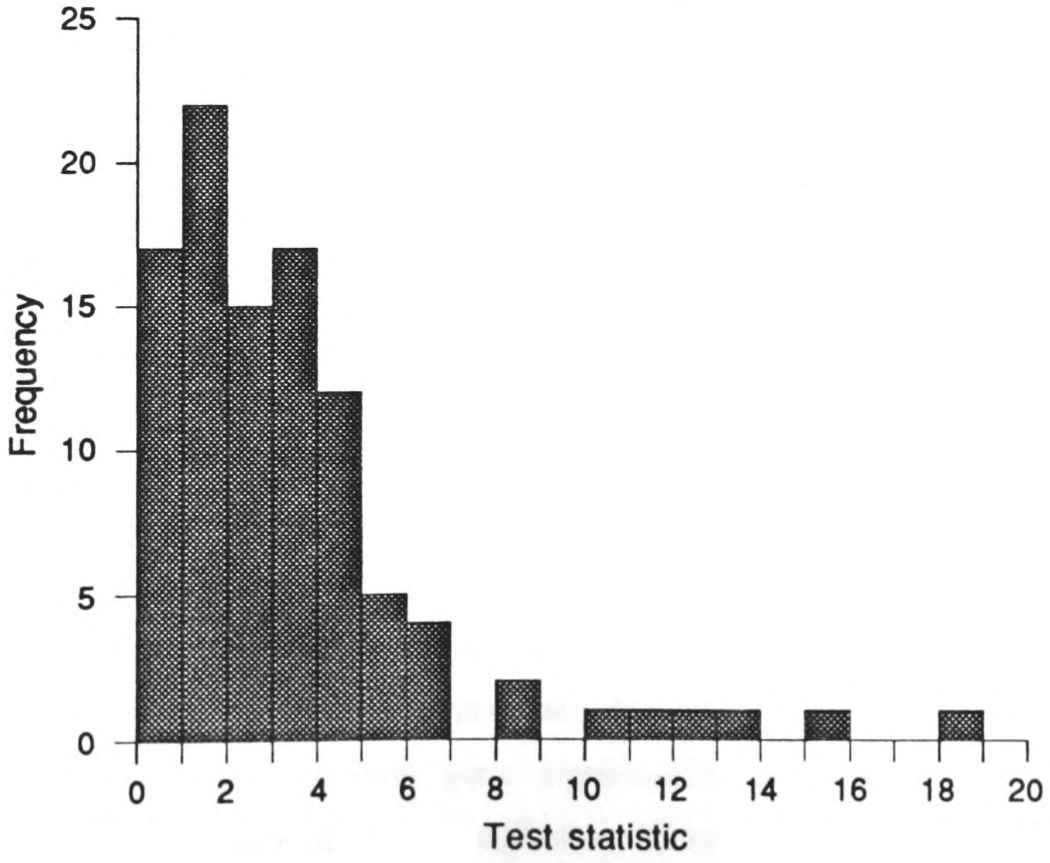
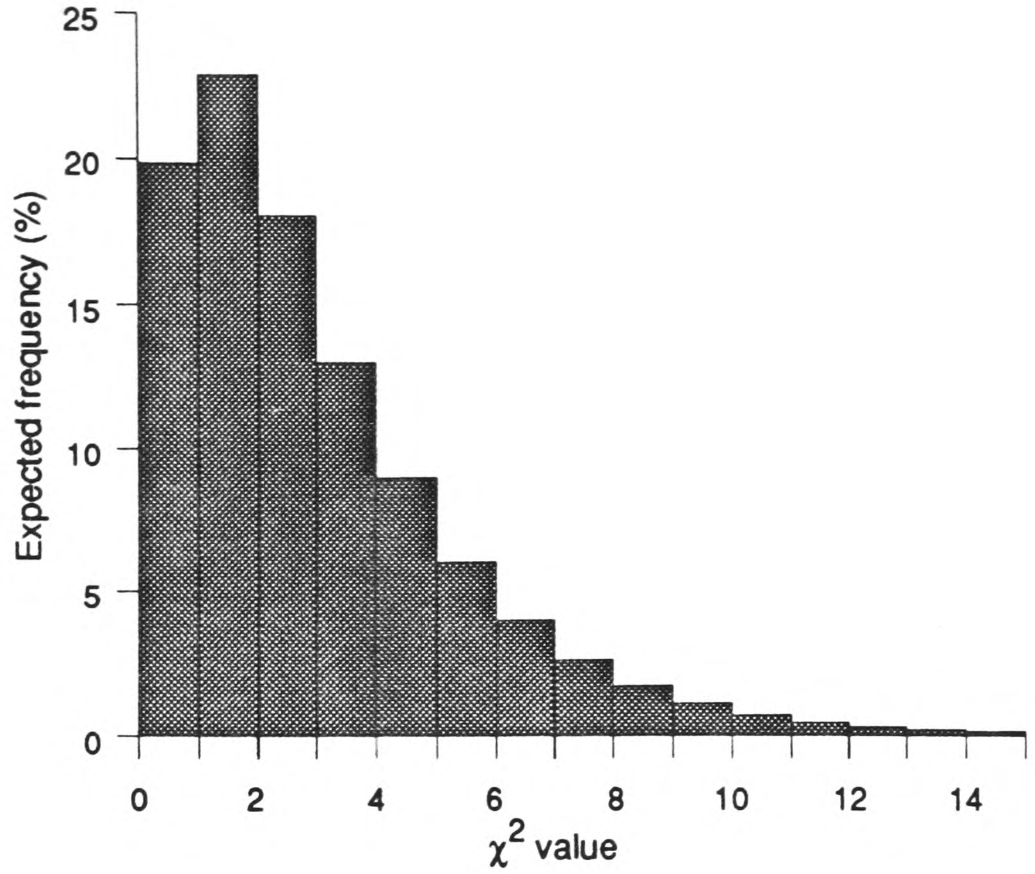


Figure 4.3 χ^2 *distribution with 3 degrees of freedom.*



The mean and variance of these distributions and number of analyses giving significant test statistics at the 5% and 1% significance levels of a χ^2 distribution with three degrees of freedom are given in table 4.4. Estimating the heritability increases the mean and variance of the distribution compared with fixing it at the expected value, also with higher polygenic variance the mean and variances of the test statistic distributions were higher. There is a linear relationship between the test statistic obtained with the heritability fixed compared with that obtained with the heritability estimated for the same set of data, the slope of the regressions being about 0.65 and the correlations 0.70.

As a test for the presence of a segregating major gene the upper tail of the distribution obtained from data simulated under a polygenic model is of interest. However with only 100 analyses good estimates of suitable 5% and 1% quantiles can not be obtained. An indication of whether the observed distribution follows a χ^2 distribution can be obtained by calculating the observed number of test statistics that fall within 10 equal classes of a χ^2 distribution, and comparing these with the expected distribution using a χ^2 test. Table 4.5 shows the distribution of observed test statistics for the four analyses compared with a χ^2 distribution with two, three and four degrees of freedom. None of the observed distributions were significantly different from a χ^2 distribution with three degrees of freedom. The distribution of the test statistic obtained analysing the data with an expected polygenic heritability of 0.4, estimated in the analyses, was also just not significantly different from a χ^2 distribution with four degrees of freedom at the 5% level. However, based on the χ^2 (9 d.f.) statistic, the distribution more closely resembled a χ^2 with three degrees of freedom.

Table 4.5 *Number of observed test statistics falling within 10% groups of a χ^2 distribution with 2, 3 or 4 degrees of freedom.*

Proportion		Analyses - polygenic heritability											
of χ^2		0.2, fixed			0.2, estimated			0.4, fixed			0.4, estimated		
distribution	Expected	2df	3df	4df	2df	3df	4df	2df	3df	4df	2df	3df	4df
0.0-0.1	10	19	10	5	26	15	6	22	10	5	20	10	8
0.1-0.2	10	16	9	7	17	11	9	15	11	6	23	9	4
0.2-0.3	10	12	10	6	13	14	9	13	11	8	13	15	6
0.3-0.4	10	13	10	8	13	6	10	8	8	8	10	15	9
0.4-0.5	10	11	9	7	5	11	8	8	10	9	10	7	16
0.5-0.6	10	5	12	6	6	12	6	12	7	6	7	10	8
0.6-0.7	10	6	10	15	7	5	14	4	6	8	5	10	8
0.7-0.8	10	5	7	11	7	6	10	13	15	9	7	7	13
0.8-0.9	10	7	8	11	4	10	7	2	9	17	3	7	11
0.9-1.0	10	6	15	24	3	10	21	3	13	24	2	10	17
χ^2(9 d.f.)		22.2	4.4	30.0	46.5	10.4	17.7	34.8	6.6	31.6	43.4	7.8	16.0

Discussion

From these 100 simulations there is no evidence to suggest that the test statistic distribution does not follow a χ^2 distribution with degrees of freedom equal to the number of parameters estimated in the mixed model but fixed in the polygenic model. That is, the distribution suggested by Wilks (1938) appears to hold when comparing the mixed and polygenic models.

In the context of segregation analysis Elsen and Le Roy (1989) interpret the degrees of freedom in Wolfe's modified likelihood ratio (Wolfe, 1971) to be equal to twice the minimum number of parameters that need to be fixed in the mixed model in order to obtain the polygenic model. In the simulations presented here only one parameter needs to be fixed, the frequency of allele A in the population (p(A)). Hence a χ^2 distribution with two degrees of freedom would be used. This is not supported by the simulations. An alternative interpretation, using twice the difference in the number of component distributions (Wolfe, 1971) would suggest testing the likelihood ratio with a χ^2 distribution with four degrees of freedom. This is also not supported by the simulation.

Le Roy *et al.* (1989) and Elsen and Le Roy (1989) also look at the likelihood ratio distribution for this likelihood (called SA in their notation). Le Roy *et al.* (1989) consider data containing 5, 10 or 20 half-sib progeny from each of 5, 10 or 20 sires, when assuming that the polygenic heritability is known and fixing it at the expected value, and just the largest data set when estimating the heritability. With fixed heritability, based on 500 simulations of each data structure, the mean of the test statistic distribution was remarkably constant, about 4.5, over all the situations. Their analyses estimate the major genotype frequencies for the sires and the allele frequency for the dams, and hence using the constraint $\mu_1 = \mu_2 = \mu_3$ the observed distribution is compared with a χ^2 distribution with four degrees of freedom (this being twice the minimum number of parameters required to be fixed in the mixed model to obtain the polygenic model). Although significantly different from this distribution, the mean of 4.5 suggests that the correct distribution would have degrees of freedom somewhere between 4 and 5, the latter also being the distribution suggested by Wilks (1938). This mean might increase if the surface was searched more thoroughly for a maximum. The large number of simulations enables the 5% and 1% quantiles to be estimated from the data. For the largest simulation (s=20, n=20) these are 11.25 and 15.65 (for 5% and 1% respectively) which are similar to values that would be obtained using a χ^2 distribution with five degrees of freedom (11.07 and 15.09). Although searching further might increase the mean of this distribution it is unlikely to affect the extreme values as any mixed model that is much more likely than the polygenic model should have been located already. With heritability estimated Elsen and Le Roy (1989) also found that increasing the heritability of the data increased the mean and variance of the test statistic distribution. Again, although the number of simulations was less, about 200, they estimated the 5% and 1% quantiles, which are similar to the values that would be obtained from a χ^2 distribution with five degrees of freedom.

Unfortunately, Elsen and Le Roy (1989) analysed different data when assuming that the polygenic heritability was known compared with when estimating it. Therefore, the effect of making the assumption is difficult to ascertain. Their results show a decrease in the mean and standard deviation of the test statistic distribution when the heritability was estimated compared with when it was assumed to be known. However, this could be because the same data sets were not being considered in the two cases.

4.4.4 Simulation Results

Power

The number of analyses in which evidence of a major gene was found is summarised in table 4.6 along with the mean and standard deviation of the test statistic. When the simulated data contained a dominant major gene, a major gene was most easily detected, with evidence for its existence being found in all the analyses, at the 1% significance level (using a χ^2 distribution with three degrees of freedom). For the three additive major genes, when the heritability was fixed, the larger the proportion of genetic variance explained by the major gene simulated the more data sets in which evidence for a major gene was found. With the heritability estimated a major gene was detected in the highest proportion of analyses when the simulated gene had a rare allele. When the heritability was assumed to be known, and the expected value from the simulation used in the analysis, a major gene was detected in more analyses than when the heritability was estimated at the same time.

Table 4.6 *Mean and standard deviation of the test statistic and the number of analyses where the test statistic was significant at the 5% and 1% significance levels of a χ^2 distribution with 3 degrees of freedom.*

Model	mean	standard deviation	5%	1%
Fixed heritability				
Add1	12.800	6.838	75	59
Add2	7.012	4.855	36	16
Dom	47.278	14.719	100	100
Rare	12.043	7.244	65	41
Estimated heritability				
Add1	5.093	3.711	20	7
Add2	4.338	3.523	15	5
Dom	41.129	12.893	100	100
Rare	6.478	4.512	33	13

Based on 100 simulations of each genetic model.

The genetic models are described in section 4.4.1.

Parameter estimates

If the estimates from the analyses are unbiassed, the mean parameter estimates over the 100 simulations should give good estimates of the population parameters, i.e. those used to simulate the data. Table 4.7 gives the expected parameter values for the four mixed models simulated. Table 4.8 gives the average results for the major gene parameters, with the higher of the two homozygote means estimated defined as AA. The variance component estimates, residual variance only when the polygenic heritability is assumed to be known and both residual and sire when the heritability is estimated, are given in table 4.9. In general the parameter estimates were in good agreement with the expected values. Using a t-test the mean parameter estimates were tested against the expected values and significant tests are indicated in the tables. However, this test assumes that the true value is known without error which is not correct in this case because, in each simulation the realised value of the parameter will be different to the expected value due to sampling. Hence the estimated standard error of the difference between the 'true' and estimated value will be larger than that used in the test and the test will give more significant values than it should. The difference between the 'true' and estimated parameter for each analysis will also differ from that used but could either increase or decrease in size. However, using this test, the mean estimate for the effect of the high scoring homozygote, when the heritability was fixed, was significantly under estimated for the three additive models and in the model containing a rare major gene the frequency of the high scoring allele was significantly over estimated. This frequency was also over estimated when the polygenic heritability was estimated. When the data contained a dominant or rare major gene the residual variance was significantly under estimated, both when the heritability was fixed and estimated, and the contribution of the major gene was over estimated by a similar amount. When estimated, the mean sire variance component was not significantly different from the expected value for the four models. For all of the models some of the analyses (20 for Add1, 7 for Add2, 8 for Dom and 9 for Rare) went to a model with the genetic component being determined by a major gene only, the sire variance being equal to zero. None of the analyses gave a polygenic model.

Table 4.7 *Expected parameter values.*

Model	p(A)	μ_{AA}	μ_{Aa}	σ^2_u	σ^2_w
Add1	0.5	20	10	5	95
Add2	0.5	20	10	10	90
Dom	0.5	20	20	5	95
Rare	0.2	20	10	5	95

Table 4.8 *Mean (and standard deviation) of major gene parameter estimates.*

Model	p(A)		μ_{AA}		μ_{Aa}	
Fixed heritability						
Add1	0.495	(0.129)	18.747	(4.645)**	9.105	(5.162)
Add2	0.486	(0.162)	18.855	(5.507)**	9.521	(5.152)
Dom	0.504	(0.048)	20.694	(3.793)	20.150	(1.751)
Rare	0.265	(0.172)**	17.911	(7.682)**	9.866	(4.566)
Estimated heritability						
Add1	0.495	(0.164)	19.240	(6.133)	9.130	(6.621)
Add2	0.480	(0.160)	19.309	(5.444)	9.755	(5.332)
Dom	0.505	(0.049)	20.493	(3.881)	20.292	(1.918)
Rare	0.256	(0.167)**	18.721	(8.143)	10.833	(8.030)

* significantly different from expected value at 5% level

** significantly different from expected value at 1% level

Although the expected values of the variance components are the same for each data set, the realised values will differ because of sampling. Values for the residual and sire variance components were obtained by analysis of variance on the simulated phenotypes with the effect of the major genotype removed. Table 4.9 gives the relationship between these parameter estimates and the estimates obtained from segregation analyses for the same set of data. Although the effect of the major gene is the same for each simulation, some variation in the allele frequency is expected, hence

Table 4.9 *Variance estimates and the correlation with, and regression on, the values estimated in the simulation for the same data set.*

		Fixed heritability			Estimated heritability		
Model		σ^2_w	σ^2_u	σ^2_{mg}	σ^2_w	σ^2_u	σ^2_{mg}
Add1	mean	94.357	4.966	50.800	93.954	5.431	50.786
	sd	11.951	-	14.274	13.823	5.213	17.128
	slope	0.803	-0.078	1.626	0.576	0.886	2.564
	r	0.288	-0.232	0.167	0.184	0.321	0.221
Add2	mean	88.703	9.856	51.365	87.963	10.178	52.302
	sd	13.823	-	17.936	13.468	6.565	18.056
	slope	0.971	-0.114	2.350	0.876	0.744	0.715
	r	0.298	-0.200	0.230	0.277	0.305	0.071
Dom	mean	90.808**	4.779	80.441**	90.052**	5.028	80.925**
	sd	8.056	-	11.173	5.028	4.072	10.847
	slope	0.768	0.014	0.940	0.877	1.016	0.878
	r	0.465	0.071	0.443	0.491	0.570	0.425
Rare	mean	90.527**	4.765	34.552*	90.091**	4.997	35.083*
	sd	10.565	-	12.621	11.257	3.802	13.789
	slope	0.895	-0.042	1.394	0.891	1.031	1.401
	r	0.356	-0.155	0.313	0.333	0.561	0.288

also in table 4.9 the major gene variance has been estimated for each data set, using the formula:

$$\sigma^2_{mg} = p(A)^2\,\mu^2_{AA} + 2p(A)p(a)\,\mu^2_{Aa} - (p(A)^2\,\mu_{AA} + 2\,p(A)p(a)\,\mu_{Aa})^2 \qquad [4.5]$$

and compared with that estimated from the segregation analysis using the same formula. With fixed heritability, the estimates of the residual and major gene variance components have a positive association with the values estimated in the simulation. The slopes of the regressions are close to one for the residual variance, although the correlations are low. The major gene variance estimated with Herm has a fairly low correlation with the value estimated in the simulation for the same data set, and except for the data containing a major gene of dominant effect, the slopes of the regressions are higher than one. As expected, there is a very low or negative association between the sire variance obtained from Herm and the value estimated from the simulation. This is because this variance is not being estimated directly, but as a fixed proportion of the residual variance. In each data set the ratio of these two variances is not the same but will differ because of sampling, and as there is more information on the residual variance this is estimated fairly well at the expense of the sire variance. With the heritability estimated, the sire variance is now being estimated directly and, generally, the correlation with, and the regression on the values estimated from the simulation are closer to one than for the other variances. The residual variance for all genetic models, except for the data containing a gene of dominant effect, is less well estimated than when the heritability was fixed.

Genotyping sires at the major locus

The probability of each sire being each genotype was calculated using equation [4.6]. Considering the true genotype for each sire, the probability of the sire being this genotype was grouped into one of three classifications. The first, if the probability of being the correct genotype was greater than 0.9, the second greater than 0.75 and the third greater than 0.5. For each analysis the percentage of correctly genotyped sires was calculated for each genotype and the total percentage correctly genotyped over all genotypes. The results are given in table 4.10 as the mean percentage correctly genotyped over the 100 simulations.

Over all genotypes, when the major gene was additive with a rare allele, at a probability of 0.9, the highest number of sires were correctly genotyped. This is because of the high proportion of the common genotype, aa, sires which were correctly identified. If the criterion for allocating a genotype to sires was that the conditional probability for that genotype had to be greater than 0.5, the highest number of sires were correctly genotyped when the simulated major gene had an allele with a dominant effect. Fewest sires were correctly genotyped for the additive major gene segregating in

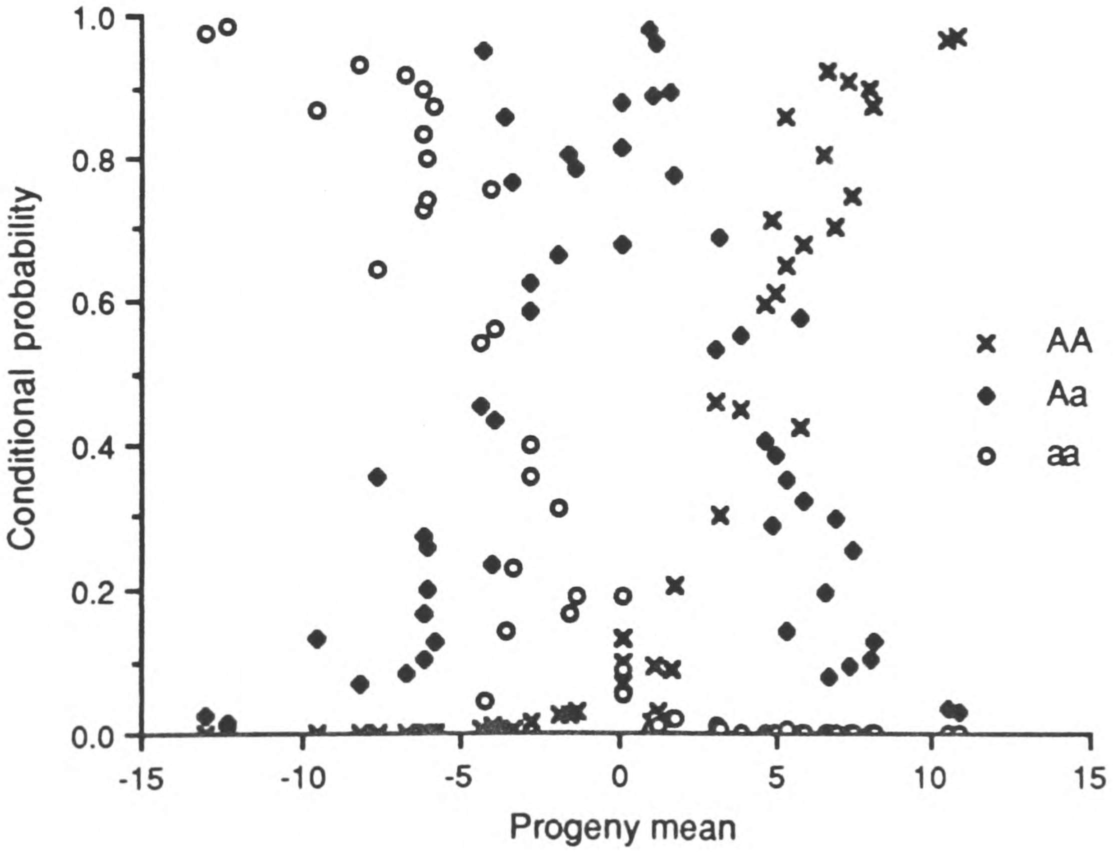
Table 4.10 Percentage of sires correctly genotyped at different values of the conditonal probability required to be assigned a genotype.

	AA			Aa			aa			Total		
Model	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5
Fixed heritability												
Add1	15.3	36.7	60.0	6.8	24.6	62.0	18.4	38.9	63.4	11.8	30.9	61.6
Add2	11.5	32.3	52.5	6.3	20.7	54.0	15.9	34.2	56.2	9.9	26.8	54.2
Dom	37.4	64.4	82.5	15.2	46.8	76.4	16.0	37.6	61.4	20.9	48.5	73.9
Rare	7.8	23.4	38.4	6.3	22.2	54.6	40.5	60.2	75.2	28.9	47.3	67.8
Estimated heritability												
Add1	23.4	43.0	60.5	12.0	28.2	56.9	24.2	42.0	59.7	17.7	35.1	58.2
Add2	14.7	33.9	52.4	8.1	23.0	53.8	18.5	36.9	57.5	12.3	28.9	54.4
Dom	40.1	65.0	8.3	16.9	47.0	76.1	19.1	38.3	61.6	23.2	48.9	73.5
Rare	13.6	25.1	37.8	10.0	25.9	53.8	44.0	63.4	78.9	32.5	50.5	69.7

a polygenic background with high heritability (0.4). Using 0.9 as the minimum value for the conditional probability for a sire to be allocated that genotype, for all models, a higher proportion of homozygous sires were correctly genotyped than heterozygous sires.

Figure 4.4 gives the relationship between the probability of the sire being each genotype and the mean of his offspring for one of the simulations with a major gene with additive effect and polygenic heritability of 0.2 (Add1). As expected, the sires whose progeny have the lowest mean have the highest probability of being genotype aa, and the sires with a high progeny mean have the highest probability of being genotype AA. At intermediate values of the progeny mean the sire could be any genotype. Figure 4.4 indicates that the conditional probability is dependent not only on the mean but that information about the distribution of phenotypes within families is taken into account.

<u>Figure 4.4</u> *Probability of the sire being each genotype, against the mean performance of the progeny of the sire.*



4.4.5 Discussion

The results given above are based on a single analysis of each data set. No attempt has been made to verify that the likelihood value obtained was the global maximum, and hence that the estimates were the ML estimates. However from previous investigation using this model (see section 4.3.2), at least with heritability fixed, the global maximum seemed relatively easy to obtain and starting from the expected estimates should mean that the maximum is near.

When the polygenic heritability was assumed to be known and fixed at the expected value, the ability of the method to detect a major gene depended on the proportion of the genetic variance which was explained by the major gene. The simulated major gene with dominant (Dom) effect accounted for 79% of the genetic variance and a major gene was detected in these data most frequently, whereas the additive major gene segregating in Add2 explained only 56% of the genetic variance and a major gene was detected in only 36 of the analyses (at the 5% significance level). When the heritability was estimated a major gene was detected more often when the simulated additive major gene had a rare allele than when a gene with alleles of the same effect as for the rare case but at equal frequency was simulated. One explanation for this is that although the gene with a rare allele contributes less to the genetic variance it causes the distribution of phenotypes to be skewed. The mean skewness of the 100 data sets containing the rare allele was 0.1185 and for the allele at a frequency of 0.5 was 0.0026 This skewed distribution can not be explained by the polygenic model and hence the mixed model is inferred. With the heritability fixed the skewness of the distribution appears to have less effect. The high detection rate for the major gene with dominant effect might also be partly explained by the data being skewed (mean skewness for the 100 data sets was -0.3171)

Le Roy *et al.* (1989) and Elsen and Le Roy (1989) give the percentage of analyses in which evidence for a major gene was found for models equivalent to Add1 and Dom (case 2 or h_{12} and case 1 or h_{11} respectively). The model for analysis, as explained previously, estimated the sire genotype frequencies and the dam allele frequency. Le Roy *et al.* (1989) compare the test statistic with 5% quantiles obtained by analysis of data simulated under a polygenic model. With fixed heritability the quantile decreases with decreasing number of sires and hence with few sires less evidence is required to detect a major gene. However with few sires and few offspring per sire the power of detection of a major gene was low. With 20 offspring from each of 20 sires the 5% quantile used was 11.25, which is similar to the value from a χ^2 distribution with five degrees of freedom.

When the simulated major gene had a dominant effect a major gene was detected in 90% of the analyses, and when the effect was additive a major gene was detected in 25% of the analyses, using the empirical quantile. When the polygenic heritability is estimated, for 20 half-sibs from each of 20 sires, the test statistic was compared with an estimated 5% quantile of 10.88 (Elsen and Le Roy, 1989). A major gene was detected in 81% of the analyses when the simulated data contained a dominant major gene segregating against a polygenic background with a heritability of 0.2 and in 86% of analyses when the simulated background heritability was 0.6. For the major gene simulated under an additive model, a major gene was detected in only 10% and 13% of the analyses with polygenic heritability of 0.2 and 0.6 respectively. The results reported here detect the major gene more often in the equivalent situations, which is probably due to the increased number of sires in the data. Elsen and Le Roy (1989) report an increase in the significant test statistics when the polygenic heritability was increased from 0.2 to 0.6. This is surprising, as the major gene will be explaining a lower proportion of the genetic variance with the high heritability. In agreement with the results presented here, Elsen and Le Roy (1989) find that estimating the heritability decreased the number of significant test statistics.

Analysing mixed model data with the polygenic heritability fixed at the value simulated, the polygenic likelihood is much less than the mixed model likelihood. This occurs in part because the fixed polygenic heritability poorly explains the total genetic variation, both major gene and polygenic. When the polygenic heritability is estimated, the difference between the polygenic and mixed model likelihoods is reduced, because an increased heritability in the polygenic model can explain some of the major gene variance. Thus the ratio of likelihoods when the heritability is estimated is smaller and this results in the major gene being detected less frequently. A corollary to this is that, if an underestimate of the polygenic heritability was used in the analyses with fixed heritability, a mixed model might be inferred, simply because the major gene can explain the additional polygenic variance.

As explained, the t-test is a poor criterion to judge the accuracy of the parameter estimates and is likely to give more significant results than it should. However, it gives an indication of any weakness in the estimation procedure. In general the mean parameter estimates were good. For the rare allele, on average, its frequency was over estimated and its effect under estimated, although there was a large variance in the estimates. On average, the major gene variance was over estimated and the residual was under estimated, both when the heritability was estimated and when it was assumed to be known. When analysing real data, there is, of course only one set of data and hence the accuracy of the estimates is important, not just a knowledge of bias. Hence we are interested in the estimates for each analysis and how these compare with the true parameter values. For the means, the variances of the estimates were large, and if just a single data set had been considered the estimate might have been some way from the correct value. When the residual and sire variance estimates were compared with the values obtained by analysis of variance on the polygenic and environmental contributions to the phenotypes for the same data set, there was a positive linear relationship, but the correlation between the ML estimates and the values estimated directly from the data was low. Hence an individual analysis may give misleading results. The parameter estimates of the major gene with an allele of dominant effect were closest to the true values with a low variance.

When estimated, the heritability was, on average (of the 100 analyses), very close to the expected value, although the variance of the estimates was high. This is in contrast to the results of Elsen and Le Roy (1989) who found that the heritability was under estimated. In some of the analyses presented here, a major gene model was obtained with the polygenic heritability equal to zero. This suggests that there is a problem in distinguishing the two sources of genetic variation. However the major gene model obtained gave a good indication as to the effect and frequency of the simulated gene.

The percentage of sires correctly genotyped was low if the criterion for classifying animals to a genotype was based on the conditional probability of that genotype being greater than 0.9. If a probability of 0.5 was required to genotype a sire, then all the models had greater than 50% of the sires correctly genotyped. However, at this probability, more sires will be incorrectly assigned to a genotype class.

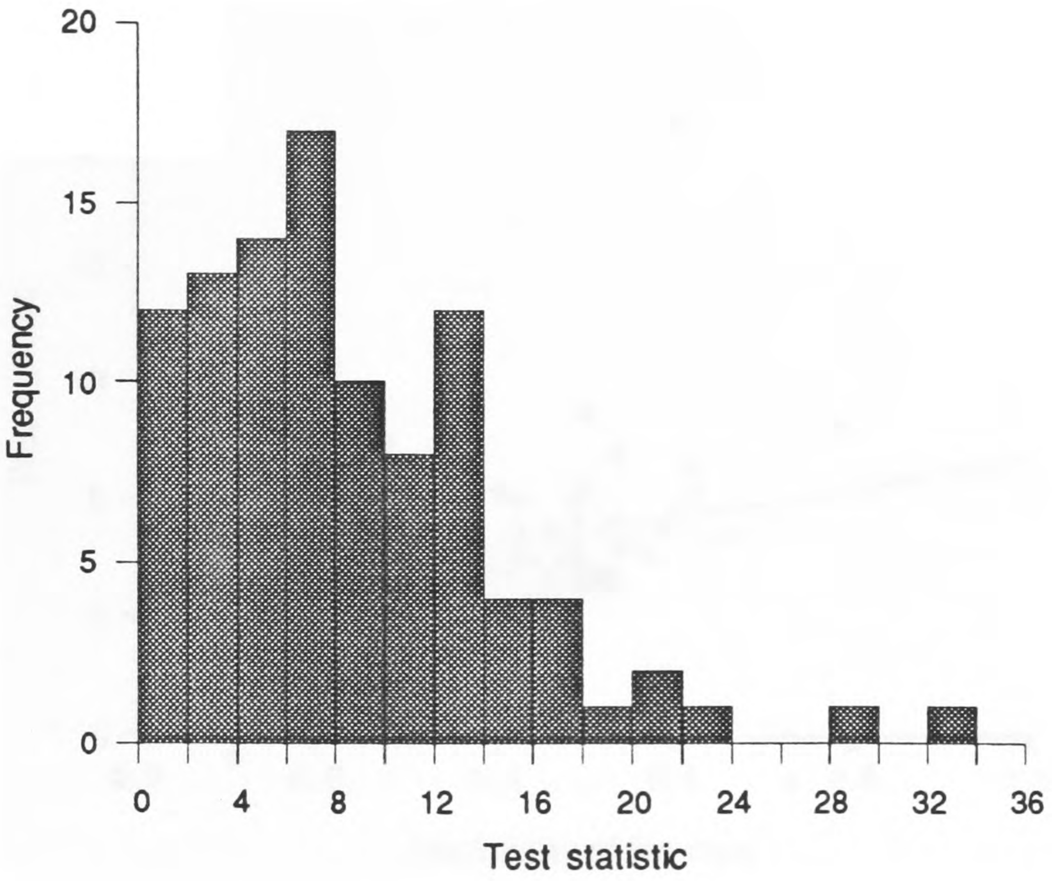
4.4.6 Incorrect estimate for the heritability.

In the analyses above it was found that a major gene was easier to detect if the polygenic heritability was fixed. In effect the major gene was being suggested to explain the extra genetic variance not explicable by the polygenic heritability. This gives concern that the method will not be robust to the assumed value of the polygenic heritability.

To investigate this, the polygenic data simulated with an heritability of 0.4 were reanalysed, this time fixing the heritability at 0.2. The analyses were the same as those carried out to confirm the test statistic distribution (section 4.4.3).

Figure 4.5 shows the test statistic distribution obtained. The mean of this distribution is 8.499 and the variance 35.77. Obviously these are much larger than expected for a χ^2 distribution with three degrees of freedom. Also 47 of the 100 simulations gave test statistics significant at the 5% level of a χ^2 distribution with three degrees of freedom and 28 significant at a 1% level. When the expected and observed frequencies falling within ten equal groups of the χ^2 distribution were compared, the χ^2 value (χ^2(9 d.f.)=295.2) was extremely significant, caused by the large number of high test statistic values.

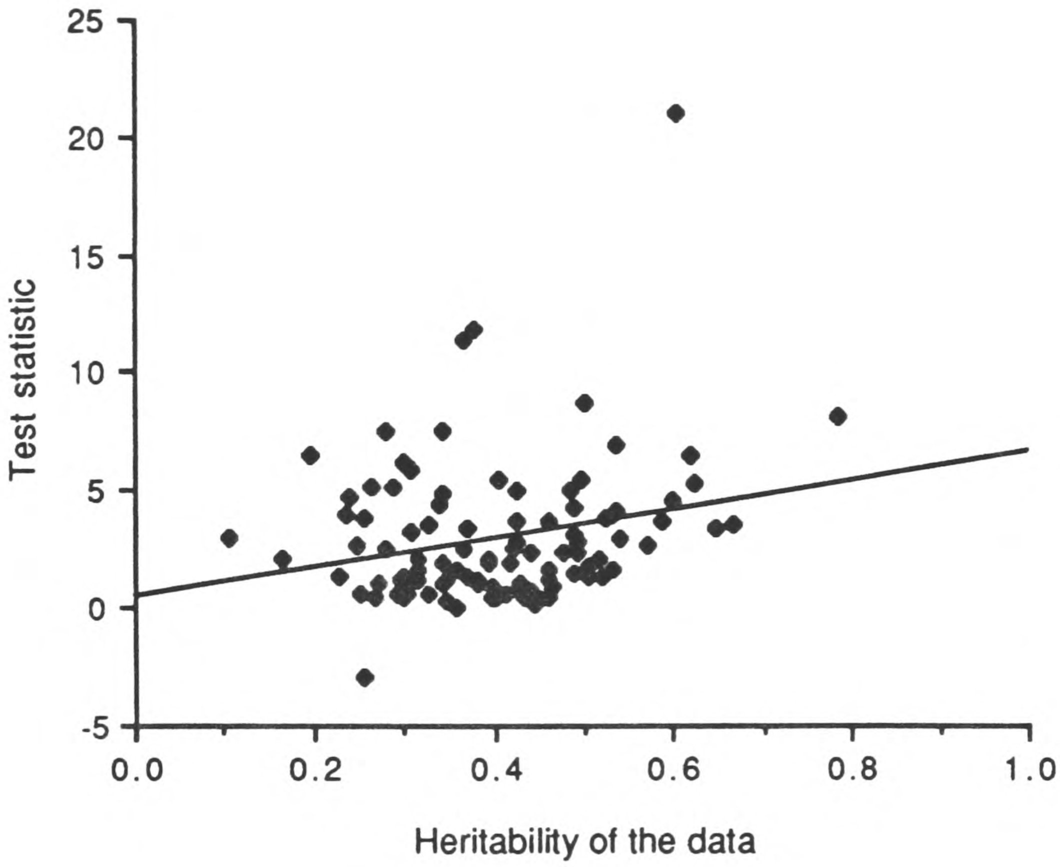
Figure 4.5 *Distribution of the test statistic from analysis of polygenic data with an expected heritability 0.4, analysed with a fixed value of 0.2.*



With the polygenic heritability assumed to be known the proportion of genetic to residual variance is fixed. If, as in this case, this is not the correct proportion, but an underestimate, then the mixed model is more likely than the polygenic, even though the data is in fact simulated under a polygenic model, because the major gene can explain some of the genetic variance which cannot be explained by the polygenic variance.

However this is an extreme case with the polygenic heritability estimate only explaining about one half of the genetic variance. Of more interest is the sensitivity of the method to small discrepancies in the value of the heritability used compared with the true polygenic heritability of the data. For the polygenic data simulated with a heritability of 0.4, although the expected value for each data set is 0.4 the true value will vary due to sampling. Figure 4.6 shows the relationship between the test statistic and the polygenic heritability of the data estimated by analysis of variance of the simulated polygenic plus environmental effects. There is a correlation of 0.23, hence, as the difference between the heritability of the data and the assumed value increases a major gene is more likely to be inferred.

Figure 4.6 *The test statistic from analyses of polygenic data with an expected heritability of 0.4, and fixed in the analyses, plotted against the heritability of the data estimated by analysis of variance.*



Finally the effect of the heritability estimate on the parameter estimates was investigated for data simulated under a mixed model. A sample of the data sets simulated with model Add1 were analysed with the heritability fixed at values from zero to one. The remaining parameters were estimated.

In the results, when possible the total variance, i.e. the sum of the major gene variance, the polygenic variance and the residual variance, was approximately equal to the total variance of the data. Hence as the polygenic heritability increased the resulting model was polygenic and as the heritability decreased the major gene effect became larger. Intermediate values gave mixed models. With high heritability estimates the total variance was over estimated, because the residual variance estimate was reasonable and the polygenic variance was calculated as a fixed proportion of the residual variance.

Figure 4.7 shows the mixed model likelihood with the polygenic likelihood (obtained when estimating the heritability) subtracted, for four simulations, plotted against the polygenic heritability assumed. The polygenic (h_p) and total heritability (h_t)of the data are marked, along with the ML estimate (h). Although simulated under the same model, the surfaces are very different. In all of these examples the maximum obtained when estimating the polygenic heritability agrees with the maximum when a grid of fixed values for the heritability were used. Hence, the maximisation process has been successful in obtaining the global maximum, although, in one of these simulations, more than one mode exists, see figure a). When the difference between the mixed and polygenic model likelihoods is greater than zero then the incorporation of a major gene is an improvement over the polygenic model. Where this difference is greater than 3.5, then this becomes a significant improvement at the 5% level. When the polygenic heritability is fixed at zero a major gene model will result. In figure b) the major gene model is the most likely, however, if the polygenic heritability had been fixed at 0.1 a significant mixed model would result, however the major gene parameters in these two models are not very different. A significant mixed model was obtained in example c). Here, the estimated polygenic heritability is higher than the total heritability of the data. In figure d) the likelihood of the data under the mixed model is an improvement over the polygenic model for all heritability values except when the polygenic heritability is set at one, but never a significant improvement. In general, it would seem that the ability of the method to correctly estimate the polygenic heritability is poor.

Figure 4.8 shows the test statistic obtained analysing the mixed model data with an additive major gene segregating and a polygenic heritability of 0.4 against the true polygenic heritability of the data, both a) for the analyses with fixed heritability and b) with the heritability estimated. There is a positive correlation between the test statistic and the true heritability when the heritability was assumed to be known in the analyses (r=0.475), which disappears when the heritability is estimated (r=0.000). Evidence for a major gene is greater when the assumed value for the polygenic heritability is an underestimate.

Figure 4.7 *Improvement of the mixed model over the polygenic likelihood for values of the heritability of 0.0 to 1.0 for four different Add1 data sets.*

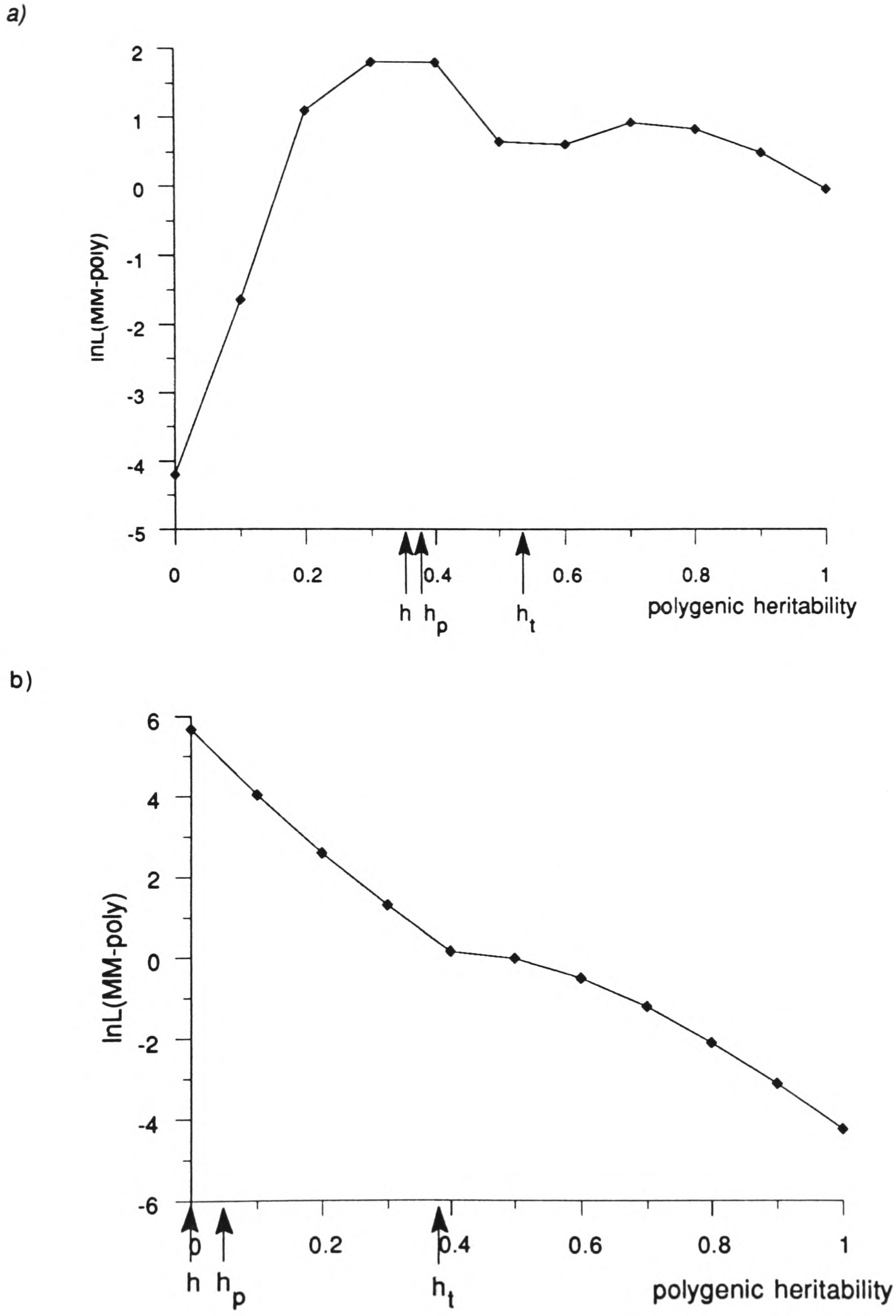
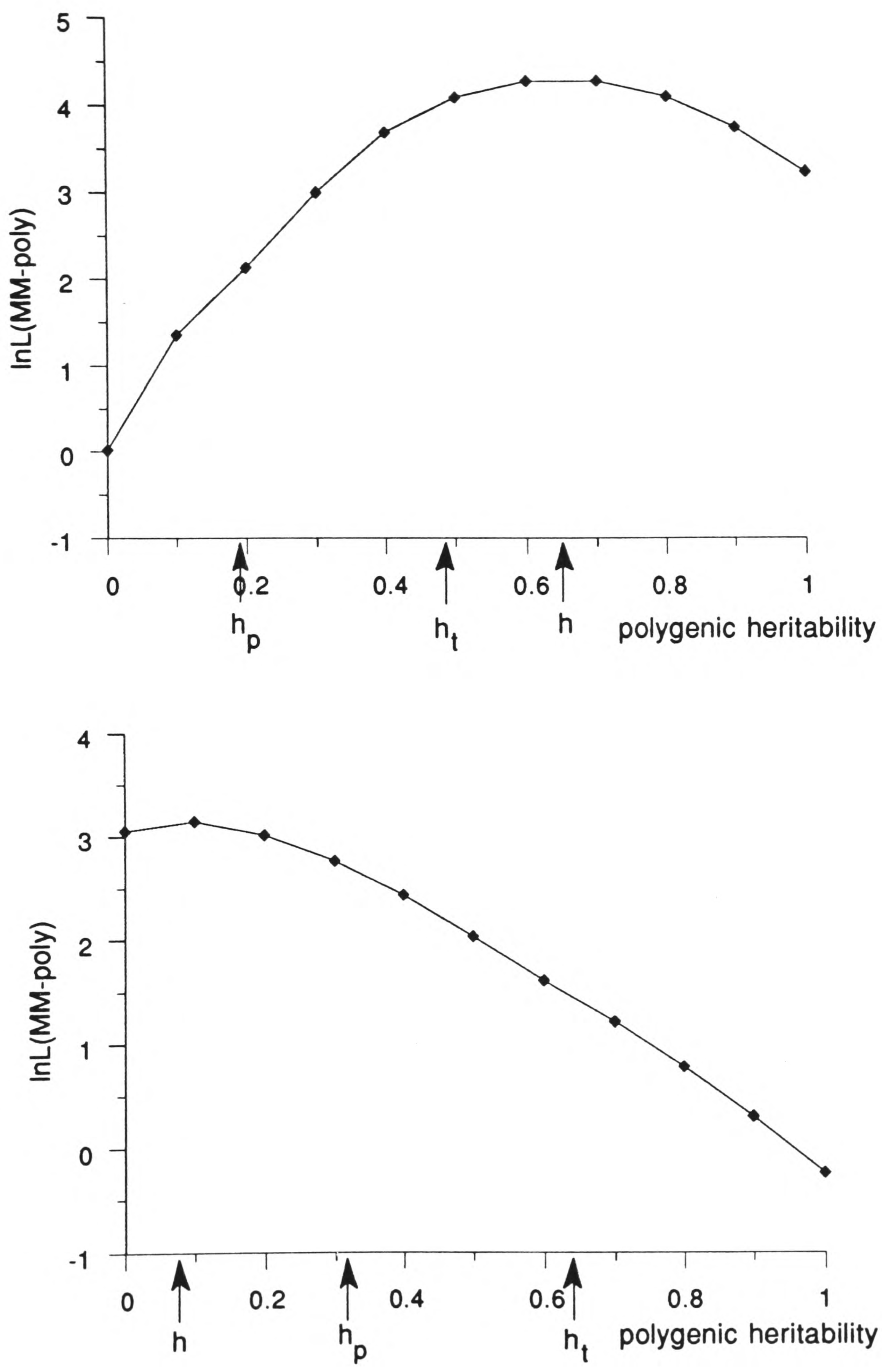


Figure 4.7 (continued)



c)

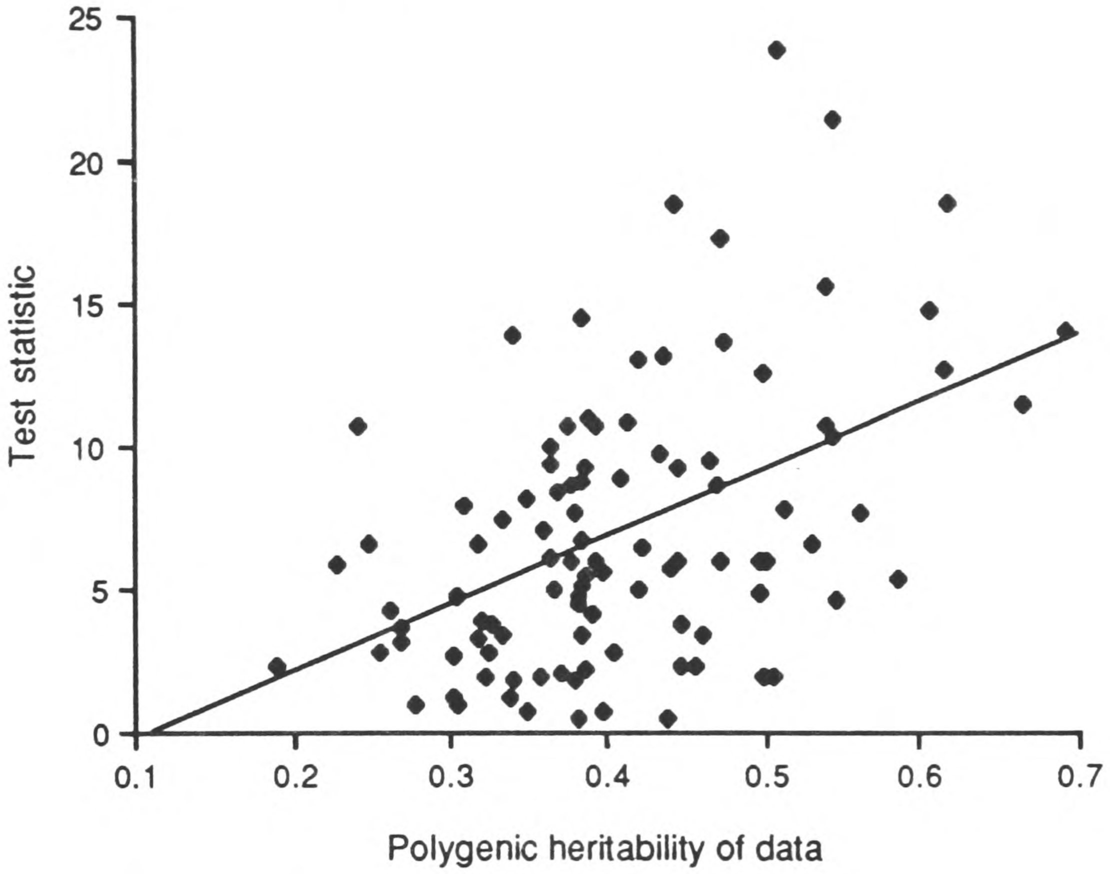
d)

Where: h is the estimated polygenic heritability from the segregation analysis.

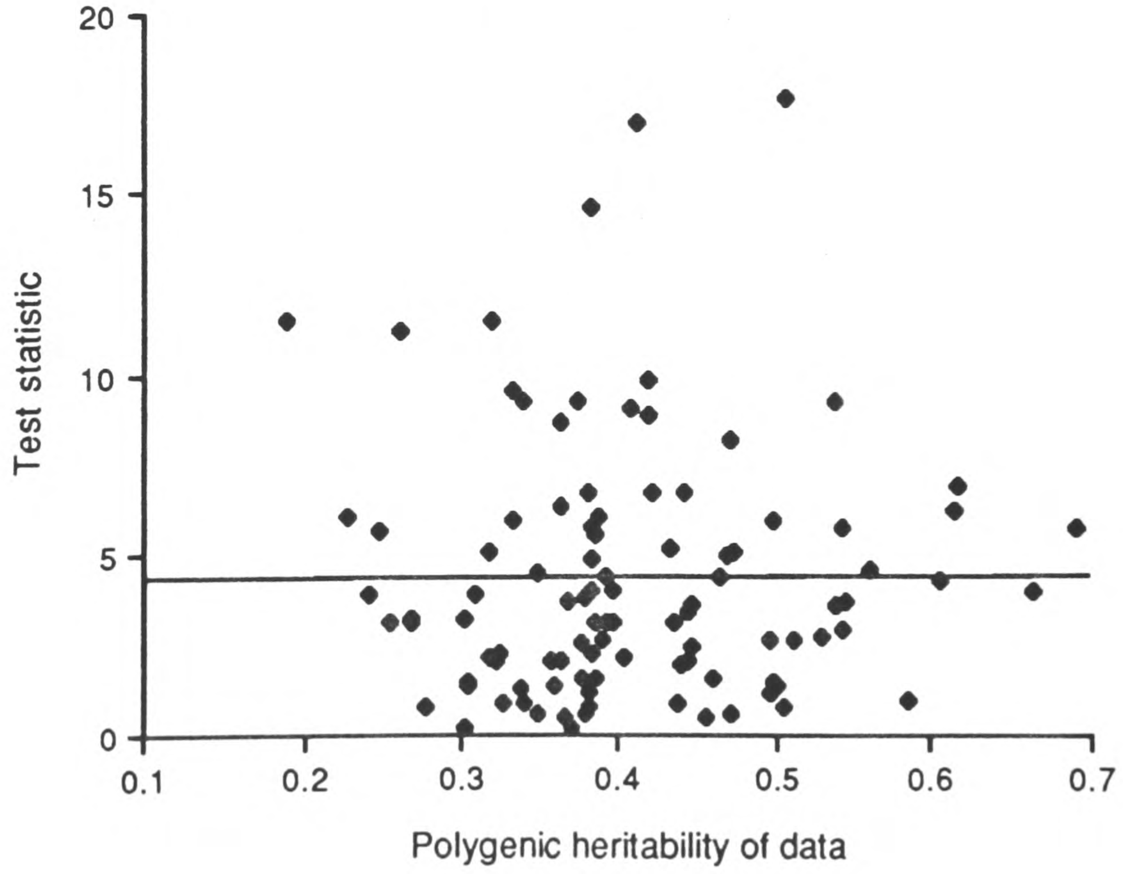
h_p and h_t are the estimated polygenic and total heritability from the simulation.

Figure 4.8 *The test statistic from the analysis of mixed model data (Add2) with expected heritability of 0.4 against the polygenic heritability of the data estimated using analysis of variance of the simulated polygenic plus environmental values.*

a) With heritability fixed in the analyses



b) With heritability estimated in the analyses



4.5 DISCUSSION

Segregation analysis has been suggested as a suitable method to detect major genes segregating within farm animal populations. The use of Hermite integration in an approximation of the mixed model likelihood makes the method feasible and with a high number of summations the approximation gives virtually the same results as the exact method, and in this study it has been used in this way. The results of the simulation study show that segregation analysis is capable of detecting a major gene and obtaining good estimates of its effect, even with the simple pedigree structure used. However, the major genes simulated had fairly large effects and further investigation is required to see how well a gene of smaller effect is detected.

In most animal breeding situations the observations on the animals would also include effects of, for example, season or herd, and these would have to be efficiently removed from the data. These could be estimated prior to the use of segregation analysis using a simple fixed effects model or, perhaps, a polygenic model. In theory, estimation of these effects at the same time as the major gene parameters should be the best method to separate the fixed and genetic effects. In this case, the computation required to estimate the extra effects might become prohibitive, although with the continual improvement in computing technology and the development of new algorithms it might become feasible. Also, the inclusion of a more complex pedigree, although providing more information on the segregation of the major gene, would increase the time taken for each likelihood evaluation.

A reduction in the number of summations may decrease the computing time without too much loss of accuracy. Using the quasi-Newton algorithm for maximisation, the likelihood has to be calculated many times and the time taken for each evaluation of the likelihood is a function of the number of summations. Another way of reducing the number of summations required might be to have a different location parameter for each genotype of each sire. With the polygenic model, good choice of location parameter reduced the number of summations required. For the mixed model, given his progeny phenotypes, a sire would be expected to have a different transmitting ability under the three major genotypes. Hence, rather than using a single estimate for each sire in the mixed model approximation, using one for each major genotype of each sire should decrease the summations required, although estimates of these might be difficult to obtain.

Ideally, to make decisions for selection, animal breeders require both the polygenic genotype and the major genotype of each animal. Neither of these are obtained immediately in the analysis and have to be estimated after maximisation. For the sire model it is relatively easy to obtain the probabilities of each major gene for each individual using the ML parameter estimates. However, for a more complex pedigree, many likelihood calculations would be involved as the genotype of an individual is dependent on the genotype of other individuals and the major genotype for all individuals should be assigned simultaneously to be the most likely combination. In the simulation study presented here, classifying sires according to their major genotype was successful for, on average, only 50% of the sires in one of the models investigated. Further investigation is required to see how useful this information could be. Obtaining an estimate of the polygenic contribution of each sire to his offspring has not been considered here. If, after segregation analysis each individual was assigned a major genotype, these could be fitted as fixed effects in a classical mixed model analysis (Henderson, 1973) and the polygenic contribution from the sire estimated or the data adjusted for these fixed effects using the estimates from Herm and just the random effects estimated.

When the polygenic heritability was assumed to be known and fixed in the analyses, a major gene was detected more frequently. The problem will be obtaining an estimate that is suitable for the data to be analysed. The method has been shown to be sensitive to the heritability estimate and if the polygenic heritability in the data is greater than the value assumed in the analysis a major gene is more likely to be inferred, if the value is smaller then that assumed a polygenic model will result.

Appendix 2

A2.1 Simulation program

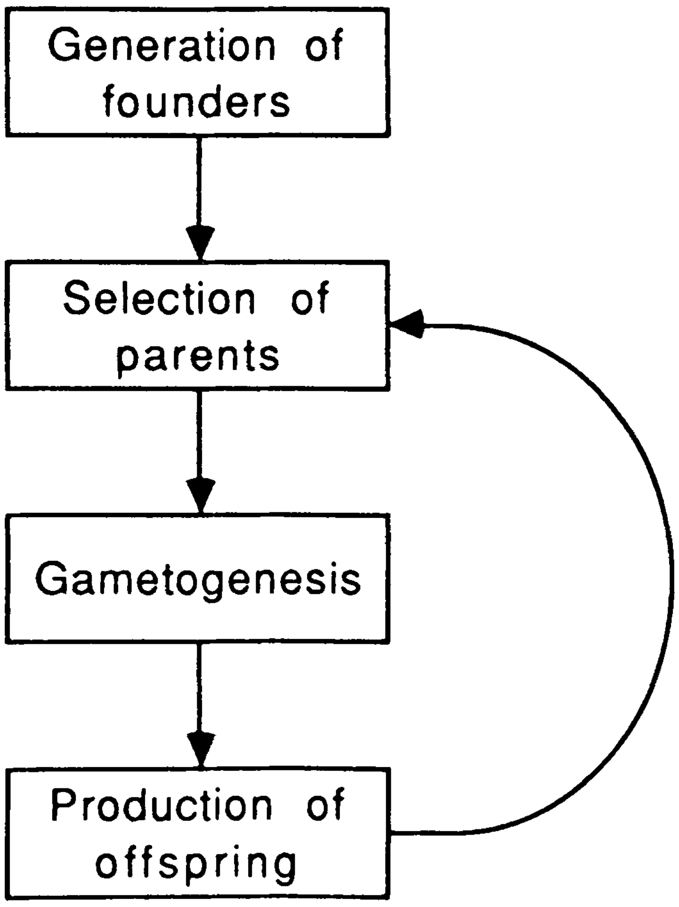
Simulated data were required in order to test the abilities of the different approximations. A Fortran program was designed to simulate a population under a genetic model which allows the effect and frequency of the genes involved to be specified.

Most computers store information as a series of words each containing a vector of fixed length of binary bits and, generally, there are functions available which perform Boolean operations on these binary vectors. Integers are stored in binary, each in a word and hence each has a unique vector. This representation of information can be utilised for genetic simulations (Fraser and Burnell, 1970).

The genetic component of each individual is composed of two words, representing homologous chromosomes and the allele at each locus is given by the vector of binary bits. Therefore at each locus there is a choice of two alleles represented by '0' or '1'. The maximum number of loci depends on the number of bits in a word, although an individual can be composed of more than two words. Hence, the genotype of each individual can be described by two integers.

Figure A2.1 illustrates the basic structure of the program and a brief description of the various stages is given below.

Figure A2.1 *A flow diagram illustrating the basic structure of the simulation program.*



A2.1.1 Founders

To generate the founders the number of loci has to be supplied, in this case restricted to a maximum of 32. At each locus there are two alleles segregating. The allele frequency and the effect of a homozygous and the heterozygous genotype at each locus is required, the remaining homozygous genotype is set to zero. The phenotype is obtained by the accumulation of the effects of the relevant genotype at each locus and the addition of a random component obtained from a normal distribution. The variance of this distribution is supplied and the mean set to zero. More complicated environmental effects can be included, for example the effect of sex, herd or year. Although for the study presented here these were ignored. The number of founders to be simulated is required along with the proportion of these to be male. Sex is attributed randomly. All individuals are simulated independent of the others.

A2.1.2 Selection of parents

Parents are selected initially from the founder population and at later generations from the offspring population, so that generations are discrete. The design is balanced and hierarchical with a given number of dams per sire and given number of offspring per mating. For this study parents were required to be randomly selected.

A2.1.3 Gametogenesis

The two homologous chromosomes are combined into a single gamete representing the effects of meiosis. This is achieved by a process called random walking (Fraser and Burnell, 1970). The recombination frequency between each adjacent loci is required. Each locus is considered in turn and depending on the recombination frequency supplied the allele from one of the two homologues is randomly chosen. This process has to be repeated on the homologues of each parent for each offspring.

A2.1.4 Production of offspring

The genotype of the offspring is specified by two gametes, one coming from the sire and the other from the dam, each gamete was represented by a word. The effect of the genotypes can be calculated in the same way as for the founders. The phenotype is obtained by the addition of any environmental effects. In this study an individual random effect simulated from a normal distribution, with the variance as specified for the founders, was the only environmental component used. Other effects as described for founders could be included. Sex was randomly allocated assuming a 50% probability of being male.

CHAPTER 5

APPROXIMATION 3 - MODAL ESTIMATION

5.1 INTRODUCTION

Animal breeders are interested in having estimates of the genetic merit of animals so that they are able to make selection decisions. For the polygenic model, methods enabling this have been investigated and are used as the standard procedure in analysis of animal breeding data. These methods also enable the simultaneous estimation of fixed effects and possible extension to include additional random effects. The use of Hermite integration (chapter 4), although having reasonable power and providing good estimates of the parameters involved in the mixed model, cannot easily be extended to include more complicated models or pedigrees. Extensions of methods for the polygenic model, to include parameters describing a major gene, have been proposed by Hoeschele (1988a and b) and Le Roy *et al.* (1989). An approximation to the mixed model likelihood based on these methods will be described and the operational characteristics investigated.

5.2 LIKELIHOODS

5.2.1 Polygenic model

Using the model described in section 2.3:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

it can be shown that for a polygenic model the following equality holds:

$$\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{1}\mu)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{1}\mu)\right] = \exp\left[-\frac{1}{2\sigma_u^2}\hat{\mathbf{u}}'\hat{\mathbf{u}}\right]\exp\left[-\frac{1}{2\sigma_w^2}(\mathbf{y}-\mathbf{1}\mu-\mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y}-\mathbf{1}\mu-\mathbf{Z}\hat{\mathbf{u}})\right] \qquad [5.1]$$

Where: $\hat{\mathbf{u}}$ is a vector of length s containing the mode of the transmitting ability distribution for each sire.

Hence, the polygenic likelihood [2.1] can be rewritten:

$$L(\text{poly}) = \frac{1}{\sqrt{(2\pi)^{sn}|\mathbf{V}|}}\exp\left[-\frac{1}{2\sigma_u^2}\hat{\mathbf{u}}'\hat{\mathbf{u}}\right]\exp\left[-\frac{1}{2\sigma_w^2}(\mathbf{y}-\mathbf{1}\mu-\mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y}-\mathbf{1}\mu-\mathbf{Z}\hat{\mathbf{u}})\right]$$

which in the non-matrix terms introduced in section 2.3 is:

$$L(\text{poly}) = \prod_{i=1}^{s} \left(\frac{2\pi\sigma_w^2}{n+\lambda} \right)^{1/2} h(\hat{u}_i) \prod_{j=1}^{n} k_o(y_{ij} \mid \mu, \hat{u}_i, \sigma_w^2)$$

with $h(\hat{u}_i)$ and $k_0(y_{ij}|\mu,\hat{u}_i,\sigma_w^2)$ as defined previously, with u_i replaced by the mode of the distribution $(\hat{u}_i)$.

The integration over all possible values of each sire's transmitting ability has been replaced and a single evaluation of the probabilities at the mode of the distribution is used, with the function suitably weighted. This is the situation considered with Hermite integration in case 2 of table 4.1.

5.2.2 Mixed model

In the same way the mixed model likelihood [2.3] can be rewritten, using equality [5.1]:

$$L(MM) = \prod_{i=1}^{s} \frac{1}{(2\pi)^{n/2} |V_i|^{1/2}} \sum_{c=1}^{m} \sum_{D=1}^{m^n} p(c)\, \text{trans}(D \mid c) \exp\left[-\frac{1}{2\sigma_u^2} \hat{u}_{icD}' \hat{u}_{icD} \right]$$
$$\exp\left[-\frac{1}{2\sigma_w^2} (y_i - 1\mu - W_D\mu_d - Z\hat{u}_{icD})'(y_i - 1\mu - W_D\mu_d - Z\hat{u}_{icD}) \right]$$

Where: $\hat{u}_{icD}$ is the mode of the distribution of transmitting ability for sire i, when he has major genotype c and the combination of major genotypes for his offspring is D.

The mode of the distribution has to be calculated for each possible combination of major genotypes for the sire and offspring, i.e. with n offspring $2^{n+1}+3^n$ sire effects are required. This becomes impracticable even with only a small number of offspring per sire, for example, with 5 offspring (n=5), 307 effects for each sire would have to be calculated and with 10 offspring, 61097 effects. Therefore an approximation to this likelihood is suggested, where a single estimate of the mode of each sire's transmitting ability distribution is calculated taking into account the possible major genotypes for the sire and offspring. The following expression is obtained for the mixed model likelihood, which will be denoted ME1:

$$L(MM) = \prod_{i=1}^{s} \left(\frac{2\pi\sigma_w^2}{n+\lambda} \right)^{1/2} h(\hat{u}_i) \sum_{c=1}^{m} p(c) \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d \mid c)\, k_d(y_{ij} \mid \mu, \mu_d, \hat{u}_i, \sigma_w^2)$$

Where: $\hat{u}_i$ is the mode of the transmitting ability distribution for sire i.

This is equivalent to Hermite integration with a single summation for each sire performed at an estimate for his polygenic breeding value.

5.3 MAXIMISATION

5.3.1 Algorithm

The likelihood has to be maximised and estimates obtained of the parameters involved. The likelihood could be maximised explicitly, using, for example, the quasi-Newton routine described in section 4.3. However, the number of parameters to be estimated has increased, and now includes an effect for each sire. As the number of evaluations of the likelihood in the quasi-Newton algorithm is a function of the number of parameters to be estimated, the process will be slow. An alternative, that makes use of well documented computing strategies, would be to obtain the partial derivatives of the log likelihood with respect to each parameter to be estimated. At a maximum the first derivatives are equal to zero, and hence by equating them to zero a series of equations is obtained which can be solved to give the maximum likelihood (ML) estimates. The equations obtained from the mixed model likelihood are given in Appendix 3.

For the polygenic model, if fixed effects are included, the familiar mixed model (i.e. fixed and random effects) equations are obtained (Henderson, 1973). These can be arranged into matrix form and for a given heritability solved directly. Using the following general polygenic model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$$

With the parameters described in 2.3 and replacing the overall mean with $\mathbf{X}\beta$.

Where: $\mathbf{X}$ is the incidence matrix for fixed effects

β is the vector of fixed effects

The first differential equations can be written, after rearrangement and simplification, as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [5.2]$$

Where: **A** contains the additive genetic relationship between the sires, assumed here to be **I**.

For the mixed model, even with fixed heritability, the equations have to be solved iteratively. The equation for each parameter contains the conditional genotype probabilities for each sire ($q_i(c)$) and for each offspring ($q_{ij}(d|c)$) (see Appendix 3) which are functions of the parameters to be estimated. However an EM (Expectation Maximisation) algorithm (Dempster *et al.*, 1977) can be used that involves estimating (E) the unknown parameters and then maximising (M) the likelihood of the data given these estimates. In the mixed model the unknown parameters are the conditional probabilities, values for which can be obtained using given values for the major genotype means, population genotype frequencies, sire effects and residual variance. For the maximisation step, these conditional probabilities can be incorporated into the equations, which can be rearranged to give matrices similar to those used for the polygenic model. With the incorporation of fixed effects these can be written as follows:

$$\begin{bmatrix} \mathbf{D}_{[a]} & \mathbf{Q}_{[a]}'\mathbf{X} & \mathbf{Q}_{[a]}'\mathbf{Y} \\ \mathbf{X}'\mathbf{Q}_{[a]} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{Q}_{[a]} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \mu_{d[a+1]} \\ \beta_{[a+1]} \\ u_{[a+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{[a]}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad [5.3]$$

Where: $\mathbf{D}_{[a]}$ is an m x m matrix containing: $\mathrm{Diag}\left[\sum_{i=1}^{s}\sum_{c=1}^{m}\sum_{j=1}^{n} q_i(c)_{[a]}q_{ij}(d|c)_{[a]}\right]$

$\mathbf{Q}_{[a]}$ is an m x sn matrix containing: $\sum_{c=1}^{m} q_i(c)_{[a]}q_{ij}(d|c)_{[a]}$

[a] refers to the iteration number, i.e. $q_i(c)_{[a]}$ is obtained using parameter estimates from iteration a.

However, additional parameters are required to obtain new values for the conditional probabilities: the population genotype frequencies and the residual variance. Equations for these are obtained in the same way, and can be solved using the values for the conditional probabilities calculated with the given parameter estimates.

The process continues, obtaining the ML parameter estimates for the given conditional probabilities, then recalculating the probabilities given the new parameter estimates, until there is virtually no change in the parameter estimates from one iteration to the next.

Information about the frequencies of the major genotypes is contained in the population frequency (p(c)) for sires and in the transmission probabilities for dams. Assuming that the population is in Hardy-Weinberg equilibrium with respect to the major genotype, information can be combined from both of these sources to obtain an equation for the allele frequency in the population (Appendix 3). Alternatively, Hoeschele (1988b) suggests estimating a frequency for each major genotype, in which case estimates can simply be calculated by taking the diagonals of **D** and dividing by the total number of progeny in the data. This assumes that the offspring are in the same frequencies as the parents but not necessarily in Hardy-Weinberg equilibrium. Le Roy *et al.* (1989) suggest allowing for a different frequency in the sire population compared with the dam's. Genotype frequencies are estimated for the sires by accumulating the conditional probabilities for each sire. For the dams, because of the use of a half-sib structure, information is only available on the allele frequency. Equations are given in Appendix 3.

Variance equations can also be obtained under different assumptions. If a reasonable estimate of the polygenic heritability is available (for example, from a previous analysis in a similar population before the incorporation of a major gene) then methods analogous to best linear unbiased prediction (BLUP) can be used. In this situation λ is known although, unlike BLUP, the residual variance component is required each iteration in order to estimate the conditional probabilities. Alternatively both the sire and residual variance components could be estimated. Equations are given in Appendix 3.

5.3.2 Initial parameter estimates

As explained above, for the EM algorithm initial estimates of the parameters are required to obtain values for the conditional probabilities. With the algorithm described, there is no guarantee that the maximum reached is the global maximum and not merely a local maximum with a more likely end point elsewhere. Hence, to reduce the chance of ending at a local maximum the initial estimates for the parameters need to be near the global maximum.

Obtaining suitable initial estimates for the sire effects is a problem. One option would be to start them from zero, this produces reasonable results when the heritability is fixed, but not if an estimate for the sire variance is required. The initial estimates have a variance of zero and the maximisation procedure has difficulties departing from this value. This could be overcome, either by supplying non-zero estimates for the sire effects or by fixing the heritability in the first few iterations of the analysis. Supplying appropriate non-zero estimates is difficult, as, of course, the progeny mean and the transmitting abilities obtained assuming a polygenic model will also contain the major gene effect of the individuals. Hence, all the difference between sires is being explained by the polygenic component, whereas some of it is due to the major gene. A high progeny mean could be caused by a high frequency of the high scoring allele in the sibship or a high polygenic contribution from the sire.

The data described in 4.3.2 was reanalysed to determine how sensitive the ME1 approximation was to the initial parameter estimates used to start the maximisation process. The five starting models used previously (4.3.2) were used and two alternatives for the initial sire transmitting ability estimates, one with the sire effects equal to zero was used for all the models and the second, for models 1 and 4, with the initial transmitting abilities equal to half the difference between the sib ship mean and the mean of all the offspring $\left(\frac{\bar{y}_{i.} - \bar{y}_{..}}{2}\right)$. When estimating a population allele frequency, all the analyses started from the polygenic model resulted in a polygenic model. For one of the data sets all the mixed model starting places resulted in the same mixed model at convergence, and for another data set model 3 resulted in a mixed model with the other initial models giving a polygenic end point. The remaining analyses all resulted in a polygenic model. With genotype frequencies being estimated for the sire and an allele frequency for the dams more analyses resulted in mixed models. Starting with non-zero estimates for the transmitting abilities did not alter the final model attained. However different models could be obtained with similar likelihood values (see table 5.1) and frequently models were obtained that made little sense genetically with, for example, the heterozygote being absent in the sires. Also, models were obtained that looked different, in that different effects for the major genotypes were suggested but on closer inspection the models are describing the same distribution, for example analysis 1 in table 5.1.

When estimating genotype frequencies for the sire and an allele frequency for the dams, the maximisation process is sensitive to the initial estimates supplied. With the major genotype means, frequencies, the sire and residual variance components fixed at their ML estimates, considering a single sire the likelihood of the phenotypes of his offspring can be calculated assuming different values for the sire's transmitting ability. Figure 5.1 shows the likelihood surface for one sire, with data simulated under a mixed model with the major gene being additive with two within major genotype standard deviations between the homozygotes. It was found that the surface could have three modes. The conditional probabilities were calculated at each value for the transmitting ability, and these three modes were found to correspond to the sire having a high probability of being one of the major genotypes. Obviously, given the data of his offspring, if the hypothesis of the sire being each major genotype was considered

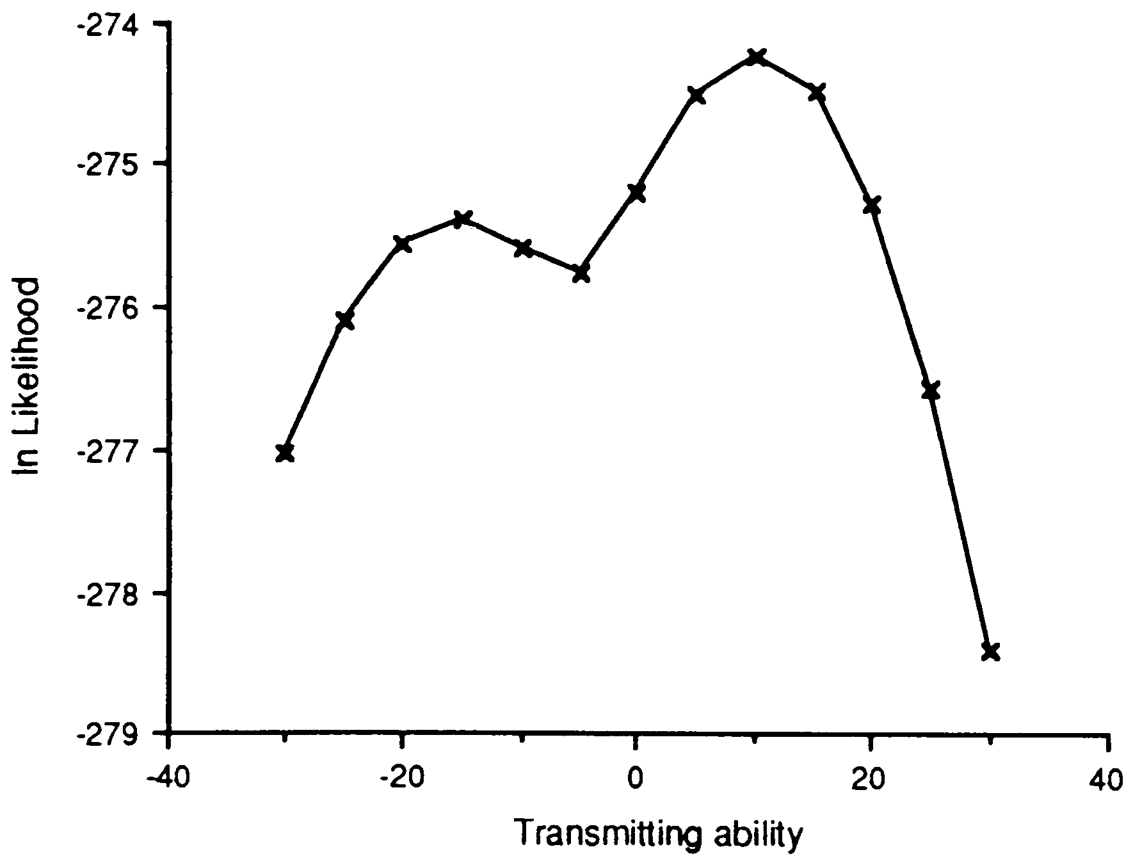
Table 5.1 *Alternative estimates obtained using different parameter values to start the maximisation process.*

	Sire			Dam						
Analysis	p(AA)	p(Aa)	p(aa)	p(A)	μ_{AA}	μ_{Aa}	μ	σ_u^2	σ_w^2	ln L
1	0.900	0.000	0.100	1.000	12.030	0.003	29.533	6.881	130.730	7.49
	0.900	0.000	0.100	0.002	12.051	12.047	29.515	6.675	126.834	7.49
2	0.511	0.000	0.489	0.379	11.157	14.167	31.770	5.846	111.077	3.85
	0.526	0.474	0.000	0.653	26.127	11.460	23.276	3.787	71.960	3.22

ln L - mixed model likelihood as a deviation from the polygenic likelihood.

separately the same estimate for the polygenic contribution of the sire to his offspring would not be obtained for each genotype. This leads to multimodality of the likelihood surface. Hence problems can be encountered when maximising the likelihood, with the sire estimate ending at the closest mode to the initial value and not necessarily at the global maximum.

Figure 5.1 *Likelihood for one sire for different values of his transmitting ability with the other parameters fixed at their ML estimates.*



5.4 SIMULATION STUDY

In order to investigate the ability of the ME1 approximation to detect a major gene and correctly estimate its effect and frequency in the population, the simulated data described in 4.4.1 were reanalysed.

5.4.1 Analyses

In the analysis, the mean effect of the low scoring homozygous genotype (μ) at the major locus and the deviation from this mean of the other two major genotype means (μ_{AA} and μ_{Aa}) were estimated. The population was assumed to be in Hardy-Weinberg equilibrium, and an allele frequency (p(A)) estimated. Also, as for Hermite integration (Herm), two different situations regarding the variance estimates were considered. For the first analysis it was assumed that the polygenic heritability was known and the residual variance estimated and for the second both the within and between sire variance components were estimated.

Initial parameter estimates

As described above, initial parameter estimates are required from which the maximisation process can start. As for the analyses using Herm these estimates will be the values used to simulate the data. The method is known to be sensitive to the initial parameter estimates provided, therefore an additional starting value was used. This alternative model explained the expected total mean and variance of the data but contained a different major gene model. For the additive models with equal allele frequency the alternative major gene explains a larger proportion of the total variance (52% compared with 33% in the simulation), for the rare gene the alternative starting point assumed the same allele effect but equal allele frequencies and for the dominant gene an additive model was assumed with the same difference between the homozygotes. For each sire the initial estimate of his transmitting ability was zero, and hence when estimating the sire variance component the heritability was fixed at its expected value until the convergence criterion (given in A3.2) was less than 0.05.

Genotyping at the major locus

As for the approximation using Hermite integration, the ability of the method to genotype individuals at the major locus is of interest. Hence the probability of each genotype for each sire is calculated. The equation [2.5] can be rewritten replacing the integration with a single estimate of the mode of the distribution:

$$q_i(c) = \frac{p(c) \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d \mid c)\, k_d(y_{ij} \mid \mu, \mu_d, \hat{u}_i, \sigma_w^2)}{\sum_{c'=1}^{m} p(c') \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d \mid c')\, k_d(y_{ij} \mid \mu, \mu_d, \hat{u}_i, \sigma_w^2)} \qquad [5.4]$$

However, the use of the EM algorithm means that this probability is already calculated each iteration and hence the values from the final iteration can be used.

5.4.2 Test statistic

For segregation analysis the test statistic used to test evidence for a major gene segregating along with polygenes is the likelihood ratio, i.e. twice the natural logarithm of the ratio of the likelihood of the data maximised under a mixed model to the likelihood maximised under the polygenic model. For data simulated under the polygenic model, the distribution of test statistics obtained is expected asymptotically to follow a χ^2 distribution with degrees of freedom equal to the number of parameters fixed in the polygenic model but estimated in the mixed model. It is not known whether this distribution holds for the ME1 approximate likelihood.

Method

To investigate the distribution of the test statistic [2.4] the data described in 4.4.3 were analysed using the ME1 approximate likelihood. As before each data set was analysed both assuming that the polygenic heritability was known and estimating the heritability. For the mixed models two sets of initial parameter estimates were used, with the major gene, additive with equal allele frequencies, explaining different proportions (50% and 13%) of the total variance in the data. The MLs of the data under the polygenic and mixed models were obtained and the test statistic calculated.

Results

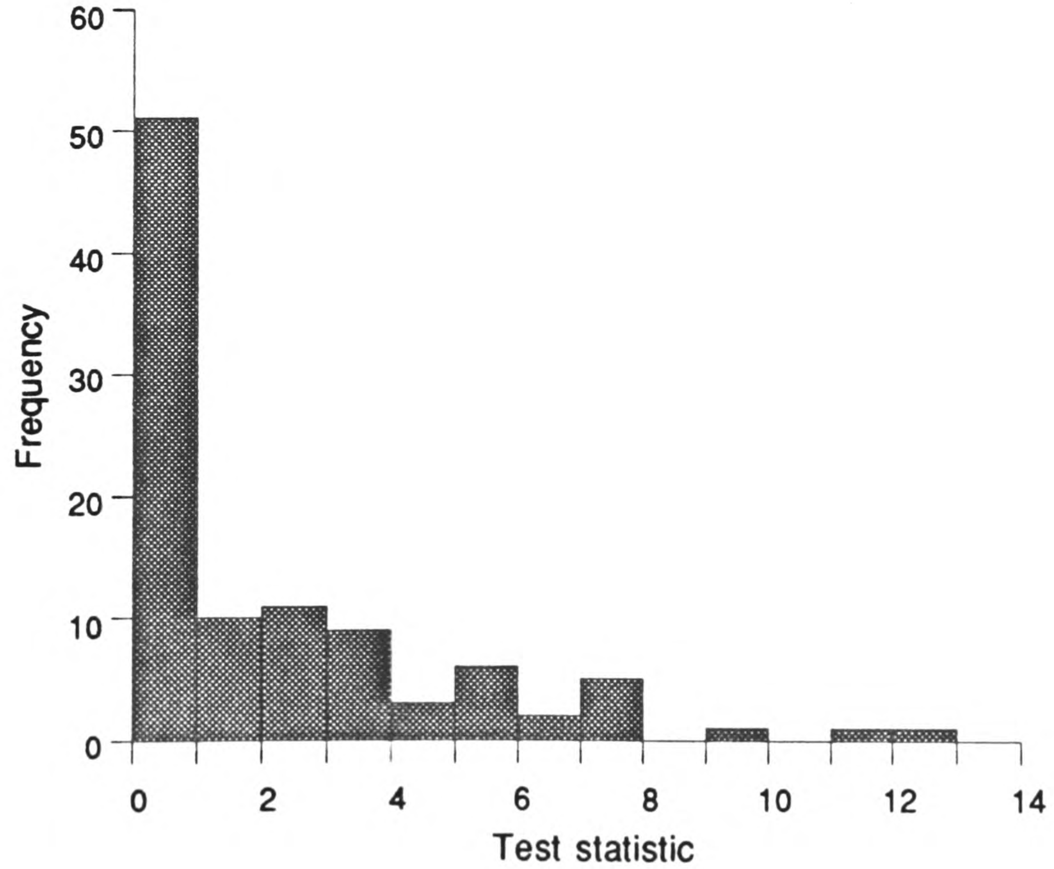
The results given are based on the mixed model analysis for each data set that resulted in the highest likelihood, without reference to the parameter estimates at this maximum. For some of the analyses the test statistic obtained was negative for the ML obtained from both initial estimates, i.e. the mixed model obtained was less likely than the polygenic model, and these test statistics have been set to zero. The mean and variance of the observed test statistic distribution for the 100 analyses are given in table 5.2. Figure 5.2 shows the distribution of test statistics for the data simulated with an expected polygenic heritability of 0.2 which was estimated in the analyses. Although 43 of the analyses gave a zero test statistic, the extreme tail of the distribution appears to be similar to a χ^2 distribution with three degrees of freedom (see figure 4.3 for the expected distribution). This observed distribution was compared to a χ^2 distribution by comparing the expected and observed number of test statistics in ten equal parts of the

χ^2 distribution with two, three and four degrees of freedom. The χ^2(9 d.f.) values obtained from this test were 146.0, 155.2 and 201.2 respectively. The high values being caused by the large proportion of analyses giving virtually zero test statistics. The other models resulted in zero test statistics for nearly all the analyses.

Table 5.2 *Mean and variance of the test statistic and the number of analyses where the test statistic was significant at the 5% and 1% significance levels of a χ^2 distribution with 3 degrees of freedom. for data simulated under a polygenic model.*

Model		Mean	Variance	no. zero	5%	1%
Expected	(χ^2 3 d.f.)	3		0	5	1
h^2=0.2	fixed	0.157	0.581	95	0	0
h^2=0.2	estimated	2.069	7.387	43	4	1
h^2=0.4	fixed	0.007	0.004	99	0	0
h^2=0.4	estimated	0.208	1.294	94	1	0

Figure 5.2 *Distribution of the test statistic from analyses of polygenic data with an expected heritability of 0.2, estimated in the analyses.*



Discussion

The test statistic distributions obtained from the 100 analyses of each model do not follow a χ^2 distribution with three degrees of freedom as expected from exact segregation analysis and supported by the results from Herm (4.4.3). It is expected that a more general model will have a higher likelihood than one in which fewer parameters are estimated, and hence that the test statistic will be greater than zero. However in this case the mixed model often ended at a polygenic model giving a test statistic of zero.

A more thorough search of the likelihood surface may locate a maximum with a likelihood greater than the polygenic likelihood, however, the test statistic obtained is unlikely to be large and hence will not have much effect on the distribution. Also the test statistics for ME1 were nearly always less than or equal to (when, with the heritability estimated, they both ended at the same major gene model) the test statistic for Herm on the same set of data.

Although the observed test statistic distributions are not χ^2 distributions, it is possible that the χ^2 distribution could provide suitable values against which the test statistic can be compared when searching for a major gene. If there is not much evidence for a major gene in the data, a polygenic model (with a test statistic of zero) results rather than a mixed model with a small value for the test statistic. Whereas, if much evidence for a major gene is present, then the test statistic follows the expected distribution. However, as the test statistic is less than the value obtained using Herm, a major gene is less likely to be detected using this approximation if the same criterion for significance is used.

Le Roy *et al.* (1989) and Elsen and Le Roy (1989) look at the test statistic distribution for this approximation (denoted ME1 and MU1 respectively). With a fixed heritability of 0.2, based on 1000 analyses, Le Roy *et al.* (1989) found that, rather than the test statistic distribution asymptoting to a χ^2 distribution as the number of sires and the number of half-sibs per sire increased, the mean of the test statistic distribution continually decreased. With 20 sires each with 20 offspring, the largest data set analysed, the mean of the distribution was 1.27 with standard deviation 2.11. The analyses estimated the major genotype frequencies for the sires and the allele frequency for the dams and hence these values are much lower than expected from the relevant χ^2 distribution (five degrees of freedom). The 5% and 1% quantiles of this distribution were estimated at 5.82 and 9.46 respectively, again much lower than expected. When estimating the heritability, for 154 analyses where the expected polygenic heritability was 0.2, a mean test statistic of 4.61 was obtained, with the estimated 5% and 1% quantiles being 11.52 and 15.83. With an expected heritability of 0.6, based on 245 simulations the mean was 3.36 and the quantiles 11.42 and 13.14. These results are in agreement with the results presented here, i.e. that estimating the heritability provides a distribution more like a χ^2 distribution and increasing the polygenic heritability of the data decreases the mean and variance of the test statistic distribution.

5.4.4 Simulation Results

Power

The results for the test statistics obtained from the analyses of the mixed model data are summarised in table 5.3. Compared with the results from Herm (table 4.5) the mean test statistic is always lower for ME1, mainly because of the high proportion of analyses resulting in a zero test statistic. With the heritability estimated the number of zero test statistics was reduced and hence the means are higher. A negative test statistic indicates that a local maximum has been reached in both analyses of the data set and a polygenic model would not be rejected. A major gene was detected most frequently when a major gene with a dominant allele was simulated. Unlike in the Herm analyses, with fixed heritability, when the data contained an additive major gene with a rare allele, a major gene was detected more often than when the simulated gene had alleles of the same effect but at equal frequency. With estimated heritability, fewer analyses ended at a polygenic model with zero test statistic, but a greater proportion resulted in a negative test statistic, having gone to a local maximum.

Also in table 5.3 are the correlations of the ME1 test statistic with, and the regressions on, the Herm test statistics for the same set of data, this time including the negative test statistics although excluding those that gave a zero test statistic. Although, generally there is a good linear relationship of the ME1 test statistic with the Herm test statistic the ME1 statistic is always lower. This is illustrated in figure 5.3 for the data simulated containing an additive major gene with equal allele frequencies segregating in a polygenic background with a heritability of 0.2. Hence, when the test statistics are compared with a χ^2 distribution with three degrees of freedom, fewer ME1 than Herm tests are significant, especially when the polygenic heritability is fixed for the additive major genes. This is supported by the test statistic distribution obtained under the null hypothesis which would suggest that the 5% and 1% quantiles from the observed distribution are lower than those used to test for significance. With the heritability estimated the correlations with the Herm test statistics were lower although the slopes of the regression lines were closer to one.

The likelihood surface for one set of data simulated with an additive major gene, with equal allele frequencies and a polygenic heritability of 0.2 (Add1) and which produced a negative ME1 test statistic is shown in figure 5.4. This was obtained by fixing the

Table 5.3 *Mean and standard deviation of the test statistic (setting negative values to zero), the number of analyses where the test statistic was significant at the 5% and 1% significance levels of a χ^2 distribution with three degrees of freedom and the regression on, and correlation with, Herm results for the same set of data.*

			no.	no.	no. significant		compared with Herm	
Model	mean	sd	zero	negative	5%	1%	slope	r
Fixed heritability								
Add1	0.578	2.380	80	10	4	1	0.698	0.850
Add2	0.000	0.000	98	2	0	0		
Dom	31.325	14.518	1	0	92	92	0.976	0.990
Rare	2.629	5.086	44	3	11	7	0.491	0.690
Estimated heritability								
Add1	3.310	5.306	5	28	13	5	1.135	0.787
Add2	1.363	8.297	63	15	8	4	1.452	0.623
Dom	37.225	13.528	0	0	99	98	0.994	0.948
Rare	4.211	5.290	17	11	18	7	0.901	0.734

Figure 5.3 *The test statistic obtained under ME1 compared with the test statistic obtained with Herm for the same set of data, for Add1 with fixed heritability .*

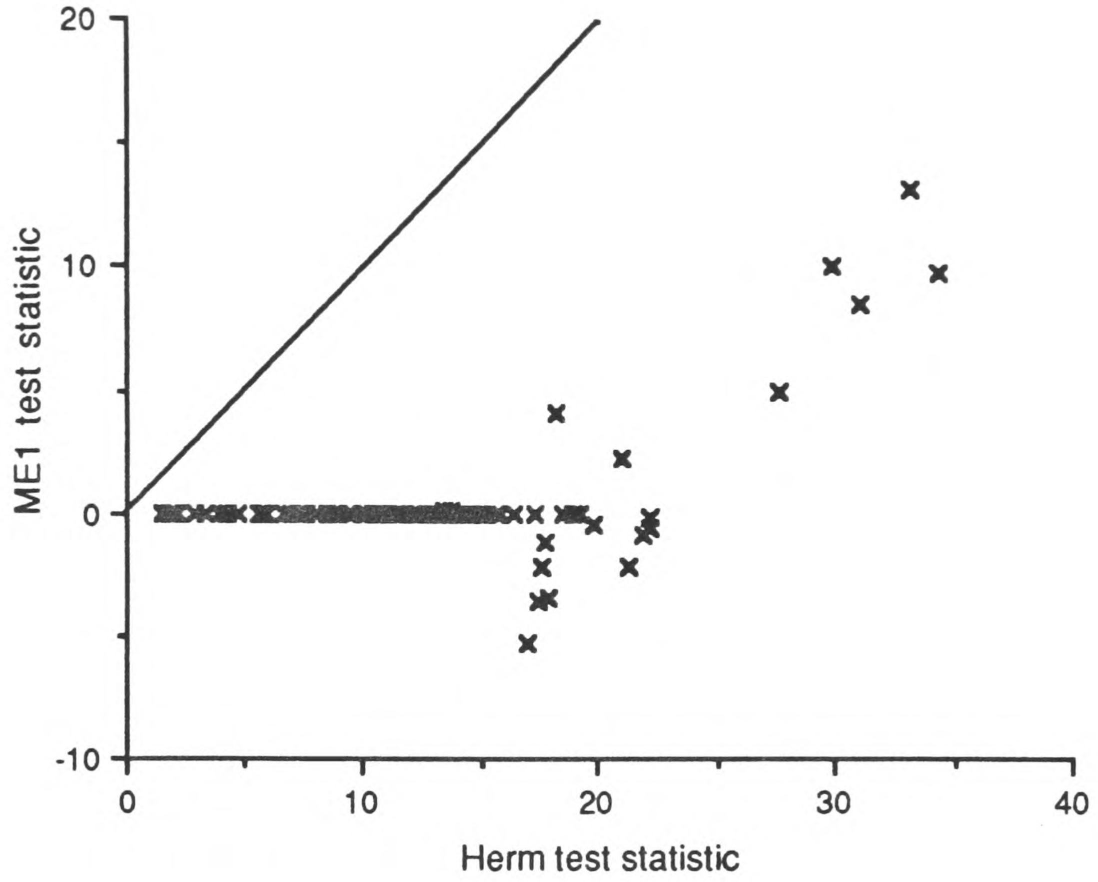
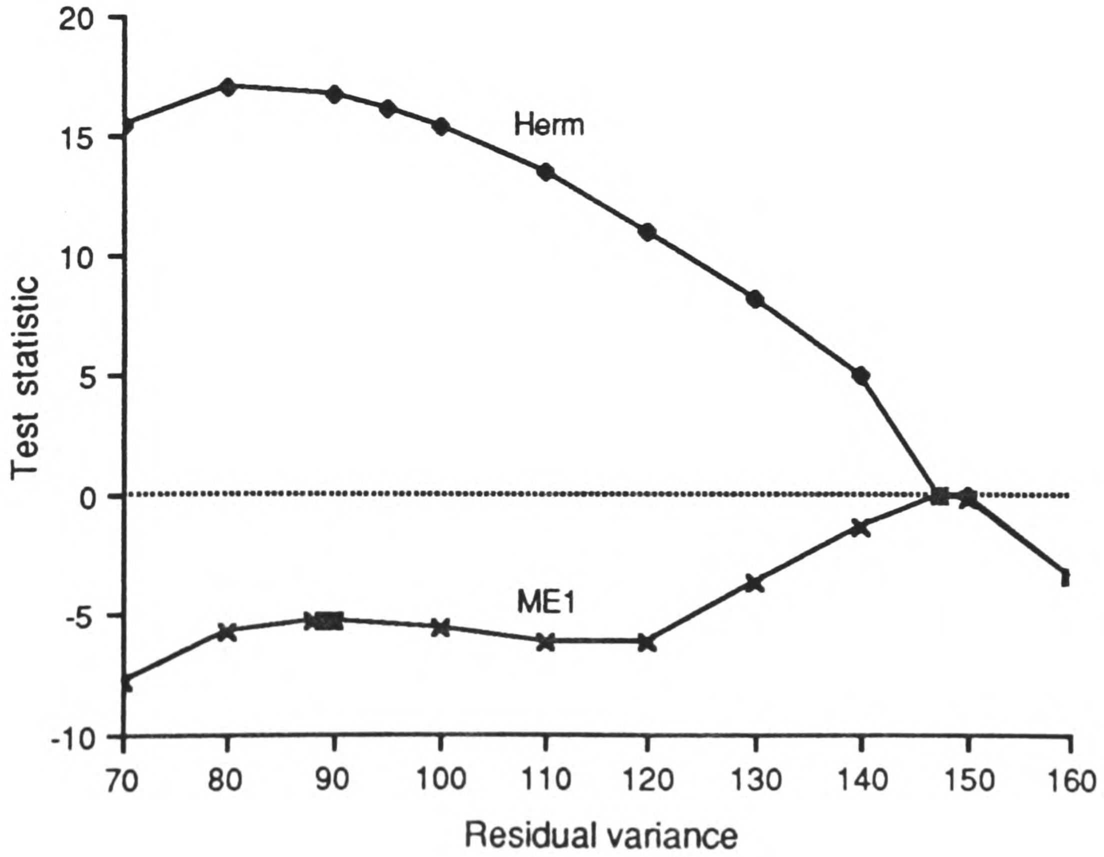


Figure 5.4 *The likelihood surface for one set of data simulated under Add1.*



residual variance over a range of values and maximising the likelihood for the remaining parameters. It can be seen that the ME1 method has a local maximum at a residual variance of about 90 and a global maximum at 147, the latter being a polygenic model. The Herm method has a global maximum at about 95.

Parameter estimates

The mean parameter estimates for the analyses with non-zero test statistics are given in table 5.4. Although the results, on average, give a reasonable indication as to the nature of the major gene segregating they are not as close tc the expected values as the results using Herm (tables 4.8 and 4.9). The high number of analyses giving a zero test statistic means that some of the results are based on rather few observations. For the analyses with non-zero test statistic the regression on, and correlation with, the Herm parameters for the same set of data are also given in table 5.4. Assuming that Herm gives the ML estimates for the exact mixed model likelihood then these are the 'best' estimates which the ME1 approximation can obtain. Table 5.5 gives the mean estimates for the variance components (with the major gene variance estimated using [4.5]) and the regression on, and correlation with, the values estimated in the simulation (by analysis of variance on the phenotypes minus the effect of the major gene for the residual and sire variance components and using [4.5] for the major gene variance) for the same set of data. The analyses giving zero test statistics and hence polygenic parameter estimates could not be included in the comparisons as the correlation of these with the Herm values will be zero. However, the mean major gene parameter estimates for Herm for those analyses that gave a zero ME1 test statistic and the mean for those that gave a non-zero ME1 test statistic are given in table 5.6.

With the heritability fixed, ME1 always over estimates the residual variance in comparison with Herm, as shown in figure 5.5, and as a consequence there is a consistent under estimation of the effects of the major gene compared to the Herm results for the same set of data. However, those simulations which resulted in a non-zero test statistic for ME1 tended to be those which had largest test statistics and produced the largest major gene estimates in Herm (table 5.6) and hence the mean major genotype means estimated using ME1 tend to be larger than the mean Herm estimates for all analyses.

When the heritability was estimated most of the analyses went to either a major gene model or polygenic model. For the data containing a dominant major gene 9 of the analyses resulted in a mixed model and for the additive gene with a rare allele only 1 analysis gave a mixed model, all the remaining non-zero test statistics were caused by a major gene model. Hence, the mean of the sire variance component for those analyses

Table 5.4 *Mean parameter estimates from analyses with non-zero test statistics, and the correlation with, and regression on, Herm estimates for the same set of data.*

Model		Fixed heritability				Estimated heritability				
		p(A)	μ_{AA}	μ_{Aa}	σ^2_w	p(A)	μ_{AA}	μ_{Aa}	σ^2_w	σ^2_u
Add1	mean	0.47	19.20	9.33	99.78	0.50	22.10	10.96	88.90	0.00
	sd	0.22	4.12	5.44	15.99	0.09	2.13	1.71	9.04	0.00
	slope	1.65	1.29	1.67	1.47	0.40	0.12	0.11	0.53	-
	r	0.88	0.87	0.90	0.91	0.71	0.33	0.40	0.76	-
Add2	mean	0.54	22.98	9.13	77.85	0.49	24.59	12.04	76.16	0.00
	sd	0.10	0.31	2.82	1.68	0.08	2.51	2.41	7.44	0.00
	slope	-	-	-	-	0.72	0.34	0.68	0.88	-
	r	-	-	-	-	0.82	0.49	0.86	0.88	-
Dom	mean	0.52	18.75	20.74	93.84	0.52	23.44	19.05	92.07	0.05
	sd	0.05	2.46	1.53	9.36	0.06	4.04	1.78	9.75	0.46
	slope	0.58	0.57	0.83	1.00	0.77	0.66	0.63	0.83	0.01
	r	0.58	0.88	0.94	0.86	0.60	0.63	0.68	0.74	0.12
Rare	mean	0.13	21.04	7.87	101.87	0.30	21.36	9.15	87.00	0.87
	sd	0.12	7.67	6.89	11.68	0.14	3.69	3.80	11.29	2.78
	slope	0.39	0.68	0.96	0.78	0.57	0.38	0.35	0.90	0.26
	r	0.48	0.61	0.57	0.77	0.62	0.73	0.41	0.88	0.33

Table 5.5 *Variance estimates and the correlation with, and regression on, the values estimated in the simulation for the same data set.*

		Fixed heritability			Estimated heritability		
Model		σ^2_w	σ^2_u	σ^2_{mg}	σ^2_w	σ^2_u	σ^2_{mg}
Add1	mean	99.781	5.252	46.427	88.898	0.000	61.318
	sd	15.994	-	19.940	9.038	0.000	10.966
	slope	1.368	0.006	4.789	0.674	0.000	1.647
	r	0.281	0.014	0.346	0.321	0.000	0.220
Add2	mean	77.851	8.650	71.370	76.155	0.000	76.639
	sd	1.680	-	6.728	7.444	0.000	12.062
	slope	-	-	-	0.245	0.000	1.890
	r	-	-	-	0.130	0.000	0.267
Dom	mean	93.840	4.939	74.454	92.066	0.046	83.745
	sd	9.362	-	12.778	9.751	0.459	11.999
	slope	0.845	0.018	1.151	1.054	0.031	0.824
	r	0.441	0.082	0.472	0.527	0.153	0.361
Rare	mean	101.874	5.362	18.252	86.996	0.873	42.207
	sd	11.678	-	13.761	11.289	2.782	14.862
	slope	1.131	-0.067	1.658	1.063	0.139	1.753
	r	0.403	-0.224	0.327	0.392	0.102	0.338

Table 5.6 *Mean (and standard deviation) of the Herm major gene parameter estimates for the ME1 analyses that gave a non-zero test statistic and for those that gave a zero test statistic.*

	non-zero test statistic				zero test statistic			
Model	t.s.	p(A)	μ_{AA}	μ_{Aa}	t.s.	p(A)	μ_{AA}	μ_{Aa}
Heritability fixed								
Add1	22.205 (6.262)	0.483 (0.117)	22.764 (2.775)	11.367 (2.944)	10.449 (4.601)	0.498 (0.133)	17.743 (4.482)	8.539 (5.448)
Add2	20.642 (4.689)	0.531 (0.058)	25.735 (0.646)	11.444 (1.354)	6.733 (4.464)	0.485 (0.163)	18.715 (5.472)	9.481 (5.196)
Dom	47.593(14.451)	0.505 (0.048)	20.725 (3.799)	20.176 (1.394)	16.107 (0.000)	0.412 (0.000)	17.618 (0.000)	17.497 (0.000)
Rare	16.002 (7.154)	0.221 (0.142)	21.569 (6.826)	10.481 (4.131)	7.005 (2.997)	0.320 (0.192)	13.256 (6.055)	9.083 (5.005)
Heritability estimated								
Add1	5.180 (3.757)	0.501 (0.160)	19.567 (5.971)	9.625 (6.002)	3.454 (2.038)	0.379 (0.209)	13.035 (6.511)	-0.260 (11.058)
Add2	6.736 (3.559)	0.483 (0.091)	22.531 (3.431)	11.459 (2.978)	2.929 (2.648)	0.480 (0.191)	17.417 (5.537)	8.754 (6.123)
Dom	41.129(12.893)	0.505 (0.049)	20.493 (3.881)	20.292 (1.918)	-	-	-	-
Rare	7.437 (4.310)	0.250 (0.146)	19.486 (7.149)	10.316 (4.488)	1.799 (1.645)	0.259 (0.213)	14.984 (11.418)	13.356 (16.957)

t.s. - test statistic

Figure 5.5 *The residual variance estimated with ME1 compared with that estimated with Herm, for Add1 with fixed heritability.*

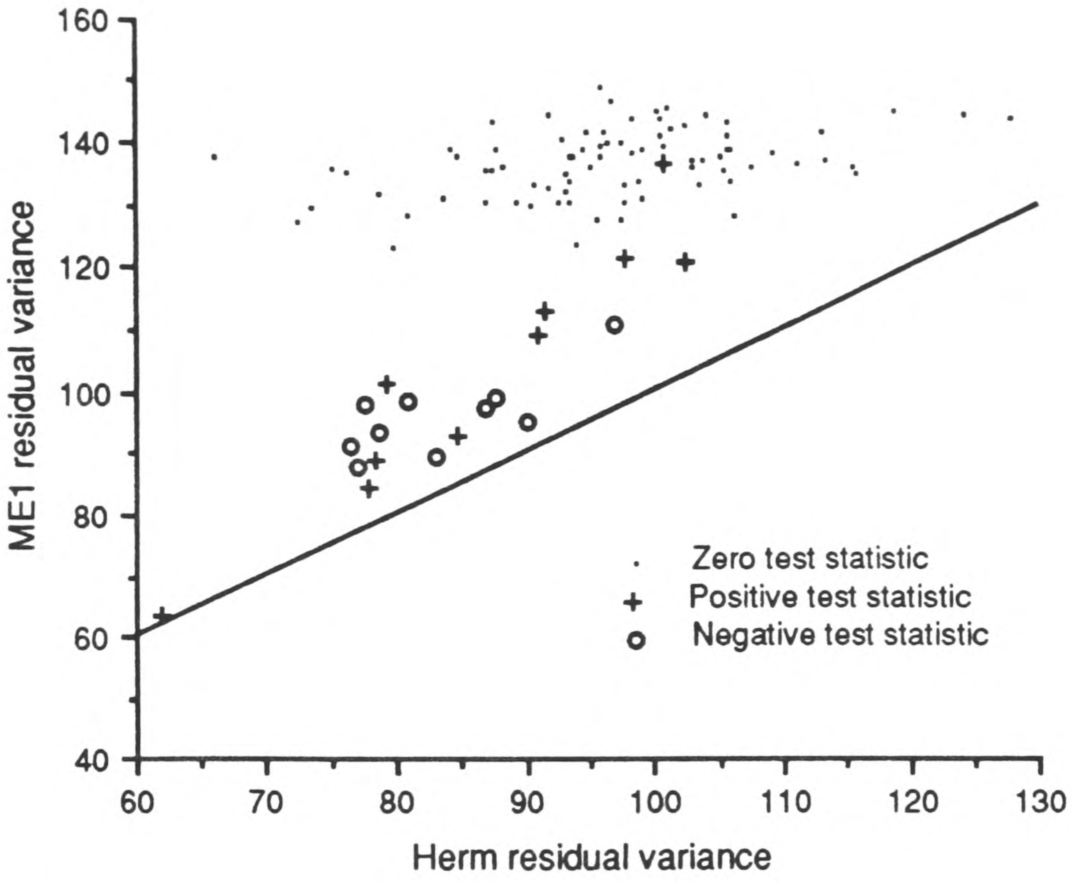
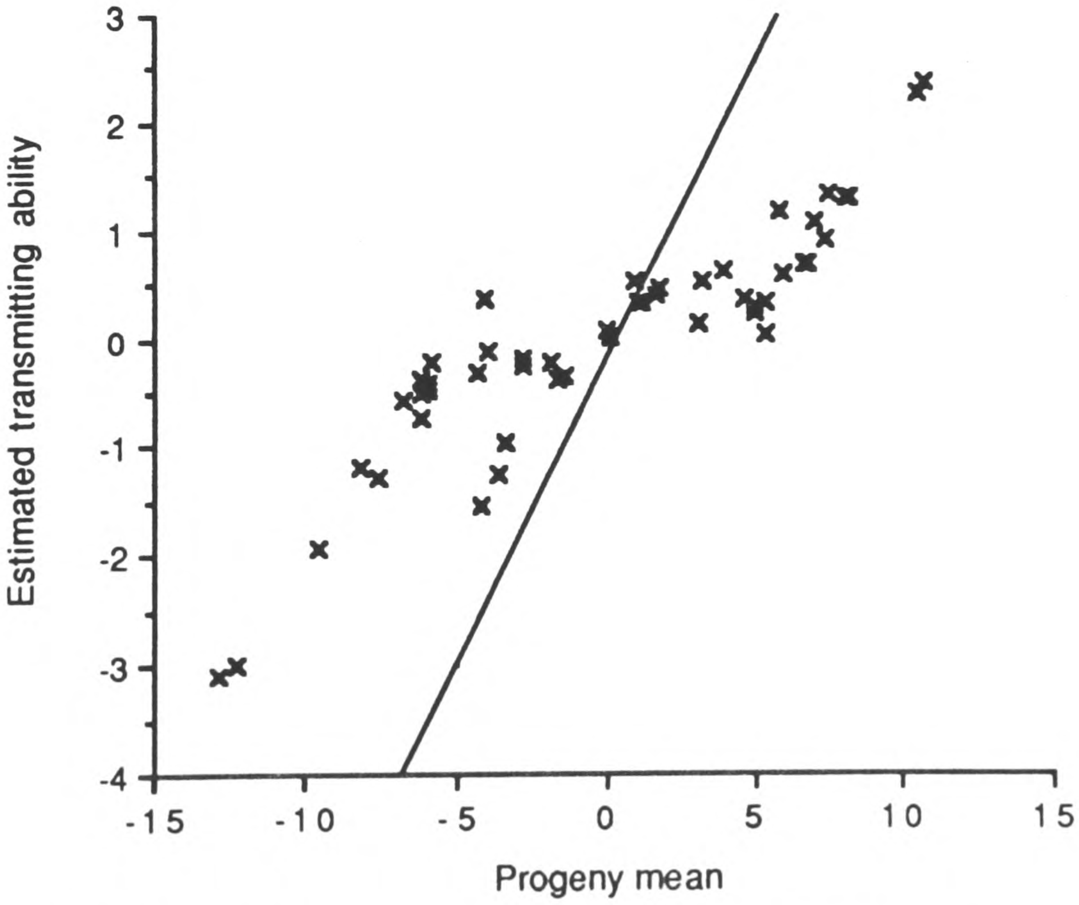


Figure 5.6 *Transmitting abilities estimated for sires from one Add1 analysis that resulted in a mixed model when analysed with fixed polygenic heritability .*



The slope indicates the expected estimates under a polygenic model.

Table 5.7 *Mean correlation (and standard deviation) of the true sire polygenic effect and the estimated effect using ME1, ignoring those that went to a major gene model.*

	fixed heritability		estimated heritability	
Model	non-zero t.s.	zero t.s.	non-zero t.s.	zero t.s.
Add1	0.469 (0.109)	0.433 (0.107)	- [a]	0.406 (0.092)
Add2	0.409 (0.057)	0.598 (0.100)	- [a]	0.591 (0.090)
Dom	0.555 (0.102)	0.503 (0.000)	0.450 (0.215)	-
Rare	0.525 (0.104)	0.503 (0.127)	0.556 (0.045)	0.541 (0.112)

t.s. - test statistic

[a] - resulted in major gene models only.

with non-zero test statistic is under estimated and the major gene parameters over estimated. The residual variance is also under estimated.

The ME1 analyses also estimated the polygenic transmitting ability of each sire and this can be compared with the true value from the simulation. The results are given in table 5.7. When the analyses resulted in a test statistic of zero, the transmitting abilities obtained are the estimates that would be obtained if a polygenic model was assumed. Hence, the effect of the major gene component is also included in these estimates. When a mixed model was obtained the major gene component should be removed from the sire effect and the correlation with the true value increased. This is supported by the results in table 5.7: when a non-zero test statistic (a mixed model) was obtained the correlations were on average higher for most models than when a zero test statistic (polygenic model) was obtained. With the heritability estimated most of the analyses where a non-zero test statistic was obtained gave major gene models with the transmitting abilities equal to zero for all sires. Figure 5.6 shows the relationship between the estimated transmitting ability and the progeny mean of sires from one Add1 data set that resulted in a mixed model. The slope indicates the transmitting abilities that would be estimated assuming the same fixed heritability but ignoring the major gene.

Genotyping the sires at the major locus

The probability of each sire being each genotype was calculated using equation [5.4]. For each sire, the probability of being the correct genotype was grouped into one of three classifications. The first, if the probability of being the correct genotype was greater than 0.9, the second greater than 0.75 and the third greater than 0.5. For each

Table 5.8 Percentage of sires correctly genotyped at different values of the conditional probability required to be assigned a genotype, for analyses that resulted in a non-polygenic model.

	AA			Aa			aa			Total		
Model	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5	>0.9	>0.75	>0.5
Fixed heritability												
Add1	28.7	50.0	61.1	16.5	34.9	51.6	36.6	61.3	69.3	30.7	25.4	46.1
Add2	48.6	59.0	66.2	38.4	52.3	56.6	36.5	65.9	69.8	41.0	58.0	63.0
Dom	42.9	71.1	86.9	17.4	50.9	75.8	18.8	38.7	59.9	24.1	52.4	74.4
Rare	5.4	10.4	15.6	11.5	17.9	26.6	58.3	87.8	94.0	42.7	63.9	70.9
Estimated heritability												
Add1	37.0	54.6	69.4	25.4	48.1	67.1	38.7	54.6	68.2	31.2	50.8	67.5
Add2	47.1	58.0	65.6	32.8	50.3	63.1	46.8	59.8	69.4	39.9	54.5	65.2
Dom	51.9	71.5	81.6	29.7	53.9	73.4	27.5	42.8	60.7	34.4	55.1	71.8
Rare	37.9	57.4	67.4	20.1	39.5	54.8	42.4	60.8	71.2	35.6	54.6	66.5

analysis the percentage of correctly genotyped sires was calculated for each genotype and the total percentage correctly genotyped over all genotypes. The results are given in table 5.8 as the mean percentage correctly genotyped over the analyses that resulted in a non-polygenic model.

The results are similar to those obtained from Herm. Over all genotypes when the heritability was assumed to be known, if the criterion for a sire being assigned to a particular genotype was that his conditional probability for that genotype was greater than 0.9, the highest number of sires were correctly genotyped for the additive major gene with rare allele. If 0.5 was taken as the criterion for classification the highest number of sires were correctly genotyped when the major gene had a dominant effect. The results from the additive model with a polygenic heritability of 0.4 are based on very few analyses, in which, presumably, evidence for the major gene was large and the sires relatively easy to genotype. The conditional probabilities are more extreme, nearer to one or zero, than those calculated using Herm, which results in more sires being correctly genotyped at probabilities of 0.75 and 0.9. However, using a criterion of 0.5 the proportion correctly genotyped is similar to the proportion correct with Herm.

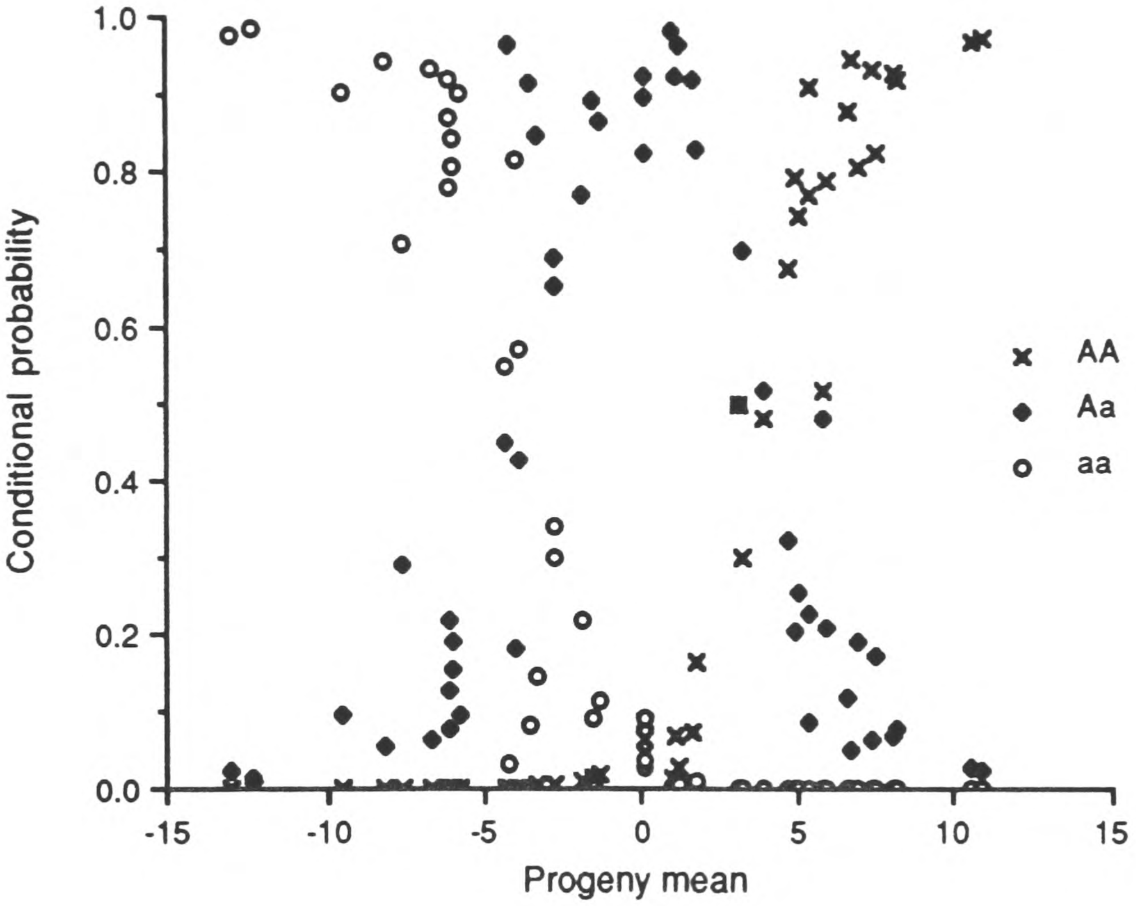
Figure 5.7 shows the probability of the sire being each genotype plotted against the mean of his progeny (a) and against his estimated transmitting ability (b) for one of the simulations with a major gene with additive effect and polygenic heritability of 0.2 (Add1) that resulted in a mixed model. Few sires have intermediate probabilities for a genotype, as suggested previously. Otherwise the distribution, when plotted against the progeny mean, is similar to that obtained from Herm, with sires whose progeny have a high mean having a high probability of being genotype AA and those with low means having a high probability of being genotype aa. The transmitting ability of the sire and his major genotype are assumed to be independent in the analysis. However, both the conditional probabilities and the transmitting ability estimate are functions of the progeny data, figure 5.7b illustrates a relationship between them.

5.4.5 Discussion

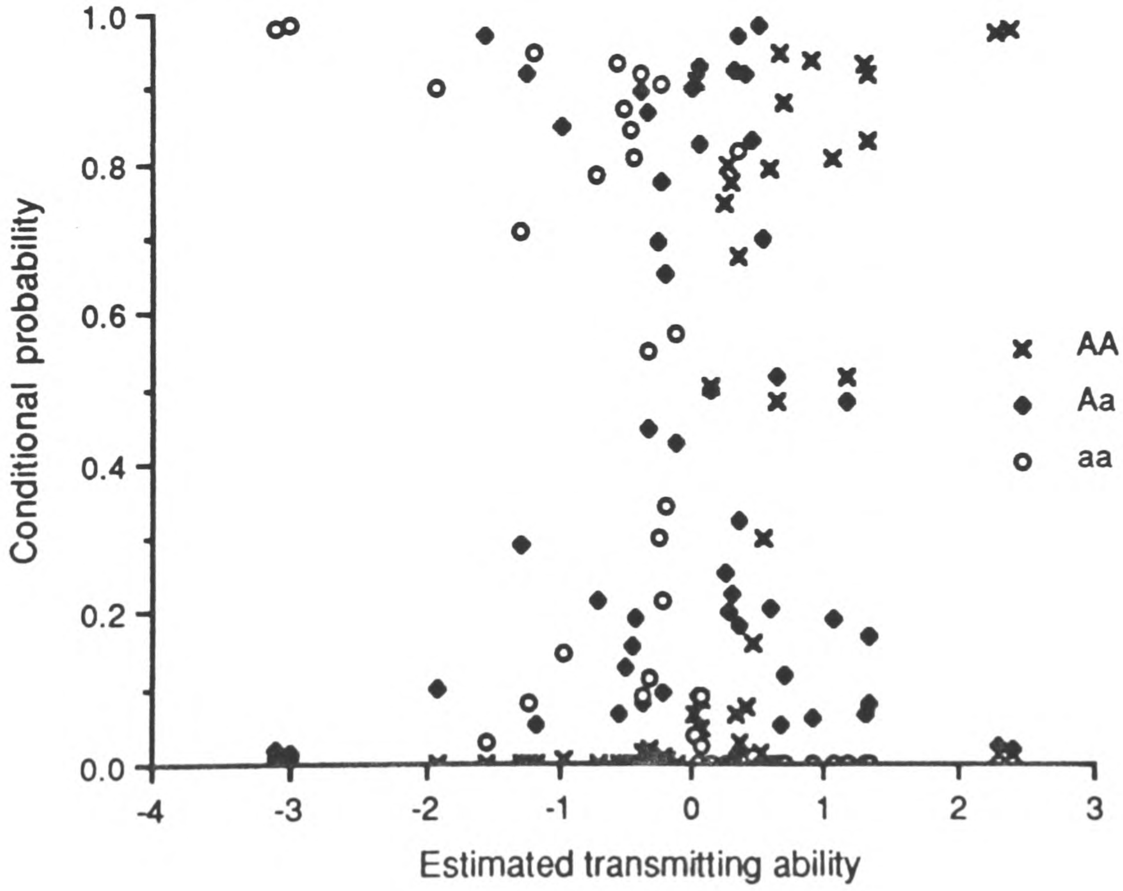
Using the ME1 approximation to the mixed model there is obviously a problem in correctly identifying a major gene, in that the test statistic obtained when the data contains a major gene is much lower than expected. Hence, comparing the test statistic to the relevant χ^2 distribution gives very few analyses where evidence for a major gene is suggested, except when the simulated major gene had a dominant allele. In this dominant case the power of the approximation was similar to the power using Herm although the mean test statistic was lower. Considering the additive major genes simulated, when one

Figure 5.7 *Probability of a sire being each genotype, for sires from an Add1 analysis, with fixed heritability that resulted in a mixed model.*

a) plotted against the mean performance of the progeny of the sire.



b) plotted against the estimated transmitting ability for each sire.



of the alleles was rare, evidence for a major gene was found in more analyses than when both alleles were at equal frequency. The resulting distribution of phenotypes for the major gene with a rare allele is skewed and this appears to have a greater influence on the detection of a major gene than the proportion of the genetic variance explained by the major gene, which is lower in this case than for the equal allele frequency situation. Reducing the value of the test statistic required to provide evidence for a major gene might increase the power without increasing the number of genes detected in data that does not contain a major gene. However, the high proportion of analyses that resulted in a polygenic model suggests that the power will never be very high. Also reasonable mixed model parameter estimates can be obtained and yet give a negative test statistic with the polygenic model being more likely.

Using an assumed value for the polygenic heritability might bias the results. If the value used was an underestimate of the true polygenic heritability in the data, a major gene may be inferred in order to explain the additional genetic variance that cannot be accounted for by the (fixed) polygenic heritability, even if it is polygenic in origin. Analysing the polygenic data simulated with an heritability of 0.4 but assuming a value of 0.2, increased the number of analyses resulting in a non-zero test statistic, and one became significant at the 5% level.

When a major gene was detected, the parameter results estimated, on average, gave a reasonable indication of the character of the major gene present in the data.

When estimating the heritability, both major gene and polygenic variation are no longer required to explain the proportion of genetic to environmental variation in the data. The ME1 approximation has difficulty distinguishing the two sources of genetic variation and suggests a model containing either all major gene or all polygenic variation. The parameter results obtained from these analyses that resulted in a major gene model, in general, gave a good indication as to the effect of the gene, whether it was additive or dominant, and the frequency of the allele in the population. The effect of the major gene was over estimated, as would be expected because of the inclusion of some of the polygenic component in this estimate.

Elsen and Le Roy (1989) give the percentage of analyses in which evidence for a major gene was found for models equivalent to Add1 and Dom (h_{12} and h_{11} respectively). The models for analysis estimate the sire genotype frequencies and the dam allele frequency. They compare the test statistics obtained from analysis of mixed model data with 5% quantiles obtained from the distribution of test statistics from analysis of polygenic data. With fixed polygenic heritability and 20 sires each with 20 half-sibs, this quantile (estimated at 5.82), is much lower than expected. Based on 100 simulations, using this criterion 80% of the analyses of data containing a dominant major gene gave significant test statistics and 21% with an additive major gene. The increased frequency of detection of the additive major gene, compared with the results presented here, is presumably because of the lower quantile used. The fewer significant results from the data containing the dominant gene could be because of the decreased sample size. With heritability estimated, the test statistics were compared with an estimated 5% quantile of 10.78, much closer to the expected value from a χ^2 distribution with five degrees of freedom. A major gene was detected in 77% of the analyses when a major gene with dominant effect was segregating in a polygenic background with heritability of 0.2 and in 67% when the polygenic heritability was 0.6. For the simulated additive major gene, a major gene was detected in 8% and 7% of analyses when the polygenic heritability was 0.2 and 0.6, respectively. These results of Elsen and Le Roy (1989) are in agreement with those presented here. A dominant major gene is easier to find than an additive gene that was simulated with the same difference in effect between the homozygous genotypes and increasing the polygenic heritability decreases the number of significant test statistics. The effect of estimating the polygenic heritability is difficult to ascertain in the study of Elsen and Le Roy (1989), as the criterion used for determining a significant test statistic when the heritability is estimated is not the same as when the heritability is assumed to be known.

Hoeschele (1988b) and Elsen and Le Roy (1989) consider the ability of the approximation to obtain reasonable parameter estimates. Hoeschele reported that the inclusion of the probabilities of transmission of the major gene from sires to offspring caused convergence to be slow or not attained. Hence, she suggests ignoring these and, in effect, making genotypes fixed effects, with each offspring having a probability of being each genotype. This probability is based on the probability of the phenotype given the major gene effect and the population major genotype frequencies, the sire's transmitting ability and any fixed effects . The inheritance of the polygenic component is as usual. Results are given based on data simulated under two different additive models both with a rare allele. An unbalanced sire model was used with 2500 records from 200 sires and 200 herd-year-seasons. Hoeschele found that the mean results based on 10 replicates of the simulations gave good estimates of the genotype frequencies and effects. However, she only estimated the variance components for one model which had an expected heritability of 0.1 and, using equations equivalent to the REML equations given in Appendix 3, found that the sire variance was, on average, twice the expected value and the major gene variance was under estimated. This result is in contrast to the results of Elsen and Le Roy (1989) who found that the major genotype means could be well estimated when the heritability was estimated, but the heritability was always under estimated and in many analyses became fixed at zero. In the results presented here, the sire variance is given as the average from those analyses which resulted in a non-polygenic model, and hence was under estimated because of the high number of analyses giving a major gene model. However the inclusion of all the results into the mean will result in an increase in the heritability, and a decrease in the major gene estimates. Hoeschele (1988b) does not report obtaining polygenic models when maximising the variance components for mixed model.

In the analyses of mixed model data, two different sets of estimates were used to start the maximisation process (see section 5.4.1). If the two analyses of the same data resulted in different maxima, the results of the analysis giving the highest likelihood were used. With fixed polygenic heritability, the analyses were not very sensitive to the initial model used, the worst case being for the data containing the major gene simulated with a rare allele. In 13 of the 100 analyses of this data, one of the initial models resulted in a polygenic model and the other a mixed model. When the polygenic heritability was estimated, the analyses were more sensitive, especially for the simulated additive major gene in a polygenic background with an heritability of 0.4 (Add2) and the one with a rare allele (Rare). In 41 of the analyses of Add2 data, starting with the expected parameter values resulted in a polygenic model and the alternative parameter values, which explained a larger proportion of the genetic variance, resulted in a major gene model.

5.5 COMPARISON OF ME1 AND EXACT LIKELIHOODS

The potential advantages of ME1 over Herm in its computation and its ease of extension to include, for example, fixed effects, suggest that further investigation of this method is merited. For this purpose the genetic model has been simplified. The polygenic component is derived as usual, half from the sire and half from the dam. However there are only two genotypes at the major locus in the offspring and the genotype of an individual depends entirely upon the major genotype of the sire, as if, for example, AA and aa sires were mated to AA females. This gives the following transmission probabilities:

		Offspring genotype	
		AA	Aa
Sire	AA	1	0
Genotype	a a	0	1

In this case, the exact mixed model likelihood can be written estimating the mode of the transmitting ability distribution for each major genotype of each sire. This is because, given the sire's major genotype, there is only one possible combination of major genotypes for his offspring, hence, in total, for each half-sib family there are only two possible combinations, giving two sire estimates, one for each major genotype for the sire. The exact and ME1 likelihoods can be written as follows:

Exact mixed model likelihood:

$$L(MM)=\prod_{i=1}^{s}\sqrt{\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)}\left(p(1)h(\hat{u}_{i1})\prod_{j=1}^{n}k_1(y_{ij}\,|\,\hat{u}_{i1},\mu_1,\sigma_w^2)+p(2)h(\hat{u}_{i2})\prod_{j=1}^{n}k_2(y_{ij}\,|\,\hat{u}_{i2},\mu_2,\sigma_w^2)\right)$$

Where the two genotypes are denoted 1 and 2, and $\hat{u}_{i1}$ and $\hat{u}_{i2}$ are the sire estimates for sire i, given that he has genotype 1 or 2.

ME1 mixed model likelihood:

$$L(MM)=\prod_{i=1}^{s}\sqrt{\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)}\left(p(1)h(\hat{u}_i)\prod_{j=1}^{n}k_1(y_{ij}\,|\,\hat{u}_i,\mu_1,\sigma_w^2)+p(2)h(\hat{u}_i)\prod_{j=1}^{n}k_2(y_{ij}\,|\,\hat{u}_i,\mu_2,\sigma_w^2)\right)$$

5.5.1 Parameter estimates

In the same way as for the three genotype model, equations for the ML estimates of the parameters can be obtained by partially differentiating the log likelihood with respect to each parameter.

Transmitting abilities

After simplification, the two estimates of the transmitting abilities for each sire from the exact likelihood can be written as follows:

$$\hat{u}_{i1}=\frac{1}{n+\lambda}\sum_{j=1}^{n}(y_{ij}-\mu_1) \qquad \hat{u}_{i2}=\frac{1}{n+\lambda}\sum_{j=1}^{n}(y_{ij}-\mu_2)$$

Hence it can be easily shown that with this model there is a constant difference between the sire effect for each genotype, this difference being a function of the difference between the major genotype effects. This is because there is no segregation of the major gene within each half-sib family with this model.

$$\hat{u}_{i1}-\hat{u}_{i2}=\frac{1}{n+\lambda}\,n(\mu_1-\mu_2)$$

Calling this difference Δ the exact likelihood can be written in terms of $\hat{u}_{i1}$ and Δ, where $\hat{u}_{i2}=\hat{u}_{i1}+\Delta$.

$$L(MM)=\prod_{i=1}^{s}\sqrt{\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)}\left(p(1)h(\hat{u}_{i1})\prod_{j=1}^{n}k_1(y_{ij}\,|\,\hat{u}_{i1},\mu_1,\sigma_w^2)+p(2)h(\hat{u}_{i1}+\Delta)\prod_{j=1}^{n}k_2(y_{ij}\,|\,\hat{u}_{i1}+\Delta,\mu_2,\sigma_w^2)\right)$$

It can be seen that when Δ is fixed at zero this gives the ME1 likelihood. With this condition the following sire estimate is obtained:

$$\hat{u}_{i0}=\frac{1}{n+\lambda}\sum_{j=1}^{n}(y_{ij}-\mu_1)+q_{i0}(2)\,\Delta$$

Where: $\hat{u}_{i0}$ is the sire estimate when Δ is equal to zero.

$q_{i0}(c)$ is the conditional probability that sire i has genotype c calculated when Δ is equal to zero.

If $q_{i0}(2)$ is close to zero, i.e. the sire has a high probability of being genotype AA, the sire effect estimated under ME1 is close to the estimate obtained under the exact model assuming that the sire is genotype AA. When $q_{i0}(2)$ is close to one, i.e. the sire has a high probability of being genotype aa, the sire estimate approaches the one estimated under the assumption of that genotype in the exact model. Otherwise the sire estimate pools the two exact estimates according to the relative probabilities of the genotypes, obtained under ME1.

Variance estimates

Assuming a fixed polygenic heritability, the following expression for the residual variance can be obtained from the exact mixed model likelihood:

$$\sigma_w^2 sn=\sum_{i=1}^{s}\sum_{c=1}^{m}q_i(c)\left(\hat{u}_{ic}^2\lambda+\sum_{j=1}^{n}(y_{ij}-\mu_c-\hat{u}_{ic})^2\right)$$

which, under this model, can be simplified to:

$$\sigma_w^2 sn = \sum_{i=1}^{s}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 + \frac{n\lambda}{n+\lambda}\sum_{i=1}^{s}(\bar{y}_{i.}-\bar{y}_{..})^2 - \frac{sn\lambda}{n+\lambda}p_1p_2(\mu_1-\mu_2)^2$$

Where: $p_1 = \dfrac{\sum_{i=1}^{s} q_i(1)}{s}$, an estimate of p(1)

i.e. the variance is composed of the within sire sum of squares and, because the heritability is fixed, a proportion of the between sire sum of squares and major gene variance, as expected.

For the ME1 likelihood, the following equation is obtained:

$$\sigma_w^2 sn = \sum_{i=1}^{s}\hat{u}_i^2\lambda + \sum_{i=1}^{s}\sum_{c=1}^{m} q_{i0}(c)\sum_{j=1}^{n}(y_{ij}-\mu_c-\hat{u}_i)^2$$

which can be rewritten as:

$$\sigma_w^2 sn = \sum_{i=1}^{s}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 + \frac{n\lambda}{n+\lambda}\sum_{i=1}^{s}(\bar{y}_{i.}-\bar{y}_{..})^2 - \frac{sn\lambda}{n+\lambda}p_1p_2(\mu_1-\mu_2)^2$$
$$+ \frac{n^2}{(n+\lambda)}(\mu_1-\mu_2)^2\left(\sum_{i=1}^{s}(q_i(1)-p_1)^2 - sp_1p_2\right)$$

This equation contains an extra component which is a function of the discrepancies between the conditional genotype probabilities for each sire and the population frequency and an additional proportion of the major gene variance. If the major gene means and population genotype frequencies are the same in the two likelihoods, the ME1 likelihood will over estimate the residual variance compared with the estimate obtained from the exact likelihood when $\sum_{i=1}^{s}(q_{i0}(1)-p_1)^2$ is larger than sp_1p_2, . This will depend on the value of the genotype frequencies in the population. For example, if $p_1 = p_2 = 0.5$ then the value of the additional component can at most (theoretically) be zero and will, in fact, always be negative and the variance under estimated compared with the value from the exact likelihood. Whereas if p_1 is closer to one or zero the expression is likely to be positive and the ME1 variance over estimated.

5.5.2 Likelihoods

Using a three genotype model, even when a major gene was identified and the parameter results were reasonable, the likelihood for ME1 was under estimated compared with the exact.

Data were simulated with 50 unrelated sires each with 20 half-sib offspring. The polygenic heritability was 0.2 and there were two major genotypes each with a frequency of 0.5 and with about $\frac{2}{3}$ phenotypic standard deviations between the means. The major genotype means (μ_1 and μ_2), population genotype frequency (p(1)) and the residual variance (σ_w^2) were fixed at their ML estimates for the exact likelihood. For each method the sire effects (u_{i1} and Δ or u_i) were estimated using an EM algorithm based on 1st derivatives as described previously (5.3.1), which results in iterating using the following equations:

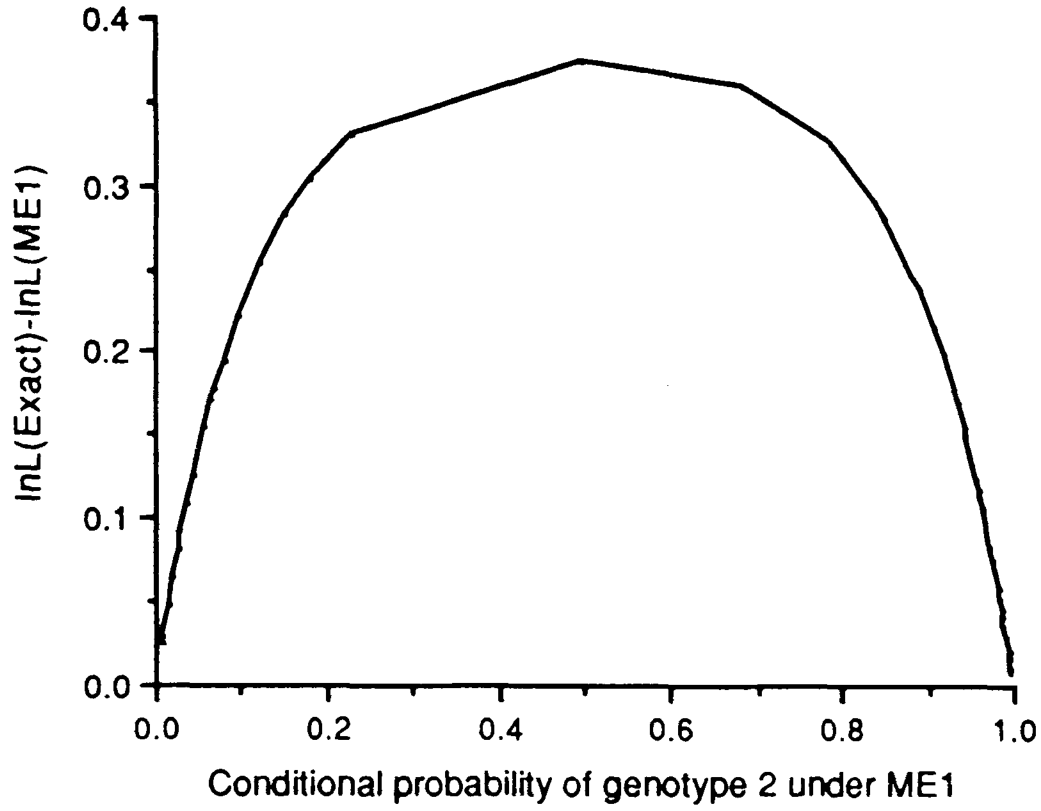
$$\frac{n+\lambda}{\sigma_w^2}\begin{bmatrix} \sum_{i=1}^{s} q_i(2) & q_1(2) & \cdots & q_s(2) \\ q_1(2) & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ q_s(2) & 0 & \cdots & 1 \end{bmatrix}\begin{bmatrix} \Delta \\ u_1 \\ \vdots \\ u_s \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{s} q_i(2)\sum_{j=1}^{n}(y_{ij}-\mu_2) \\ q_1(1)\sum_{j=1}^{n}(y_{1j}-\mu_1)+q_1(2)\sum_{j=1}^{n}(y_{1j}-\mu_2) \\ \vdots \qquad \vdots \\ q_s(1)\sum_{j=1}^{n}(y_{sj}-\mu_1)+q_s(2)\sum_{j=1}^{n}(y_{sj}-\mu_2) \end{bmatrix}\frac{1}{\sigma_w^2} \tag{5.5}$$

When maximised under these conditions the two methods give different log likelihoods, with the data being more likely when maximised under the exact model than when maximised under ME1. The difference between the likelihoods can be explained in terms of the sire effects, Δ and the conditional sire genotype probabilities ($q_i(c)$) which are different in the two methods as a consequence of the other changes.

Figure 5.8 gives the difference between the exact and ME1 log likelihoods for each sire plotted against his conditional probability of being genotype 2 using the ME1 likelihood ($q_{i0}(2)$). When $q_{i0}(2)$ is equal to one or zero the exact and approximate models give the same likelihood value, as then, effectively, the genotype of the sire is known. Otherwise the approximation always under estimates the likelihood. The largest difference between the likelihoods is when $q_{i0}(2)$ is equal to 0.5 as then the sire estimate under

ME1 is half-way between the two estimates under the exact model. Investigation is required to see what is causing the difference.

Figure 5.8 *The difference between the exact and the ME1 log likelihoods for each sire plotted against his conditional probability.*



Explaining the difference between the likelihoods

One way of considering this would be to see what would have to be added to the ME1 likelihood in order for it to equal the exact. Using a Taylor expansion it is possible to approximate the exact likelihood in terms of the ME1 likelihood. A first approximation would be to assume that the difference can be explained in terms of Δ. Ignoring terms beyond a quadratic the following equation is obtained.

$$\ln L(\hat{\Delta}) \approx \ln L(0) + \hat{\Delta}\frac{\partial \ln L(0)}{\partial \Delta} + \hat{\Delta}^2 \frac{1}{2}\frac{\partial^2 \ln L(0)}{\partial \Delta^2} \qquad [5.6]$$

Where: $\ln L(\hat{\Delta})$ is the exact ln likelihood.

$\ln L(0)$ is the ME1 ln likelihood.

$\hat{\Delta}$ is the ML estimate for Δ obtained from the exact likelihood.

Approximation 1. Initially the partial 2nd derivatives of the ME1 likelihood with respect to Δ and u_i can be approximated by assuming that the conditional sire probabilities are constant, i.e. that $\frac{\partial q_i(c)}{\partial \Delta}$ is equal to zero. The model can then be considered as linear and hence the matrix of 2nd derivatives is equal to minus the coefficient matrix from equation [5.5]. Absorbing the sire effects gives a value for $\frac{\partial^2 \ln L(0)}{\partial \Delta^2}$ taking account of the change in u_i caused by the change in Δ. Incorporating this and the 1st derivative into equation [5.6] gives the following approximation for the exact likelihood:

$$\ln L(\hat{\Delta}) \approx \ln L(0) + \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s} q_{i0}(1)\, q_{i0}(2)$$

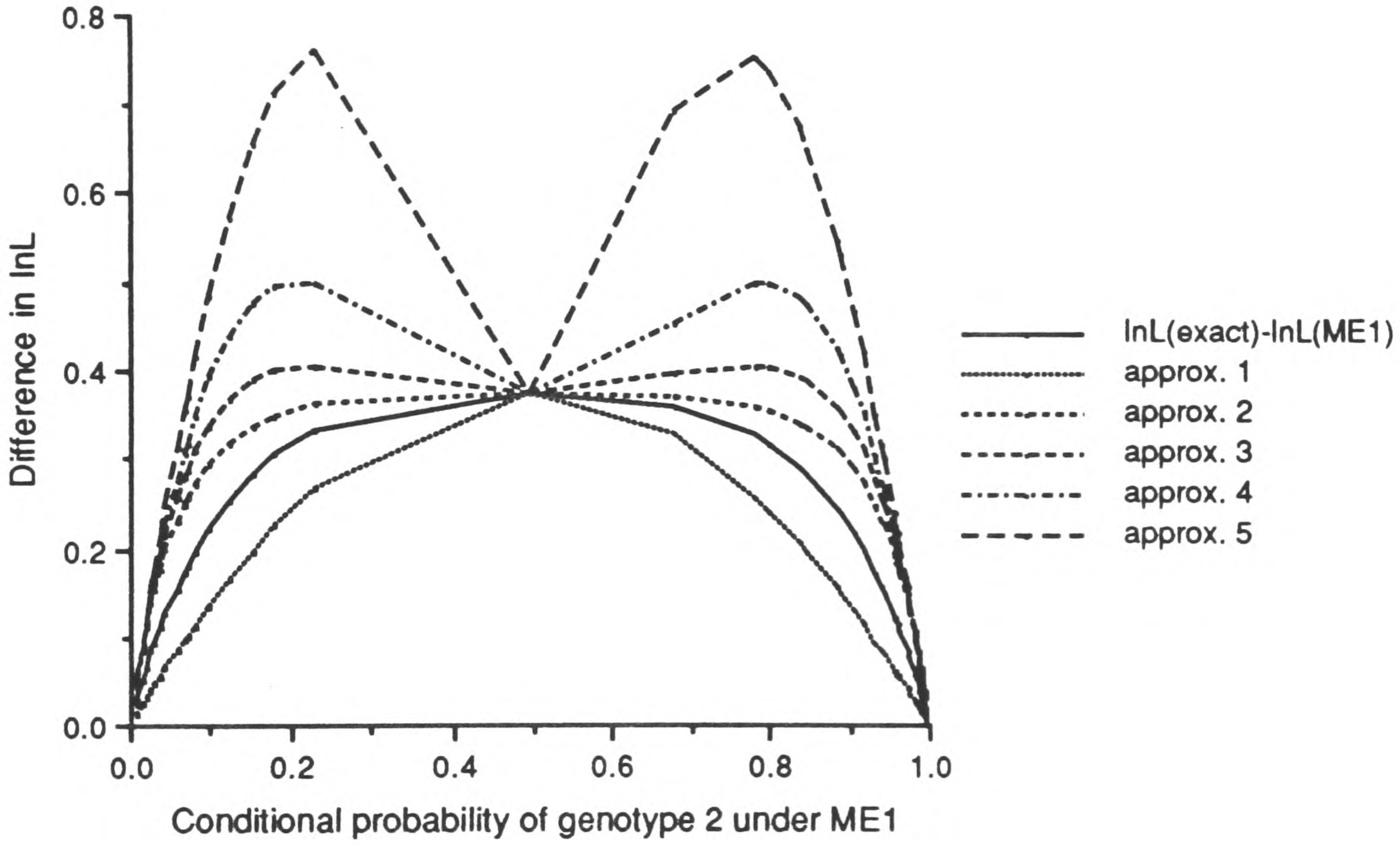
Figure 5.9 shows how well this approximation (approx. 1) explains the difference between the exact and the ME1 likelihood. When $q_{i0}(2)$ is equal to one or zero the approximation adds nothing to the ME1 likelihood and hence it still has the same value as the exact likelihood. When $q_{i0}(2)$ is equal to 0.5 the difference between the likelihoods can be totally explained by the approximation. At all other values of $q_{i0}(2)$ the approximation under estimates the difference.

Approximation 2. Alternatively, it might be easier to consider the difference between the likelihoods by expanding about the exact likelihood to obtain an expression for the approximate likelihood. Making the same approximation for the 2nd differentials as for approximation 1, the following equation is obtained:

$$\ln L(0) \approx \ln L(\hat{\Delta}) - \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s} q_{i\Delta}(1)\, q_{i\Delta}(2)$$

This is similar to approximation 1 except that the conditional genotype probabilities are obtained from the maximum of the exact likelihood rather than the ME1 likelihood. This approximation is shown in figure 5.9 as approx. 2. It can be seen that for all sires, except when their conditional probability is equal to 1, 0.5 or 0, the approximation over estimates the difference between the likelihoods. The conditional probabilities estimated under the ME1 likelihood are more extreme (closer to one or zero) than those estimated using the exact likelihood with the same estimates for the major gene effect and variance components. Hence, $q_{i0}(1)q_{i0}(2)$ is less than $q_{i\Delta}(2)q_{i\Delta}(2)$ and the approximated value of the difference between the likelihoods is smaller for approximation 1 than approximation 2.

Figure 5.9 *Five approximations of the difference between the exact and approximate mixed model likelihood.*



See text for explanations of the approximations.

It is not clear from these equations what assumption has been made about the change in the sire effects that occurs with the change in Δ. Rather than absorbing the sire equations a Taylor expansion for more than one variable can be used. For example expanding about lnL(0):

$$\ln L(\hat{\Delta}) \approx \ln L(0) \quad + \begin{bmatrix} \hat{\Delta} & \delta u_1 & \ldots & \delta u_s \end{bmatrix} \begin{bmatrix} \frac{\partial \mathrm{LnL}(\hat{\Delta})}{\partial \Delta} \\ \frac{\partial \ln L(\hat{\Delta})}{\partial u_1} \\ \vdots \\ \frac{\partial \ln L(\hat{\Delta})}{\partial u_s} \end{bmatrix}$$

$$+\frac{1}{2} \begin{bmatrix} \hat{\Delta} & \delta u_1 & \ldots & \delta u_s \end{bmatrix} \begin{bmatrix} \frac{\partial^2 \mathrm{LnL}(\hat{\Delta})}{\partial \Delta^2} & \frac{\partial^2 \mathrm{LnL}(\hat{\Delta})}{\partial \Delta \partial u_1} & \ldots & \frac{\partial^2 \mathrm{LnL}(\hat{\Delta})}{\partial \Delta \partial u_s} \\ \frac{\partial^2 \ln L(\hat{\Delta})}{\partial u_1 \partial \Delta} & \frac{\partial^2 \ln L(\hat{\Delta})}{\partial {u_1}^2} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial^2 \mathrm{LnL}(\hat{\Delta})}{\partial u_s \partial \Delta} & 0 & \ldots & \frac{\partial^2 \mathrm{LnL}(\hat{\Delta})}{\partial {u_s}^2} \end{bmatrix} \begin{bmatrix} \hat{\Delta} \\ \delta u_1 \\ \vdots \\ \delta u_s \end{bmatrix}$$

[5.7]

It can be shown that absorbing the sire part of the matrix of 2nd derivatives is equivalent to using $-\left(\frac{\partial^2 \ln L(0)}{\partial {u_i}^2}\right)^{-1} \left(\frac{\partial^2 \ln L(0)}{\partial u_i \partial \Delta}\right)$ for the change in the sire effect which accompanies a change in Δ. Hence, in the above situations, where the 2nd derivatives have been approximated, absorbing sires has been equivalent to using $q_{i0}(2)\hat{\Delta}$, in approximation 1, and $q_{i\Delta}(2)\hat{\Delta}$, in approximation 2. Figure 5.10 shows the change in the estimate of the sire's transmitting ability caused by a change in Δ for three sires who have a conditional probability of being genotype 1, calculated under ME1, of approximately 1.00, 0.75 and 0.50. When the conditional probability is equal to 0.5 the change in u_i is linear because the conditional probability is constant over all values of Δ. Hence, with this conditional probability, the approximations can explain the difference between the exact and ME1 likelihoods because the correct change in the sire's transmitting ability with a change in Δ can be used. Also at extreme conditional probabilities, there is little change in the value of the probabilities with a change in Δ. However, with intermediate values the change in the transmitting ability is not linear with a linear change in Δ because the conditional probability is not constant, but becomes nearer to zero or one as Δ decreases. Figure 5.11 illustrates the assumptions about the change in the transmitting ability of a sire with a change in Δ for a sire with a high probability of being genotype 1.

Figure 5.10 *Estimates of sire transmitting abilities for three sires at different values of Δ*

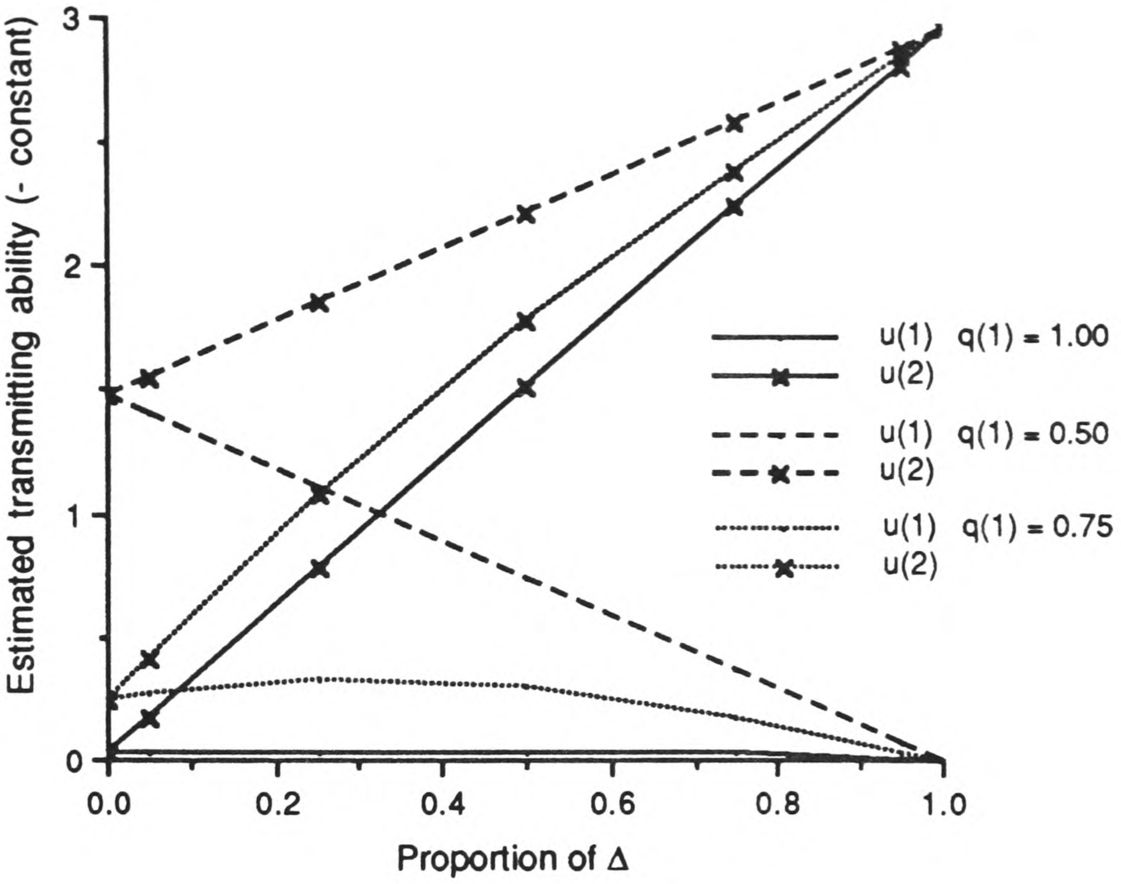
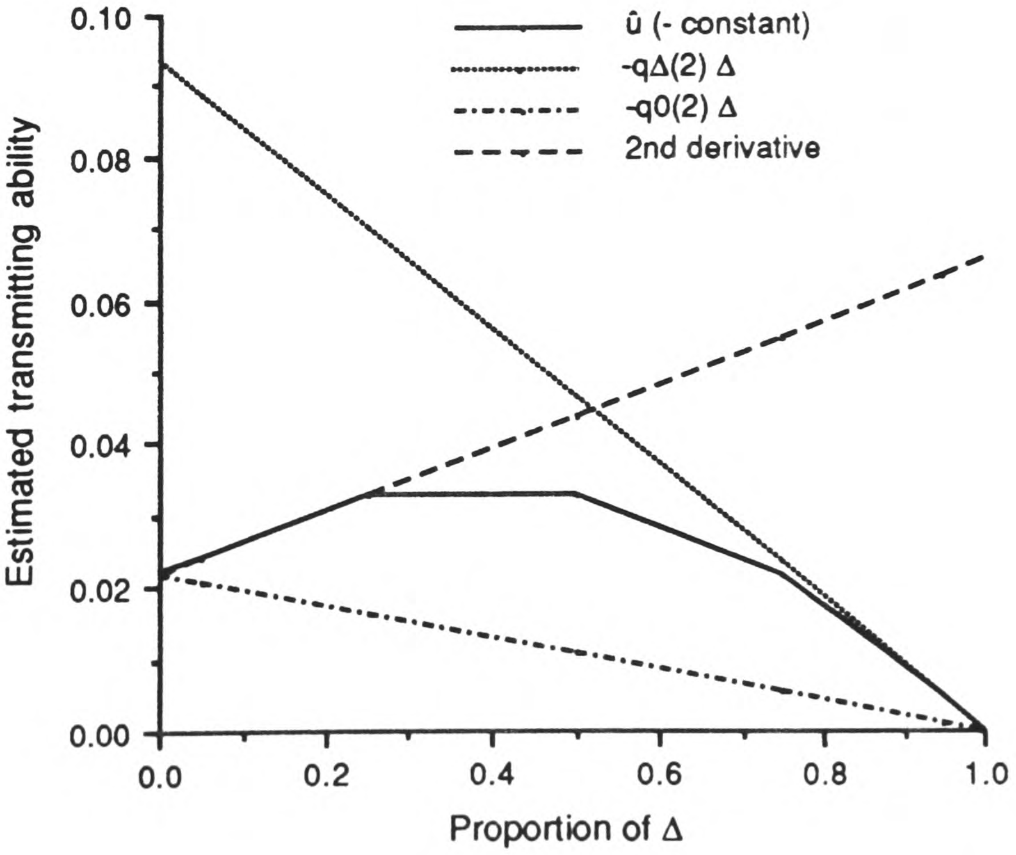


Figure 5.11 *Approximations to the change in the transmitting ability that accompanies a change in Δ.*



Approximation 3. It can be easily shown that the total change in u_i when Δ is fixed at zero compared with when it is estimated is equal to $q_{i0}(2)\hat{\Delta}$. This has been used when expanding about the ME1 likelihood but not for the exact likelihood. An approximation for the ME1 likelihood in terms of the exact using $q_{i0}(2)\hat{\Delta}$ as the change in sire estimates and approximating the 2nd derivatives as before is as follows:

$$\ln L(0,\hat{u}_{i0}) \approx \ln L(\hat{\Delta},\hat{u}_{i\Delta}) - \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s}(q_{i\Delta}(1)\,q_{i\Delta}(2) + (q_{i\Delta}(2)-q_{i0}(2))^2)$$

This is shown as approx. 3 in figure 5.9. For most values of $q_{i0}(2)$ this approximation over estimates the difference between the exact and the ME1 likelihoods to a larger extent than approximation 2.

As the conditional probabilities are functions of the other parameters and hence, the model not linear, an improvement in the approximations might be achieved by using the true 2nd derivatives, i.e. including the derivatives of these probabilities with respect to Δ and u_i. The matrix of 2nd derivatives is given in figure 5.12. Under the exact model, incorporating the additional derivative of the conditional probabilities with respect to Δ and u_i into the 2nd differentials causes no change. This is because under this model the equations for u_i and Δ can be simplified so as not to incorporate $q_i(c)$.

Approximation 4. Incorporating the true 2nd derivatives into the approximation for the exact likelihood in terms of the ME1 likelihood [5.7] and using the observed change in the sire estimates ($q_{i0}(2)\hat{\Delta}$) gives the following approximation:

$$\ln L(\hat{\Delta},\hat{u}_{i\Delta}) \approx \ln L(0,\hat{u}_{i0}) + \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s}\left(q_{i0}(1)\,q_{i0}(2)\left(1+\frac{n+\lambda}{\sigma_w^2}\hat{\Delta}^2\,(q_{i0}(1)-q_{i0}(2))^2\right)\right)$$

Which is shown as approx. 4 in figure 5.9. This approximation over estimates the difference between the exact and ME1 likelihoods and is worse than previous approximations, especially for values of $q_{i0}(2)$ around 0.25 and 0.75 .

Approximation 5. Alternatively the 2nd derivatives could be used for the change in the transmitting ability. Absorbing the sire effects is equivalent to using $-\left(\frac{\partial^2 \ln L(0)}{\partial u_i^2}\right)^{-1}\left(\frac{\partial^2 \ln L(0)}{\partial u_i \partial \Delta}\right)\delta\Delta$ which gives the following equation for the change in u_i:

Figure 5.12 *Matrix of partial 2nd derivatives with respect to Δ and u_i, for the mixed model likelihood with two modes.*

$$\begin{bmatrix} \frac{(n+\lambda)}{\sigma_w^2}\sum_{i=1}^{s} q_i(2) - \sum_{i=1}^{s} q_i(1)q_i(2)t1_it1_i & \frac{(n+\lambda)}{\sigma_w^2} q_1(2) - q_1(1)q_1(2)t1_1t2_1 & \cdots & \frac{(n+\lambda)}{\sigma_w^2} q_s(2) - q_s(1)q_s(2)t1_st2_s \\ \frac{(n+\lambda)}{\sigma_w^2} q_1(2) - q_1(1)q_1(2)t1_1t2_1 & \frac{(n+\lambda)}{\sigma_w^2} - q_1(1)q_1(2)t2_1t2_1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \frac{(n+\lambda)}{\sigma_w^2} q_s(2) - q_s(1)q_s(2)t2_st2_s & 0 & \cdots & \frac{(n+\lambda)}{\sigma_w^2} - q_s(1)q_s(2)t1_st2_s \end{bmatrix}$$

Where:

$$t1_i = \left(-(u_i + \Delta)(n+\lambda) + \sum_{j=1}^{n} (y_{ij} - \mu_2) \right) \frac{1}{\sigma_w^2} \qquad t2_i = (-\Delta(n+\lambda) + (\mu_1 - \mu_2)) \frac{1}{\sigma_w^2}$$

When: $\Delta = \hat{\Delta}$ $\quad t1_i = 0 \qquad t2_i = 0$

$\Delta = 0$

$$t1_i = q_i(1)\, n\, (\mu_1 - \mu_2) \frac{1}{\sigma_w^2} \qquad t2_i = n\, (\mu_1 - \mu_2) \frac{1}{\sigma_w^2}$$

$$\delta u_{i0} = \left(\frac{\frac{n+\lambda}{\sigma_w^2} q_{i0}(2) - q_{i0}(1) q_{i0}(1) q_{i0}(2) \left(\frac{n(\mu_1+\mu_2)}{\sigma_w^2} \right)^2}{\frac{n+\lambda}{\sigma_w^2} - q_{i0}(1) q_{i0}(2) \left(\frac{n(\mu_1+\mu_2)}{\sigma_w^2} \right)^2} \right) \delta\Delta$$

This should give a better estimate of the change in u_i near Δ equals 0 but will not give the correct total change in u_i. It is indicated by a dashed line in figure 5.11. When this equation for the change in u_i is incorporated into [5.7] the following approximation for the exact likelihood is obtained:

$$\ln L(\hat{\Delta}, \hat{u}_{i\Delta}) \approx \ln L(0, \hat{u}_{i0}) + \frac{(n+\lambda)}{2\sigma_w^2} \hat{\Delta}^2 \sum_{i=1}^{s} q_{i0}(2) \left(2q_{i0}(1) + \frac{(n+\lambda)}{\sigma_w^2} \hat{\Delta}^2 q_{i0}(1)^3 - 1 \right)$$

$$- \frac{\hat{\Delta}^2}{2} \sum_{i=1}^{s} \left(\frac{\left(-\frac{(n+\lambda)}{\sigma_w^2} q_{i0}(2) \left(1 - \frac{(n+\lambda)}{\sigma_w^2} q_{i0}(1)^2 \right) \right)^2}{-\frac{(n+\lambda)}{\sigma_w^2} \left(1 - \frac{(n+\lambda)}{\sigma_w^2} \hat{\Delta}^2 q_{i0}(1) q_{i0}(2) \right)} \right)$$

Which is approx. 5 in figure 5.9.

Discussion

Attempting to refine the approximation for the difference between the exact and ME1 likelihood by including the true second differentials did not improve it. Approximating the change in the sire effect with the second derivatives caused the difference between the likelihoods to be over estimated to a large extent at conditional genotype probabilities of around 0.25 and 0.75. This could be because in some situations (see fig 5.11) using the 2nd derivatives actually suggests a change in u_i in the wrong direction, i.e. increasing the estimate with an increase in Δ, whereas the estimated value of u_i when Δ is equal to $\hat{\Delta}$ is lower than when Δ is zero. The best approximations were obtained using the approximate 2nd derivatives and absorbing sire effects.

To obtain the exact difference between the likelihoods the estimate for u_i and the 2nd derivatives would have to be continually updated as Δ changed from zero to $\hat{\Delta}$. This is not feasible as at each stage new estimates of the conditional sire probabilities would have to be obtained.

From the above approximations, it can be seen that fixing Δ at zero causes a term to be excluded from the log likelihood. A large component of this term (present in all the above equations) is a function of the difference between the major gene means (Δ) and the conditional sire probabilities.

$$\frac{(n+\lambda)}{2\sigma_w^2}\hat{\Delta}^2\sum_{i=1}^{s} q_i(1)q_i(2)$$

Using this model with two genotypes, but assuming that we have prior knowledge that the difference between means is zero, or, more precisely that the difference follows a normal distribution with mean zero and with variance σ_m^2 would give the following log likelihood:

$$\ln L(\text{prior}) = \ln L(\hat{\Delta}) - \frac{(\mu_1-\mu_2)^2}{2\sigma_m^2}$$

It can be shown that the ME1 likelihood can be written in a similar form. Δ is a function of the difference between the two major genotype means and hence Δ divided by its variance can be rewritten in terms of the major genotype means and the variance of the difference between them.

The inverse of the matrix of 2nd derivatives can be used to give the variance of the parameter estimates. For Δ, using the 2nd differentials from the exact model the variance is:

$$\mathrm{Var}(\hat{\Delta}) = \left(\frac{n+\lambda}{\sigma_w^2}\sum_{i=1}^{s} q_i(1)q_i(2)\right)^{-1}$$

Therefore:

$$\frac{\hat{\Delta}^2}{\mathrm{var}(\hat{\Delta})} = \frac{(n+\lambda)}{2\sigma_w^2}\hat{\Delta}^2\sum_{i=1}^{s} q_i(1)q_i(2)$$

Which is the equation given for the difference between the exact and ME1 likelihoods. Also:

$$\frac{\hat{\Delta}^2}{\mathrm{var}(\hat{\Delta})} = \frac{\frac{n^2}{(n+\lambda)^2}(\mu_1-\mu_2)^2}{\mathrm{var}\left(\frac{n}{n+\lambda}(\mu_1-\mu_2)\right)}$$

$$= \frac{(\mu_1-\mu_2)^2}{\mathrm{var}(\mu_1-\mu_2)}$$

Therefore, it is suggested that estimating a single sire effect is similar to using the exact likelihood with prior knowledge that the difference between the means is zero. If there is a lot of evidence in the data to suggest that this is not the case, the major gene parameters will be estimated and the mixed model likelihood will be more likely than the polygenic model. However with less evidence for a difference between the means, the prior knowledge that there is no difference will outweigh the data evidence and a polygenic model will be suggested. From the results of the simulation study already presented, it seems that the evidence from the data has to be very strong before a non-polygenic model is obtained.

5.6 DISCUSSION

If the genotypes of all the individuals in the pedigree were known the ME1 approximation would be the same as the exact mixed model likelihood. Otherwise, the approximation pools the information on major genotypes to obtain an estimate of the sire's transmitting ability. The likelihood can then be calculated assuming independence of the offspring given the sire's genotype. The exact likelihood effectively calculates the likelihood of all combinations of major genotype and weights them according to the frequency of the alleles at the major locus in the population.

When a major gene was detected the estimates obtained for its effect and frequency were reasonable, but the very low power of detection suggests that the method will not be very useful. The ability of the method to estimate the polygenic heritability was low, the estimate being zero in the majority of analyses.

There are several advantages of this method. It can easily be extended, in theory, to include fixed effects and more complicated relationships. Hoeschele (1988b) suggests the extension of the animal model to include major gene effects. This would allow the use of more generations and take account of selection, although may be infeasible for large data sets. Also, as the effect of the major gene increases, or as data allowing more precise genotyping accumulates, the ME1 likelihood becomes exact. Therefore other pedigree structures, such as an animal model, might improve the approximation.

Estimates of the polygenic transmitting ability for each sire are obtained immediately from the analysis and the conditional probability of the animals to be each major genotype. This information is useful to select the required animals and optimise genetic improvement. However, the transmitting ability is a pooled estimate including information from all three genotypes for the sire. Whereas, in reality, of course, the sire has only one major genotype. Pooling information over the three possible major genotypes for the sire seems to be equivalent to assuming that there is prior information that a gene with large effect is not segregating in the population.

Appendix 3

A3.1 EM algorithm for maximisation

The log of the ME1 approximation to the mixed model likelihood can be written as follows:

$$\ln L = \sum_{i=1}^{s} \left\{ \frac{1}{2}\ln(2\pi\sigma_w^2) - \frac{1}{2}\ln(n+\lambda) + \ln\left(\sum_{c=1}^{m} p(c)h(u_i) \prod_{j=1}^{n} \sum_{d=1}^{m} \text{trans}(d|c)k_d(y_{ij}|\mu_d,u_i,\sigma_w^2) \right) \right\}$$

This can be differentiated with respect to each of the parameters to be estimated. At a maximum the 1st derivatives will be equal to zero and hence by equating the derivatives for each parameter to zero and solving them for the unknown parameters, the maximum likelihood (ML) estimates are obtained.

After rearrangement and simplification the following equations are obtained from the 1st derivatives:

Major genotype means

$$\mu_d = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n} q_{ij}(d|c)\,(y_{ij}-\mu-u_i)}{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n} q_{ij}(d|c)}$$

Polygenic transmitting ability for sires

$$u_i = \frac{\sum_{c=1}^{m}\sum_{j=1}^{n}\sum_{d=1}^{m} q_i(c)q_{ij}(d|c)\,(y_{ij}-\mu-\mu_d)}{n+\lambda}$$

Genotype frequencies

Assuming that the population is in Hardy-Weinberg equilibrium the following equation is obtained for the allele frequency in both the sire and dam population:

$$p = \frac{\sum_{i=1}^{s}\left\{\sum_{c=1}^{m} q_i(c)\sum_{j=1}^{n} q_{ij}(1|c) + q_i(3)\sum_{j=1}^{n} q_{ij}(2|3) + 2q_i(1) + q_i(2)\right\}}{sn - \sum_{i=1}^{s} q_i(2)\sum_{j=1}^{n} q_{ij}(2|2) + 2s}$$

Where $\frac{\sum_{i=1}^{s}(2q_i(1)+q_i(2))}{2s}$ would be the estimate for the allele frequency based on information from the sire and the remaining information is from the dam.

However, if this assumption is relaxed and genotype frequencies estimated in the sires and an allele frequency in the dams, the following equations are obtained:

$$\text{freq}(c)\,(\text{sires}) = \frac{\sum_{i=1}^{s} q_i(c)}{s}$$

$$p\,(\text{dams}) = \frac{\sum_{i=1}^{s}\left\{\sum_{c=1}^{m} q_i(c)\sum_{j=1}^{n} q_{ij}(1\,|\,c) + q_i(3)\sum_{j=1}^{n} q_{ij}(2|3)\right\}}{sn - \sum_{i=1}^{s} q_i(2)\sum_{j=1}^{n} q_{ij}(2|2)}$$

Variance components

If the polygenic heritability is assumed to be known, then the following equation for the residual variance is obtained:

$$\sigma_w^2 = \frac{\sum_{i=1}^{s}\left(u_i^2\lambda + \sum_{c=1}^{m} q_i(c)\sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d\,|\,c)\,(y_{ij} - \mu - \mu_d - u_i)^2\right)}{sn}$$

Otherwise equations for both the residual and sire variance components are as follows:

$$\sigma_w^2=\frac{\sum_{i=1}^{s}\sum_{c=1}^{m}q_i(c)\sum_{j=1}^{n}\sum_{d=1}^{m}q_{ij}(d\,|\,c)\,(y_{ij}-\mu-\mu_d-u_i)^2}{sn-s+\frac{s\lambda}{n+\lambda}}$$

$$\sigma_u^2=\frac{\sum_{i=1}^{s}u_i^2}{s-\frac{s\lambda}{n+\lambda}}$$

The equations have been written as functions of $q_i(c)$ and $q_{ij}(d|c)$, where:

$q_i(c)$ is the conditional probability of genotype c for sire i.

$$q_i(c)=\frac{p(c)\prod_{j=1}^{n}\sum_{d=1}^{m}\text{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)}{\sum_{c'=1}^{m}p(c')\prod_{j=1}^{n}\sum_{d=1}\text{trans}(d\,|\,c')\,k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)}$$

$q_{ij}(d|c)$ is the conditional probability that offspring j from sire i has genotype d given that his sire has genotype c.

$$q_{ij}(d\,|\,c)=\frac{\text{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)}{\sum_{d=1}\text{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,u_i,\sigma_w^2)}$$

These equations cannot be solved directly as $q_i(c)$ and $q_{ij}(d|c)$ are functions of the parameters to be maximised. However an iterative algorithm for maximisation can be obtained, where the conditional probabilities are calculated using the parameter estimates obtained in the previous iteration. These values for $q_i(c)$ and $q_{ij}(d|c)$ can be substituted into the equations given above to give new estimates for the parameters. Using the new estimates for the major genotype means and frequencies, the sire transmitting abilities and the residual variance the conditional probabilities can be recalculated. The process continues until there is virtually no change in the parameter estimates from one iteration to the next.

The equations for the major genotype means (μ_d) and sire transmitting abilities (u_i) (and fixed effects, if included) can be solved simultaneously by rearranging them into matrix form. Matrices are obtained which are similar to those for the classical mixed model for a polygenic model with fixed and random effects but extended to include the major gene component. They are given in equation [5.3]. The frequencies are functions of the conditional probabilities only and hence can be calculated immediately. The variances can then be estimated using the new estimates for the means and transmitting abilities, and the conditional probabilities based on the parameters from the previous iteration.

Obviously to start the maximisation process initial parameter estimates are required to calculate the conditional probabilities and hence to set up the matrices and estimate the parameters. The likelihood cannot be assumed to have a single maximum and so starting values close to the global maximum are required. Hoeschele (1988a) suggests a method based on quantile-quantile plots to obtain initial estimates.

Convergence is attained when there is no change in the parameter estimates obtained in iteration [a] compared with those obtained from iteration [a-1]. The parameter estimates in iteration [a] are the maximum likelihood estimates, and the likelihood calculated using these parameters is the maximum likelihood. Convergence is assumed when, for example, $\sqrt{\frac{(\theta_{[a]}-\theta_{[a-1]})'(\theta_{[a]}-\theta_{[a-1]})}{\dim(\theta)}} < 10^{-5}$, where θ is a vector of parameter estimates of dimension $\dim(\theta)$.

A3.2 REML variance estimates

The ML equations for the variance components are biassed as they do not take account of the degrees of freedom lost due to the use of the data to estimate fixed effects. With a large number of observations and no fixed effects, except the mean, in the data the bias will be small and reasonable results will be obtained. However if fixed effects are also being estimated the degrees of freedom by which the relevant sums of squares are divided should be reduced to take account of this. Hence methods such as REML (restricted ML) and Marginal ML (which reduces to REML under the assumption of normality) have been suggested.

For the polygenic model, the taking account of this reduction in the degrees of freedom can be achieved by incorporating additional terms. One is the determinant of the variance-covariance matrix for the estimates of the fixed effects in the likelihood. This variance matrix is the inverse of the coefficient matrix from [5.2] after absorption of the sire effects. Another additional term is the determinant of the matrix **X'X**, which is consant for a given fixed effect structure, and lastly the constant in terms of π is reduced by a function of the rank of **X'X**. Using the model described in 5.3.1, the following equation is obtained:

$$\ln L(\text{poly}) = -\frac{sn-t}{2}\ln(2\pi) - \frac{1}{2}\ln(|\mathbf{V}|) - \frac{1}{2}\ln(|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|) + \frac{1}{2}\ln(|\mathbf{X}'\mathbf{X}|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

Where: variance(**y**) = $\mathbf{V} = \mathbf{ZAZ}'\sigma_u^2 + \mathbf{I}\sigma_w^2$

Differentiating this likelihood with respect to the variance components and equating the derivatives to zero gives the REML variance component equations, which can be written as follows:

$$\sigma_w^2 = \frac{(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu})'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu})}{sn - s - t + \lambda\,\mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}_{zz})}$$

$$\sigma_w^2 = \frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{s - \lambda\,\mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}_{zz})}$$

Where: $\mathbf{C}_{zz}$ is the s x s part of the inverted coefficient matrix of [5.2] referring to **u**.
t is the dimension of **X'X**.

For the mixed model, equations can be derived analogously to those for the polygenic model. When the major genotypes of all the individuals is known, the model becomes linear and the ME1 approximation gives the exact likelihood. With this restriction the genotype effects can be fitted as fixed effects and the matrix **D** can be considered in the form **X'X**. Where **X** is an sn x m design matrix containing a 1 when for the genotype of each individual and a zero otherwisre. Hence, the denominators, which are obtained from the derivatives of $\ln|\mathbf{V}| + \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$, will be the same expressions as for the polygenic model although the rank of the matrix will have increased because of the additional fixed effects. The inverse of the coefficient matrix will contain additional terms for the major genotypes. The numerators are obtained from the derivative of the rest of the likelihood with respect to the relevant variance component. For the residual variance, this will contain a summation over the major genotype for each individual, with each summation weighted by the probability of that genotype. As the sire effects are not conditional on the major genotype the conditional probabilities disappear and the equation for the sire variance component looks the same as that obtained from the polygenic model. Assuming that sires are unrelated the following equations are obtained:

$$\sigma_{\mathbf{w}}^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d \mid c)\,(y_{ij} - \mu - \mu_d - u_i)^2}{sn - s - m + 1 + \lambda\,\mathrm{tr}(\mathbf{C}_{\mathbf{zz}})}$$

$$\sigma_u^2 = \frac{\sum_{i=1}^{s} u_i^2}{s - \lambda\,\mathrm{tr}(\mathbf{C}_{\mathbf{zz}})}$$

Where: $\mathbf{C}_{\mathbf{zz}}$ is the s x s part of the inverted coefficient matrix of [5.3] referring to **u**.

Additional fixed effects can be incorporated in the same way as the major gene means, reducing the denominator by the rank of the matrix **X'X** and adjusting the phenotypes for the relevant effects.

CHAPTER 6

APPROXIMATION 4 - ESTIMATING A SIRE EFFECT FOR EACH MAJOR GENOTYPE

6.1 INTRODUCTION

The performance of the ME1 approximation was poor, especially in its detection of a major gene. However, the method has some advantages over Herm, such as speed of computation, ease of incorporation of fixed effects and immediate estimation of polygenic transmitting abilities. The ME1 likelihood is exact if the genotype of all individuals is known. Otherwise, when estimating the mode of each sire's transmitting ability distribution, information is pooled from all possible major genotypes of the sire and of his offspring. It has been suggested that this is equivalent to assuming that there is prior information that a major gene is not present, and this contributes to the poor performance of the approximation.

Given the phenotypes of his offspring, a sire would be expected to have contributed different polygenic effects depending on the major genotype being suggested for the sire. Hence, an approximation is proposed where the transmitting ability for each sire is estimated for each major genotype. As described before, the exact mixed model likelihood involves estimating a sire effect for each combination of major genotypes for the offspring and sire. Therefore this approximation is exact if the major genotype of all offspring is known as the uncertainty of the major genotype of the sire is accounted for correctly but not for the offspring. If the offspring genotypes are unknown, rather than considering all possible combinations of major genotype for the half-sibs, information about the genotypes for each individual given the sire's genotype is pooled. This approximation has been suggested independently by Elsen and Le Roy (1989).

6.2 LIKELIHOODS

Estimating the mode of each sire's transmitting ability distribution under the hypotheses of each major genotype for the sire, so that three estimates are obtained gives the following approximation to the mixed model likelihood (ME3) for the same half-sib family structure described previously in section 2.3:

$$L(MM)=\prod_{i=1}^{s}\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)^{1/2}\sum_{c=1}^{m}p(c)\,h(\hat{u}_{ic})\prod_{j=1}^{n}\sum_{d=1}^{m}\text{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,\hat{u}_{ic},\sigma_w^2) \qquad [6.1]$$

Where: $\hat{u}_{ic}$ is the mode of sire i's transmitting ability distribution given that he has genotype c.

$h(\hat{u}_{ic})$ and $k_d(y_{ij}|\mu,\mu_d,\hat{u}_{ic},\sigma_w^2)$ are as defined before but now the transmitting ability of the relevant genotype (c) is used.

6.3 MAXIMISATION

Algorithm

The ME3 likelihood is a function of the major genotype means and frequencies and polygenic and environmental variance components, as are the Herm and ME1 likelihoods, and also each sire's transmitting ability under the three hypotheses of the sire being each genotype. This means that an extra 3s (3(number of sires)) parameters are required to be estimated compared with Herm. Hence the quasi-Newton algorithm described in 4.3 would be slow, as the number of likelihood evaluations required depends on the number of parameters to be estimated. However the EM algorithm described in 5.3 for ME1 can easily be extended to include the extra sire effects. The partial 1st derivatives of the log likelihood with respect to each parameter to be estimated are derived and, by equating these to zero a series of equations can be obtained. These can be solved iteratively to yield the maximum likelihood (ML) estimates. Equations for the parameters are given in Appendix 4.

For the sire effects and major genotype means matrices can be set up similar to those used in ME1 to solve the equations.

$$\begin{bmatrix} D_{[a]} & \sum_{c=1}^{m} Q_{c[a]}'X & Q_{1[a]}'Z & \cdot & Q_{3[a]}'Z \\ \sum_{c=1}^{m} X'Q_{c[a]} & X'X & X'Zq_{1[a]} & \cdot & X'Zq_{3[a]} \\ Z'Q_{1[a]} & q_{1[a]}Z'X & q_{1[a]}(Z'Z+I\lambda) & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ Z'Q_{3[a]} & q_{3[a]}Z'X & 0 & \cdot & q_{3[a]}(Z'Z+I\lambda) \end{bmatrix} \begin{bmatrix} \mu_{d[a+1]} \\ \beta_{[a+1]} \\ \hat{u}_{1[a+1]} \\ \cdot \\ \hat{u}_{3[a+1]} \end{bmatrix} = \begin{bmatrix} \sum_{c=1}^{m} Q_{c[a]}'y \\ X'y \\ q_{1[a]}Z'y \\ \cdot \\ q_{3[a]}Z'y \end{bmatrix} \qquad [6.2]$$

Where: $\mathbf{D}_{[a]}$ is an m x m matrix containing: $\mathrm{Diag}\left[\sum_{i=1}^{s}\sum_{c=1}^{m}\sum_{j=1}^{n} q_i(c)_{[a]} q_{ij}(d|c)_{[a]}\right]$

$\mathbf{Q}_{c[a]}$ is an m x sn matrix containing: $q_i(c)_{[a]} q_{ij}(d|c)_{[a]}$

$\mathbf{q}_{c[a]}$ is an s x s matrix containing: $\mathrm{Diag}\left[q_i(c)\right]$

Equations for the frequencies can be obtained under the different assumptions described in 5.3, and these equations are discussed in Appendix 4. Likewise the variance equations can be obtained under the two different assumptions considered previously, either assuming prior knowledge of the polygenic heritability or estimating this along with the major gene parameters. The ML equations are given in the Appendix 4. If fixed effects are also being estimated, the variance estimates need to take account of the degrees of freedom used to estimate these effects from the data. However, problems are encountered when obtaining REML estimates because the sire part of the coefficient matrix is repeated for each major genotype of the sire and hence the same information reused for different sire estimates. The derivation of REML variance estimates is given in Appendix 4.

Initial parameter estimates

Initial estimates for the parameters are required so that the conditional probabilities can be calculated and the matrices [6.2] set up. As for ME1 there is no guarantee that a global maximum is attained and hence the initial estimates need to be near the global maximum.

By estimating three sire effects, one for each major genotype, multimodality of the likelihood surface, as observed with ME1 (figure 5.1), should no longer be a problem. Hence the final estimates should be less sensitive to the initial values used. When estimating the polygenic heritability non-zero estimates for the sires are required, which can be obtained by fixing the heritability for the first iteration, as described for ME1.

6.4 SIMULATION STUDY

In order to investigate the ME3 approximation, to observe its ability to detect a major gene and estimate its effect and frequency in the population, the simulated data described in 4.4.1 were reanalysed.

6.4.1 Analyses

The analysis was the same as that for the Herm and ME1 approximations, that is, the mean effect of the low scoring homozygote at the major locus (μ) was estimated, along with the deviation of the other two major genotype means from this mean ($\mu(AA)$ and $\mu(Aa)$). The population was assumed to be in Hardy-Weinberg equilibrium and an allele frequency (p(A)) was estimated. First, each data set was analysed assuming that the polygenic heritability was known and fixing it at the expected value so that only the residual variance is estimated and then repeated estimating the polygenic heritability from the data. As before ML variance estimates were calculated.

The initial parameter estimates were the same as those described in 5.4.1, with the initial sire effects for all three genotypes being zero and the heritability fixed initially when estimating both variance components.

Genotyping at the major locus

The probability of each sire being each genotype, based on his offspring's phenotypes, is calculated to see how good the method is at allocating sires to a genotype class. The equation [2.5] can now be written as:

$$q_i(c) = \frac{p(c)\, h(\hat{u}_{ic}) \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d \mid c)\, k_d(y_{ij} \mid \mu, \mu_d, \hat{u}_{ic}, \sigma_w^2)}{\sum_{c'=1}^{m} p(c')\, h(\hat{u}_{ic'}) \prod_{j=1}^{n} \sum_{d=1}^{1} \mathrm{trans}(d \mid c')\, k_d(y_{ij} \mid \mu, \mu_d, \hat{u}_{ic'}, \sigma_w^2)}$$

As for ME1, these probabilities are calculated each iteration as part of the maximisation process and the values from the final iteration can be used.

6.4.2 Test statistic

It has been shown that the test statistic distribution obtained by analysis of polygenic data with the ME1 approximation does not follow a χ^2 distribution (see section 5.4.2). Investigation is required to see whether the test statistic obtained using the ME3 mixed model approximation follows this expected distribution.

Method

To investigate the test statistic distribution the data described in 4.4.3 were analysed using the ME3 approximation to the mixed model. The analyses were the same as those described in 5.4.2, with the test statistic being calculated as given in [2.4] using the ME3 mixed model likelihood.

Results

Figures 6.1 and 6.2 show the distribution of the test statistic for the 100 simulations with the expected heritability equal to 0.2 and 0.4 respectively. If the two sets of starting values gave different results, the highest likelihood has been used. The mean and variance of these observed distributions are given in table 6.1 along with the number of simulations giving a significant value at the 5% and 1% significance levels when tested against a χ^2 distribution with three degrees of freedom. All negative test statistics have been set to zero.

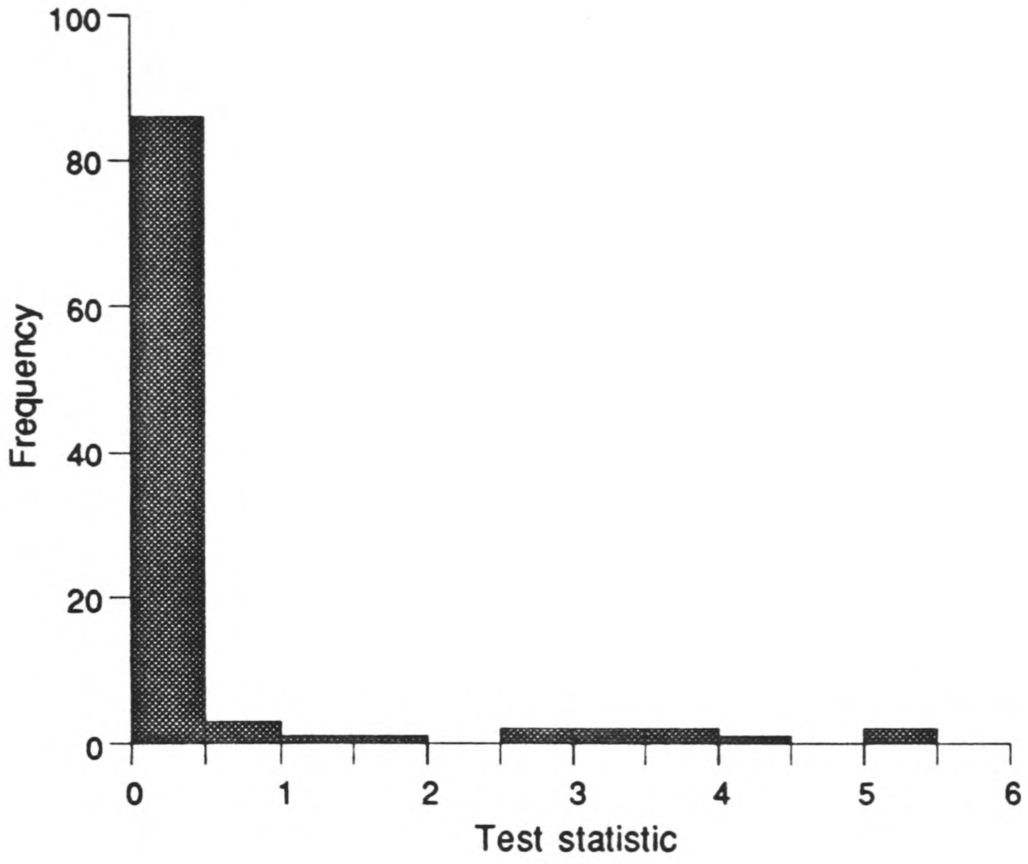
Compared with the results using the ME1 approximation fewer analyses converged to a polygenic model, as shown by the decreased number of zero test statistics. Hence, the mean of the test statistic distribution is higher than observed for ME1, although lower than expected for a χ^2 distribution with three degrees of freedom. The distribution most resembles a χ^2 distribution when the heritability is estimated, although all of the observed distributions are significantly different from a χ^2 distribution with two, three or four degrees of freedom. With a heritability of 0.2, estimated in the analyses, the distribution is closest to that expected. However, even in this case when comparing the observed and expected number of test statistics falling in ten equal parts of the χ^2 distribution with two, three and four degrees of freedom the χ^2(9 d.f.) values obtained were 38.0, 42.9 and 73.6 respectively. The high values being caused by the large proportion of zero test statistics.

Table 6.1. *Mean and variance of the test statistic and the number significant when compared with a χ^2 distribution with 3 degrees of freedom.*

Model		Mean	Variance	no. zero	5%	1%
Expected	(χ^2 3 d.f.)	3		0	5	1
h^2=0.2	fixed	0.414	1.313	78	0	0
h^2=0.2	estimated	2.657	7.616	20	4	1
h^2=0.4	fixed	0.300	2.375	89	1	1
h^2=0.4	estimated	1.107	8.072	79	6	3

Figure 6.1 *Distribution of the test statistic from analyses of polygenic data with an expected heritability of 0.2.*

a) With the polygenic heritability assumed to be known in the analyses.



b) With the polygenic heritability estimated in the analyses.

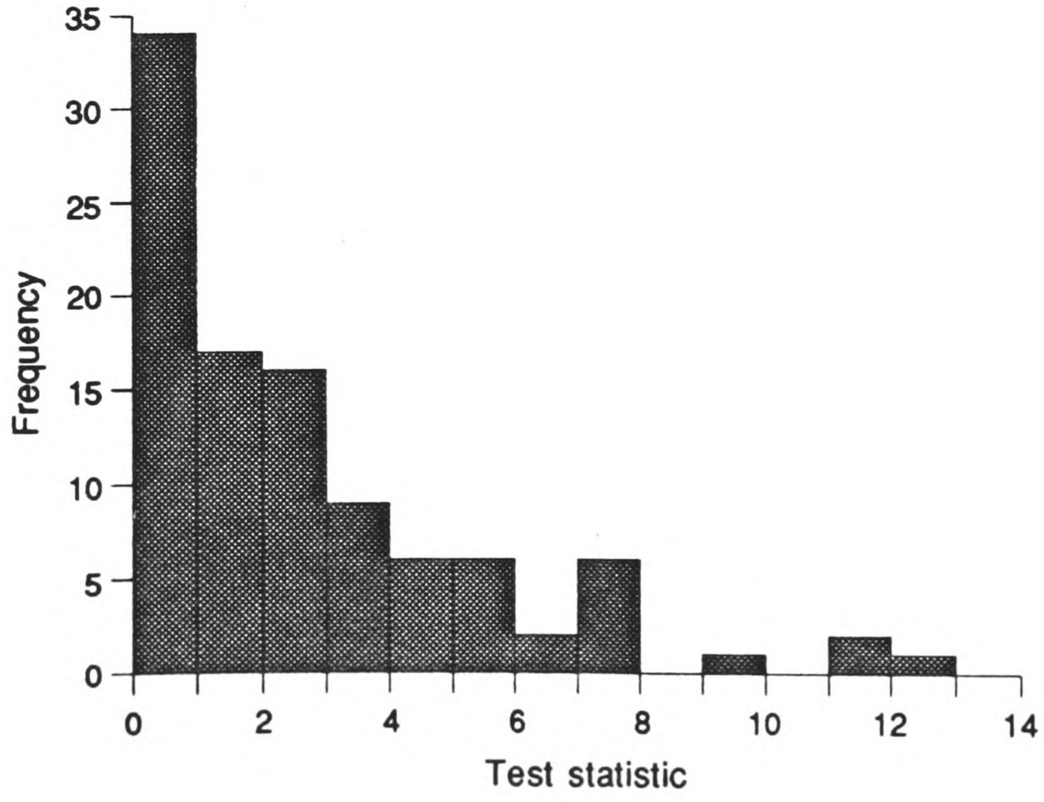
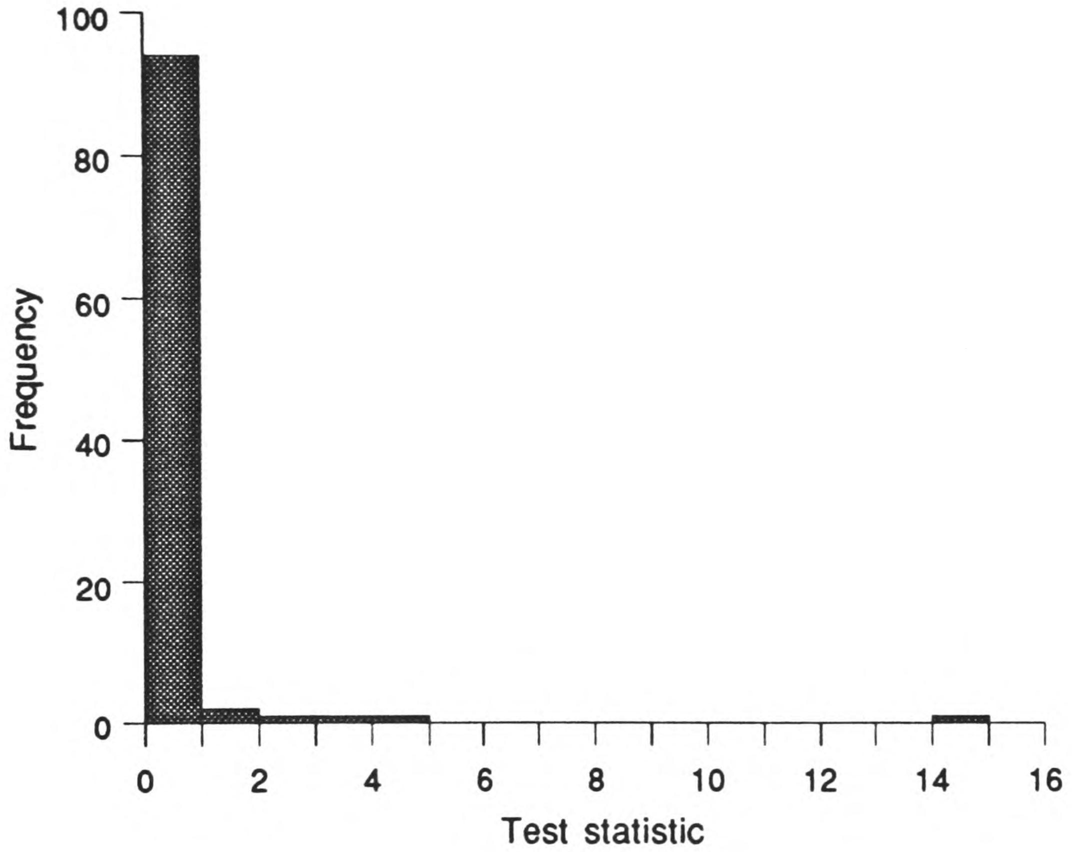
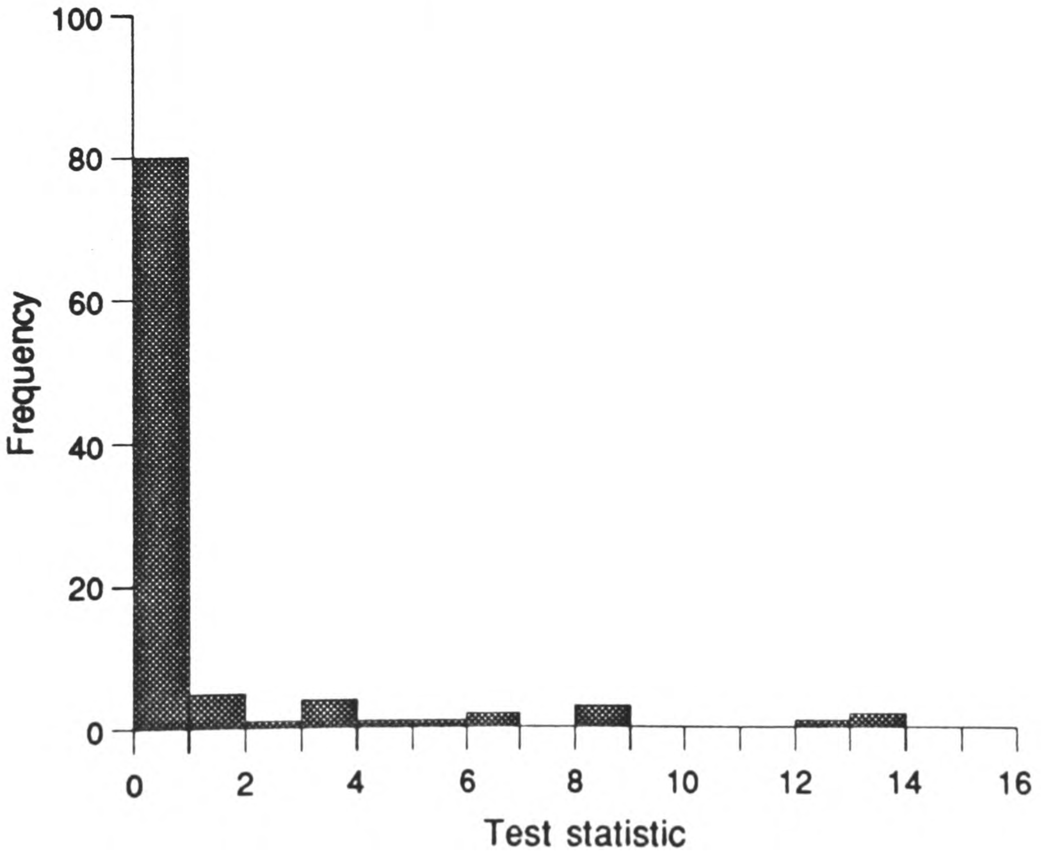


Figure 6.2 *Distribution of the test statistic from analyses of polygenic data with an expected heritability of 0.4.*

a) With the polygenic heritability assumed to be known in the analyses.



b) With the polygenic heritability estimated in the analyses.



Discussion

The test statistic distributions obtained from the 100 analyses of each model do not follow a χ^2 distribution with three degrees of freedom. However, the observed distributions have fewer zero values than obtained using ME1. When testing for the presence of a major gene the extreme tail of the distribution is of interest, that is, a value is required against which the test statistic can be compared that would incorrectly suggest a major gene in just 5% or 1% of analyses. With the heritability estimated the χ^2 distribution appears to provide a reasonable criterion.

These results are in agreement with those given by Elsen and Le Roy (1989) for this approximation, denoted MU3 in their notation. In their analyses five more parameters were estimated in the mixed model than in the polygenic model, as genotype frequencies for the sire population and an allele frequency for the dam population were estimated. Considering the situation with the polygenic heritability estimated in the analyses, based on 318 analyses with an expected heritability of 0.2 the mean test statistic was 3.91 (standard deviation 3.30) and when the expected heritability was 0.6 the mean test statistic from 244 analyses was 3.44 (standard deviation 3.75). Both of these are lower than expected from a χ^2 distribution with five degrees of freedom. They also estimated the 5% and 1% quantiles for these two distributions and obtained estimates of 10.61 and 15.11, respectively, when the heritability was 0.2 and 11.08 and 17.11 when the heritability was 0.6. With a heritability of 0.2 these estimates are similar to the values that would be obtained from a χ^2 distribution with five degrees of freedom (11.07 and 15.09). With a higher expected heritability the 1% quantile was over estimated. As found here, a χ^2 distribution might provide a suitable criterion when the expected heritability is low (0.2). With a higher heritability, the results of Elsen and Le Roy (1989) suggest that too many spurious major genes will be detected.

6.4.4 Simulation results

The results are based on the analysis of each data set that gave the highest likelihood.

Power

The results for the test statistic obtained from the analyses of mixed model data are summarised in table 6.2. Also given are the regression on and the correlation with Herm results for the same set of data, including the negative test statistics but ignoring those that went to zero. The mean test statistic is always lower than the mean Herm value, however it was higher than that obtained using the ME1 approximation especially

Table 6.2 *Mean and standard deviation of the test statistic (setting negative values to zero), the number of analyses where the test statistic was significant at the 5% and 1% significance levels of a χ^2 distribution with three degrees of freedom and the regression on, and correlation with, Herm results for the same set of data.*

			no.	no.	no. significant		compared with Herm	
Model	mean	sd	zero	negative	5%	1%	slope	r
Fixed heritability								
Add1	6.698	4.959	5	0	33	14	0.718	0.972
Add2	2.366	2.692	57	2	3	1	0.371	0.621
Dom	38.255	14.306	0	0	99	97	0.968	0.996
Rare	7.454	8.632	3	0	38	19	0.803	0.969
Estimated heritability								
Add1	3.336	4.430	10	20	13	5	1.034	0.862
Add2	1.715	4.867	64	7	9	5	1.076	0.919
Dom	37.341	13.437	0	0	99	99	0.996	0.955
Rare	4.585	4.418	15	2	18	7	0.895	0.883

when the heritability was fixed because of the decrease in the number of zero test statistics. Fewer analyses resulted in negative test statistics than with ME1, suggesting that convergence to local maxima is less of a problem. The likelihood surface for the same set of data illustrated in figure 5.4, which produced a negative test statistic for ME1, is shown in figure 6.3. It can be seen that the local maximum, present when analysing with the ME1 approximation is no longer present when using the ME3 approximation.

A major gene was easiest to detect when the simulated data contained a major gene with an allele of dominant effect. Evidence for its existence was found in 99 of the analyses both when an assumed value for the heritability was used and when the heritability was estimated. Considering the three additive major genes simulated, a major gene was detected in the highest number of analyses when one of the alleles was rare. When estimating the heritability the test statistic decreased, as observed using Herm but not ME1, and hence the number of analyses giving a significant test statistic decreased. Also when the heritability was estimated the improvement of the ME3 approximation over ME1 was reduced. This is because many of the non-zero test statistics resulted from analyses ending at major gene models which are the same for the two methods.

When comparing the results from the same data sets the ME3 likelihood is always less than the Herm likelihood when a mixed model is obtained using Herm. Hence when compared with a χ^2 distribution fewer are significant. However there is a good linear relationship with the Herm results, with both the slope of the regression and the correlation being close to one. This is illustrated in figure 6.4 for the data simulated under Add1, which also shows the improvement of the ME3 over the ME1 likelihood. When a mixed model is obtained for ME3 the likelihood is always greater than that obtained using ME1.

Parameter estimates

Table 6.3 gives the mean parameter estimates for those analyses that gave a non-zero test statistic for the four mixed models simulated. The highest scoring homozygote at the major locus was defined as being genotype AA. Also given are the regression of the estimates on, and their correlation with the Herm estimates for the same set of data. The mean estimates for the variance components (with the major gene variance estimated using [4.5]) and their regression on, and correlation with the values from the simulation (estimated by analysis of variance on the polygenic plus environmental effects for the residual and sire variance components and using equation [4.5] for the major gene variance) for the same set of data for those analyses giving non-zero test statistics are given in table 6.4.

Figure 6.3 *The likelihood surface for one set of data simulated under Add1.*

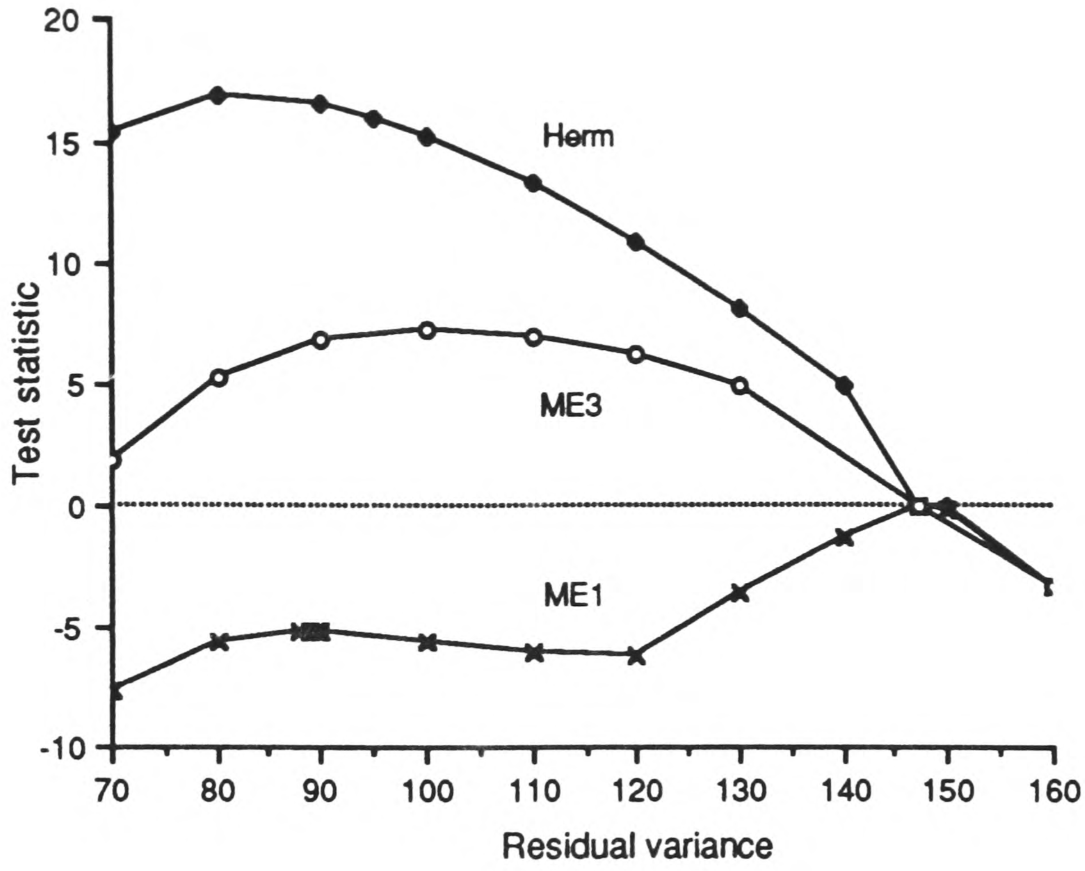


Figure 6.4 *The test statistic obtained under ME3 and ME1 compared with the test statistic obtained with Herm for the same set of data, for Add1 with fixed heritability.*

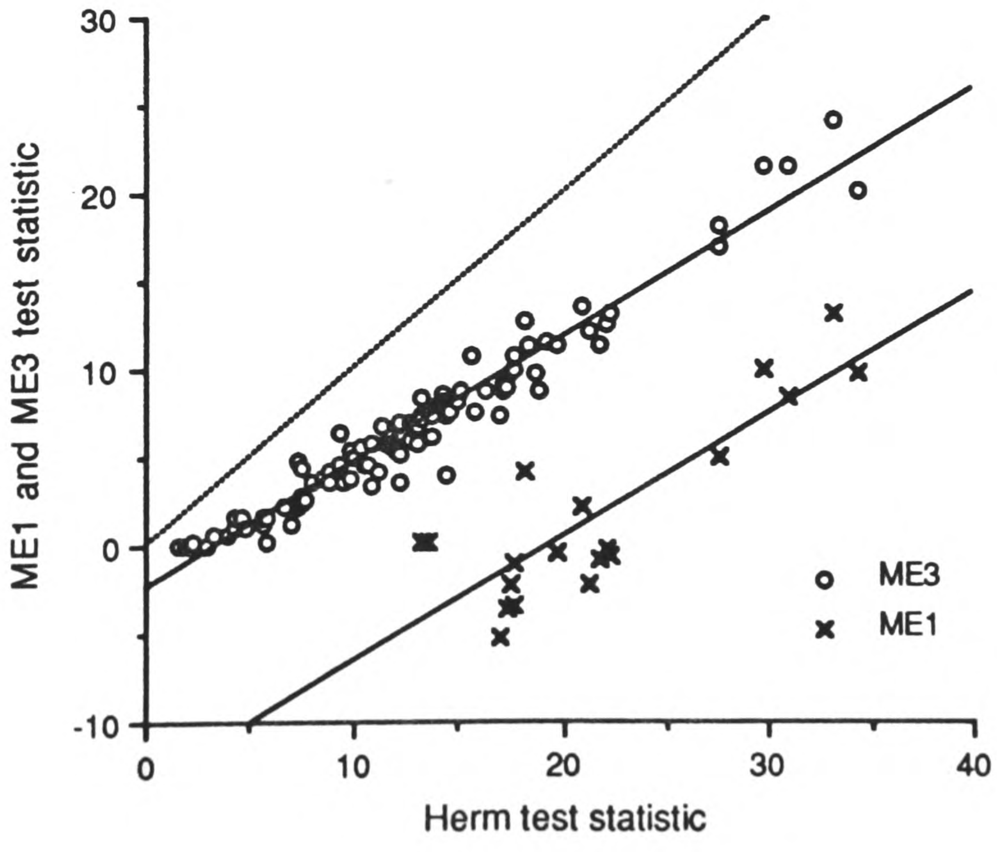


Table 6.3 *Mean parameter estimates from analyses with non-zero test statistics, and the correlation with, and regression on, Herm estimates for the same set of data.*

		Fixed heritability				Estimated heritability				
Model		p(A)	μ_{AA}	μ_{Aa}	σ^2_w	p(A)	μ_{AA}	μ_{Aa}	σ^2_w	σ^2_u
Add1	mean	0.49	16.22	8.03	108.99	0.51	21.86	10.82	90.30	0.25
	sd	0.18	3.68	3.66	11.88	0.10	2.46	2.18	11.18	1.69
	slope	1.14	0.65	0.60	0.86	0.50	0.27	0.27	0.63	0.07
	r	0.82	0.70	0.78	0.83	0.70	0.55	0.67	0.71	0.17
Add2	mean	0.48	16.02	7.08	111.03	0.50	23.16	11.79	82.22	0.76
	sd	0.26	5.84	5.00	11.99	0.12	2.39	3.63	9.21	4.58
	slope	1.62	0.49	0.71	0.58	0.78	0.46	0.58	0.94	0.62
	r	0.81	0.28	0.50	0.48	0.86	0.66	0.78	0.96	0.73
Dom	mean	0.51	20.24	19.80	95.26	0.51	22.72	19.26	93.06	0.80
	sd	0.04	2.50	1.63	8.10	0.05	3.75	1.65	13.44	9.21
	slope	0.66	0.61	0.90	0.95	0.93	0.70	0.67	0.78	0.34
	r	0.82	0.93	0.97	0.95	0.86	0.73	0.79	0.74	0.60
Rare	mean	0.21	18.38	8.48	100.52	0.28	20.93	9.24	90.40	2.03
	sd	0.15	6.90	4.05	10.24	0.15	4.98	3.65	10.79	4.15
	slope	0.70	0.80	0.59	0.81	0.49	0.44	0.39	0.66	0.84
	r	0.81	0.86	0.64	0.84	0.47	0.67	0.48	0.65	0.71

Table 6.4 *Variance estimates and the correlation with, and regression on, the values estimated in the simulation for the same data set.*

		Fixed heritability			Estimated heritability		
Model		σ^2_w	σ^2_u	σ^2_{mg}	σ^2_w	σ^2_u	σ^2_{mg}
Add1	mean	108.990	5.736	33.303	90.298	0.249	59.836
	sd	11.882		13.810	11.183	1.692	14.110
	slope	0.915	-0.094	2.142	0.657	0.120	3.159
	r	0.328	-0.279	0.233	0.257	0.128	0.329
Add2	mean	111.027	12.336	22.902	82.219	0.764	66.107
	sd	11.988		16.144	9.212	4.581	14.926
	slope	0.655	-0.201	-0.129	0.625	0.105	2.66
	r	0.220	-0.434	-0.040	0.277	0.054	0.360
Dom	mean	95.257	5.014	74.790	93.061	0.801	81.815
	sd	8.096		11.060	13.437	9.209	11.898
	slope	0.873	0.006	0.955	0.918	0.520	0.876
	r	0.526	0.033	0.454	0.486	0.513	0.388
Rare	mean	100.520	5.291	22.441	90.396	2.025	37.527
	sd	10.241		10.527	10.794	4.152	15.667
	slope	1.527	-0.045	0.942	0.817	0.581	1.608
	r	0.634	-0.174	0.252	0.306	0.273	0.296

With fixed heritability the residual variance is always over estimated by ME3 in comparison with the estimate from Herm, and hence there is a consistent underestimate of the effect of the major gene. However, in general there is a good linear relationship between the estimates from both methods. This is shown in figure 6.5 which also shows that although, on average, ME3 gives a higher estimate of the residual variance compared with ME1, considering the analyses individually the ME3 estimate is sometimes higher and sometimes lower. The estimates obtained from the analyses of data containing the major gene with dominant effect are very close to the expected values, with the residual variance estimate closer than when using Herm. When the residual variance and major gene variance estimates are compared with the estimates from the simulation for each set of data there is a positive linear relationship. The correlations were high for both variances when the simulated major gene had a dominant effect, and for the residual variance when the simulated gene had a rare allele. There is a poor relationship between the sire variance component estimated in the segregation analysis and estimated from the simulation because this is not being estimated directly in the analyses but as a fixed proportion of the residual variance.

When the polygenic heritability is estimated, in general, the residual variance is under estimated and the major gene effect over estimated compared with the expected values. This is because of the large number of analyses resulting in a major gene model, with the sire variance equal to zero. However this still results in an overestimate of the residual variance compared with the Herm estimate for the data containing a gene with dominant effect and about the same estimate when a rare, additive gene is present. Only 2 of the Add1 analyses, 1 of the Add2, 4 of the Dom and 21 of the Rare analyses resulted in a mixed model, the remainder of the non-zero test statistics being caused by a major gene model. In general, when the estimates are compared with the expected values for the same data set, the slope of the regression and the correlation is lower than that obtained with fixed heritability. The sire variance component, which is now being estimated directly, has a positive linear relationship with the estimate from the simulation for the same set of data.

Table 6.5 gives the mean Herm major gene estimates for those simulations which gave a non-zero ME3 test statistic and for those that gave a zero ME3 test statistic. The analyses that gave a zero test statistic when analysing with ME3 were those with a low test statistic from Herm, which, on average, were not significant. These analyses also resulted in lower estimates for the major genotype means when using Herm than those which ended at a non-polygenic model in ME3. The allele frequencies using Herm are generally less well estimated for the ME3 zero test statistic group and have a larger variance.

Figure 6.5 *The residual variance estimated with ME3 compared with that estimated with Herm for Add1 data analysed with fixed heritability.*

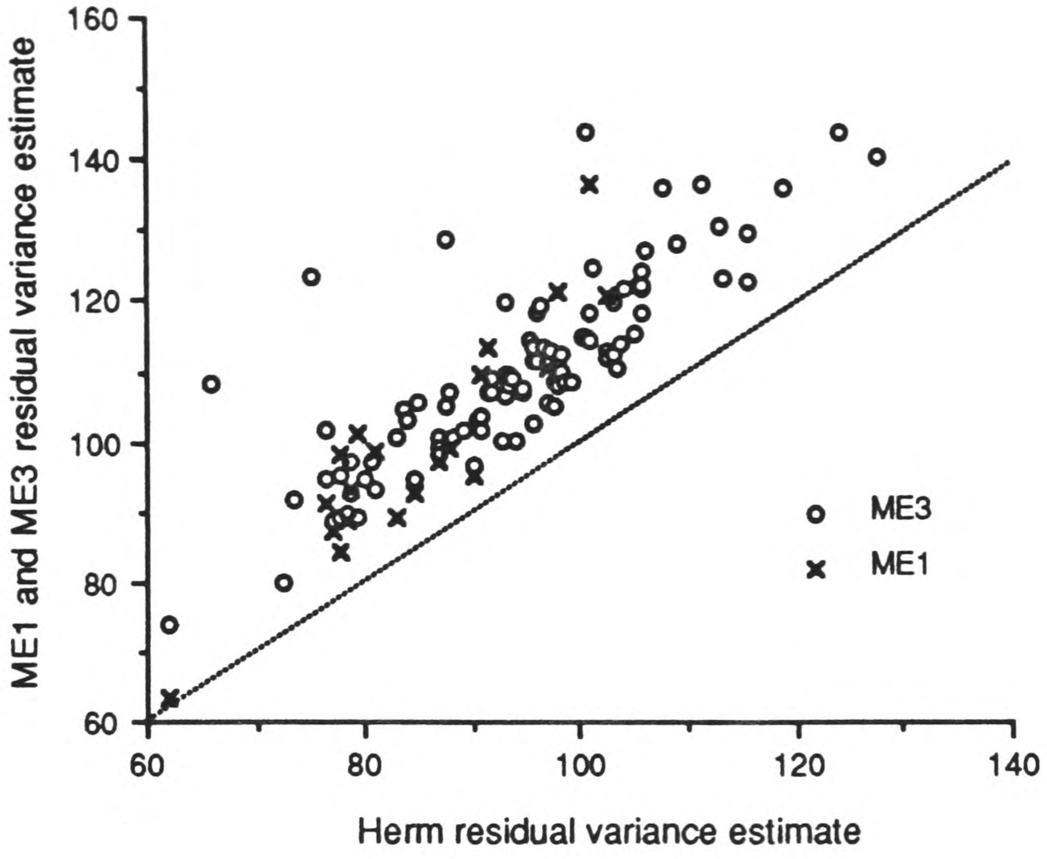


Figure 6.6 *Comparison of the transmitting abilities estimated using ME1 and using ME3 against the progeny mean for a data set simulated under Add1 analysed with fixed heritability.*

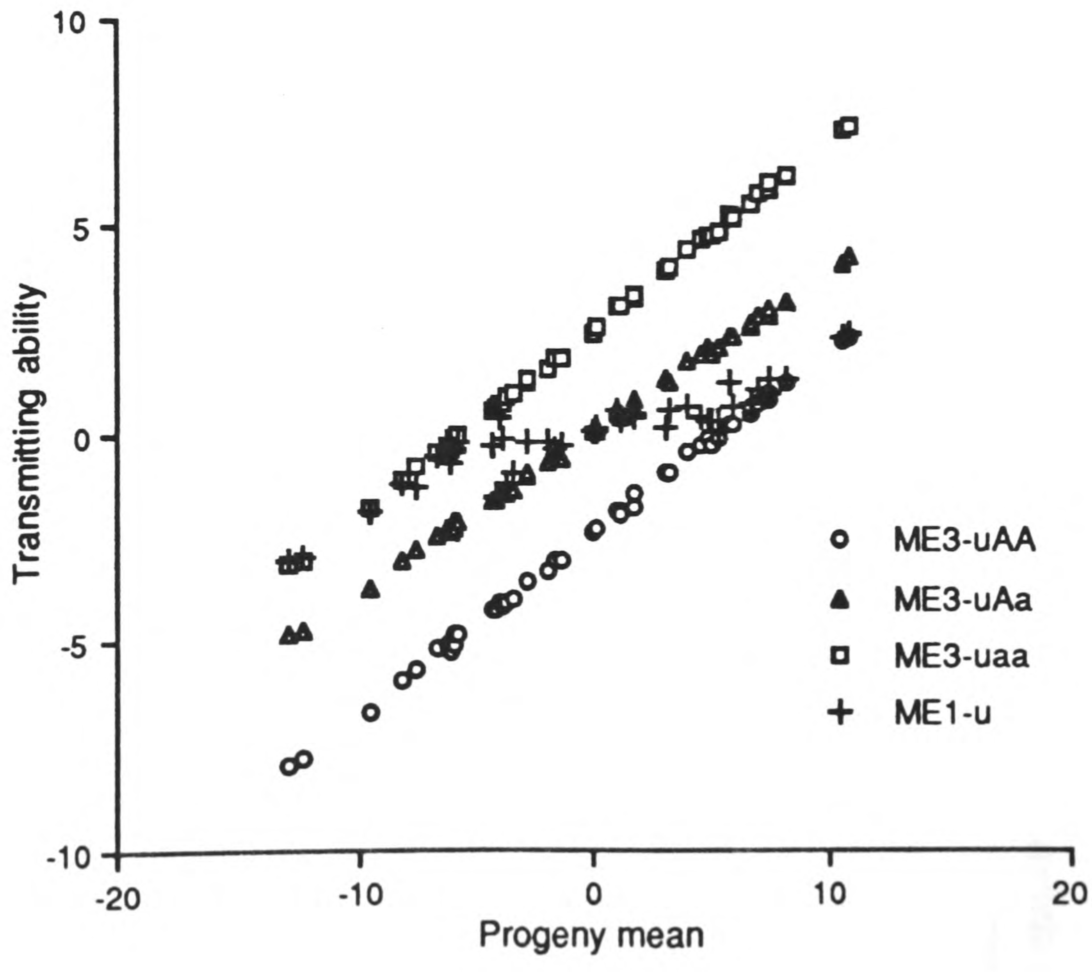


Table 6.5 *Mean (and standard deviation) of the Herm major gene parameter estimates for the ME3 analyses that gave a non-zero test statistic and for those that gave a zero test statistic.*

	non-zero test statistic				zero test statistic			
Model	t.s.	p(A)	μ_{AA}	μ_{Aa}	t.s.	p(A)	μ_{AA}	μ_{Aa}
Heritability fixed								
Add1	13.362 (6.546)	0.496 (0.129)	19.296 (3.979)	9.251 (4.776)	2.117 (0.468)	0.473 (0.151)	8.316 (4.347)	6.315(10.640)
Add2	10.943 (4.473)	0.507 (0.132)	21.977 (3.286)	10.203 (3.522)	4.046 (2.419)	0.471 (0.181)	16.500 (5.690)	9.006 (6.083)
Dom	47.278(14.719)	0.504 (0.048)	20.694 (3.792)	20.149 (1.751)	-	-	-	-
Rare	12.367 (7.111)	0.262 (0.172)	18.335 (7.390)	9.979 (4.415)	1.572 (1.079)	0.346 (0.187)	4.204 (3.012)	6.207 (8.699)
Heritability estimated								
Add1	5.333 (3.794)	0.483 (0.141)	19.433 (5.013)	9.460 (5.465)	2.931 (1.650)	0.602 (0.289)	17.504(12.708)	6.162(13.264)
Add2	6.600 (4.158)	0.507 (0.132)	21.526 (3.471)	10.659 (4.892)	3.066 (2.308)	0.466 (0.174)	18.063 (5.957)	9.246 (5.536)
Dom	41.129(12.893)	0.505 (0.049)	20.493 (3.881)	20.292 (1.918)	-	-	-	-
Rare	7.251 (4.357)	0.246 (0.142)	19.040 (7.686)	10.292 (4.501)	2.098 (2.423)	0.283 (0.236)	16.910(10.502)	13.896(17.960)

t.s. - test statistic

Figure 6.6 shows the relationship between the transmitting abilities estimated under ME1 and those estimated under the hypothesis of each major genotype. When the progeny mean is low the sire has a high conditional probability of being genotype aa (see figure 5.7), and hence the ME1 estimate equals the estimates obtained using ME3 under the hypothesis of that major genotype. Likewise, with a high progeny mean the ME1 estimate approaches that estimated under ME3 assuming that the sire has major genotype AA. At intermediate progeny means, when the conditional probability from the ME1 analyses for the heterozygote is high the ME1 estimates are the same as the heterozygote from the ME3 analyses and otherwise they are pooled estimates of the ME3 transmitting abilities depending on the conditional genotype probabilities.

Genotyping sires at the major locus

The probability of each sire being each major genotype was calculated using [5.3]. Considering the correct genotype for each sire, the probability of being that genotype was grouped into three classifications. The first, if the probability was greater than 0.9, the second greater than 0.75 and the third greater than 0.5. For each analysis the percentage of correctly genotyped sires of each genotype was calculated and the total percentage correctly genotyped over all genotypes. The results are given in table 6.6 as the mean percentage correctly genotyped over the analyses that resulted in a non-polygenic model.

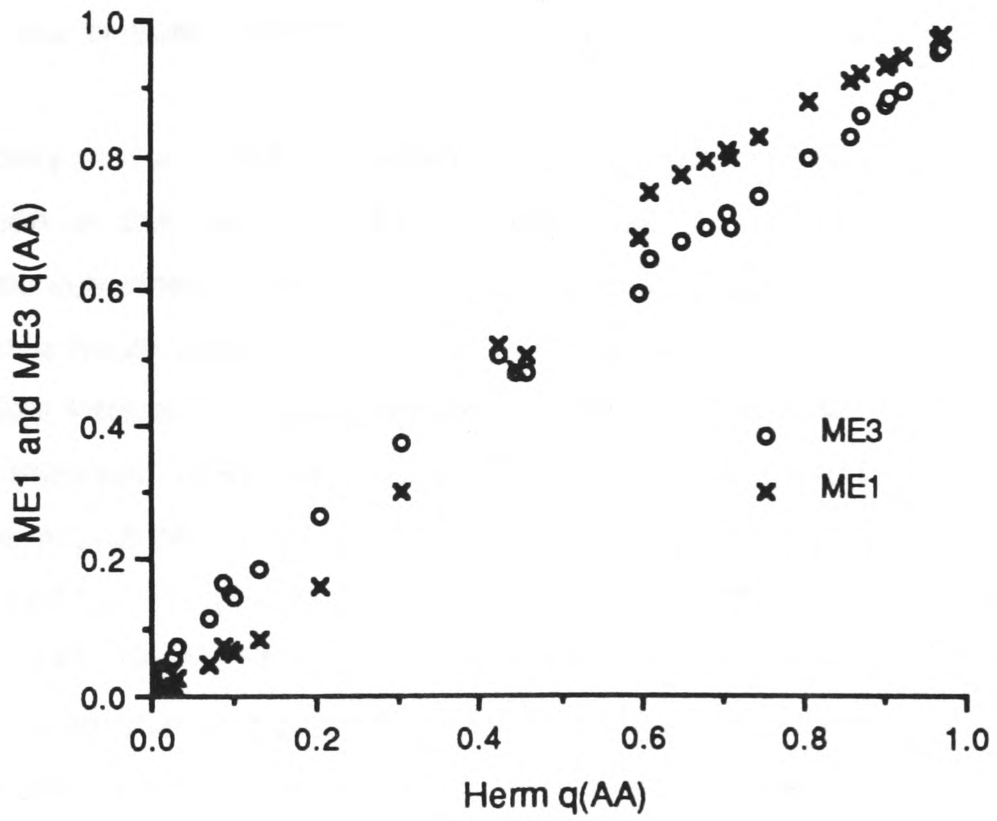
The results are similar to those obtained using Herm (see table 4.9). Over all genotypes, if the criterion for a sire to be assigned to a particular genotype was that his conditional probability for that genotype was greater than 0.9, the highest proportion of sires was correctly genotyped for the additive major gene with rare allele. If 0.5 was taken as the criterion, the highest number of sires correctly genotyped occurred when the major gene had a dominant effect. With fixed polygenic heritability the ME3 analyses correctly identified the genotype of a similar or lower proportion of sires compared with Herm for each situation, except for the common genotype of the analyses containing a rare major allele. Also using the ME3 approximation a lower proportion were correctly identified than with ME1, this could be because the ME1 results are based on fewer analyses, in which the evidence for a major gene was large and hence the sires easier to genotype. The conditional probabilities for ME1 tend to be more extreme than those estimated using ME3. This is shown in figure 6.7 where the conditional probability of a sire being the high scoring genotype estimated using ME1 and ME3 are plotted against the probability estimated using Herm for data simulated under Add1 analysed with fixed heritability. For the probability of being heterozygous the ME1 distribution is even more extreme.

Table 6.6 Percentage of sires correctly genotyped at different values of the conditional probability required to be assigned a genotype, for analyses that resulted in a non-polygenic model.

	AA			Aa			aa			Total		
Model	≥0.9	≥0.75	≥0.5	≥0.9	≥0.75	≥0.5	≥0.9	≥0.75	≥0.5	≥0.9	≥0.75	≥0.5
Fixed heritability												
Add1	12.2	29.8	55.4	1.5	9.5	50.8	15.0	35.9	58.8	8.0	21.4	54.2
Add2	14.3	27.7	42.3	0.6	3.9	33.4	18.1	30.6	47.8	8.6	17.3	39.9
Dom	37.9	67.0	85.6	12.4	43.9	76.4	15.0	36.3	61.5	19.3	47.4	74.8
Rare	3.5	11.1	29.5	2.7	12.1	44.0	48.5	69.0	84.2	33.0	49.4	70.0
Estimated heritability												
Add1	35.4	53.9	68.8	24.5	46.6	65.9	36.7	54.7	68.2	29.8	49.9	66.8
Add2	39.4	55.3	64.4	24.9	42.4	56.6	40.5	52.5	62.5	31.8	47.7	59.8
Dom	49.5	70.3	82.0	26.8	52.5	74.4	26.9	43.1	62.1	32.4	54.2	72.8
Rare	30.8	48.9	58.2	15.6	33.7	51.2	43.3	63.1	74.1	34.5	53.8	66.7

When the heritability is estimated the results from ME3 are similar to those from ME1, because of the high proportion of analyses that resulted in the same major gene model. In general a higher proportion of sires were correctly genotyped than using Herm especially at high probabilities (0.9 and 0.75).

Figure 6.7 *The conditional probability for a sire being genotype AA ($q_i(AA)$) estimated using ME1 and ME3 plotted against the estimates using Herm for Add1 data with fixed heritability.*



6.4.5 Discussion

Considering the ability of the methods to detect a major gene, the ME3 approximation to the mixed model likelihood is an improvement over the ME1 approximation. With fixed heritability, using ME3, major genes were detected in more analyses than with ME1 and there was less of a problem with negative and zero test statistics. However, the mixed model likelihood was lower than that calculated with Herm, which results in a lower test statistic and, hence, fewer analyses giving significant results. When the polygenic heritability was estimated in the analyses, the difference between the polygenic and mixed model likelihoods decreased compared with the situation when the heritability was fixed and, hence, a major gene was detected less frequently. This is because an increased heritability in the polygenic model can explain some of the major gene variance which cannot be explained when the heritability is fixed at the expected value for the polygenic component. This was observed for the analyses with Herm. It suggests that if an underestimate of the polygenic heritability was used in the analyses with fixed heritability, a mixed model might be inferred, simply because the major gene can explain some of the additional polygenic variance. Analyses of the polygenic data with an expected heritability of 0.4 were repeated, this time assuming a value of 0.2 for the heritability. The number of analyses in which evidence for a major gene was found increased to 19 at the 5% significance level and 7 at the 1% level. The mean of the test statistic distribution is 4.48, higher than expected with three degrees of freedom.

The ability of the method to detect a major gene seemed to be more dependent on the distribution of the data than the proportion of the genetic variance it explained. When the data contained an additive major gene with a rare allele, which caused the data to be skewed, a major gene was detected more frequently than when the equivalent gene with equal allele frequencies was simulated. A major gene was detected most frequently, in 99% of the analyses, when the simulated gene had a dominant effect. This gene both explained the largest proportion of the variance and caused the data to be skewed.

Elsen and Le Roy (1989) give the percentage of analyses in which evidence for a major gene was found for models equivalent to Add1 and Dom (h_{12} and h_{11}, respectively) with the polygenic heritability estimated. The model for analysis, as explained previously, estimated the sire genotype frequencies and the dam allele frequency. They compare the test statistic with a 5% quantile obtained by analysis of data simulated under a polygenic model. The value of the quantile used was 10.82 which is similar to the expected value from a χ^2 distribution with five degrees of freedom (11.07). From 100 analyses of data containing 20 half-sibs from each of 20 sires, a major gene was detected in 9 analyses when the simulated gene was additive and 77 when it had a dominant effect. These are lower than the results reported here, especially for the gene with dominant effect, which could be due to the lower number of individuals simulated. When the background heritability was increased (to 0.6) Elsen and Le Roy (1989) found that the number of significant analyses decreased (to 3 and 67 for the additive and the dominant major gene, respectively) as reported here.

On average, most of the parameter results obtained using the ME3 mixed model likelihood were in reasonable agreement with the expected values when those analyses resulting in a polygenic model were ignored. The variance of the estimates was large and hence individual analyses might give misleading results. With the heritability fixed the residual variance was over estimated and the effect of the major gene under estimated. When the heritability was estimated most of the non-zero test statistics were the result of a major gene model with the polygenic heritability at zero.

Each analysis was repeated, using the expected parameter estimates to start the maximisation process and using alternative values where the major gene explained a higher proportion of the variance. The results reported are based on the analysis giving the highest likelihood. With fixed heritability, the analyses were not very sensitive to the initial parameter values used. In only 5 of the 400 analyses were different final models obtained, and in each case two different mixed models were obtained, which might be caused by convergence not being attained. When the heritability was estimated, the analyses were more sensitive, the worst case being for the simulated major gene with a rare allele (Rare), where 22 of the 100 analyses resulted in a different model. Some of these were caused by parameters going to bounds in the analyses and one giving, for example, a major gene model and the other a mixed model. Other differences were caused by different parameter estimates for mixed or major gene models. The highest likelihood could result from either starting model. For Add2 18 of the 100 analyses gave two different models, with the expected parameter estimates giving a polygenic model and the alternative values giving a mixed model.

6.5 USE OF REML VARIANCE EQUATIONS

The ML equations for the variance estimates used in the simulation study are biassed, because they do not take account of the degrees of freedom lost through estimation of the fixed effects. For the mixed model although fixed effects have not been estimated in the analysis, the major genotype means have been estimated and this has not been taken into account in the variance equation. REML equations have been derived in Appendix 4 treating these major genotype means as fixed effects.

Using the ML equations, when the heritability was estimated, the mixed model analyses often resulted in a model with the genetic component composed of only a major gene or only polygenes. A mixed model was rarely obtained. When a major gene model results the sire variance has converged to zero. REML takes account of the estimation of the major genotype means and, hence, of some of the uncertainty caused by not knowing the major genotype of the individuals. It was hoped that this might improve the ability of the approximation to determine the correct model of inheritance. However, as stated in section 2.3.3, the mixed model and polygenic likelihoods obtained with REML are not directly comparable within a statistical framework.

Table 6.7 *Comparison of ML and REML parameter estimates - means of five selected data sets.*

Model	ML model	Method	p(A)	μ_{AA}	μ_{Aa}	μ_{aa}	σ_u^2	σ_w^2
Dom	Mixed	ML	0.509	50.710	50.679	31.001	6.467	99.378
		REML	0.511	50.439	50.715	30.998	7.339	99.921
Dom	Major gene	ML	0.510	51.981	50.089	31.452	0.000	91.422
		REML	0.510	51.208	50.074	31.481	0.000	92.296
Add1	Polygenic	ML	0.423	41.277	41.265	41.258	17.891	128.225
		REML	0.471	41.275	41.264	41.258	18.415	128.885

ML model - resulting model using ML

To consider the effects of using the REML equations some of the data were reanalysed and the resulting parameter estimates from ML and REML compared. Several different situations were considered: five data sets simulated with a dominant major gene that resulted in a mixed model using ML and five that resulted in a major gene model with ML, and five, simulated with an additive model, that resulted in a polygenic model with ML.

Table 6.7 gives the mean results from the analyses of each situation for ML and REML. The use of REML did not alter the resulting model, analyses which ended in a major gene model with ML also resulted in a major gene model with REML. The estimates of the variance components increased which caused slight changes in the other parameters.

Without fixed effects, the use of the REML equations has little effect on the results. However, these REML equations were obtained assuming that the major genotype of the offspring were known, whereas, in fact, there is uncertainty in the classification of animals to genotype classes. Some of this uncertainty is accounted for in the trace of the inverse of the coefficient matrix, however, the equations might be improved by taking further account of this additional variance.

6.6 INCORPORATION OF 2ND DERIVATIVES

The EM algorithms for ML described for both the ME1 and ME3 methods (see section 5.3, Appendix 3, section 6.3 and Appendix 4) can be obtained from the 1st partial derivatives of the log likelihood with respect to each of the parameters to be estimated. For a linear set of equations the algorithm given is equivalent to the Newton-Raphson algorithm (shown in equation [4.5]) because the coefficient matrix from the mixed model equations obtained using 1st derivatives [5.2] is equal to minus the matrix of 2nd partial derivatives (called the observed information matrix). For example, with the following simple model:

$$y = X\beta + e$$

The log likelihood of y given β is:

$$\ln L = \text{const} + \left(-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta) \right)$$

and the first derivative of this with respect to β is:

$$\frac{\partial(\ln L)}{\partial\beta} = X'V^{-1}(y - X\beta)$$

which can be written as before:

$$(X'V^{-1}X)\,\beta = X'V^{-1}y$$

the 2nd derivative is:

$$\frac{\partial^2(\ln L)}{\partial\beta'\partial\beta} = -X'V^{-1}X$$

which is minus the coefficient matrix given above. Hence the Newton-Raphson algorithm can be written as:

$$\begin{aligned}\beta_a &= -\,(-X'V^{-1}X)^{-1}(X'V^{-1}y - X'V^{-1}X\,\beta_{a-1}) + \beta_{a-1}\\ &= (X'V^{-1}X)^{-1}X'V^{-1}y\; - \beta_{a-1}+\beta_{a-1}\\ &= (X'V^{-1}X)^{-1}X'V^{-1}y\end{aligned}$$

Which is the 1st derivative equation used so far.

However, the mixed model likelihood does not give a series of linear equations and hence incorporation of the 2nd derivatives with respect to these parameters into the maximisation routine might improve the rate of convergence. The EM algorithm described is equivalent, for the major genotype means and sire transmitting abilities, to a Newton-Raphson routine assuming that the conditional probabilities for the sire and offspring are fixed and not functions of these parameters.

6.6.1 2nd derivatives

Differentiating the vector of 1st partial derivatives with respect to each parameter gives the matrix of 2nd derivatives. For the major genotype means and sire effects these can be written in the series of matrices shown in figure 6.8.

6.6.2 Maximisation

The first and second partial derivatives with respect to the sire transmitting abilities and the major genotype means can be incorporated into the Newton-Raphson algorithm. Here the Newton-Raphson algorithm will be used only for the major genotype means and sire transmitting abilities and the equations obtained from the EM algorithm will be used for the variances and the allele frequency.

The Newton-Raphson algorithm takes account of the slope and shape of the likelihood surface and hence, in some situations will converge faster than the EM algorithm. Everitt (1984) illustrated this when estimating the parameters for a mixture of two normal distributions. For a single parameter, the effect of the algorithm is that each iteration provides the point at which the tangent to the likelihood curve at the previous point cuts the x-axis and if the likelihood surface is parabolic the maximum will be reached in a single iteration. In regions of negative curvature the algorithm will successfully converge to the maximum of this region. At a point of inflection the second derivative

Figure 6.8 *Partial second derivatives with respect to the major genotype means and the transmitting abilities for sires for the ME3 likelihood.*

$$-\sum_{i=1}^{s}\sum_{j=1}^{n}\sum_{c=1}^{m}\sum_{d=1}^{m} q_i(c)q_{ij}(d\,|\,c)w_{icd}w_{icd}'\left(\frac{1}{\sigma_w^2}-\left(\frac{y_{ij}-w_{icd}\theta}{\sigma_w^2}\right)^2\right)$$

$$-\sum_{i=1}^{s}\sum_{j=1}^{n}\sum_{c=1}^{m} q_i(c)\,w_{ijco}w_{ijco}'$$

$$+\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\left(w_{ioco}-\frac{u_{ic}}{\sigma_u^2}\right)\left(w_{ioco}-\frac{u_{ic}}{\sigma_u^2}\right)'$$

$$-\sum_{i=1}^{s}(w_{iooo}-u_{io})\,(w_{iooo}-u_{io})'$$

$$-\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\frac{1}{\sigma_u^2}\,w_{ico}w_{ico}'$$

Where: w_{iod} is a vector of length (mn+m) containing a one for the cth genotype of the ith sire and for genotype d and zeros otherwise.

w_{iod} is a vector of length (mn+m) containing a one for the cth genotype of the ith sire and zeros otherwise.

$$w_{jco} = \sum_{d=1}^{m} q_{ij}(d\,|\,c)\left(\frac{y_{ij}-w_{icd}\theta}{\sigma_w^2}\right)w_{icd}$$

$$w_{ioco} = \sum_{j=1}^{n} w_{jco}$$

$$w_{iooo} = \sum_{c=1}^{m} q_i(c)\,w_{ioco}$$

$$u_{io} = \sum_{c=1}^{m} q_i(c)\,\frac{u_{ic}}{\sigma_u^2}\,w_{ico}$$

Note that the first derivatives can be written as follows:

$$\sum_{i=1}^{s}\sum_{j=1}^{n}\sum_{c=1}^{m}\sum_{d=1}^{m} q_i(c)q_{ij}(d\,|\,c)w_{icd}\left(\frac{y_{ij}-w_{icd}\theta}{\sigma_w^2}\right)-\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)w_{ico}\frac{u_{ic}}{\sigma_u^2}$$

$$=\sum_{i=1}^{s} w_{iooo}-\sum_{i=1}^{s} u_{io}$$

will be zero and therefore the correction to the previous estimate will be infinite, near a point of inflection the second derivative will be very small leading to new estimates wildly out. Beyond the turning point in a region of positive curvature the process will lead away from the maximum. Hence, the process is sensitive to the initial parameters used to initiate the maximisation process. To improve this it may be worthwhile transforming to a new parameter which gives a likelihood curve that is more nearly parabolic (Edwards, 1972).

For the ME3 likelihood the surface is not continuously convex and hence problems are encountered when attempting to maximise the likelihood using the Newton-Raphson algorithm. Even when the expected values from the simulation were used as the initial parameter estimates, convergence was not, generally, attained. One means of overcoming this problem is to use the equations based on the 1st derivatives (EM algorithm) and then include the 2nd derivatives when a region of negative curvature around the maximum has been attained.

To see the effect of combining the two maximisation routines the algorithm described above with Newton-Raphson for the means and transmitting abilities and EM for the frequency and variance and the EM algorithm for all the parameters were used. The polygenic heritability was fixed at the expected value. The data contained an additive major gene and was known to converge to a mixed model with the EM algorithm. Different numbers of iterations with the first derivatives were tried before incorporation of the second derivatives. The results are given in table 6.8, with the convergence criterion given after the last EM iteration and measured as: $\sqrt{\frac{(\theta_{[a]}-\theta_{[a-1]})'(\theta_{[a]}-\theta_{[a-1]})}{\dim(\theta)}}$, where θ is a vector of parameter estimates for the major genotype means and sire transmission probabilities of dimension dim(θ) and [a] is the iteration number and the time for convergence to be attained when the criterion was less than 10^{-5}.

It can be seen that the EM algorithm converges fast initially but then slows down and, using this convergence criterion, the criterion increases before finally reaching convergence. Even using just two EM iterations enabled the maximum to be found with the Newton-Raphson algorithm. The use of the second derivatives decreases the number of iteration required, from 1164 to 949, the lowest in this sample. However, each iteration is more complicated and hence, in this example, there was no benefit in terms of the time taken to reach convergence.

Table 6.8 *Number of iterations required for convergence to be attained, using a combination of EM and Newton-Raphson algorithms.*

Number of iterations		Convergence	
EM	Newton-Raphson	criterion	No. secs.
2	952	0.2643	3517
10	939	0.0165	3501
50	910	0.0042	3479
100	871	0.0009	3435
200	786	0.0007	3310
500	537	0.0012	3054
700	384	0.0014	2970
1000	151	0.0003	2738
1164	0	-	2537

Where the convergence criterion is the value of the criterion after 1st derivative iterations and the time is the number of seconds taken for convergence to be attained.

Incorporation of the partial second derivatives with respect to the residual variance and allele frequency have not been investigated, but might reduce the number of iterations further. However, each iteration would become even more complex and take longer to compute and, hence, the time taken to reach convergence might not be improved.

6.6.3 Standard errors of the estimates

Estimates of the standard errors of the estimates can be obtained from the inverse of minus the matrix of 2nd derivatives or the inverse of the observed information matrix.

$$\mathrm{var}(u - \hat{u}) = C_{zz}\,\sigma^2_w$$

and $\mathrm{var}(\hat{b}) = C_{xx}\,\sigma^2_w$

where C is the inverse of minus the matrix of 2nd derivatives.

Alternatively the expected second derivatives can be used.

For a linear model the inverse of the coefficient matrix from the first derivative equations can be used. In general it is not computationally feasible to invert this matrix and methods of approximating the error variances have been used. Often, the surface is assumed to be quadratic around the ML estimate and the prediction error variances calculated form the second derivatives of this curve (Smith and Graser, 1986). For a non-linear model the advantage of the Newton-Raphson algorithm is that the second derivatives are calculated and can be used to obtain the variances and covariances of the estimates of the parameters.

For the mixed model the prediction error variances from the second derivatives will be taking account of the change in the conditional probabilities, rather than assuming them to be fixed.

6.6.4 Use of the 2nd derivatives in the variance equations

Using a Bayesian argument Foulley *et al.* (1987) suggest the use of the 2nd derivatives in the variance equations. They consider a situation where the sire of some individuals in the sample is unknown. Information from genotyping using, for example, major histocompatibilty markers or from knowledge of management procedures enables sires to be suggested as the true sire each with an associated probability. Hence, this is a similar problem to allocating major genotypes to individuals. The major genotypes can be considered as fixed effects which are assigned to each individual with a probability, dependent on the genotype frequencies of the parents and the transmission probabilities. In this case the uncertainty is connected to the fixed effects rather than the random effects. Also, the data has been used to estimate the frequency of the major genotypes whereas for the sires the prior probability of paternity has come from external sources.

Foulley *et al.* (1987) take the expected value of the sums of squares assuming that the posterior probability of being the sire of any individual is a constant and not dependent on the parameter estimates. The expression for which the expectation is required contains a quadratic in θ and hence as the expectation is taken with respect to the conditional distribution $f(\theta|\sigma^2_e, \sigma^2_u, y)$ the variance of $(\theta|y)$ is required. This can be approximated by the inverse of the matrix of 2nd derivatives, which will be taking account of the uncertainty of which offspring belongs to which sire.

A similar idea could be pursued incorporating the uncertainty of the major genotypes of the offspring by using the 2nd derivatives.

6.7 DISCUSSION

The likelihood calculated using the ME3 approximation is exact if the genotypes of all the offspring are known. Otherwise the approximation pools the information from the major genotypes of the offspring conditional on the sire major genotype to obtain an estimate of the transmitting ability of the sire for each major genotype. The likelihood can then easily be calculated as the offspring genotypes are independent given the genotype of the sire.

The ME3 approximation is an improvement over ME1, in that evidence for a major gene is found more frequently when a major gene exists in the data. It is not as powerful as Herm. When a major gene was detected the parameter results were in good agreement with the expected values although the ability of the method to correctly estimate the polygenic heritability was low, the estimate being zero in the majority of cases.

When the polygenic heritability was assumed to be known and fixed in the analyses a major gene was detected more frequently. However, as observed with Herm, the method is not robust to the value of the heritability assumed and if this assumed value is an underestimate of the true polygenic heritability in the data a major gene is more likely to be inferred.

Estimates of the transmitting abilities for each sire are obtained under the hypothesis of each major genotype. To make use of this information, and the major gene information, the major genotype of the sire is required. The relevant polygenic component could then be assumed for the sire. Alternatively, if the sire cannot be allocated a major genotype (because the conditional probabilities of being each genotype are similar) these transmitting abilities could be combined, weighted by the probability of the major genotype. This would give estimates equivalent to those obtained from ME1. In the worst case, on average, only 40% of the sires were correctly genotyped when the criterion for allocation to a major genotype was that the conditional probability of being that genotype was greater than 0.5.

Although the power of ME3 was lower than Herm it has some advantages, in particular the faster speed of computation and the ease of inclusion of fixed effects, without increasing the time for convergence to a great extent. However, the extension of ME3 to more complicated pedigrees, for example including relationships between the parents would increase the complexity of calculation significantly. For example, the inclusion of full-sibs but ignoring any further relationships between progeny, would mean that the breeding value of both the sire and dam under the different major genotypes would have to be considered at the same time, giving nine estimates of the contribution of the parental polygenes to the offspring instead of the three being considered presently. In general all combinations of major genotype for related individuals which are parents in the pedigree would have to be considered and with each combination the breeding value of the individuals calculated. Depending on the species concerned, this might reduce the calculations significantly compared with the exact likelihood, where all combinations of major genotype for all related individuals is required. However, the computation required might still be prohibitive for large pedigrees.

Appendix 4

A4.1 EM Algorithm for maximisation

The log of the ME3 approximation to the mixed model likelihood can be written as follows:

$$\ln L = \sum_{i=1}^{s} \left\{ \frac{1}{2}\ln(2\pi\sigma_w^2) - \frac{1}{2}\ln(n+\lambda) + \ln\left(\sum_{c=1}^{m} p(c)h(\hat{u}_{ic}) \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d|c)k_d(y_{ij}|\mu,\mu_d,\hat{u}_{ic},\sigma_w^2) \right) \right\}$$

Differentiating this with respect to each parameter to be estimated and equating the derivatives to zero gives the following equations for the parameters:

Major genotype means

$$\mu_d = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n} q_{ij}(d|c)\,(y_{ij}-\mu-\hat{u}_{ic})}{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n} q_{ij}(d|c)}$$

Polygenic transmitting ability for each major genotype for sires

$$\hat{u}_{ic} = \frac{q_i(c) \sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d|c)\,(y_{ij}-\mu-\mu_d)}{q_i(c)\,(n+\lambda)}$$

Genotype frequencies

An equation can be obtained for the allele frequency in both the sire and dam population assuming that the population is in Hardy-Weinberg equilibrium. Alternatively this assumption can be relaxed and genotype frequencies estimated in the sires and an allele frequency in the dams. The equations obtained, under the different assumptions, are identical to those given for the ME1 approximation in Appendix 3.

Variance components

If the polygenic heritability is assumed to be known, then the following equation for the residual variance is obtained:

$$\sigma_w^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\left(\hat{u}_{ic}^2\lambda + \sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d\,|\,c)\,(y_{ij}-\mu-\mu_d-\hat{u}_{ic})^2\right)}{sn}$$

Otherwise, equations for the residual and sire variance components are as follows:

$$\sigma_w^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d\,|\,c)\,(y_{ij}-\mu-\mu_d-\hat{u}_{ic})^2}{sn - s + \frac{s\lambda}{n+\lambda}}$$

$$\sigma_u^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\,\hat{u}_{ic}^2}{s - \frac{s\lambda}{n+\lambda}}$$

As for ME1 the equations have been written as functions of $q_i(c)$ and $q_{ij}(d|c)$, where:

$q_i(c)$ is the conditional probability of genotype c for sire i.

$$q_i(c) = \frac{p(c)\prod_{j=1}^{n}\sum_{d=1}^{m}\mathrm{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,\hat{u}_{ic},\sigma_w^2)}{\sum_{c'=1}^{m} p(c')\prod_{j=1}^{n}\sum_{d=1}^{m}\mathrm{trans}(d\,|\,c')\,k_d(y_{ij}\,|\,\mu,\mu_d,\hat{u}_{ic'},\sigma_w^2)}$$

$q_{ij}(d|c)$ is the conditional probability that offspring j from sire i has genotype d given that the sire has genotype c.

$$q_{ij}(d\,|\,c) = \frac{\mathrm{trans}(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,\mu_d,\hat{u}_{ic},\sigma_w^2)}{\sum_{d'=1}^{m}\mathrm{trans}(d'\,|\,c)\,k_{d'}(y_{ij}\,|\,\mu,\mu_{d'},\hat{u}_{ic},\sigma_w^2)}$$

The maximisation process is the same as described in Appendix 3. The equations for the major genotype means (μ_d) and transmitting abilities for each major genotype

(u_{ic}) (and fixed effects if included) can be solved simultaneously by arranging them into matrix form. The matrices are similar to those for ME1, although now the sire part of the matrix is larger with three lines for each sire, one for each genotype. The matrices are given in equation [6.2].

A4.2 REML variance estimates

The ML equations described above give biassed estimates of the true variances as they do not take account of the degrees of freedom lost through estimation of the fixed effects. REML has been developed for the polygenic model as described in section 2.3.3.

The ME3 approximation to the mixed model likelihood is exact for the sire model when the major genotypes of all the offspring are known. With this assumption and assuming a general model with different design matrix ($\mathbf{Z}_{ic}$) and variance-covariance matrix ($\mathbf{V}_{ic}$) for each genotype of each sire the likelihood can be written as, using the notation defined in section 2.2:

$$\ln L(MM) = -\frac{sn}{2}\ln(2\pi) + \sum_{i=1}^{s} \ln\left(\sum_{c=1}^{m} \frac{p(c)}{|\mathbf{V}_{ic}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{W}_D\mu_c)'\mathbf{V}_{ic}^{-1}(\mathbf{y}_i - \mathbf{W}_D\mu_c) \right] \right)$$

Assuming that sires are unrelated the coefficient matrix from the first derivatives is:

$$\begin{bmatrix} \mathbf{W}_D'\mathbf{W}_D & q_i(1)\mathbf{W}_D'\mathbf{Z}_{i1} & q_i(2)\mathbf{W}_D'\mathbf{Z}_{i2} & q_i(3)\mathbf{W}_D'\mathbf{Z}_{i3} \\ q_i(1)\mathbf{Z}_{i1}'\mathbf{W}_D & q_i(1)(\mathbf{Z}_{i1}'\mathbf{Z}_{i1}+\lambda) & 0 & 0 \\ q_i(2)\mathbf{Z}_{i2}'\mathbf{W}_D & 0 & q_i(2)(\mathbf{Z}_{i2}'\mathbf{Z}_{i2}+\lambda) & 0 \\ q_i(3)\mathbf{Z}_{i3}'\mathbf{W}_D & 0 & 0 & q_i(3)(\mathbf{Z}_{i3}'\mathbf{Z}_{i3}+\lambda) \end{bmatrix}$$

Assuming that the posterior probability of each genotype for each sire is not a function of the parameters but fixed, the variance-covariance matrix can be obtained from the first derivatives. Absorption of the sire effects gives the following matrix:

$$\left(\mathbf{W}_D'\mathbf{W}_D - \sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\, \mathbf{W}_D'\mathbf{Z}_{ic}(q_i(c)(\mathbf{Z}_{ic}'\mathbf{Z}_{ic} + \lambda))^{-1}\mathbf{Z}_{ic}'\mathbf{W}_D q_i(c) \right) \frac{1}{\sigma_w^2}$$

which will be denoted: $\mathbf{W}_D'\mathbf{V}_*^{-1}\mathbf{W}_D$

The mixed model likelihood can be written:

$$\begin{aligned}\ln L(MM) = & -\frac{sn-m}{2}\ln(2\pi) - \frac{1}{2}\ln\left| W_D' \mathcal{V}_*^{-1} W_D \right| + \frac{1}{2}\ln\left| W_D' W_D \right| \\ & + \ln\left(\sum_{i=1}^{s}\sum_{c=1}^{m} \frac{p(c)}{\left| V_{ic} \right|^{1/2}} \exp\left[-\frac{1}{2}(y_i - W_D\mu_c)' \mathcal{V}_{ic}^{-1} (y_i - W_D\mu_c) \right] \right)\end{aligned}$$

Differentiating this with respect to the two variance components, σ_w^2 and σ_u^2, again assuming that the conditional frequencies for each sire are constants and not functions of the variances, and equating the derivatives to zero gives the following equations for the variance components:

$$\sigma_w^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c) \sum_{j=1}^{n}\sum_{d=1}^{m} q_{ij}(d \mid c)\,(y_{ij} - \mu - \mu_d - \hat{u}_{ic})^2}{sn - s - m + 1 + \sum_{i=1}^{n}\sum_{c=1}^{m} \{q_i(c)\,\lambda\,\mathrm{tr}[C_{ic}] - \lambda\,\mathrm{tr}[(Z_{ic}'Z_{ic} + \lambda)^{-1}]\,(1 - q_i(c))\}}$$

$$\sigma_u^2 = \frac{\sum_{i=1}^{s}\sum_{c=1}^{m} q_i(c)\,\hat{u}_{ic}^2}{s - \sum_{i=1}^{s}\sum_{c=1}^{m} \{q_i(c)\,\lambda\,\mathrm{tr}[C_{ic}] - \lambda\,\mathrm{tr}[(Z_{ic}'Z_{ic} + \lambda)^{-1}]\,(1 - q_i(c))\}}$$

Where: C_{ic} is the element of the inverse of the coefficient matrix for sire i with genotype c.

$(Z_{ic}'Z_{ic} + \lambda)$ is a scalar with the value $n+\lambda$, hence $\mathrm{tr}[(Z_{ic}'Z_{ic} + \lambda)^{-1}]$ is $(n+\lambda)^{-1}$.

Generally $Z_{i1} = Z_{i2} = Z_{i3} = Z_i$ and $V_{i1} = V_{i2} = V_{i3} = V_i$ and the equations can be simplified.

Fixed effects can be included in the same way as the major genotype means, reducing the denominator by the matrix X'X (see section 2.3.3) and adjusting the phenotypes for the relevant effects.

CHAPTER 7

SUMMARY AND CONCLUDING REMARKS

7.1 SUMMARY

Most traits of economic importance in animal breeding are quantitative, that is they are controlled by a number of loci which cannot be identified and manipulated independently by classical animal breeding techniques. The aim of this work has been to develop statistical methods to identify the existence of genes of large effect that contribute variation to the trait of interest.

Work carried out on human populations indicated that segregation analysis was the most appropriate method, being more powerful than other methods suggested, giving parameter estimates and having a fairly wide application. However, the likelihood of the general model (i.e. the mixed model, which allows both polygenic and major genetic variance) cannot be calculated in a reasonable length of time, even for pedigrees of moderate size, and hence approximations to this likelihood have been considered.

The use of Hermite integration to replace exact integration gives a value for the likelihood that is a very good approximation, and in the work presented here this method has been used as a baseline against which the other approximations have been compared. Two other approximations have been considered that are extensions of methods already used in animal breeding theory, ME1 and ME3. PAP, an approximation used by human geneticists, was investigated but not compared with the others because it was thought to be too difficult to extend to be suitable for animal breeding situations.

Simulated data has been used to compare the approximations. Two different situations have been considered. The first assumed that an accurate estimate of the polygenic heritability for the trait in question was already available, for example from previous analyses of data from populations in which a major gene was not segregating. In the second situation, the polygenic heritability was estimated. The distribution of the test statistic (i.e. twice the difference in the natural logarithms of the MLs of the two models being compared) obtained from analysis of polygenic data was investigated. In comparisons of the mixed and polygenic models, with a single allele frequency to describe the genotype frequencies, the test statistic distribution is expected asymptotically to follow a χ^2 distribution with three degrees of freedom. Using Hermite integration the distribution of test statistics was not significantly different from this expected distribution both with fixed and estimated polygenic heritability. The distributions obtained with ME1 and ME3 were significantly different from a χ^2 distribution because of the high number of zero test statistics. However, when the polygenic heritability was estimated the ME3 approximation gave about the correct number of significant results at the 5% and 1% levels, hence the χ^2 distribution might provide a reasonable criterion to test for a major gene in this situation.

Assuming a well estimated polygenic heritability, the methods were more powerful and the parameter estimates closer to the expected values than when the heritability was estimated in the analyses. With the polygenic heritability fixed, however, there is a problem that a major gene may be suggested to explain any additional genetic variance that is not already accounted for, even if it is polygenic in origin. Segregation analysis may not be robust even to small deviations from the true value. Hence the polygenic heritability needs to reflect the true polygenic heritability in the sample being analysed.

An important comparison of the three approximations is the power, i.e. whether a major gene, when it is present and segregating, can be detected. For the genetic models considered here, both when the heritability was estimated and when it was assumed to be known in the analyses, Herm was the most powerful approximation. The largest differences between the results from Herm and the other approximations were for the simulated additive major genes when a value for the polygenic heritability was assumed. Also in this situation the improvement in the power of the ME3 approximation over that of ME1 was greatest. For example, when the simulated data contained a gene with equal allele frequencies explaining 71% of the genetic variance with a polygenic heritability of 0.2 (Add1), Herm detected a major gene in 75 of the 100 analyses, ME3 in 33 and ME1 in only 4. Using Herm, a major gene was detected least frequently when the simulated major gene was additive and explained only 56% of the genetic variance, with a polygenic heritability of 0.4 (Add2). In this case, when the heritability was estimated, in only 15 of the 100 analyses was there significant evidence for a major gene. When the major gene explained the largest proportion of the genetic variance (79% in Dom), the three approximations gave similar results in terms of the number of significant analyses (almost 100% of analyses) both for the analyses with fixed and for those with estimated heritability. However, the mean test statistic for ME3 was lower than that with Herm and the mean ME1 test statistic was even lower. The ability of the methods to identify a major gene is also affected by distribution of the data. When the heritability was estimated, evidence for a major gene was found more frequently when the distribution of phenotypes was skewed. This is because a polygenic model cannot explain a non-normal distribution of phenotypes but the addition of a major gene can account for the skewness.

The parameter estimates obtained from the methods are also of interest, and are required to determine the best use of the major gene. On average, the parameter estimates obtained using Herm were in good agreement with the expected values. There was a weak, positive correlation between the variance components estimated in the analyses and the estimates of the variances actually generated in the simulation. The highest correlations were obtained for the estimated sire variance when the simulated major gene had an allele with dominant effect (Dom; r=0.57) and when the major gene was additive with a rare allele (Rare; r=0.56). The estimates of all parameters had greater genetic variance over the 100 analyses when the polygenic heritability was estimated. In the two models where the major gene caused a skewed distribution (i.e. Rare and Dom), its contribution to the variance, was, on average, over estimated using Herm and the residual variance under estimated to a similar extent. When the heritability was fixed, the ME1 and ME3 analyses, on average, over estimated the residual variance and consequently under estimated the effect of the major gene compared with the results from Herm. When the heritability was estimated, there were problems distinguishing the simulated major gene and polygenic variation using ME1 and ME3, and these analyses usually resulted in a major gene model. However, in this case, the parameters describing the major gene gave a good indication as to the action and frequency of the simulated major gene, although, on average, over estimated its effect. When a model involving a major gene was suggested, in all situations there was a strong linear relationship between the parameter estimates under Herm and those from ME1 and ME3.

The use of major genes, once identified, depends on the ability to genotype individuals at the major locus. In this work, sires were allocated genotypes dependent on the phenotypes of their offspring. When a mixed model resulted the three approximations were similar in their ability to correctly genotype the sires and, on average, at least 50% of the sires could be correctly genotyped when the criterion for allocation to a genotype was that the conditional probability of being that genotype was greater than 0.5. However, unfortunately the number of sires incorrectly genotyped at this probability is not known. The offspring have not been genotyped at the major locus, but, in general, the genotyping would be much less accurate then for the sires. The benefit of knowing the major genotype of just a proportion of animals, including some incorrectly genotyped has not been investigated, however, obviously the requirement is for a high number of sires correctly genotyped with few incorrect.

For selection to be efficient the major genotype and polygenic genotype of each individual is required. The polygenic merit of the sires is obtained directly from ME1, although the estimate involves pooling over the three possible major genotypes for the sire. ME3 gives an estimate for the transmitting ability for each major genotype of each sire. If the sire can be accurately genotyped at the major locus, then the transmitting ability estimated under that genotype would be the required value, otherwise the estimates would have to be pooled depending on the conditional probability of each genotype, resulting in the same equation as used in ME1. Herm does not estimate the transmitting ability as the likelihood is calculated as a summation over many different values for this parameter. However, some indication of the transmitting ability could be obtained from the likelihood of the half-sib family for each of the values used in the summation. Alternatively, the transmitting abilities could be estimated after maximisation using the ML parameter estimates, although the same problem will exist as for ME3; that is which major genotype to consider for each sire, or whether to pool the estimate over all genotypes.

An important consideration is the time taken for the analyses. The time taken can be broken into two components, the time for each iteration or function evaluation, and the number of iterations required for convergence to be attained. A single iteration of the EM algorithm for ME3 takes about 50% more time than an iteration for ME1. The number of iterations was extremely variable depending on the simulated genetic model and whether a mixed or polygenic model was attained at convergence. On average, ME3 required more iterations (about 460) than ME1 (about 200) to reach convergence. With the heritability fixed, ME3 took much longer to converge than when the heritability was estimated, except when the simulated major gene had an allele with dominant effect (Dom). With this dominant model nearly all the analyses resulted in a major gene and the ME3 approximation took fewer iterations than ME1. Herm has been programmed using a quasi-Newton routine with a different convergence criterion to that used for the EM algorithm, and, hence, is not directly comparable with the results from ME1 and ME3. Each function evaluation takes about 15 times as long as an iteration for ME1. However, the number of function evaluations required for convergence was much less (about 170) than the number of iterations for ME1 and ME3, with much less variation over the 100 analyses for each situation.

In the analyses, the expected values for the parameters were known and these could be used as the initial values from which the maximisation process starts. Hence, the analyses are likely to be starting near the expected global maximum. When analysing real data the true values of the parameters are unknown and several starting values should be tried. Using an allele frequency, Herm was not very sensitive to the initial starting values. In the main simulation study two different starting values were used for ME1 and ME3. ME3 resulted in the same model from these starting values more often than ME1. With fixed polygenic heritability the methods were not very sensitive, the worst case being for ME1 when the simulated major gene was additive with a rare allele (Rare), when in 13 of the 100 analyses different final models were obtained, one of these models being polygenic and the other a mixed model. With the heritability estimated the methods were more sensitive, especially for the models where the major gene explained a lower proportion of the genetic variance. In the worst case, for the major gene segregating in a polygenic background with heritability of 0.4 (Add2), 41 of the ME1 analyses resulted in two different end models, the expected parameter values always giving the polygenic model and the alternative starting values a major gene model.

7.2 FUTURE CONSIDERATIONS

7.2.1 Segregation analysis.

The work presented here has used a sire model, however, the use of a more complex model, especially an animal model incorporating all genetic relationships, might be advantageous. For a polygenic model, incorporating relationships can account for the effects of selection. In the case of the mixed model, selection affects not only the polygenic component, but also the frequency of the major genotypes. The sire model assumes that the major genotype frequency is constant over time. Indicating relationships between parents will allow for a change in the major genotype frequency, the frequency estimated being the frequency in the base population. Selection might also create a relationship between the polygenic and major genotypes, which are assumed to be independent in the base population. Of the approximations examined here, the ME1 likelihood is the easiest to extend to an animal model making use of algorithms already used for the polygenic model. Herm and ME3 likelihoods become much more computationally demanding even with only simple additions to the pedigree structure. For example, including full-sibs, means that both the sires and dams have to be considered simultaneously, therefore, rather than summing over the three genotype options for a sire, there has to be a summation over the nine combinations for the sire and dam and for ME3 the breeding values for the sires and dams have to be calculated in these nine different situations. Further investigation is required, for example it might be possible to include the weighted mean for spouse when considering the three genotype frequencies for one of the parents. Incorporating more individuals for ME1 might improve the approximation because there will be more information to correctly genotype the animals and if the genotype of all individuals is known, the approximation becomes exact.

Fixed effects have been ignored in this comparison of the approximations, however, generally these are required in animal breeding data and it would be better if they could be estimated simultaneously with the effects of the major genotypes. The ME1 and ME3 approximations can easily be extended to incorporate these effects in theory, as shown in equations [5.3] and [6.2], respectively. Using the quasi-Newton algorithm for maximisation for the Herm approximation means that the incorporation of fixed effects would significantly reduce the speed of computation. If fixed effects are incorporated the variance estimates ought to take account of this and the associated degrees of freedom adjusted accordingly. REML has been developed for this reason, although the resulting mixed model likelihood is not directly comparable with the polygenic likelihood.

Unlike the other approximations considered, PAP is able to make use of the complete pedigree. This makes it comparatively slow for the two generation model. The complexity of the approximation as programmed makes it difficult to extend, although similar approximations could be considered.

Further problems that have not yet been addressed include the use of repeat records, which might improve the separation of genetic and environmental effects, the effect of non-normality caused by non-genetic effects and the best policy to use when analysing non-normal data and the analysis of threshold or meristic traits, such as litter size. The inclusion of additional random effects, for example, an effect for common environment, in the model also requires study.

7.2.2 Use of linked markers.

Recently, there has been much interest in creating genome maps of farm animals using markers. If these maps become a reality then the use of markers linked to loci affecting quantitative traits would be useful for the genetic improvement of these animals (Geldermann, 1975). When this project was started, it was thought that it would be some time before there would be enough markers mapped in farm animal species and even longer before animals were automatically typed for these marker genotypes. Recently, however, interest in this area has increased, and with proposed collaboration between laboratories (for example, Haley *et al.* 1990) the idea of complete genome maps becomes feasible.

The use of markers to detect genes of large effect (often termed quantitative trait loci or QTLs in this situation) is of great potential value, because it allows easy manipulation of the major gene once found (marker assisted selection (MAS), for example, Soller and Beckmann, 1982) and provides a possible route to its isolation using reverse genetics (for example, Orkin, 1986). To make use of markers, the existence of linkage between a locus controlling a quantitative trait of interest and a locus at which it is easy to determine the genotype is required. Segregation must be occurring at both loci and then genotypes at the marker locus can be associated with phenotypes or genotypes at the QTL.

The idea of using markers linked to a trait of interest is not new, however, initially the marker loci used were blood groups or controlled coat colour (Neimann-Sørensen and Robertson, 1961). To be useful a marker needs to be polymorphic, preferably with multiple alleles as matings are most informative when the parents are heterozygous, each with different alleles. In this situation, the parent transmitting each allele in the offspring can be identified and the offspring will be segregating within the family, so that the alleles can be associated with an effect.

Molecular markers now give access to abundant genetic polymorphisms, some of which are ideally suited to be genetic markers. Restriction fragment length polymorphisms (RFLPs) are often caused by the loss or addition of a cleavage site for a specific restriction enzyme. When the DNA is digested with this enzyme, fragments of different length will be obtained depending on the restriction sites present. These RFLPs suffer from the disadvantage that they are diallelic, because they are scored as a presence or absence of the restriction site, and hence show, at most, 50% heterozygosity. However, to improve the information contained in these markers, multiple site polymorphisms could be considered, where several markers which are located very close to each other are treated as a haplotype (Georges *et al.* 1990). The practical manipulations required, however, possibly become cumbersome, with the use of many restriction enzymes.

More recently a second type of possible marker has been identified which have been termed, variable number of tandem repeat loci (VNTRs; Nakamura, 1987). Some sequences of DNA are highly repetitive and polymorphism exists in the number of sequence repeats. This variation can be observed when digesting with a restriction enzyme that cuts at a site located to either side of the series of repeats. A large number of alleles exist, each allele having a different number of copies of the repeat sequence and hence a different fragment length. These can be observed by hybridisation on a gel using the repeat sequence as a probe. One type of VNTR are known as minisatellites (Jeffreys, 1985). Under reduced stringency conditions the probe hybridises to several or many loci containing the repeat (or very similar repeat). The large number of alleles found at these loci means that the different sites are unlikely to have the same number of repeats and hence many bands will be found along the length of the gel. Although providing a lot of information, these DNA fingerprints are difficult to analyse because of the difficulties in determining allelism. Site-specific VNTR loci, obtained by using a unique sequence located to one side of one set of the repeats as a probe, are more amenable to analysis.

Microsatellites are VNTRs with a short repeat sequence, usually two bases, for example, TG. Such sequences seem to be distributed throughout the genome and are very common (Weber and May, 1989). They can be located initially using a two base pair repeat as a probe and then a unique sequence to both ends of the repeats can be obtained to use as a primer for the polymerase chain reaction (PCR) for DNA synthesis followed by polyacrylamide gel electrophoresis to determine allele sizes. Like minisatellites they can have many alleles and thus be very heterozygous and their ubiquity and ease of use means that they are rapidly becoming the marker of choice for mapping studies.

In order for any markers to be of use they need to be associated with loci which control traits of interest. Initially the markers can be placed on a genetic map using linkage studies. There are many computer programs written to enable this, and methods now exist that consider several marker loci at one time rather than just two (for example, Lathrop *et al.*, 1984). Much work has already been carried out in humans and the mouse has been fairly well mapped. The high synteny between humans, mice and farm animals means that information already obtained from these genomes should be useful for farm animals. Measurements on traits of interest can be obtained on the same animals as the marker genotypes and these can be analysed to see if there is an association between any of the markers and the trait, if so these trait loci can also be mapped and the effect of such loci estimated. There have been several theoretical studies recently, looking at alternative methods of identifying linkage of a marker with a QTL and estimating the position and effect of the QTL, for example, Weller (1986), Luo and Kearsey (1989) and Lander and Botstein (1989). All of these authors suggest using ML methods or approximations to them. The first two methods are based on crosses between inbred lines, however, they might also be suitable for situations in animal breeding when crosses between different lines are considered, where the initial parents could be assumed to be homozygous for different alleles at the loci concerned. Lander and Botstein (1989) introduced interval mapping, where flanking markers are considered, which should improve the ability to detect QTLs.

The information and markers found from these methods can be used in a similar way to major genes found using segregation analysis, except that further information is available. For example, using the markers, it is now possible to follow the progression of the QTL and select only those individuals with the desired allele (or more precisely, select only those individuals with the relevant marker genotypes, which are associated with the quantitative trait of interest) (Soller and Beckmann, 1982). Using microsatellites enough DNA can be obtained from embryos to type them for their marker genotype. This could be used to distinguish superior animals from their sibs at birth, which, by classical methods would have the same predicted merit because they have the same ancestral information (Kashi *et al.*, 1990). It is also possible to introgress genes into another breed by following the markers and excluding all unwanted background effects while selecting the genotype required. In animals, work has mainly concentrated on selection within families. Being natural out breeders there is much variation between families and hence there is a problem of establishing the phase of the linkage between the marker and QTL, which will differ between families. Each generation, recombination can occurs between the marker and the locus of interest and there is a gradual breakdown of linkage disequilibrium. With tight linkage recombination between the loci of interest will occur less frequently and hence the phase of linkage does not have to be reassessed each generation.

A few studies using linked markers to detect QTLs have been carried out in animal populations. Paterson *et al.* (1988) used RFLPs to identify QTLs effecting three traits, which jointly determine the yield of tomato paste, using interval mapping (Lander and Botstein, 1989). They found evidence for regions of large effect on four to six chromosomes for each trait. However, these regions could be composed of a single gene or several linked genes effecting the same trait and further rounds of recombination would be necessary to distinguish these. Georges *et al.* (1990) used DNA fingerprints in cattle and found a candidate for a marker linked to the bovine muscular hypertrophy gene.

Segregation analysis requires only the pedigree and phenotypes for the individuals for the trait of interest. Making use of linked markers requires more work initially to create the genome map and then to use this information animals have to be typed for their marker genotype as well as for their phenotype. Also, the phase of linkage within each family might have to be confirmed subsequent generations. However, the use of linked markers to detect major genes should be more powerful, as there is more information available because use is made of mean differences between marker genotypes and not only the second and higher order statistics used by segregation analysis. This also means that the method should be less sensitive to the distribution of phenotypes and thus to departures from normality. At present, there have been a limited number of studies investigating the use of markers to detect genes in animal populations, especially when there are several genes effecting the trait or, as used in this study, a single gene with large effect segregating against a background of many small genes. Also there has been little work looking at the parameter estimates obtained and the accuracy of the predicted positions of the genes found. The use of markers has been more thoroughly investigated in plants, although here there is the advantage of easy manipulation to produce, for example, inbred lines. Some of the findings of plant breeders will be transferable to animals but in addition further work is required to examine the application of the methodology to animals in more detail.

7.3 CONCLUSIONS

To summarise the results obtained here, segregation analysis is capable of detecting a major gene and successfully estimating its effect and frequency in the population, even with the simple pedigree structure considered here. Approximations to the mixed model are required to make computation of the likelihood practicable. The use of classical animal breeding techniques, as used in ME3, allows the easy extension to include fixed effects, the estimation of breeding values and the addition of extra random effects. Using an estimate for the polygenic heritability, although if accurate gives a more powerful analysis, can give misleading results, suggesting spurious major genes to explain polygenic variation not explained by the fixed heritability. When the polygenic heritability was estimated, the genetic component of the resulting models was often composed only of a major gene, the polygenic heritability estimate being zero. In general, Hermite integration gives the best results and as computers improve the use of this method for more complex models would become feasible. However, as genetic marker maps are developed, and genotyping of individuals for markers becomes progressively easier, it is likely that segregation analysis methods based on the phenotype alone will be superseded by methods incorporating more information.

REFERENCES

Arendonk, J.A.M.van, Smith, C. and Kennedy, B.W. 1989. Method to estimate genotype probabilities at individual loci in farm livestock. *Theoretical and Applied Genetics* **78**: 735-740.

Bonney, G.E. 1984. On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *American Journal of Medical Genetics* **18**: 731-749.

Bonney, G.E. 1986. Regressive logistic models for familial disease and other binary traits. *Biometrics* **42**: 611-625.

Bulfield, G. 1985. The potential for improvement of commercial poultry by genetic engineering techniques. In: Hill, W.G., Manson, J.M., Dewitt, D. eds. *Poultry genetics and breeding.* British Poultry Science Ltd., Longman, Harlow, pp 37-46.

Cannings, C., Thompson, E.A. and Skolnick, M.H. 1976. The recursive derivation of likelihoods on complex pedigrees. *Advances in Applied Probability* **8**: 622-625.

Cannings, C., Thompson, E.A. and Skolnick, M.H. 1978. Probability functions on complex pedigrees. *Advances in Applied Probability* **10**: 26-61.

Castle, W.E. 1921. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. *Science* **54**: 223.

Cockerham, C.C. 1986. Modifications in estimating the number of genes for a quantitative character. *Genetics* **114**: 659-664.

Comstock, R.E. and Enfield, F.D. 1981. Gene number estimation when multiplicative genetic effects are assumed - growth in flour beetles and mice. *Theoretical and Applied Genetics* **59**: 373-379.

Demenais, F.M. and Bonney, G.E. 1989. Equivalence of the mixed and regressive models for genetic analysis. I. Continuous traits. *Genetic Epidemiology* **6**: 597-617.

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**: 523-542.

Eaves, L.J. 1983. Errors of inference in the detection of major gene effects on psychological test scores. *American Journal of Human Genetics* **35**:1179-1189.

Edwards, A.W.F. 1972. Likelihood. Cambridge University Press, London.

Elsen, J.M. and Le Roy, P. 1989. Simplified versions of segregation analysis for detection of major genes in animal breeding data. Paper presented at: 40th Annual Meeting of E.A.A.P., Dublin.

Elsen, J.M., Vu Tien Khang, J. and Le Roy, P. 1988. A statistical model for genotype determination at a major locus in a progeny test design. *Génétique, Sélection, Evolution* **20**: 211-226.

Elston, R.C. 1979. Major locus analysis for quantitative traits. *American Journal of Human Genetics* **31**: 655-661.

Elston, R.C. 1990. A general linkage method for the detection of major genes. In: D. Gianola and K. Hammond. eds. *Advances in Statistical Methods for Genetic Improvement of Livestock*. Springer-Verlag, Berlin pp 495-506.

Elston, R.C. and Stewart, J. 1971. A general model for the genetic analysis of pedigree data. *Human Heredity* **21**: 523-542.

Elston, R.C. and Stewart, J. 1973. The analysis of quantitative traits for simple genetic models from parental, F1 and backcross data. *Genetics* **73**: 695-711.

Everitt, B.S. 1981. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research* **16**: 171-180.

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.

Everitt, B.S. and Hand, D.J. 1981. Finite mixture distributions. Chapman and Hall, London.

Fain, P.R. 1978. Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. *Annals of Human Genetics* **42**: 109-120.

Falconer, D.S. 1981. Introduction to quantitative genetics, 2nd edition. Longman, London.

Famula, T.R. 1986. Identifying single genes of large effect in quantitative traits using best linear unbiased prediction. *Journal of Animal Science* **63**:68-76.

Felsenstein, J. 1973. Estimation of number of loci controlling variation in a quantitative character. *Genetics* **74** (supplement part 2): s78-s79.

Fisher, R.A., Immer, F.R. and Tedin, O. 1932. The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics* **17**: 107-124.

Foulley, J.L., Gianola, D. and Planchenault, D. 1987. Sire evaluation with uncertain paternity. *Génétique, Sélection, Evolution* **19**: 83-102.

Fraser, A. and Burnell, D. 1970. Computer models in genetics. McGraw-Hill, Inc., New York.

Geldermann, H. 1975. Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. *Theoretical and Applied Genetics* **46**: 319-330.

Georges, M., Lathrop, M., Hilbert, P., Marcotte, A., Schwers, A., Swillens, S., Vassart, G. and Hanset, R. 1990. On the use of DNA finger prints for linkage studies in cattle. *Genomics* **6**: 461-474.

Georges, M., Mishra, A., Sargeant, L., Steele, M. and Zhao, X. 1990. Progress towards a primary DNA marker map in cattle. In: *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production* **XIII**: 107-112.

Gibson, J.P., Jansen, G.B. and Rozzi, P. 1990. The use of κ-casein genotypes in dairy cattle breeding. In: *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production* **XIV**: 163-166.

Go, R.C.P., Elston, R.C. and Kaplan, E.B. 1978. Efficiency and robustness of pedigree segregation analysis. *American Journal of Human Genetics* **30**: 28-37.

Graser, H.-U., Smith, S.P. and Tier, B. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of Animal Science* **64**: 1362-1370.

Haley, C.S., Archibald, A., Andersson, L., Bosma, A.A., Davies, W., Fredholm, M., Geldermann, H., Groenen, M., Gustavsson, I., Ollivier, L., Tucker, E.M. and Van de Weghe, A. 1990. The pig gene mapping project - PiGMaP. In: *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production* **XIII**: 67-70.

Hammond, K. and James, J.W. 1970. Genes of large effect and the shape of the distribution of a quantitative character. *Australian Journal of Biological Science* ***23***: 867-876.

Hammond, K. and James, J.W. 1972. The use of higher degree statistics to estimate the number of loci which contribute to a quantitative character. *Heredity* **28**: 146-147.

Hanrahan, J.P. and Owen, J.B. 1985. Variation and repeatability of ovulation rate in Cambridge ewes. *Animal Production (Abstract)* **40**: 529.

Hanset, R. and Michaux, C. 1985a. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. I. Experimental data. *Génétique, Sélection, Evolution* **17**: 359-368.

Hanset, R., Michaux, C. 1985b. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. II. Population data. *Génétique, Sélection, Evolution* **17**: 369-386.

Harvey, W.R. 1977. User's guide for LSML76. The Ohio State University, Columbus.

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Society* **72**: 320-338.

Hasstedt, S.J. 1982. A mixed-model likelihood approximation on large pedigrees. *Computers and Biomedical Research* **15**: 295-307.

Hasstedt, S.J. 1989. Pedigree Analysis Package. Revision 3.0. Department of Human Genetics, University of Utah Medical Center,

Hasstedt, S.J. and Cartwright, P.E. 1981. Pedigree Analysis Package. Revision 2.0. Department of Medical Biophysics and Computing, University of Utah Medical Center, Technical Report No. 18.

Healy, P.J., Sewell, C.A., Nieper, R.E., Whittle, R.J. and Reichmann, K.G. 1987. Control of generalised glycogenosis in a Brahman herd. *Australian Veterinary Journal* **64**: 278-280.

Henderson, C.R. 1973. Sire evaluation and genetic trends. In: *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. J.L. Lush*, ADSA and ASAS. pp10-41.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**: 69-83.

Hildebrand, F.B. 1974. Introduction to Numerical Analysis. (International Series in Pure and Applied Mathematics). McGraw-Hill Inc., New York.

Hoeschele, I. 1988a. Statistical techniques for detection of major genes in animal breeding data. *Theoretical and Applied Genetics* **76**: 311-319.

Hoeschele, I. 1988b. Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theoretical and Applied Genetics* **76**: 81-92

Howell, J. McC., Dorling, P.R., Cook, R.D., Robinson, W.F., Bradley, S. and Gawthorne, J.M. 1981. Infantile and late onset form of generalised glycogenosis type II in cattle. *Journal of Pathology* **134**: 266-277.

Jeffreys, A.J.,Wilson, V. and Thein, S.L. 1985. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**: 67-73.

Jinks, J.L. and Towey, P. 1976. Estimating the number of genes in a polygenic system by genotype assay. *Heredity* **37**: 69-81.

Kammerer, C.M., MacCluer, J.W. and Bridges, J.M. 1984. An evaluation of three statistics of structured exploratory data analysis. *American Journal of Human Genetics* **36**: 187-196.

Karlin, S. and Williams, P.T. 1981. Structured exploration data analysis SEDA for determining mode of inheritance of quantitative traits. II. Simulation studies on the effect of ascertaining families through high-valued probands. *American Journal of Human Genetics* **33**: 282-292.

Karlin, S., Carmelli, D. and Williams, R. 1979. Index measures for assessing the mode of inheritance of continuously distributed traits. I. Theory and justifications. *Theoretical Population Biology* **16**: 81-106.

Karlin, S., Williams, P.T., Carmelli, D. 1981. Structured exploratory data analysis SEDA for determining mode of inheritance of quantitative traits. I Simulation studies on effect of background distributions. *American Journal of Human Genetics* **33**: 262-281.

Kashi, Y., Hallerman, E. and Soller, M. 1990. Marker-assisted selection of candidate bulls for progeny testing programmes. *Animal Production* **51**: 63-74.

Lalouel, J.M. 1979. GEMINI - - a computer program for optimization of a nonlinear function. Department of Medical Biophysics and Computing, University of Utah, Technical Report No. 14.

Lalouel, J.M. and Morton, N.E. 1981. Complex segregation analysis with pointers. *Human Heredity* **31**: 312-321.

Lalouel, J.M., Rao, D.C., Morton, N.E. and Elston, R.C. 1983. A unified model for complex segregation analysis. *American Journal of Human Genetics* **35**: 816-826.

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. *Genetics* **99**: 541-553.

Lander, E.S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185-199.

Lathrop, G.M., Lalouel, J.M., Julier, C. and Ott, J. 1984. Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Science USA* **81**: 3443-3446.

Le Roy, P., Elsen, J.M. and Knott, S. 1989. Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genetics, Selection, Evolution.* **21**: 341-357.

Luo, Z.W. and Kearsey, M.J. 1989. Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. *Heredity* **63**: 401-408.

MacCluer, J.W., Wagener, D.K. and Spielman, R.S. 1983. Genetic analysis workshop. I: Segregation analysis of simulated data. *American Joumal of Human Genetics* **35**: 784-792.

Mackay, T.F.C. 1985. Transposable element-induced response to artificial selection in *Drosophila melanogaster. Genetics* **111**: 351-374.

MacLean, C.J., Morton, N.E. and Lew, R. 1975. Analysis of family resemblance. IV. Operational characteristics of segregation analysis. *American Journal of Human Genetics* **27**: 365-384.

MacLean, C.J., Morton, N.E., Elston, R.C. and Yee, S. 1976. Skewness in commingled distributions. *Biometrics* 32:695-699.

Maki-Tanila, A. 1982. The validity of the heritability concept in quantitative genetics. Ph.D. Thesis, University of Edinburgh.

Mather, K. and Jinks, J.L. 1982. Biometrical genetics, 3rd edition. Chapman and Hall, London.

Matthysse, S., Lange, K. and Wagener, D.K. 1979. Continuous variation caused by genes with graduated effects. *Proceedings of the National Academy of Science USA* **76**: 2862-2865.

Mayo, O., Hancock, T.W. and Baghurst, P.A. 1980. Influence of major genes on variance within sibships for a quantitative trait. *Annals of Human Genetics* **43**: 419-421.

Mayo, O., Eckert, S.R. and Nugroho, W.H. 1983. Properties of the major gene index and related functions. *Human Heredity* **33**: 205-212.

McLachlan, G.J. and Basford, K.E. 1987. Mixture models: Inference and applications to clustering. Marcel Dekker, Inc., New York.

McPhee, C.P. and Reichmann, K.G. 1990. A genetic analysis of lysosomal enzyme activities in Brahman cattle. *Australian Journal of Agricultural Research* **41**: 205-211.

Merat, P. 1968. Distributions de frequences, interpretation du determinisime genetique des characteres quantitatifs et recherche de "genes majeurs". *Biometrics* **24**: 277-293.

Merat, P. and Ricard, F. H. 1974. Etude d'un gene de nanisme lie aus sexe chez la poule: importance de l'etat d'engraissement et gain de poids chez l'adulte. *Annales de Génétique et de Sélection Animale* **6**: 211-217.

Meyer, K. 1988. DFREML - A set of programs to estimate variance components under an individual animal model. *Journal of Dairy Science* **71** (supplement 2) : 33-34.

Meyer, K. 1989. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genetics, Selection, Evolution* **21**: 317-340.

Morton, N.E. and MacLean, C.J. 1974. Analysis of family resemblance. III. Complex segregation of quantitative traits. *American Journal of Human Genetics* **26**: 489-503.

Morton, N.E., Williams, W.R. and Lew, R. 1982. Trials of structured exploratory data analysis. *American Journal of Human Genetic* **34**: 489-500.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Kumlin, E. and White, R. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622.

Neimann-Sørensen, A. and Robertson, A. 1961. The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agriculturae Scandinavica* **11**: 163-196.

Numerical Algorithms Group 1988. The NAG Fortran Library Manual - Mark 13. NAG Ltd.

O'Donald, P. 1971. The distribution of genotypes produced by alleles segregating at a number of loci. *Heredity* **26**: 233-241.

Orkin, S.H. 1986. Reverse genetics and human disease. *Cell* **47**: 845-850.

Paterson, A.H., Lander, E.S., Hewitt, J.D., Peterson, S., Lincoln, S.E. and Tanksley, S.D. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**: 721-726.

Patterson, H.D. and Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545-554.

Pearson, K. 1904. Mathematical contributions to the theory of evolution. XII. On a generalised theory of alternative inheritance, with special reference to Mendel's laws. *Philosophical Transactions of the Royal Society of London* **A203**: 53-86.

Penrose, L.S. 1969. Effects of additive genes at many loci compared with those at a set of alleles at one locus in parent-child and sib correlations. *Annals of Human Genetics* **33**: 15-21.

Piper, L.R. and Bindon, B.M. 1982. Genetic segregation for fecundity in Booroola Merino sheep. In: Barton, R.A., Smith, W.C. eds. *Proceedings of the World Conference on Sheep and Beef Cattle Breeding.* Dunmore Press, Palmerston North, New Zealand. Vol 1. pp 395-400.

Piper, L.R., Bindon, B.M. and Davis, G.H. 1985. The single gene inheritance of the high litter size of the Booroola Merino. In: Land, R.B., Robinson, D.W. eds. *Genetics of reproduction in sheep.* Butterworths, London, pp 115-125.

Reichmann, K.G., Twist, J.O., McKenzie, R.A. and Rowan, K.J. 1987. Inhibition of bovine α-glucosidase by *Castanospermum australe* and its effect on the biochemical identification of heterozygotes for generalised glycogenosis type II (Pompe's disease) in cattle. *Australian Veterinary Journal* **64**: 274-276.

Roberts, R.C., Smith, C. 1982. Genes with large effects - theoretical aspects in livestock breeding. *Proceedings of the 2nd World Congress of Genetics Applied Livestock Production.* Garsi, Madrid **6**: 420-438.

Robertson, A. 1977. The non-linearity of offspring-parent regression. In: Pollak, E., Kempthorne, O. and Bailey, T.B.,Jr eds. *Proceedings of the International Conference on Quantitative Genetics.* Iowa State Univ Press, Ames, pp 297-304.

Rollins, W.C., Tanaka, M., Nott, C.F.G. and Thiessen, R.B. 1972. On the mode of inheritance of double-muscled conformation in bovines. *Hilgardia* **41**: 433-456.

Searle, S.R. 1971. Linear models. John Wiley and Sons, Inc., New York. p 418.

Searle, S.R. 1979. Notes on variance component estimation: A detailed account of maximum likelihood and kindred methodology. Paper BU-673M, Biometrics Unit, Cornell University, Ithaca, New York.

Selby, S.H. 1970. Handbook of tables for mathematics. 4th edition. The Chemical Rubber Company, Cleveland, Ohio, pp 894-895.

Shaw, R.G. 1987. Maximum likelihood approaches applied to quantitative genetics of natural populations. *Evolution* **41**: 812-826.

Shrimpton, A.E. 1981. The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster.* Ph.D. Thesis, University of Edinburgh.

Smith, C. 1967. Improvement of metric traits through specific genetic loci. *Animal Production* **9**: 349-358.

Smith, C. and Bampton, P.R. 1977. Inheritance of reaction to halothane anaesthesia in pigs. *Genetical Research* **29**: 287-292.

Smith, C. and Webb, A.J. 1981. Effects of major genes on animal breeding strategies. *Zeitschrift für Tierzüchtung und Züchtungsbiologie* **98**: 161-169.

Smith, S.P. and Graser, H.-U. 1986. Estimating variance components in a class of mixed models by restricted maximum likelihood. *Journal of Dairy Science* **69**: 1156-1165.

Soller, M. and Beckmann, J.S. 1982. Restriction fragment length polymorphisms and genetic improvement. *Proceedings of the 2nd World Congress of Genetics Applied Livestock Production.* Garsi, Madrid **6**:396-404.

Soller, M. and Beckmann, J.S. 1985. Restriction fragment length polymorphisms and animal genetic improvement. In: Leng, R.A., Barker, J.S.F., Adams, D.B. and Hutchinson, K.J. eds. *Reviews in rural science, 6. Biotechnology and recombinant DNA Technology in the Animal Production Industries* pp 10-18.

Thoday, J.M. 1961. Location of polygenes. *Nature* **191**: 368-370.

Thomas, A. 1986a. Approximate computation of probability for pedigree analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **3**: 157-166.

Thomas, A. 1986b. Optimal computation of probability functions for pedigree analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **3**: 167-178.

Thompson, J.N. and Thoday, J.M. 1979. Synthesis: polygenic variation in perspective. In: Thompson, J.N. and Thoday, J.M. eds. *Quantitative genetic variation.* Academic Press, New York, pp 295-301.

Titterington, D.M., Smith, A.F.M. and Makov. U.E. 1985. Statistical analysis of finite mixture distributions. John Wiley and Sons, Chichester.

Weber, J.L. and May, P.E. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**: 388-396.

Weller, J.I. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**: 627-640.

Wilks, S.S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**: 60-62.

Wolfe, J.H. 1971. A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Naval Personnel and Training Research Laboratory, Technical Bulletin, STB 72-2, San Diego.

Wright, S. 1952. The genetics of quantitative variability. In: Reeve, E.C.R., Waddington and C.H. eds. *Quantitative inheritance.* Her Majesty's Stationary Office, London, pp 5-41.

Published Paper.

© Elsevier/INRA

Original article

Comparison of four statistical methods for detection of a major gene in a progeny test design

P. Le Roy[1], J.M. Elsen[1] and S. Knott[2]

[1] *Institut National de la Recherche Agronomique, centre de recherches de Toulouse, station d'amélioration génétique des animaux, Auzeville 31326 Castanet-Tolosan Cedex, France*

[2] *AFRC, IAPGR, Edinburgh Research Station, Roslin, Midlothian, EH 25 9PS, UK*

(received 17 August 1988, accepted 23 January 1989)

Summary – In livestock improvement it is common to design a progeny test of sires in order to estimate their breeding values. The data recorded for these estimate are useful for the detection of major genes. They are the $n.m$ performances Y_{ij} of m progeny j of n sires i. These data need to be corrected for the polygenic influence of the sire on its progeny (sire i effect U_i). Four statistical tests of the segregation of a major gene are compared. The first (l_{SA} for "segregation analysis") is the classical ratio of the likelihoods under H_0 (no major gene) and H_1 (a major gene is segregating). The parameters describing the population (means and standard deviations within genotype) are estimated by maximizing the marginal likelihood of the Y_{ij}. The other statistics studied are approximations of this l_{SA} statistic where the sire i effect (U_i) is considered as a fixed effect (l_{FE} statistic) or, following Elsen *et al.* (1988) and Höschele (1988), where the parameters, and U_i, are estimated maximizing the joint likelihood of U_i and Y_{ij} (l_{ME1} and l_{ME2} statistics). Simulation studies were done in order to describe the distribution of these statistics. It is shown that l_{SA} and l_{ME1} are the most powerful test, followed by l_{ME2}, whose relative loss of power ranged between 20 and 40%, depending on the H_1 case studied, when 400 progeny are measured ($n = m = 20$). The segregation analysis, based on direct maximization of the likelihood, required 30 times more computation time than the l_{ME} test using an EM algorithm.

major gene – segregation analysis – statistical test

Résumé – Comparaison de quatre méthodes statistiques pour la détection d'un gène majeur dans un test sur descendance. *Il est fréquent, en sélection, de tester sur descendance, des mâles, afin d'estimer leur valeur génétique. Les données recueillies dans ce but peuvent être utilisées afin de mettre en évidence un gène majeur. Elles sont constituées des* n.m *performances* $\mathrm{Y_{ij}}$ *de* m *descendants* j *de* n *mâles* i. *Ces données doivent être corrigées pour l'effet polygénique du père (*$\mathrm{U_i}$*) sur ses descendants. Quatre tests statistiques de mise en évidence d'un tel gène majeur sont comparés. Le premier (*$\mathrm{l_{SA}}$ *pour "segregation analysis") est le rapport classique des vraisemblances sous* $\mathrm{H_0}$ *(pas de gène majeur) et sous* $\mathrm{H_1}$ *(existence d'un gène majeur). Les paramètres caractéristiques de la population (moyennes et écarts types intragénotype) sont estimés en maximisant la vraisemblance marginale des* $\mathrm{Y_{ij}}$. *Les autres statistiques de tests sont des approximations de* $\mathrm{l_{SA}}$ *pour lesquelles, soit l'effet pèrt* $\mathrm{U_i}$ *est considéré comme un effet fixé (test* $\mathrm{l_{FE}}$*) soit, comme proposé par Elsen* et al. *(1988) et Höschele (1988), les paramètres, et* $\mathrm{U_i}$, *sont obtenus en maximisant la vraisemblance conjointe des* $\mathrm{Y_{ij}}$ *et des* $\mathrm{U_i}$ *(test* $\mathrm{l_{ME1}}$

et l_{ME2}*). Nous avons réalisé des simulations afin de décrire les distributions de ces tests.* l_{SA} *et* l_{ME1} *sont les tests les plus puissants, suivi par* l_{ME2}*, dont la perte relative de puissance varie entre 20 et 40% selon l'hypothèse* H_1 *étudiées, quand 400 descendants sont mesurés* ($n = m = 20$)*. L'analyse de ségrégation, réalisée par maximisation directe de la vraisemblance, demande 30 fois plus de temps de calcul que les tests* l_{ME} *réalisés l'aide d'un algorithme EM.*

gène majeur – analyse de ségrégation – test statistique

INTRODUCTION

In recent years, several genes having major effects on commercial traits have been identified. The dwarf gene in poultry (Mérat & Ricard, 1974), the halothane sensitivity gene in pigs (Ollivier, 1980), the Booroola gene in sheep (Piper & Bindon, 1982), or the double muscling gene in cattle (Ménissier, 1982) are notable examples.

These discoveries, as well as improvement of transgenic techniques, have stimulated interest in new techniques for detection of single genes. Various tests have been described concerning livestock (Hanset, 1982). Their general principle is that the within family distribution of the trait depends on the parents' genotypes, and therefore varies from one family to another. These methods involve simple computations but are not powerful. Concurrently, segregation analysis in complex pedigrees was developed in human genetics (Elston & Stewart, 1971) by comparing the likelihoods of the data under different trait transmission models. These methods are much more powerful than the previous ones, but involve much computation. They require numerical simplification to deal with the population structure of farm animals. Additionally, the known properties of the test statistics, a likelihood ratio test, are only asymptotic, which raises the question of their validity when applied to samples of limited size.

In livestock improvement it is common to use progeny tests where males are mated to large numbers of females. Concentrating on this simple family structure the present paper tries to give some elements of a solution to the problems of simplification and validity. Four methods are compared on simulated data.

METHODS

The four methods considered rely upon the same information structure and the same type of test statistics.

Experimental design

The data are simulated according to a hierarchical and balanced family structure: one sample consists of n sire families ($i = 1, ...n$) with m mates per sire ($j = 1, ...m$) and one offspring per dam. Sires and dams are assumed to be unrelated. Only offspring are measured, with one Y_{ij} datum per animal.

Models and notations

Models

The Y_{ij} performances are considered under the two following models:

General hypothesis (H_1): "mixed inheritance"

In this model a monogenic component is added to the assumed polygenic variation.

When two alleles A and a are segregating at a major locus, three genotypes are possible (AA, Aa, aa) which we shall respectively denote 1, 2, 3. Sires are of genotype $s(s = 1, 2, 3)$ with probability P_s. Dams transmit to their offspring allele A with a probability q and allele a with a probability $1 - q$. Conditional on its genotype $t(t = 1, 2, 3)$, the ijth progeny has the performance Y_{ij}^t. The following linear model can be formulated.

$$Y_{ij}^t = \mu_t + U_i + E_{ij}$$

Where μ_t is the mean value of the performances of genotype t progeny.

U_i is the sire i random effect, assumed to be independent of the genotype t and normally distributed with a mean 0 and a variance σ_u^2.

E_{ij} is the residual random effect, assumed to be independent of the genotype t and normally distribued with a mean 0 and a variance σ_e^2.

U_i and E_{ij} are assumed to be independent.

Concerning production traits of livestock, the proportion of variance explained by polygenic effects has been generally estimated in many populations. Thus, we shall assume known *a priori* the heritability of the trait, h^2, defined as:

$$h^2 = 4\sigma_u^2 \mid (\sigma_u^2 + \sigma_e^2)$$

so that sires are assumed to be unselected.

The model thus defined on seven parameters:

$$\mu_1, \mu_2, \mu_3, \sigma_e, q, p_1, p_2 \ (p_3 = 1 - p_1 - p_2)$$

This hypothesis (H_0): "polygenic inheritance".

Null subhypothesis, to be tested against the general model, is fixed by $\mu_1 = \mu_2 = \mu_3 = \mu_0$:

$$Y_{ij} = \mu_0 + U_i + E_{ij}$$

Where μ_0 is the general mean of the performances. U_i and E_{ij} have the same definition as under H_1.

Matrix notation

Let S be the vector of the genotypes of the n males $\mathbf{S} = (S_1, \ldots, S_i, \ldots, S_n)$ and $s = (s_1, \ldots s_i, \ldots s_n)$ one realization of S.

$\mathbf{Y}_i$ be the vector of the m performances of the ith sire's progeny: $\mathbf{Y}_i = (Y_{il}, \ldots T_{ij}, \ldots Y_{im})$, and y_i the vector of realizations of $\mathbf{Y}_i$.

$\mathbf{T}_i$ the vector of order m of the genotypes at the major locus of the ith sire's progeny: $\mathbf{T}_i = (T_{i1}, \dots T_{ij}, \dots T_{im})$. Three realizations being possible for T_{ij}, 3^m different realizations t_i of $\mathbf{T}_i$ are possible. Prob $(\mathbf{T}_i = t_i|s_i)$ is the probability of the realization of the genotypes vector $t_i = (t_{il}, \dots t_{ij}, \dots t_{im})$ when sire i is of genotype s_i.

$\boldsymbol{\mu}$ the vector of genotype means:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

Given $\mathbf{E}_i$, the vector of order m of residuals, the vector $\mathbf{Y}_i$ can be written under H_0:

$$\mathbf{Y}_i = \mathbf{X}.\mu_0 + \mathbf{Z}.U_i + \mathbf{E}_i$$

where $\mathbf{X}$ and $\mathbf{Z}$ are two matrices of order $m \times 1$, whose elements all equal 1, under H_1:

$$\mathbf{Y}_i = \mathbf{X}_{iti}.\boldsymbol{\mu} + \mathbf{Z}.U_i + \mathbf{E}_i$$

where $\mathbf{X}_{iti}$ is the $m \times 3$ incidence matrix for the fixed effects of the model, when the realization of the genotypes of the sire i progeny is $\mathbf{t}_i$.

The $\mathbf{V}_i$ covariance matrix for the performances Y_{ij} of the sire i family is:

$$\mathbf{V}_i = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 + \sigma_e^2 \end{pmatrix} = \mathbf{V}{=}\mathbf{Z}.\mathbf{D}.\mathbf{Z}'{+}\mathbf{R}$$

with $\mathbf{D} = \sigma_u^2$ and $\mathbf{R}$ the diagonal $m \times m$ matrix $\mathbf{R}= \sigma_e^2 \cdot \mathbf{I}_m$.

General expression of the likelihood ratio test (LR test)

The test statistic is based on the ratio of the likelihoods under $H_0(M_0)$ and under $H_1(M_1)$, or an estimate of this ratio. In practice the test statistic considered is: $l = -2.\log\ (M_0/M_1)$. With our notation, and given the preceding hypothesis, M_0 is:

$$M_0 = \prod_{i=1}^{n} f_0(y_i)$$

with

$$f_0(y_i) = \frac{1}{\sqrt{2\pi^m|\mathbf{V}|}} \exp\ (-\frac{1}{2}\ (y_i - \mathbf{X}_{\mu 0})'\ \mathbf{V}^{-1}\ (y_i - \ \mathbf{X}_{\mu 0}))$$

and M_1 is:

$$M_1 = \prod_{i=1}^{n} f_1(y_i)$$

with

$$f_1(y_i) = \sum_{s_i=1}^{3} p_{s_i} \sum_{t_i} Prob(\mathbf{T}_i = t_i|s_i) \; \frac{1}{\sqrt{2\pi^m|\mathbf{V}|}}$$

$$exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu}))$$

The four proposed methods are all based on the two following equalities:

$$\frac{1}{\sqrt{2\pi^m|\mathbf{V}|}} \; exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - X_{it_i}\boldsymbol{\mu})) =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_u^2}} \; exp\left(-\frac{1}{2}\left(\frac{u_i}{\sigma_u}\right)^2\right) \; \frac{1}{\sqrt{2\pi^m|\mathbf{R}|}}$$

$$exp(-\frac{1}{2} \; (y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}u_i)'\mathbf{R}^{-1} \; (y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}u_i))du_i \qquad (1)$$

and:

$$exp(-\frac{1}{2} \; (y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})) =$$

$$exp\left(-\frac{1}{2}\left(\frac{\widehat{u}_i}{\sigma_u}\right)^2\right) \; exp(-\frac{1}{2} \; (y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}\widehat{u}_i)'\mathbf{R}^{-1} \; (y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}\widehat{u}_i)) \qquad (2)$$

Where $\widehat{u}_i$ is the mode of the distribution of U_i given $\mathbf{Y}_i$ and the genotypes $\mathbf{t}_i$. Formula (2) results from the equality of mode and expectation for symetrical distributions.

Definition and interests of the four proposed methods

The differences between the four methods concern the sire effects.

First method: SA

In the SA method ("segregation analysis", Elston 1980), we consider without simplification the model and the test statistic as they were defined above. The likelihoods under H_1 and H_0 are calculated using equality (1) and taking account of:

$$Prob(\mathbf{T}_i = t_i \mid s_i) = \prod_{j=1}^{m} Prob(T_{ij} = t_{ij} \mid s_i)$$

Then:

$$M_{1SA} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \int_{-\infty}^{\infty} k(u_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \cdot k_{t_{ij}}(y_{ij} \mid u_i)du_i$$

with:

$$k(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\, exp\left(-\frac{1}{2}\left(\frac{u_i}{\sigma_u}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\, h(u_i)$$

$$k_{t_{ij}}(y_{ij} \mid u_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_{t_{ij}} - u_i}{\sigma_e}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, h_{t_{ij}}(y_{ij} \mid u_i)$$

and:

$$M_{0\mathrm{SA}} = \prod_{i=1}^{n} \int_{-\infty}^{\infty} k(u_i) \prod_{j=1}^{m} k_0(y_{ij} \mid u_i)\, du_i$$

with:

$$k_0(y_{ij} \mid u_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_0 - u_i}{\sigma_e}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, h_0(y_{ij} \mid u_i)$$

The well known asymptotic properties of the LR test under H_0 are the main advantage of this method. If some regularity conditions hold, the test statistic l is asymptotically distributed according to a central χ^2 with d degrees of freedom, d being the number of parameters with fixed value under H_0 (Wilks, 1938). However, in the particular context of testing a number of components in a mixture, the regularity conditions are not satisfied since the mixing proportions p_1 and p_2 have the value zero under H_0, which defines the boundary of the parameter space.

Studying mixtures of m-normal distributions, Wolfe (1971) suggested that the distribution of the LR test is proportional to a χ^2 distribution with $2d$ degrees of freedom. The proportionality coefficient c should be $c = (n-1-m-1/2g_2)/n$ where n represents the sample size, and g_2 the number of components in the mixture under H_1. If these results hold in our case, when the number or sires is very large, l_{SA} should have a χ^2 distribution with 4 degrees of freedom.

The problem with this method is that it requires heavy computation: a complex function of the Y_{ij} must be integrated n times for each estimation of l_{SA}.

Second and third methods: ME

These methods ("modal estimation" of the sire effect U_i), use the equation (2).

Under H_0, the likelihood may be written as follows:

$$M_{0\mathrm{ME1}} = \frac{1}{\sqrt{2\pi^m \mid \mathbf{V} \mid}} \prod_{i=1}^{n} h(\widehat{u}_i) \prod_{j=1}^{m} h_0(y_{ij} \mid \widehat{u}_i) \tag{3}$$

Under H_1, the equality (2) leads to

$$M_{1\mathrm{ME1}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} P_{s_i} \sum_{t_i} Prob(\mathbf{T}_i = t_i \mid s_i)\ .$$

$$\frac{1}{\sqrt{2\pi^m |\mathbf{V}|}}\, h(\widehat{u}_i) \prod_{j=1}^{m} \cdot h_{t_{ij}}(y_{ij} \mid \widehat{u}_{it_i})$$

However. the sums over the vectors $\mathbf{t}_i$ for each sire make this computation practically impossible as soon as m is larger than a few units ($3^5 = 243$, $3^{10} = 59\,049$).

Thus, following Elsen *et al.* (1988) we suggest the approximation

$$M_{1\text{ME}1} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \frac{1}{\sqrt{2\pi^m |\mathbf{V}|}} h(\widehat{u}_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \, . \, h_{t_{ij}}(y_{ij} \mid \widehat{u}_i) \tag{4}$$

Where $\widehat{u}_i$ is the distribution mode of U_i conditional on Y_i, whatever the genotypes s_i and $\mathbf{t}_i$ are. The statistic $l_{\text{ME}1} = -2\log(M_{0\text{ME}1}/M_{1\text{ME}1})$ is no longer an LR test but an approximation lacking the asymptotic properties described above. However we hope that this statistic which requires much less computation will nonetheless retain the power of the first proposed.

An alternative to this second method is to estimate the likelihood $M_{0\text{SA}}$ and $M_{1\text{SA}}$ directly by:

$$M_{0\text{ME}2} = \prod_{i=1}^{n} k(\widehat{u}_i) \prod_{j=1}^{m} k_0(y_{ij} \mid \widehat{u}_i) \tag{5}$$

$$M_{1\text{ME}2} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} k(\widehat{u}_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \, . \, k_{t_{ij}}(y_{ij} \mid \widehat{u}_i) \tag{6}$$

where $\widehat{u}_i$ is defined as above.

As stated by Höschele (1988) this "approximation will be close to l_{SA} only if the likelihood is very peaked ($m \to \infty$) with most of its probability mass concentrated over a small region about the ML estimates".

Fourth method: FE

The method (fixed effect of the sires), does not consider the *a priori* information contained in the heritability of the trait. The u_i sire effects are assumed to be fixed, and become supplementary parameters which need to be estimated. The likelihood ratio may be written:

$$I_{\text{FE}} = -2\log\frac{M_{0\text{FE}}}{M_{1\text{FE}}}$$

with:

$$M_{0\text{FE}} = \prod_{i=1}^{n} \prod_{j=1}^{m} k_0(y_{ij} \mid u_i)$$

and:

$$M_{1\text{FE}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \, . \, k_{t_{ij}}(y_{ij} \mid u_i)$$

This method has the advantage of its computational simplicity, while retaining the well known asymptotic properties of the LR test. However, there may be an important loss of power, due to the loss of information on the polygenic variation.

The comparisons

Three problems were studied:

Distributions of the statistics under H_0

We have just mentioned uncertainties concerning the asymptotic distributions (χ^2 with 4 degrees of freedom for l_{SA} and l_{FE} if Wolfe's (1971) approximation is valid, no known property for l_{ME}). Furthermore these distributions are unknown in samples of limited size. In order to estimate these distributions, samples were simulated under H_0 (500 samples for SA, 1 000 for FE and ME) with different numbers of sires ($n = 5$, 10, 20) and of progeny per sire ($m = 5$, 10, 20). The test statistics l_{SA}, l_{ME1}, l_{ME2} and l_{FE} were calculated for each sample. The estimated distributions obtained were used to test the convergences to χ^2 distributions. They also helped determine boundaries for critical regions in samples of a limited size. We used the Harrel and Davis (1982) method to estimate quantiles at 5 and 1% and their jackknife variance as defined by Miller (1974). These simulations were based on a heritability of 0.2.

Comparisons of the powers

By using the table of the critical regions thus obtained for each family structure, we have been able to compare the powers of the tests. These powers depend not only on the number and size of the families in the sample but also on the values of the parameters (μ, σ_e, p_1, p_2, q) which characterize the major gene segregating in the population.

For each of the 9 family structures described above, three H_1 hypotheses were considered, each with a simulation of 100 samples. All these populations are assumed to follow the Hardy Weinberg law. The differences between the three H_1 hypotheses lie in the mean effects of the genotypes (expressed in standard deviation units) and the frequency of the allele A.

Case 1: complete dominance and equal allele frequencies

$$\mu_1 = \mu_2 = 0, \ \mu_3 = 2 \text{ and } q = 0.5.$$

Case 2: additivity, equal allele frequencies

$$\mu_1 = 0, \ \mu_2 = 1, \ \mu_3 = 2, \text{ and } q = 0.5$$

Case 3: Complete dominance, recessive allele rare

$$\mu_1 = \mu_2 = 0, \ \mu_3 = 2 \text{ and } q = 0.9$$

The power of the tests was measured by the percentage of H_0 rejection.

Algorithms and cost of calculations

The methods must also be compared on the basis of how much computation they require. The calculations described above were made using the quadrature and optimization subroutines of the NAG fortran library. In order to maximize the likelihoods of the sample we used a Quasi-Newton algorithm in which the derivatives are estimated by finite differences.

The same algorithm was used for the four methods, giving results of a similar degree of precision. However, various algorithms can be used to estimate the maximum likelihood of the parameters. In the ME and FE tests, the first derivatives have a simple algebraic form and the maximum likelihood solutions are reached by zeroing the first derivatives (with respect to each of the parameters) of the logarithm of the likelihood. Under H_1 the corresponding system of equations can be solved iteratively, but not directly, by using for instance the EM algorithm defined by Dempster *et al.* (1977): see appendix.

This is the algorithm we used for the ME2 test in order to obtain more extensive information on critical region: 5, 10, 20, and 40 sires, 5, 10, 20 and 40 progenies/sire, heritability of 0, 0.2, 0.4.

RESULTS AND DISCUSSION

Comparison of the four methods

Tables I to IV show the main characteristics of the distributions of the 4 test statistics: mean, standard deviation, 5% and 1% empirical quantiles and percentage of replicates beyond the 5% and 1% quantiles of a χ^2_4. Table V shows their powers.

First, we can note that for the number of progeny increases, the mean distributions as the four test statistics decrease (except l_{SA} between $m = 5$ and $m = 10$ for $n = 5$).

The fact that l statistics distributions converge toward a χ^2 with 4 degrees of freedom cannot be confirmed since all the distributions of l, but one (segregation analysis with 5 sires and 5 progenies/sire), are significantly different from a χ^2 using a χ^2 test of fit. Moreover, the scaled statistics $(2E(l)/\text{var}\ (l)).\ l$ are also significantly different from a χ^2. It must be emphasized that the samples studied are far from the conditions of validity of Wolfe's approximation which requires that $n > 10.m$ (Everitt, 1981). The l_{SA} statistics show a notable stability as the family size varies, whereas for l_{FE} the statistics only reaches an asymptote as m, the number of progeny per sire increases. As regards the l_{ME} statistics, the results are totally different.

The mean and standard deviation of the l_{ME1} statistic decreases when the number of sires or progeny per sire increases. It appeared that the distribution of this l_{ME1} statistic becomes very peaked near zero. It must be noticed that this pattern is close to the asymptotic distribution of the LR test of a mixture of 2 known distributions in unknown proportion studied by Titterington *et al.* (1985). These authors found that, under H_0 (only one component) the LR test "is 0 with a probability 0.5 and, with the same probability, is distributed as a χ^2 with one degree of freedom". On the other hand, for a given number of progeny, the mean of the l_{ME2} distribution increases with the number of sires. The fewer the progeny, the greater the increase.

The calculation of the power (Table V) shows some important facts: very low power of the four statistics for low nomber of sires and/or progeny, clear superiority of the segregation analysis and first of the modal estimation method whatever

Table I. Results of the simulations under H_0 for the l_{SA} statistic: means (μ), standard deviations (σ), 5% (s_5) and 1% (s_1) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% (r_5) and 1% (r_1) quantiles of a χ^2_4.

Number of							
sires *(n)*	*progeny* *(m)*	μ	σ	s_5	s_1	r_5	r_1
5	5	4.42	2.92	9.74 (0.44)	14.91 (1.87)	5.71	0.98
	10	4.48	3.00	9.99 (0.33)	14.30 (0.69)	6.04	1.59
	20	4.44	3.17	10.64 (0.45)	14.36 (0.49)	7.59	1.86
10	5	4.71	3.16	10.86 (0.19)	14.69 (0.27)	8.36	1.91
	10	4.47	3.20	10.50 (0.32)	14.26 (0.61)	7.62	1.66
	20	4.36	3.15	10.50 (0.46)	14.31 (1.16)	7.39	1.14
20	5	4.74	3.36	11.10 (0.27)	15.51 (0.74)	8.87	1.94
	10	4.42	3.25	10.75 (0.55)	15.15 (0.83)	7.38	1.75
	20	4.14	3.45	11.25 (0.51)	15.65 (1.21)	7.89	2.17

Table II. Results of the simulations under H_0 for the l_{MEi} statistic: means (μ), standard deviations (σ), 5% (s_5) and 1% (s_1) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% (r_5) and 1% (r_1) quantiles of a χ^2_4.

Number of							
sires *(n)*	*progeny* *(m)*	μ	σ	s_5	s_1	r_5	r_1
5	5	4.61	3.52	11.01 (0.39)	15.96 (0.83)	8.8	2.4
	10	3.65	3.09	9.51 (0.29)	14.54 (0.74)	4.8	1.4
	20	2.92	2.95	8.47 (0.39)	12.86 (0.74)	3.5	0.8
10	5	3.83	3.20	10.31 (0.20)	15.69 (1.20)	7.6	1.5
	10	2.77	2.94	8.71 (0.25)	12.59 (0.58)	3.6	0.7
	20	2.10	2.63	7.36 (0.32)	12.80 (1.39)	1.9	1.1
20	5	2.75	3.04	8.55 (0.31)	14.14 (1.09)	3.2	1.2
	10	1.81	2.52	6.97 (0.29)	11.14 (0.80)	1.7	0.4
	20	1.27	2.11	5.82 (0.34)	9.46 (0.46)	0.9	0.0

Table III. Results of the simulations under H_0 for the l_{ME2} statistic: means (μ), standard deviations (σ), 5% (s_5) and 1% (s_1) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% (r_5) and 1% (r_1) quantiles of a χ^2_4.

Number of		μ	σ	s_5	s_1	r_5	r_1
sires (n)	*progeny (m)*						
5	5	12.28	4.41	20.27 (0.13)	25.03 (0.32)	71.4	35.5
	10	9.71	4.09	16.90 (0.20)	21.69 (0.59)	48.5	17.5
	20	7.60	4.14	15.61 (0.30)	19.79 (0.36)	27.9	10.1
10	5	17.28	5.27	26.81 (0.32)	32.90 (0.54)	95.4	77.2
	10	13.52	5.11	22.94 (0.25)	27.36 (0.51)	78.0	48.0
	20	9.36	4.85	18.54 (0.31)	23.51 (0.49)	43.8	19.0
20	5	26.47	6.66	38.24 (0.41)	44.63 (0.78)	99.9	99.1
	10	19.56	6.49	31.15 (0.36)	37.21 (0.86)	96.7	82.9
	20	12.17	5.94	23.02 (0.41)	29.30 (0.58)	63.4	36.8

Table IV. Results of the simulations under H_0 for the l_{FE} statistic: means (μ), standard deviations (σ), 5% (s_5) and 1% (s_1) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% (r_5) and 1% (r_1) quantiles of a χ^2_4.

Number of		μ	σ	s_5	s_1	r_5	r_1
sires (n)	*progeny (m)*						
5	5	9.62	5.13	18.93 (0.19)	24.42 (0.47)	46.0	21.3
	10	6.26	4.28	14.30 (0.23)	19.13 (0.54)	20.5	6.87
	20	4.64	3.86	12.33 (0.28)	16.72 (0.53)	11.3	3.85
10	5	12.27	6.68	24.03 (0.30)	30.89 (1.27)	63.2	39.6
	10	7.19	5.40	17.31 (0.31)	22.41 (0.59)	30.3	13.9
	20	4.22	3.99	12.17 (0.30)	17.68 (0.89)	10.5	3.50
20	5	16.20	9.61	32.79 (0.47)	42.90 (1.22)	73.2	59.8
	10	7.86	6.82	20.72 (0.25)	28.55 (1.00)	34.8	20.6
	20	3.69	3.84	11.35 (0.36)	16.29 (0.58)	9.15	2.81

these numbers, with respectively a 90% and a 80% power in the best case (though involving only 400 animals), very poor performance of the l_{FE} statistic, intermediate power for l_{ME2}.

Thus knowledge of heritability is a substantial advantage and gives a reason to prefer the l_{ME} statistics against the l_{FE}, which requires similar amounts of computation.

Table V. Results of the simulations under H_1 powers of the 4 tests for a 5% first type error (percentage of H_0 rejection) and their 5% confidence intervals between brackets. Comparisons of different family structures and parameters values.

Number of		*CASE 1*				*CASE 2*				*CASE 3*			
sires	*progeny*												
(n)	*(m)*	*SA*	*ME1*	*ME2*	*FE*	*SA*	*ME1*	*ME2*	*FE*	*SA*	*ME1*	*ME2*	*FE*
	5	9 (6-14)	5 (2-11)	7 (4-11)	2 (1-5)	3 (1-8)	6 (3-12)	3 (1-8)	3 (1-9)	9 (4-18)	5 (2-11)	7 (3-15)	4 (1-12)
5	10	17 (12-24)	14 (9-22)	9 (5-16)	8 (4-15)	11 (6-19)	5 (2-11)	8 (4-15)	3 (1-9)	7 (3-15)	5 (2-11)	6 (3-13)	6 (2-14)
	20	24 (17-39)	30 (22-40)	18 (12-27)	16 (10-25)	9 (5-16)	9 (5-16)	8 (4-15)	1 (0-6)	19 (12-29)	21 (14-30)	14 (8-23)	7 (3-15)
	5	17 (12-24)	14 (9-22)	12 (7-20)	7 (3-14)	7 (3-14)	6 (3-12)	7 (3-14)	2 (1-8)	5 (2-12)	8 (4-15)	3 (1-9)	2 (0-9)
10	10	27 (21-23)	41 (32-51)	16 (11-21)	9 (5-14)	8 (4-16)	11 (6-19)	3 (1-9)	1 (0-6)	19 (12-28)	18 (12-27)	7 (3-14)	1 (0-7)
	20	54 (45-63)	60 (50-69)	38 (29-48)	24 (17-33)	15 (10-22)	16 (10-24)	5 (2-10)	1 (0-5)	34 (25-43)	30 (22-40)	16 (10-25)	11 (6-20)
	5	26 (19-34)	27 (19-36)	18 (12-26)	7 (3-13)	9 (5-16)	11 (6-19)	6 (3-13)	2 (1-6)	15 (10-22)	18 (12-27)	4 (2-9)	1 (0-6)
20	10	51 (42-61)	48 (38-58)	27 (19-36)	7 (3-13)	21 (14-30)	18 (12-27)	7 (4-14)	7 (3-14)	33 (25-42)	37 (28-47)	13 (8-20)	8 (4-16)
	20	90 (83-94)	80 (71-87)	72 (62-80)	56 (46-65)	25 (19-33)	21 (14-30)	15 (9-24)	9 (5-16)	48 (38-57)	62 (52-71)	34 (26-43)	31 (23-41)

The comparison of powers in hypothesis H_1 is also interesting: it is much more difficult to detect an additive major gene (case 2) than a dominant one (case 1) even with the segregation analysis which is 3 to 4 times less powerful in case 2 than in case 1. In comparison with the isofrequent case, the third case shows a 50% loss of power: with measurements made on a small population, very few individuals if any, belong to the high mean distribution.

The computation requirements have been estimated, on a 3083 IBM computer, by the CPU time needed for the evaluation of the statistics under H_0. Ten replicates of a sample of 10 sires and 10 progenies per sire used 640 s for the l_{SA} statistic, 142 s for the l_{FE} statistic and 48 s for the l_{ME} statistics. Using the EM algorithm instead of the direct maximization of l_{ME} with the NAG subroutines decreases the time requirements to 20 s only. Thus, the proposed simplified tests l_{ME} are 30 times as fast as the segregation analysis.

Tables of quantiles

Although theoretical works are still needed in order to describe the asymptotic behaviour of the l_{SA}, l_{ME1} and l_{FE} tests, one can use, as a first approach, the quantiles given in our tables for larger populations since this will produce an overestimation of the first type error. On the contrary, some more calculations are needed for the l_{ME2} test.

The 5 and 1% points for this statistic are given in figures 1 to 3 depending on the heritability (0.0, 0.2, 0.4). Each figure gives these points for varying numbers of sires and progeny per sire.

Note that when the heritability is 0., the sire effect is not defined and, thus, that the $u_i[a+1]$ terms disappear from the equations given in the appendix.

The results of Table III are confirmed: the quantile estimates increase with the number of sires n (for a given number of progeny per sire, m) and decrease when the number of progeny per sire increases. Two other results must be noticed:

- given n and m, the lower the heritability, the greater the quantiles.

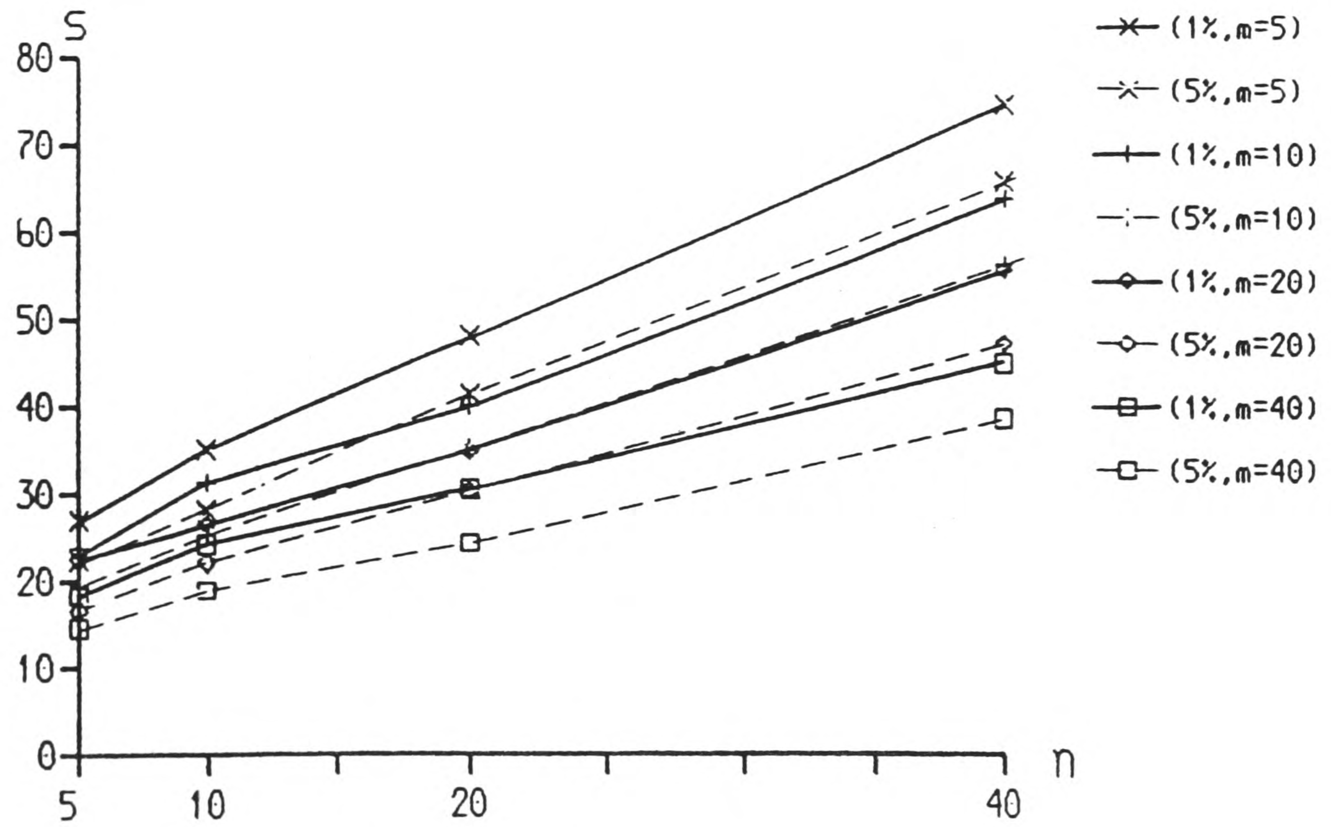


Fig. 1. 5% and 1% quantiles of the l_{ME2} test statistic for varying family structures ($h^2 = 0$).

- on the variation range studied for m, the number of progeny per sire, the increase of the quantiles is nearly linear with n (number of sires) allowing some extrapolations for higher values of this number.

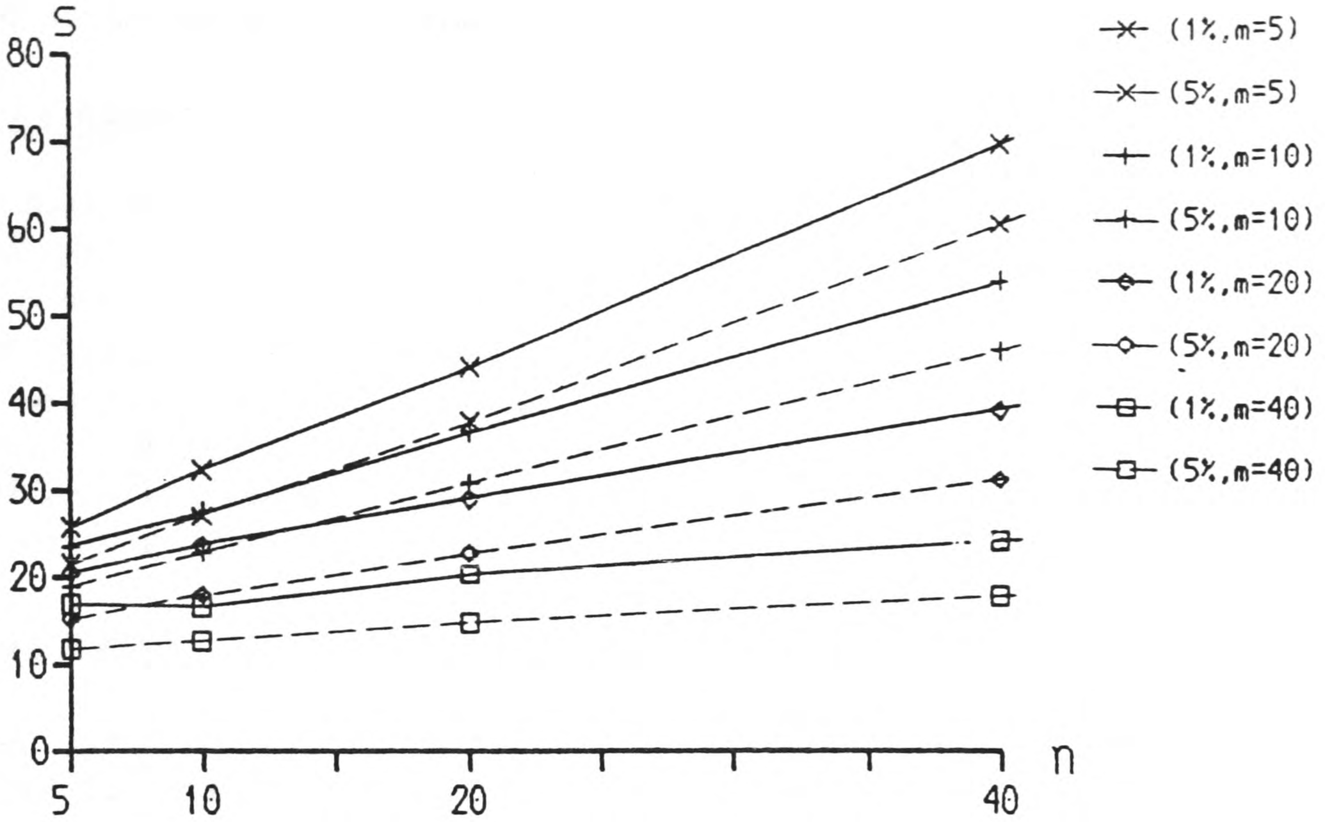


Fig. 2. 5% and 1% quantiles of the l_{ME2} test statistic for varying family structures ($h^2 = 0.2$).

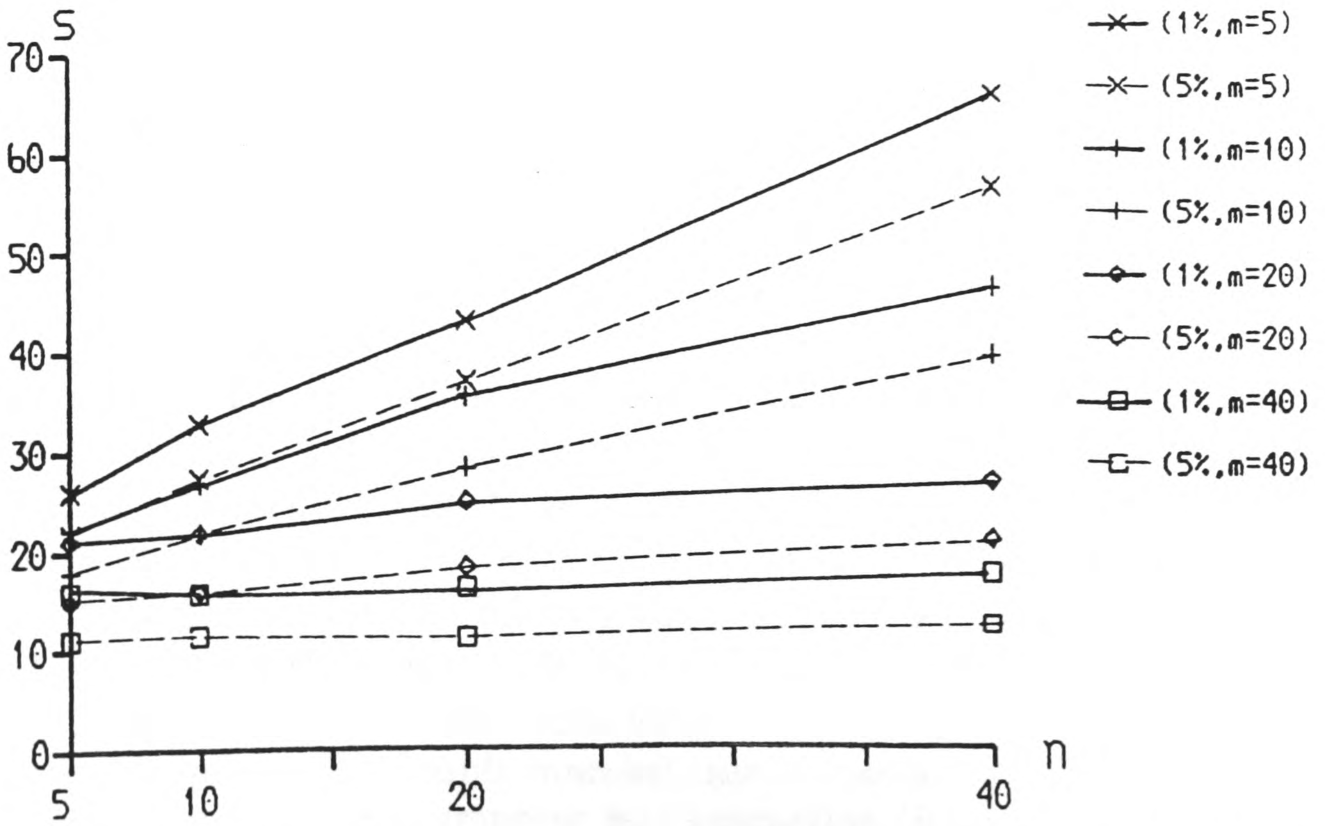


Fig. 3. 5% and 1% quantiles of the l_{ME2} test statistic for varying family structures ($h^2 = 0.4$).

Finally, the jackknife standard deviation of the estimated quantile varies, for the 5% case, between 0.23 and 0.89, with a mean value of 0.52 and, for the 1% case, between 0.39 and 1.65 with a mean value of 0.92. These errors could explain the observed deviations of the plotted curves from smoothness.

CONCLUSIONS

On the four statistical tests studied, the "segregation analysis" method is, as expected, the most powerful. Applied on a large scale, this test requires a great deal for computation. The "modal effect" method requires much less computation than the segregation analysis and shows practically no loss of power for the first version and a limited loss of power (diminishing as soon as the sample size is sufficient) for the second version. Unfortunately, the asymptotic distribution of this last statistic is unknown. The tables of quantiles we obtained by simulation permit the utilization of this test for typical sample sizes and for various heritability values.

REFERENCES

Dempster A.P., Laird N.M. & Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc., Series B* 39, 1-38

Elsen J.M., Vu Tien Khang J. & Le Roy P. (1988) A statistical model for genotype determination at a major locus in a progeny test design. *Genet. Sel. Evol.* 20, 211-226

Elston R.C. (1980) Segregation analysis. *In: Current developments in anthropological genetics* (Mielke J.H. & Crawford M.H. eds), 1, Plenum Publishing Corporation, New York, 327-354

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

Everitt B.S. (1981) A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivar. Behav. Res.* 16, 171-180

Hanset R. (1982) Major genes in animal production, examples and perspectives: cattle and pigs. *2nd world congress on genetics applied to livestock production, Madrid, 4-8 oct., 1982*, 5, Editorial Garsi, Madrid, 439-453

Harrel F.E. & Davis C.E. (1982) A new distribution-free quantile estimator. *Biometrika* 69, 635-640

Höschele I. (1988) Statistical techniques for detection of major genes in animal breeding data. *Theor. Appl. Genet.* 76, 311-319

Ménissier F. (1982) Present state of knowledge about the genetic determination of muscular hypertrophy or the double-muscled trait in cattle. *In: Muscle hypertrophy of genetic origin and its use to improve beef production* (King J.W.B. & Ménissier F. eds), Martinus Nijhof, The Hague, 387-428

Mérat P. & Ricard F.H. (1974) Etude d'un gène de nanisme lié au sexe chez la poule: importance de l'état d'engraissement et gain de poids chez l'adulte. *Ann. Génét. Sél. Anim.* 6, 211-217

Miller R.G. (1974) The Jackknife. A review, *Biometrika* 61, 1-15

Ollivier L. (1980) Le déterminisme génétique de l'hypertrophie musculaire chez le porc. *Ann. Génét. Sél. Anim.* 12, 383-394

Piper L.R. & Bindon B.M. (1982) The *Booroola Merino* and the performance of medium *non-peppin* crosses at Armidale. *In: The* Booroola Marino, (Piper L.R., Bindon B.M. & Nethery R.D. eds), CSIRO, Melbourne, 9-20

Titterington D.M., Smith A.F.M. & Makow U.E. (1985) *Statistical analysis of finite mixture distributions.* Wiley, New York

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

Wolfe J.H. (1971) A Monte Carlo study of the sampling distribution of the likelihood ratio for mixture of multinormal distributions. *Tech. Bull.*, STB 72-2, Naval Personnel and Training Research Laboratory, San Diego

APPENDIX

Application of the EM algorithm to the estimation of the test statistic l_{ME} under H_1

The EM algorithm is an iterative procedure. Each of its iterations consists of two steps E (Expectation) and M (Maximization). In our calculations we have considered that convergence is obtained when, a being the iteration number, the following inequality is satisfied:

$$|l_{\mathrm{ME}}[a+1] - l_{\mathrm{ME}}[a]| < 10^{-6}|l_{\mathrm{ME}}[a]|$$

Step E of the ath iteration consists of estimating posterior probabilities of the observations

$$q_i(s_i)[a+1] = Prob(S_i = s_i \mid \mathbf{Y}_i,\ u_i[a])$$

$$q_{ij}(t_{ij} \mid s_i)[a+1] = Prob(T_{ij} = t_{ij} \mid S_i = s_i,\ \mathbf{Y}_i,\ u_i[a])$$

$$q_{ij}(t_{ij})[a+1] = Prob(T_{ij} = t_{ij} \mid \mathbf{Y}_i,\ u_i[a])$$

These probabilities are estimated using the ath iteration values of $\sigma_e[a]$, $q[a]$, $u_i[a]$ $(i = 1, ..., n)$, $\mu_t[a]$ $(t = 1, 2, 3)$ and $p_s[a]$ $(s = 1, 2, 3)$. The following quantities are calculated successively:

$$k_{t_{ij}}(y_{ij} \mid u_i)[a+1] = \frac{1}{\sqrt{2\pi}\,\sigma_e[a]}\ exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_{t_{ij}}[a] - u_i[a]}{\sigma_2[a]}\right)^2\right)$$

$$q_{ij}(t_{ij} \mid s_i)[a+1] = \frac{Prob(T_{ij} = t_{ij} \mid s_i)k_{t_{ij}}(y_{ij} \mid u_i[a+1])}{\Sigma_{t'_{ij}} Prob(T_{ij} = t'_{ij} \mid s_i)k_{t'_{ij}}(y_{ij} \mid u_i[a+1])}$$

$$q_i(s_i)[a+1] = \frac{p_{s_i}[a]\Pi_j(\Sigma_{t_{ij}} Prob(T_{ij} = t_{ij} \mid s_i)k_{t_{ij}}(y_{ij} \mid u_i[a+1]))}{\Sigma_{s'_i} p_{s'_i}[a]\Pi_j(\Sigma_{t_{ij}} Prob(T_{ij} = t_{ij} \mid s'_i)k_{t_{ij}}(y_{ij} \mid u_i[a+1]))}$$

$$q_{ij}(t_{ij}[a+1] = \sum_{s_i} q_i(s_i)[a+1] \,.\, q_{ij}(t_{ij} \mid s_i)[a+1]$$

$l_{\mathrm{ME1}}[a+1]$ is calculated as in (3) and (4), and $l_{\mathrm{ME2}}[a+1]$ is calculated as in (5) and (6).

Step M of the ath iteration

Given the previous posterior probabilities, the distribution parameters are obtained by annulling the derivatives of $l_{\mathrm{ME}}[a+1]$ with respect to these parameters. We then get:

for $t = 1, 2, 3$

$$\mu_t[a+1] = \frac{\Sigma_i\ \Sigma_j\ q_{ij}(t)[a+1] \,.\, (y_{ij} - u_i[a])}{\Sigma_i\ \Sigma_j\ q_{ij}(t)[a+1]}$$

for $i = 1, ..., n$

$$u_i[a+1] = \frac{\Sigma_j\ \Sigma_t\ q_{ij}(t)[a+1] \,.\, (y_{ij} - \mu_t[a+1])}{\sigma_e^2\ \sigma_u^{-2} + \Sigma_j\ \Sigma_t\ q_{ij}(t)[a+1]}$$

$$\sigma_e^2[a+1] = \frac{\sigma_e^2\ \sigma_u^{-2}\ \Sigma_i\ u_i^2[a+1] + \Sigma_i\Sigma_j\Sigma_t\ q_{ij}(t)[a+1] \,.\, (y_{ij} - u_i[a+1] - \mu_t[a+1])^2}{nm}$$

the denominator being $n(m+1)$ for the l_{ME2} test.

$$p_{s_i}[a+1] = \frac{\Sigma_i q_i(s_i)[a+1]}{n}$$

$$q[a+1] = \frac{\Sigma_i\Sigma_j q_{ij}(1)[a+1] + \Sigma_i(q_i(3)[a+1] \,.\, \Sigma_j q_{ij}(2|3)[a+1])}{nm - \Sigma_i(q_i(2)[a+1] \,.\, \Sigma_j q_{ij}(2|2)[a+1])}$$