Mathematical Statistics
Stockholm University

On Computation of p-values in Parametric Linkage Analysis

Azra Kurbasic and Ola Hössjer

Research Report 2003:18

ISSN 1650-0377

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2003:18**,
http://www.math.su.se/matstat

On Computation of p-values in Parametric Linkage Analysis

Azra Kurbasic*and Ola Hössjer†

October 2003

Abstract

Parametric linkage analysis is usually used to find chromosomal regions linked to a disease (phenotype) that is described with a specific genetic model. This is done by investigating the relations between the disease and genetic markers, that is, well-characterized loci of known position with a clear Mendelian mode of inheritance. Assume we have found an interesting region on a chromosome that we suspect is linked to the disease. Then we want to test the hypothesis of no linkage versus the alternative one of linkage. As a measure we use the maximal lod score $Z_{\max}$. It is well known that the maximal lod score has asymptotically a $(2\ln 10)^{-1} \times (\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1))$ distribution under the null hypothesis of no linkage when only one point (one marker) on the chromosome is studied. In this paper, we show, both by simulations and theoretical arguments, that the null hypothesis distribution of $Z_{\max}$ has no simple form when more than one marker is used (multipoint analysis). In fact, the distribution of $Z_{\max}$ depends on the number of families, their structure, the genetic model, marker denseness, and marker informativity. This means that a constant critical limit of $Z_{\max}$ leads to tests associated with different significance levels. Because of the above-mentioned problems, from the statistical point of view a p-value is more desirable measure of significance than the maximal lod score.

Keywords Linkage analysis, lod score distribution, pointwise/genomwide p-value.

*Centre for Mathematical Sciences/Department of Mathematical Statistics and Department of Oncology, Lund University, Sweden.

†Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ola@math.su.se.
Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

1 Introduction

The aim of linkage analysis is to infer the position (locus) along one or several chromosomes, of a gene underlying or contributing to a certain trait, often related to a certain disease. Based on trait phenotype and DNA marker data from a number of families, this is done by estimating the relative positions along the chromosome(s). The DNA marker data give information about occurrence of crossovers (switching between grandpaternal and grandmaternal transmission of alleles) at the regions close to the trait locus during meioses; markers cosegregate with the trait phenotypes in the families. Statistically, linkage analysis can be formulated as an hypothesis testing problem for testing the null hypotesis (H_0) that the trait locus is unlinked to the chromosomal region(s) of interest against the alternative (H_1) that it is located on one of the chromosomes. One can use either parametric or nonparametric methods. The parametric methods assume the genetic model (mode of inheritance and trait allele frequency) to be known. It is traditionally based on logarithms of likelihood ratios, so called lod scores, for testing H_0 against H_1. The nonparametric methods are typically based on allele sharing statistics and do not require knowledge of the genetic model. Both parametric and nonparametric linkage analysis use the maximal linkage score $Z_{\max}$ (either based on lod scores or allele sharing statistics) as measure of strength of evidence for linkage. It is well known that parametric linkage analysis is more powerful than nonparametric when the underlying genetic model is known and true. On the other hand nonparametric linkage analysis is more robust against misspecification of the genetic model.

This paper adresses theoretical and practical aspects of parametric linkage analysis when we have a likelihood ratio based score, *lod score*. It is statistically important that the H_0 distribution of the test statistic $Z_{\max}$ is (asymptotically) independent of a number of nuisance parameters such as genetic model, pedigree structures, marker data informativity and trait phenotypes. Otherwise, there is no natural correspondence between p-values and $Z_{\max}$, rendering statistical conclusions more difficult to draw when knowledge of the genetic model can be questioned. It is well known that the maximal lod scores have a well defined asymptotic limit distribution when only one marker is studied (two-point linkage analysis), cf. Ott (1999). In this papper we show, by simulation and theoretical arguments, that the situation is completely different for maximal lod scores with many markers (multipoint linkage analysis). The reason for this discrepency is that the parameter space for the parameter of interest is the recombination fraction θ in two-point analysis and map position x in multipoint analysis. In the former case the parameter space is connected and in the latter case not. In fact, the H_0 parameter is an isolated point for multipoint analysis.

We also briefly discuss some alternatives to lod scores with asymptotic H_0 distributions that are independent of nuisance parameters for multipoint analysis. These include extensions of affected pedigree (AMP) methods, (Weeks & Lange 1988, Fimmers et al. 1989, Whittemore & Halpern 1994 and Kruglyak et al. 1996) and mod scores (Risch 1984, Clerget-Darpoux et al. 1986 and Whittemore 1996).

2 Parametric Linkage Analysis

2.1 One Marker

Let ψ be the genetic model parameters (disease allele frequency and penetrance parameters) and θ the recombination fraction between the marker and disease locus. Disease allele frequency is the population frequency of the disease susceptibility allele and penetrance is the conditional probability that an individual is affected given the genotype. The recombination fraction is a measure of distance between two loci and is defined as the probability of occurrence of recombination, cf. Ott (1999) and Sham (1998). Recombination is a result of an odd number of crossovers between two loci. Furthermore, let Y and M denote the collection of disease phenotypes and marker data, respectively. Then the lod score

$$Z(\theta;\psi) = \log_{10} \frac{P(Y,M|\theta,\psi)}{P(Y,M|0.5,\psi)}$$

is used for testing $H_0 : \theta = 0.5$ against alternatives $\theta < 0.5$. The composite hypotesis testing problem uses the alternative $H_1 : \theta \in [0, 0.5)$. The total parameter space can be depicted as

with ◯ indicating H_0. Thus H_0 is a boundary point of the parameter space. The maximal lod score

$$Z_{\max}(\psi) = \sup_{0\le\theta\le 0.5} Z(\theta;\psi) = (2\ln 10)^{-1} 2\ln \frac{\sup_\theta P(Y,M|\theta,\psi)}{P(Y,M|0.5,\psi)} \tag{1}$$

has asymptotically a $(2\ln 10)^{-1}\times(\frac{1}{2}\chi^2(0)+\frac{1}{2}\chi^2(1))$ distribution under H_0 as the number of pedigrees (sets of relatives with known family structure) grows and when the genetic model is correctly specified. This is asymptotically independent of pedigree structure, marker informativity and disease phenotypes, although for finite samples the distribution will usually depend on such quantities to some extent, cf. Section 4.4 and 4.6 in Ott (1999). Mixture of χ^2 distributions typically arise for log likelihood ratios under the null hypotesis when the H_0 parameter is a boundary point of a connected parameter space, cf. Self & Liang (1987). In (1) the same asymptotic limit distribution also appears when ψ is misspecified, cf. Dudoit & Speed (1999) and Williamson & Amos (1990). Hence the significance level

$$\alpha(T) = P_{H_0}(Z_{\max}(\psi) \ge T) \tag{2}$$

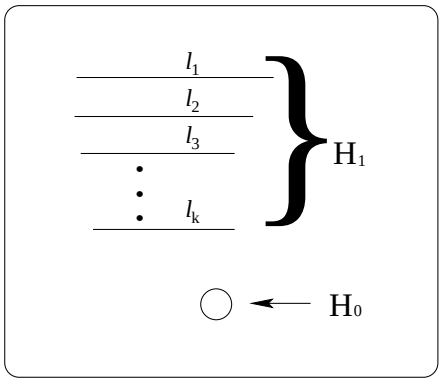
is asymptotically independent of the nuisance parameter ψ. It is common practice to use a lod score as a measure of significance rather than a p-value, cf. Ott (1999) and Sham (1998). However, because of (1), there is asymptotically a one-to-one correspondence between $Z_{\max}$ and the p-value $\alpha(Z_{\max})$.

2.2 Multipoint Linkage Analysis

Simultaneous analysis of several linked markers (multipoint analysis) is preferable to the single marker analysis when at least two markers are available on a chromosome. The genetic map distance between the loci is in units of Morgans (1 Morgan = 100 centiMorgans). It is defined as the expected (average) number of crossovers occuring on a single chromosome between two loci, cf. Ott (1999). The parameter of interest in multipoint analysis is the disease locus x, and we write

$$Z(x;\psi) = \log_{10} \frac{P(Y,M|x,\psi)}{P(Y,M|\infty,\psi)} \tag{3}$$

where $x = \infty$ corresponds to H_0. The parameter space is no longer connected but can be depicted as



where K is the number of chromosomes and l_i the genetic length of chromosome i. The maximal lod score

$$Z_{\max}(\psi) = max_{x \in H_1} Z(x;\psi) \tag{4}$$

does no longer converge to a limiting distribution under H_0 as the number of pedigrees grows. The reason is that the parmeter space is no longer connected. This implies that the maximal lod score (4), as opposed to (1), has a degenarated limit distribution at $-\infty$ as the number of families $N \longrightarrow \infty$. For multipoint linkage analysis the significance (2) for a given treshold T depends heavily on ψ, the number of pedigrees, their structure, and the disease and marker phenotypes.

3 Distribution of Lod Scores

3.1 Inheritance Vectors

Before continuing the discussion we will give some definitions. Individuals whose parents are members of the pedigree are designated as *nonfounders* and those whose parents are members of the pedigree as *founders*. The number of founders in the pedigree is denoted by f and the number of nonfounders by $n - f$, where n is the number of individuals in the pedigree. We also assume that founders are unrelated and carry $2f$ alleles that are not IBD (identical by descent). IBD means that alleles are inherited from a common ancestor. The inheritance process can then be seen as the distribution of the founder's alleles among the nonfounders. At one locus x in the genome the inheritance pattern can be represented by a *binary inheritance vector*, $v(x)$, defined as $v(x) = (p_1, m_1, p_2, m_2, ..., p_{n-f}, m_{n-f})$. The coordinates of the inheritance vector p_i and m_i describe the outcome of the paternal and maternal meioses. They are set to 0 or 1 if the i^{th} nonfounder's allele at position x originate from a grandfather or a grandmother. The probability distribution over possible inheritance vectors given marker data is referred to as the *inheritance distribution*. In the absence of any genotype information, the probability distribution of $v(x)$ is uniform over all $2^{2(n-f)}$ possible inheritance vectors ($P_{uniform}$). This is a consequence of Mendel's law of segregation. More details about inheritance vectors can be found in Kruglyak et al. (1996).

For two-point analysis (one marker and disease), we define the inheritance vector slightly differently. Let $v(\theta)$ be the inheritance vector of a locus at recombination fraction θ from the single marker. Althought this locus is not unique (there are often two loci having the same recombination fraction to the marker), the marker gives exactly the same information about both of these inheritance vectors. Hence, as we will seen in the next subsection, this definition make sense when defining two-point lod scores.

3.2 Some Properties of Lod Scores

We first consider one single pedigree. For notational simplicity, we omit the genetic model parameters ψ in the notation. Kruglyak et al. (1996) define both parametric and nonparametric linkage analysis using inheritance vectors. A scoring function that depends on the inheritance vector v and the observed phenotypes in the pedigree Y is specified. It is a measure of compatibility between v and Y, i.e. the extent to which the phenotype vector Y can be explaind by an inheritance vector v at the disease locus. In parametric linkage analysis, when the inheritance vector is known, $P(Y|v)$ is the likelihood of observed phenotypes Y in the pedigree conditioned on the inheritance vector v. The scoring function for one family with m meioses is

$$S(v) = \frac{P(Y|v)}{\sum_{w \in V} P_{uniform}(w) P(Y|w)} = \frac{P(Y|v)}{\sum_{w \in V} 2^{-m} P(Y|w)}. \tag{5}$$

For two-point analysis, we define the distribution

$$P_\theta(w) = P(v(\theta) = w|M),$$

which quantifies the information the single marker yields at a recombination fraction θ away from it. Then the two-point lod score $Z(\theta)$ and likelihood ratio $LR(\theta)$ can be rewritten (Morton 1995) as

$$Z(\theta) = \log_{10} LR(\theta) = \log_{10} \sum_w S(w) P_\theta(w).$$

For multipoint analysis we let

$$P_x(w) = P(v(x) = w|M)$$

quantify the information that the marker data gives about inheritance at locus x. Then the multipoint lod score and likelihood ratio can be written as

$$Z(x) = \log_{10} LR(x) = \log_{10} \sum_w S(w) P_x(w). \tag{6}$$

The total multipoint lod score for N families, finally, is obtained by summing the individual family lod scores,

$$Z(x) = \sum_{i=1}^{N} Z_i(x), \tag{7}$$

where Z_i is the i^{th} family score. For two-point analysis, (7) remains true if we replace x by θ.

The two extreme cases of marker informativity are perfect marker data ($LR(x) = S(v(x))$ for one pedigree) and no marker information at all ($P_x = P_{uniform}$ for one pedigree). If we let E and $LR_p(x)$ denote the expectation and the likelihood ratio for perfect marker data, respectively, we see from (6) that

$$LR(x) = E(LR_p(x)|M).$$

From this and the fact that the H_0 distribution of $v(x)$ (in a population of many pedigrees) is $P_{uniform}$ we get

$$E_{H_0}(LR(x)) = E_{H_0}(LR_p(x)) = 1 \tag{8}$$

$$Var_{H_0}(LR(x)) \leq Var_{H_0}(LR_p(x)) \quad \text{when the variances exist} \tag{9}$$

$$Z(x) = 0, \quad \text{with no marker information} \tag{10}$$

$$E_{H_0}(Z_p(x)) \leq E_{H_0}(Z(x)) \leq 0, \tag{11}$$

where $Z_p(x) = \log LR_p(x)$ is the lod score for perfect marker data. Of these equations, (9) and (11) can be deduced from Jensen's inequality, cf. Royden (1968). We also conjecture that in most cases

$$Var_{H_0}(Z(x)) \leq Var_{H_0}(Z_p(x)). \tag{12}$$

Although we have no general proof of (12), it is a natural consequence of (9) in most situations.

For two-point analysis, there is no marker information at $\theta = 0.5$ ($Z(0.5) = 0$), whereas perfect marker information can arise only at $\theta = 0$ if the marker is fully polymorphic. Hence, $Z_{\max}$ is derived quite differently in two- and multipoint analysis. In the former case $Z(\theta)$ will be negative for most values of θ away from 0.5 under H_0 but $Z_{\max}$ is never negative. For multipoint analysis we maximize a function $Z(x)$ whose mean value under H_0 is negative for all x (unless there is no marker information somewhere), and this often implies that $Z_{\max}$ is negative as well.

4 Calculating Genomwide Significance Levels

In this section we describe methods for approximating the genomwide significance level $\alpha(T)$ for multipoint lod scores. The most straightforward method is to use Monte Carlo simulations, that is

$$\alpha(T) \approx \frac{1}{N_R} \sum_{i=1}^{N_R} I(Z^i_{max} \geq T) \tag{13}$$

where Z^i_{max} are independent copies of $Z_{\max}$ under H_0 and N_R is the number of generated copies. Methods for simulating Z^i_{max} are described for example in Section 9.7 of Ott (1999). The Monte Carlo method is very general, but can sometimes be slow, especially for large pedigrees and low marker informativity.

For perfect marker data we can use analytical methods based on Gaussian extreme value theory to approximate $\alpha(T)$ as described for example by Lander & Kruglyak (1995) and Ängquist & Hössjer (2003). Notice first that $Z(x)$ as well as family scores $Z_i(x)$ are stationary processes under H_0 when marker information is perfect. We start assuming $P(Z(x) = -\infty) = 0$ at all x under H_0. This is typically true for most genetic models except for those with complete penetrance and no phenocopies. Let μ_i and σ_i denote the mean and standard deviation of the i^{th} family score $Z_i(x)$ under H_0. Then $\mu = \sum_{i=1}^N \mu_i$ and $\sigma = \sqrt{\sum_{i=1}^N \sigma_i^2}$ are the mean and and standard deviation of $Z(x)$ under H_0. Further

$$Z(x) = \mu + \sigma \tilde{Z}(x), \tag{14}$$

where $\tilde{Z}(x) = \sum_{i=1}^N \sigma_i \tilde{Z}_i(x) / \sqrt{\sum_{i=1}^N \sigma_i^2}$ and $\tilde{Z}_i(x) = (Z_i(x) - \mu_i)/\sigma_i$. Notice that $\tilde{Z}$ and all $\tilde{Z}_i$ are stationary processes under H_0 with mean zero and variance one. For large sample sizes the central limit theorem implies that $\tilde{Z}$ approaches a Gaussian process with mean zero and unit variance. With $\tilde{T} = (T - \mu)/\sigma$ the transformed threshold, and $\tilde{Z}_{max} = max_x \tilde{Z}(x)$ we get

$$\alpha(T) = P_{H_0}(\tilde{Z}_{max} \geq \tilde{T}) \approx 1 - e^{-\mu(\tilde{T})}. \tag{15}$$

This formula was intoduced by Lander & Kruglyak (1995). The quantity $\mu(\tilde{T})$ approximates the average number of exceedances of $\tilde{Z}$ over the threshold $\tilde{T}$. It can

be written as

$$\mu(\tilde{T}) = \alpha_{pt}(T) \cdot C \tag{16}$$

where $\alpha_{pt}(T)$ equals or approximates the pointwise significance level

$$P_{H_0}(Z(x) \geq T) = P_{H_0}(\tilde{Z}(x) \geq \tilde{T}) \tag{17}$$

and C is a correction factor that accounts for multiple testing. It depends, among other quantities, on the threshold T, the number of chromosomes K, and the total genomic length $l_1 + l_2 + ... + l_K$. A formula for $\mu(\tilde{T})$ based on the assumption of a Gaussian $\tilde{Z}$ process was given by Lander & Kruglyak (1995). Their formula was generalized to arbitrary pedigrees and corrected for non-Gaussianity in Änqquist & Hössjer (2003). We refer to these two methods as the normal and adjusted normal approximations, respectively.

In case $\varepsilon = P(Z(x) = -\infty) > 0$, we proceed by defining (14) as well as μ_i and σ_i conditionally on the event $Z(x) > -\infty$. The procedure is similar except that we replace $\mu(\tilde{T})$ by $(1-\varepsilon) \cdot \mu(\tilde{T})$ to account for the fact that the average number of exceedances of the threshold $\tilde{T}$ is reduced by a factor $(1-\varepsilon)$.

5 Results

Standardized lod scores $\tilde{Z}(x)$ turned out very useful when calculating p-values in parametric linkage analysis. We calculated genomwide p-values using Monte Carlo simulations (13), the normal approximation formula (15) and adjusted normal approximation (Ängquist & Hössjer, 2003). Figures 6-13 display calculated genomwide p-values as a function of the threshold T. We used four different pedigree structures and seven different genetic models, cf. Figure 5 and Table 1. For simplicity of interpretation we considered data sets with N pedigrees of the same kind. It is obvious from the figures that there is a substantial variation of the p-value depending on the number of the families in the data, the family structure, marker denseness and informativity, and the genetic model. Yet there is a rule of thumb: A higher number of the families in the study/larger families/stronger genetic models give a flatter p-value function $\alpha(T)$. Vice versa, a lower number of families/smaller families/ weaker genetic models in the study give a higher p-value for small thresholds T and then a more rapidly decreasing curve $\alpha(T)$ as T increases. This is because in general $E_{H_0}(Z_{max})$ decreases and $Var_{H_0}(Z_{max})$ increases by the increased extent of information in data (in the special case of perfect versus imperfect marker information, see equations (11) and (12) for the pointwise version of this phenomenon).

It follows from (15) and (16) that the genomwide p-value at least for perfect marker data is approximately a function of the pointwise p-value $\alpha_{pt}(T)$ and the multiple testing factor C. Even though C increases the p-value by orders of magnitude, it tends to be less sensititive than $\alpha_{pt}(T)$ with respect to variations in informativity in the data set. For this reason, it is instructive to analyse how $\alpha_{pt}(T)$ depends on the informativity of the data set. As opposed to $\alpha(T)$, it can easilly be calculated

exactly for perfect marker data, see Appendix A for details. Figures 14-15 show some examples of exactly calculated pointwise p-values. We can see the same kind of pattern for these as for genomwide p-values.

An advantage of pointwise p-values is that they are easier to analyze theoretically than genomwide ones. In fact, we have found that the simple model

$$Z(x) \in^{H_0} \Gamma(a, b, c) \tag{18}$$

captures several of the essential properties of lod scores surprisingly well. Here $\Gamma(a, b, c)$ is a gamma distribution $\Gamma(a, b)$ translated to have left endpoint at c. We interpret a as the amount of information (strength of genetic model, pedigree size, and amount of marker information) per family and b is proportional to the effective number of families in the data set. The constant $c = c(a, b)$ is chosen so that (8) holds. It turns out that (9)-(12) are satisfied by (18), as shown in Appendix B. Furthermore, both $\mu_0 = E_{H_0}(Z(x))$, $\mu_1 = E_{H_1}(Z(x))$ and $\sigma_0^2 = Var_{H_0}(Z(x))$ are proportional to b, see Figure 1 and Appendix B for more details. Given the treshold

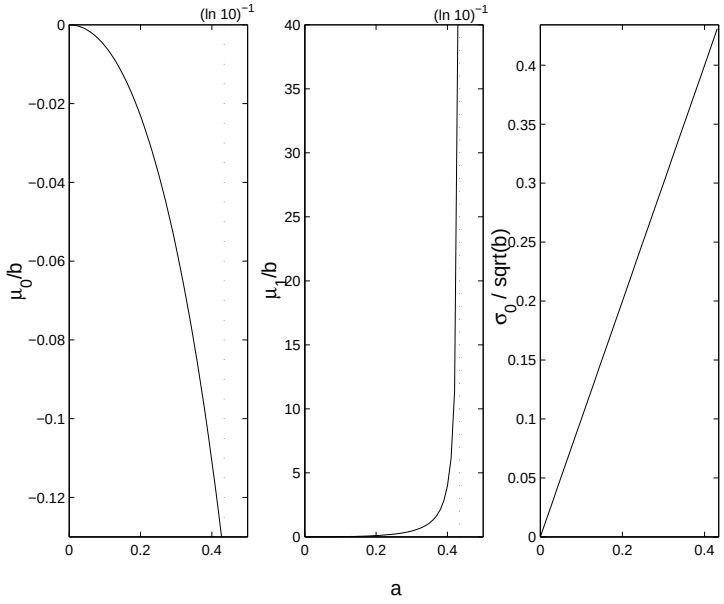


Figure 1: Expected lod score and standard deviation for model (18) as function of a.

T the pointwise p-value is

$$P_{H_0}(Z > T) = P(aX + c > T) = P(X > \frac{T - c}{a}) \tag{19}$$

where $X \in \Gamma(1, b)$, see Figure 2.

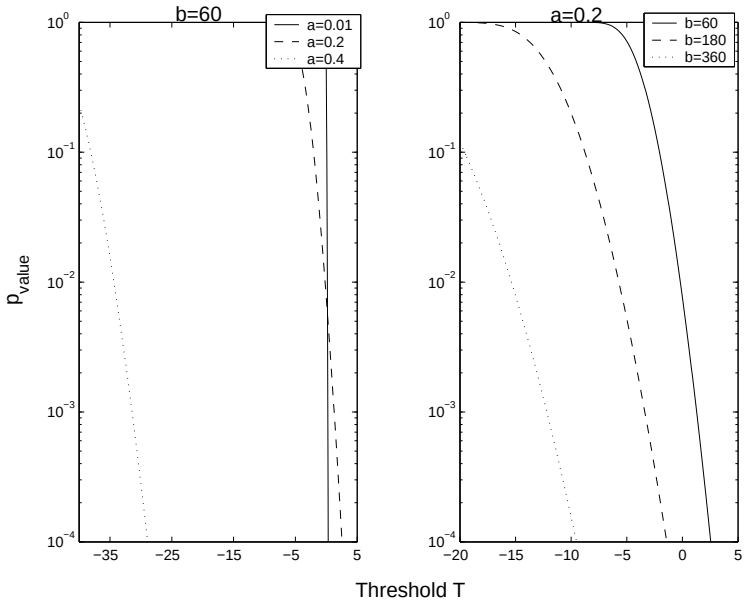


Figure 2: P-value in (19) as a function of the threshold T and different values of a and b.

6 Summary and Other Approaches

An essential issue in this paper was to find out and explain the behaviour of p-values based on maximal lod scores. For multipoint linkage analysis we have found, both by extensive simulations and theoretical arguments, that the H_0 distribution of pointwise and maximal lod scores depend heavily on genetic model parameters, number and structure of pedigrees, phenotypes, and marker informativity. For this reason the p-value is a more appropriate performance measure than the maximal lod score itself. We also introduced a simplified analytical model which captures several fundamental properties of lod scores.

It is possible to use alternative approaches in parametric linkage analysis than traditional lod scores whose $Z_{\max}$ distribution under H_0 are less sensitive to variations of nuisance parameters. One possibility is to define a mod score (Risch 1984, Clerget-Darpoux et al. 1986, Whittemore 1996)

$$Z(x) = \sup_{\psi} Z(x;\psi) \tag{20}$$

by maximizing (3) over a predefined set of genetic model parameters. In this case the parameter (x,ψ) contains both the locus x and genetic model parameters ψ, of which the latter are nuisance parameters. Let ψ_0 be a set of genetic model parameters corresponding to no genetic effect at the disease locus. For a chromosome $[0,l]$ of length l, the null hypothesis is no longer formulated as $x = \infty$ but rather as a disease locus $x \in [0,l]$ with no genetic component ($\psi = \psi_0$). This parameter

space, illustrated in Figure 3, is connected if the genetic model parameter space ψ is connected. As a result, the distribution of $Z(x)$ under H_0 is asymptotically free from nuisance parameters. Depending on the set of genetic models $\{\psi\}$ locally around ψ_0 and whether or not ψ_0 is at the boundary of $\{\psi\}$, the limit distribution can be either a χ^2 distribution or a mixture of χ^2 distributions, cf. Self & Liang (1987), Rotnitzky et al. (2000), and Hössjer (2003b). Hence, at least the pointwise p-value $\alpha_{pt}(T)$ is asymptotically free from nuisance parameters.

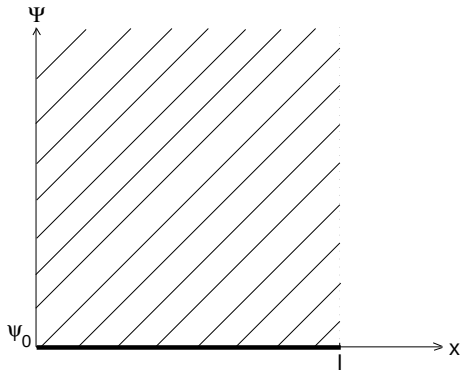


Figure 3: The total parameter space corresponds to $[0, l] \times [0, \infty)$ and H_0 to the line $[0, l] \times \{0\}$.

The nonparametric linkage method in Genhunter (Kruglyak et al. 1996) is based on an allele sharing score function $S(v)$ and NPL score

$$Z(x) = \sum_w S(w) P_x(w), \tag{21}$$

for one pedigree with P_x as defined in Section 3. The score function S is standardized so that $Z(x)$ has zero mean and unit variance under H_0 when the marker information is perfect. This method originally goes back to the affected pedigree method of Weeks & Lange (1988), although these authors focused on identity by state (IBS) rather than IBD sharing. It is also possible to define S for genetic models with other phenotypes than affected/nonaffected, cf. Whittemore (1996), McPeek (1999), and Hössjer (2001, 2003b), and use it for parametric linkage analysis. In fact, a large class of scores (21) can be obtained by differentiating the lod score (3) with respect to ψ at ψ_0. In that case, $Z(x)$ can be interpreted as a score test version of the mod score (20), which is a profile likelihood curve with ψ being the profiled set of parameters.

The H_0 distribution of $Z(x)$ in (21) is asymptotically normal distributed and quite insensitive to variation of genetic model parameters. In fact, it is equivalent to the standardized score function $\tilde{Z}$ in Section 4 for lod scores. A modified version

of (21) suggested by Kong & Cox (1997) makes the H_0 distribution less sensitive to variation in marker informativeness as well. As a result, the pointwise p-value $\alpha_{pt}(T)$ is asymptotically free of nuisance parameters. However, the same is not true for the genomwide p-value $\alpha(T)$, since the effective amount of multiple testing depends on for example pedigree structure, see Feingold et al. (1993), Lander & Kruglyak (1995), and Ängquist & Hössjer (2003).

A Calculating Theoretical Pointwise p-value

Let $Z_i(x)$ be the lod score for a family i and $Z(x)$ the total lod score. Further let $\varepsilon_i = P(Z_i(x) = -\infty)$ and $\varepsilon = P(Z(x) = -\infty)$. Then $\varepsilon = 1 - \prod_{i=1}^{N}(1-\varepsilon_i)$. Let F_i and F denote the distribution function of the lod score for family i and the total lod score, respectively. The distribution under H_0, of both Z_i and Z, is a mixture of a point mass $-\infty$ and a proper distribution, that is $F_i(x) = (1-\varepsilon_i)\cdot G_i(x)$ and $F(x) = (1-\varepsilon)\cdot G(x)$. One can calculate the distribution function G by convolution

$$G = G_1 * G_2 * \cdots * G_N.$$

We computed the distribution function G by using an approximation of the exact distribution based on linear binning and mixtures of uniform distributions, see Kruglyak et al (1996) and Appendix C in Ängquist & Hössjer (2003). Pointwise p-value for a threshold T is then

$$\alpha_{pt}(T) = P_{H_0}(Z(x) \geq T) = (1-\varepsilon)\cdot(1-G(T)). \tag{22}$$

The value of ε depends both on the assumed genetic model, family structure, and the number of families N. Hence the pointwise p-value also depends on these factors. When $\varepsilon = 1$, i.e. $\varepsilon_i = 1$ for at least one i, the distribution for Z under the null hypotesis H_0 is a one point distribution at $-\infty$. If $0 < \varepsilon < 1$ then the p-value is given as in (22) and some examples are given in Figure 4.

When $\varepsilon = 0$, that is, when we have a genetic model with phenocopies then

$$F(x) = G(x).$$

Phenocopies (also called sporadic cases) are individuals who are affected not owing to genetic predisposition at the locus under study. In the genetic model occurrence of phenocopies means a nonzero penetrance for nonsusceptible genotypes. More details about phenocopies can be found in Ott (1999) and Sham (1998). The pointwise p-value in absence of phenocopies is

$$\alpha_{pt}(T) = P_{H_0}(Z(x) \geq T) = 1 - G(T).$$

B Model for Lod Score Distribution

Perfect Data

Let $LR = LR(x)$ be the likelihood ratio and $Z = Z(x) = \log_{10}(LR(x))$ the lod score. Equation (8) is true if the following is fulfilled in model (18)

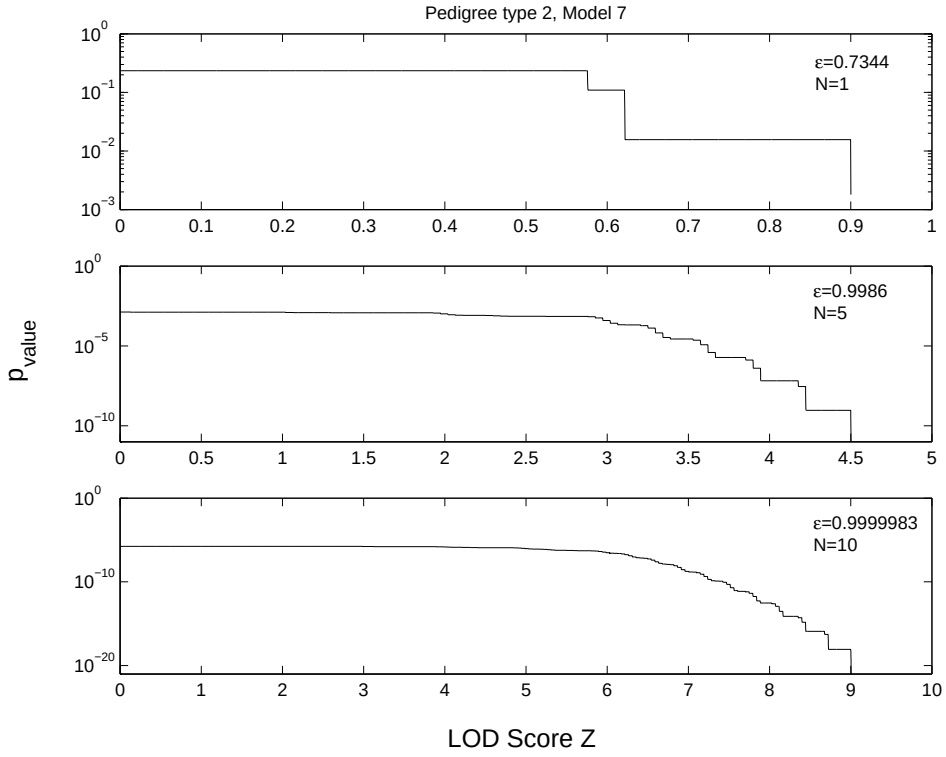


Figure 4: Comparisons between theoretical pointwise p-values for three different sample sizes N (number of families) of pedigree type 2 and model 7. For more details see Figure 5 and Table 1, respectively.

$$\begin{aligned} 1 = E_{H_0}[e^{\ln 10 Z}] &= M(\ln 10) = e^{c \ln 10}(1 - a \ln 10)^{-b} \\ \iff c &= \frac{b \ln(1 - a \ln 10)}{\ln 10} =: c(a, b) \end{aligned} \tag{23}$$

where

$$M(t) = E_{H_0}(e^{tZ}) = e^{ct}(1 - at)^{-b}$$

is a moment generating function for Z under H_0. Because of (23) we have a model with two free parameters. The first one $b > 0$ is proportional to the effective number of families in data, and the second one $0 < a < (\ln 10)^{-1}$ corresponds to the strength of the genetic model, (where $a = 0$ means no genetic effect).
The expectation of the lod score under the null hypothesis is

$$\mu_0 = E_{H_0}(Z) = c + ab = b(a + \frac{\ln(1 - a \ln 10)}{\ln 10}) < 0.$$

This inequality corresponds to (11) for the pointwise multipoint lod score. Further we have

$$\sigma_0 = a\sqrt{b}$$

and

$$\begin{aligned}
\mu_1 &= ELOD = E_{H_1}(Z) = \int z dP_{H_1}(z) = \int zLR(z) dP_{H_0}(z) \\
&= \int ze^{\ln 10 z} dP_{H_0}(z) = E_{H_0}(Ze^{\ln 10 Z}) = M'(\ln 10) \\
&= M(\ln 10)(c + \frac{ab}{1 - a\ln 10}) = b(\frac{\ln(1 - a\ln 10)}{\ln 10} + \frac{a}{1 - a\ln 10}).
\end{aligned}$$

Imperfect Data

Let $a \in (0, \frac{1}{\ln 10})$ be a given genetic model. We define incomplete marker information as a weakened genetic model ϵa, where $0 \leq \epsilon < 1$. Let Z and Z_p be the lod score for incomplete and complete data, respectively. Likelihood ratios are then defined as $LR = 10^Z$ and $LR_p = 10^{Z_p}$. Note that the properties (9) - (11) must always be fulfilled. Our model satisfies (9) - (11) and (12) as well. It follows from the fact that $Z \in^{H_0} \Gamma(\epsilon a, b, c(\epsilon a, b))$ and

$$\begin{aligned}
Var_{H_0}(LR) &= E_{H_0}(LR^2) - E_{H_0}(LR)^2 \\
&= E_{H_0}(e^{2\ln 10 Z}) - 1^2 \\
&= M_Z(2\ln 10) - 1 \\
&= e^{c(\epsilon a, b) 2\ln 10}(1 - 2\epsilon a \ln 10)^{-b} - 1 \\
&= \frac{(1 - \epsilon a \ln 10)^{2b}}{(1 - 2\epsilon a \ln 10)^b} - 1 < \frac{(1 - a\ln 10)^{2b}}{(1 - 2a\ln 10)^b} - 1 = V_{H_0}(LR_p).
\end{aligned}$$

This assumes that both variances exist, that is $0 < a < (2\ln 10)^{-1}$. In the inequality we used $V_{H_0}(LR) = (1-x)^{2b}/(1-2x)^b - 1$, where $x = \epsilon \ln 10$, which is an increasing function of x.

Acknowledgement

This research is sponsored by the National Research School in Genomics and Bioinformatics. We whish to thank Pär-Ola Bendhal for discussing and suggesting improvments on the manuscript.

References

Claus, E.B., Rish, N.J., Thompson, W.D. (1990). Age at onset of familial risk of breast cancer. *Am. J. Epidemiol.* **131**: 961-967.

Feingold, E., Brown, P.O. and Siegmund, D. (1993). Gaussian Models for Genetic Linkage Analysis Using Complete High Resolution Maps of Identity by Descent. *Am. J. Hum. Genet.* **53**: 234-251.

Fimmers, R., Seuchter, S.A., Neugebauer, M., Knapp, M., Baur, MP., (1989). Identity by descent analysis using all genotype solutions. *In: Elston R.C., Spence M.A. Hodge S.E. MacCluer J.W. (eds) Multipoint mapping and linkage based on affected pedigree members: genetic Analysis Workshop 6. Alan R. Lis, New York.*

Dudoit, S. and Speed, T.P. (1999). A Score Test for Linkage Using Identity by Descent Data from Sibships. *Annals of Statistics* **27**: 943-986.

Clerget-Darpoux, F. Bonaïti-Pellié, C. and Hoches, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42**: 393-399.

Hössjer O. (2001). Determining Inheritance Distributions via Stochastic Penetrances. *Preprints in Mathematical Sciences, Centre for Mathematical Sciences, Lund University. To appear in Journal of The American Statistical Association.*

Hössjer O. (2003a). Assessing Accuracy in Linkage Analysis by means of Confidence Regions. *Genet Epidemiology* **25(1)**: 59-72.

Hössjer O. (2003b). Conditional Likelihood Score Function in Linkage analysis. *Preprint 2003:10, Mathematical Statistics, Stockholm University.*

Kong, A. and Cox, N.J. (1997). Allele-Sharing Models: LOD Scores and Accurate Linkage Tests. *Am. J. Hum. Genet.* **61**: 1179-1188.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and Nonparametric Linkage Analysis: A unified Multipoint Approach. *Am. J. Hum. Genet.* **58**: 1347-1363.

Lander, E.S. and Kruglyak, L. (1995). Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results. *Nature Genetics* **11**: 241:247.

McPeek M.S. (1999). Optimal Allele-Sharing Statistics for Genetic Mapping Using Affected Relatives. *Genetic Epidemiology* **16**: 225-249.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Gennet.* **7**: 277-318.

Ott, J. (1999). Analysis of Human Genetic Linkage, Third Edition. *Johns Hopkins University Press.*

Risch, N. (1984). Segregation Analysis Incorporating Genetic Markers. I. Single-Locus Models with an application to type I Diabetes. *Am. J. Hum. Genet.* **36**: 363-386.

Rotnitzky, A., Cox, D.R., Bottai, M. and Robins, J. (2000). Likelihood-based Inference with Singular Information Matrix. *Bernoulli* **6**: 243-284.

Royden H.L. (1968). Real analysis, 2d ed. *Macmillan Publishing, New York.*

Self, S.G. and Liang, K.Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82**: 605-610.

Sham, P. (1998). Statistics in Human Genetics. *Arnold, a member of the Hodder Headline Group.*

Weeks, D.E. Lange, K. (1988). The affected-pedigree-member method of linkage analysnis. *Am. J. Hum. Genet.* **42**: 315-326.

Whittemore A. C. (1996). Genome Scanning for Linkage: An Overview. *Am. J. Hum. Genet.* **59**: 704-716.

Whittemore, A. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50**: 118-127.

Williamson, J.A. and Amos, C.I. (1990). On the asymptotic behaviour of the estimate of the recombination fraction under the null hypotesis of no linkage, when the model is misspecified. *Genetic Epidemiology* **7**: 309-318.

Ängquist, L. and Hössjer, O. (2003). Improving the Calculation of Statistical Significance in Genome-Wide Scans. *Preprints in Mathematical Sciences, Centre for Mathematical Sciences, Lund University.*

List of Tables

Table 1: Summary of details about the models. Affection status locus type is autosomal dominant disease. The penetrance value f_i is the conditional probability that an individual is affected given the genotype with i disease alleles at the disease locus.

Model	Disease Allele Frequency	Penetrance Values f_0	f_1	f_2
1	0.1	0.001	0.5	0.8
2	0.1	0.2	0.5	0.8
3	0.0033	Age dependent penetrance*		
4	0.1	Age dependent penetrance*		
5	0.0033	0.001	0.5	0.8
6	0.0033	0.2	0.5	0.8
7	0.1	0	1	1

*Penetrance values from the standard genetic model for breast cancer derived by Claus et al. (1991), see Table 2.

Table 2: Penetrance values used in the study for models three and four. Seven age groups $\times$ two disease classifications, affected and unaffected, were used. Unaffected males and individuals with unknown affection status were assigned to liability class one. Disease and normal allele at the disease locus are denoted by d and D, respectively. They give rise to three different genotypes.

		Penetrance of Genotype		
Liab. Class	Age Group	dd	Dd	DD
	Cumulative risk for unaffected females			
1	<30.....	0.00009	0.008	0.008
2	30-39.....	0.00146	0.083	0.083
3	40-49.....	0.0083	0.269	0.269
4	50-59.....	0.0210	0.469	0.469
5	60-69.....	0.0390	0.616	0.616
6	70-79.....	0.0610	0.724	0.724
7	$\geq$80.....	0.0820	0.801	0.801
	Density for affected females			
8	<30.....	0.00002	0.00167	0.00167
9	30-39.....	0.00026	0.01276	0.01276
10	40-49.....	0.00112	0.02305	0.02305
11	50-59.....	0.00137	0.01711	0.01711
12	60-69.....	0.00226	0.01260	0.01260
13	70-79.....	0.00218	0.00908	0.00908
14	$\geq$80.....	0.00213	0.00654	0.00654

List of Figures

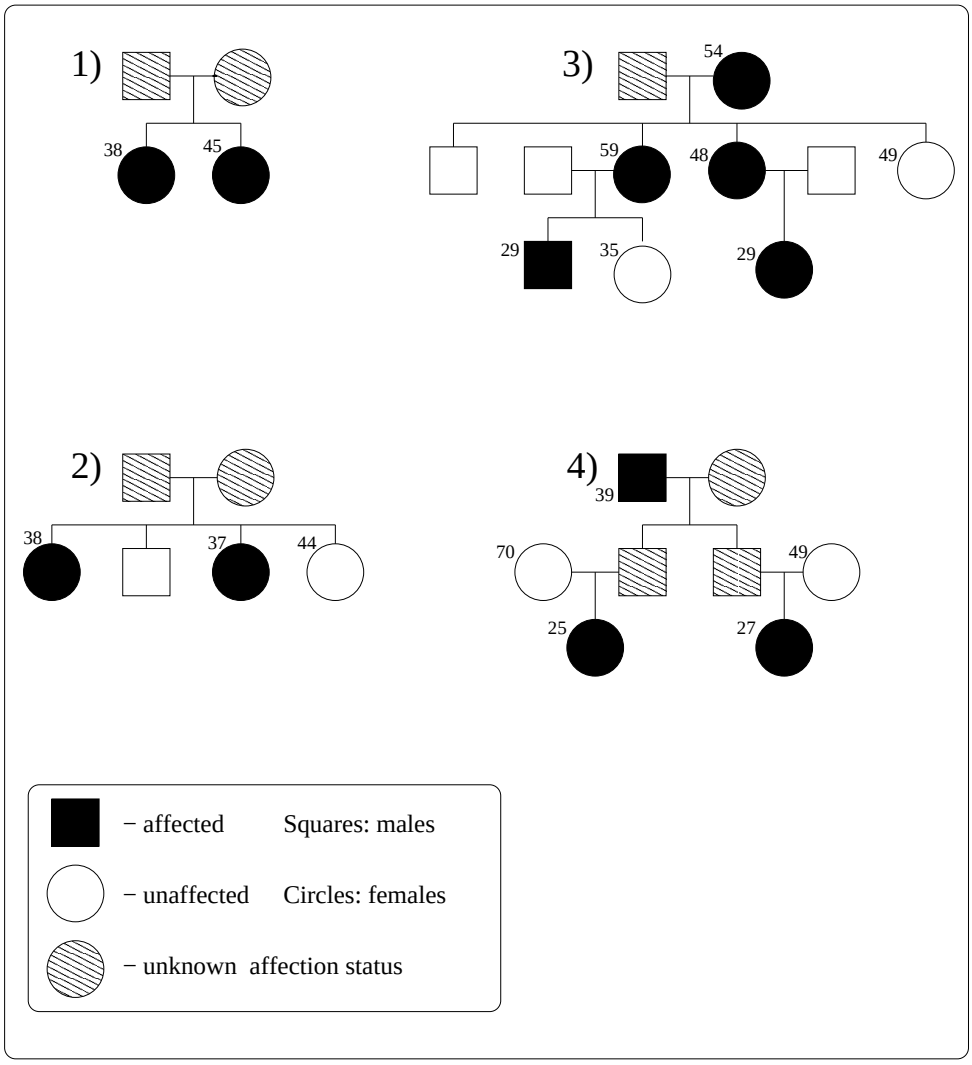


Figure 5: Pedigrees used in the study. Each individual is assigned to one of 14 liability classes, indicated in the figure, depending on age and affection status.

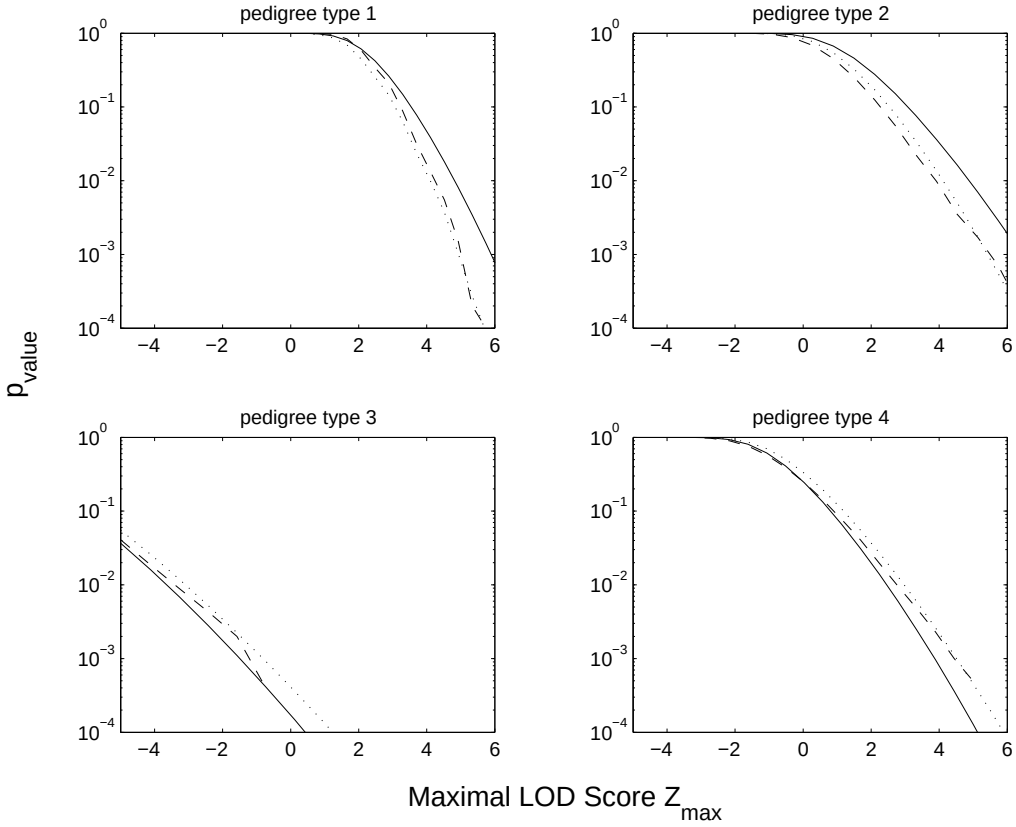


Figure 6: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 1 and 60 families for each pedigree type. Marker data is perfect.

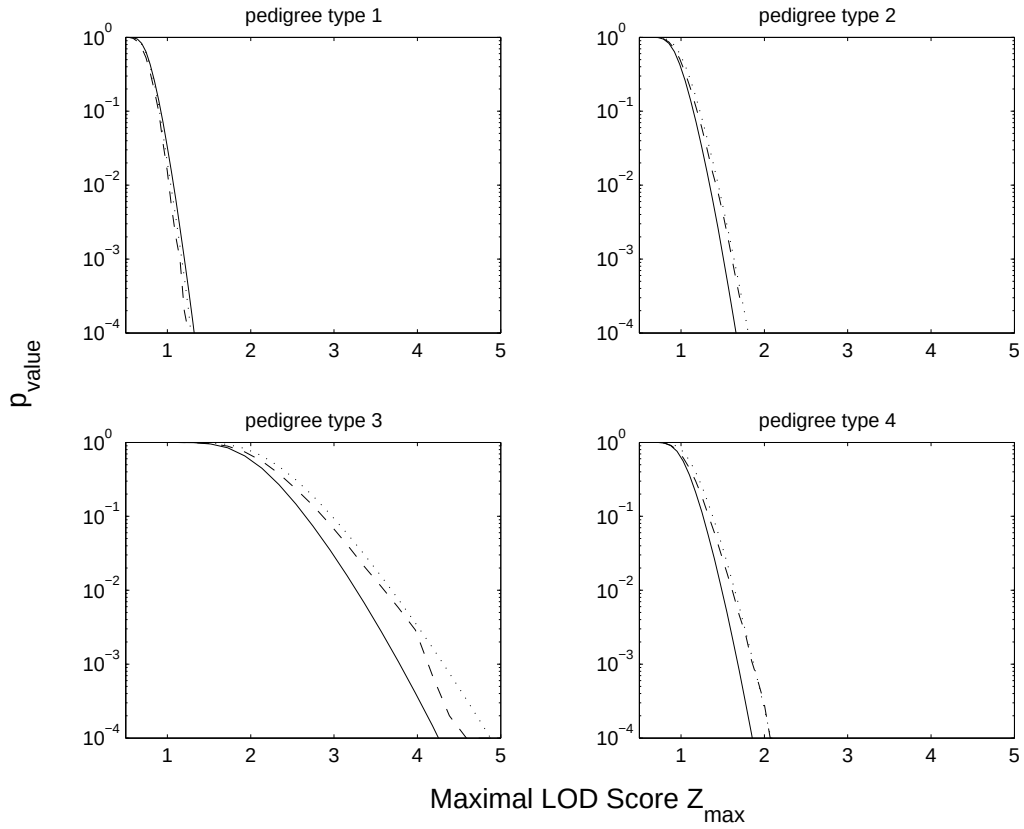


Figure 7: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 2 and 60 families for each pedigree type. Marker data is perfect.

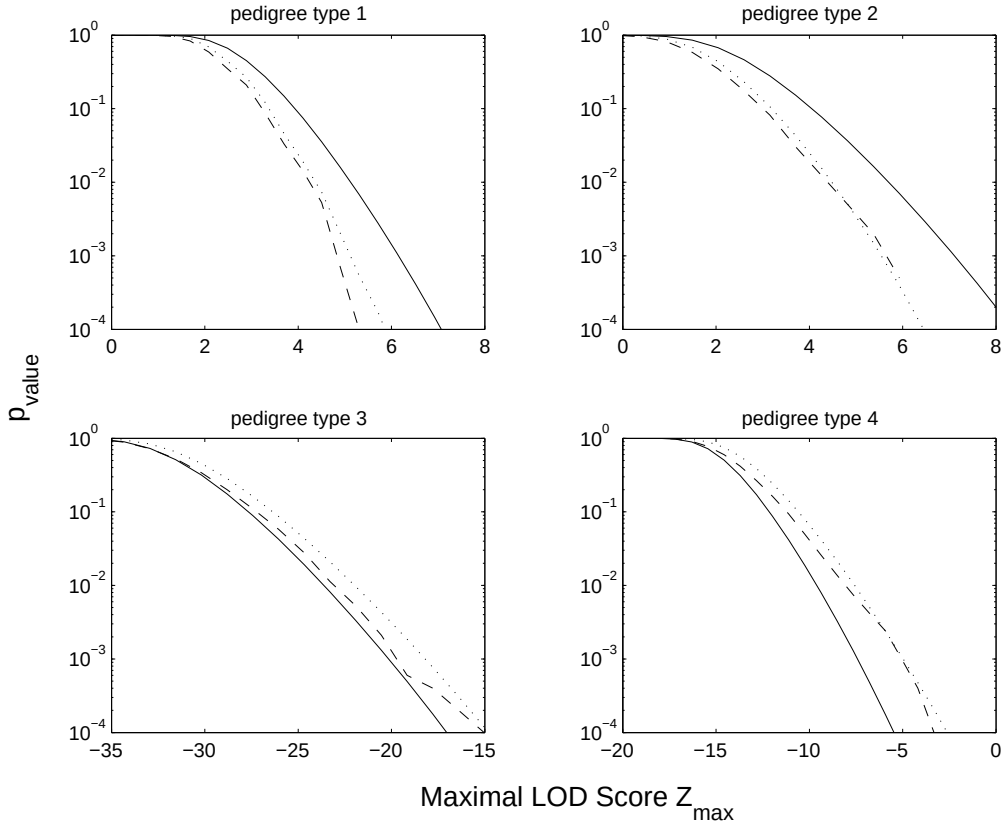


Figure 8: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 3 and 60 families for each pedigree type. Marker data is perfect.

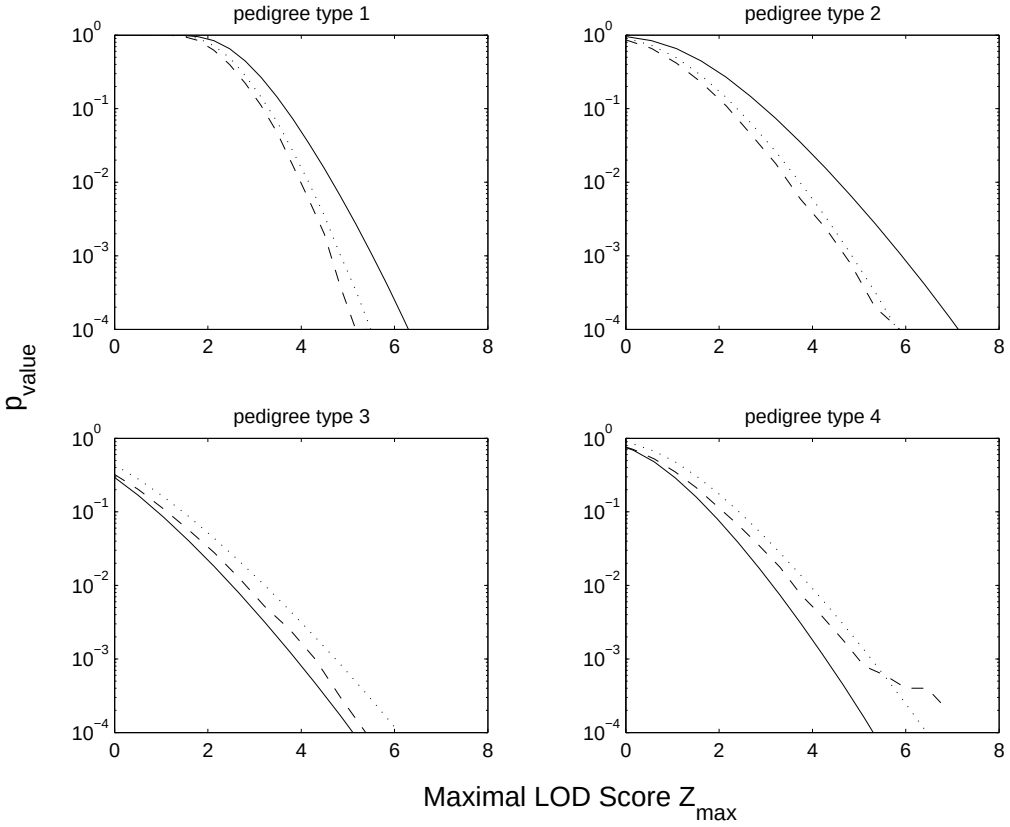


Figure 9: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 4 and 60 families for each pedigree type. Marker data is perfect.

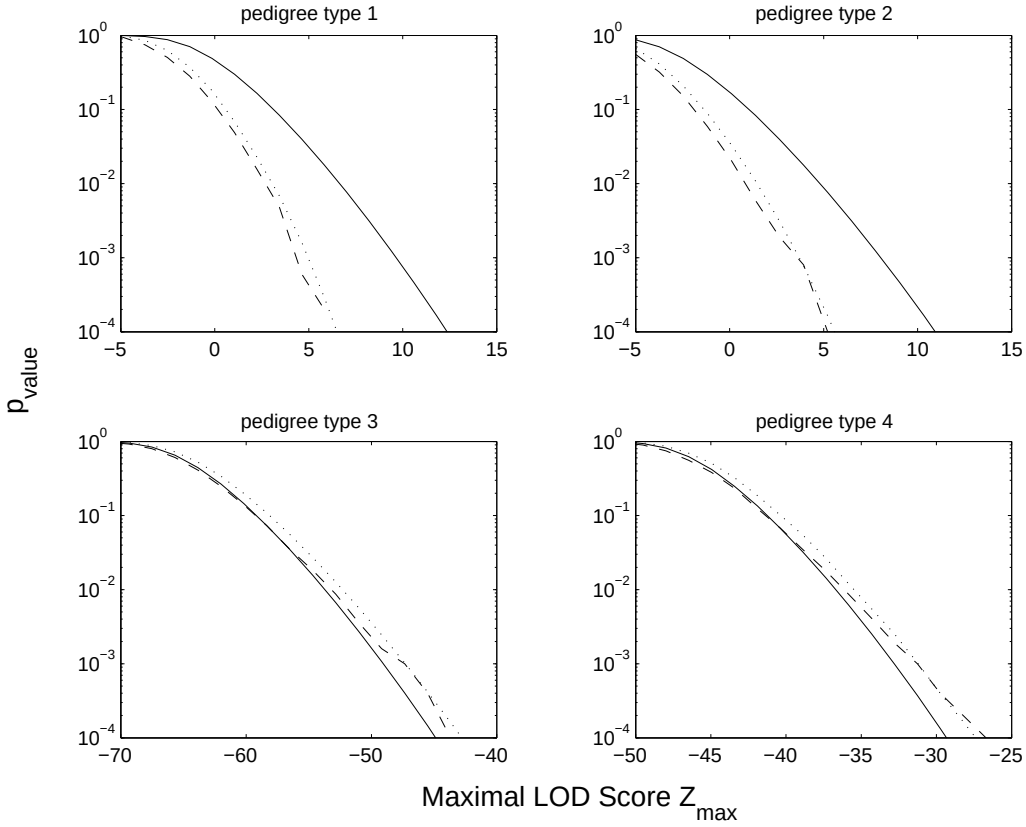


Figure 10: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 5 and 60 families for each pedigree type. Marker data is perfect.

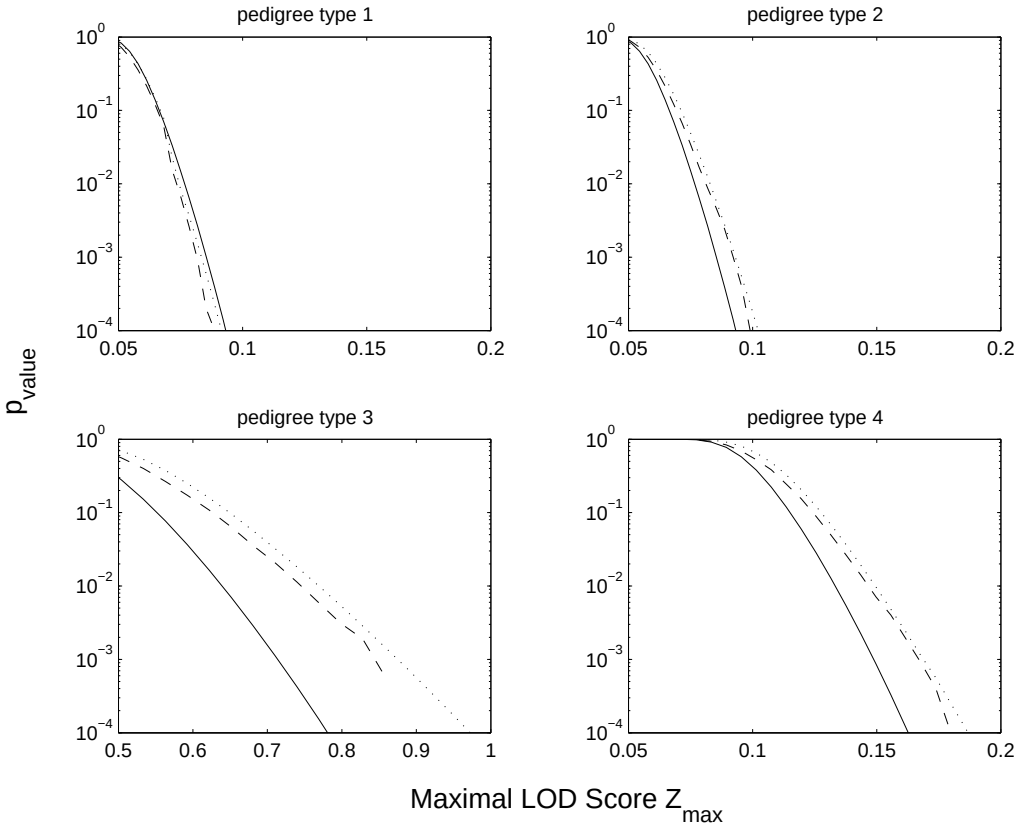


Figure 11: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 6 and 60 families for each pedigree type. Marker data is perfect.

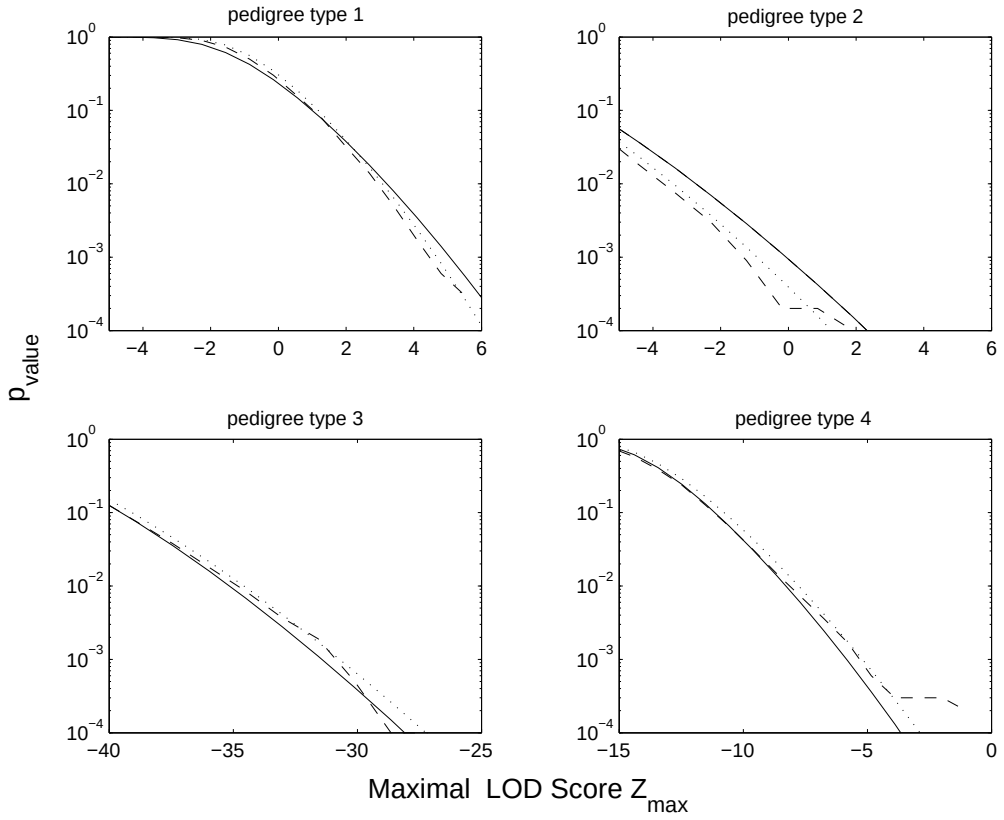


Figure 12: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 1 and 180 families for each pedigree type. Marker data is perfect.

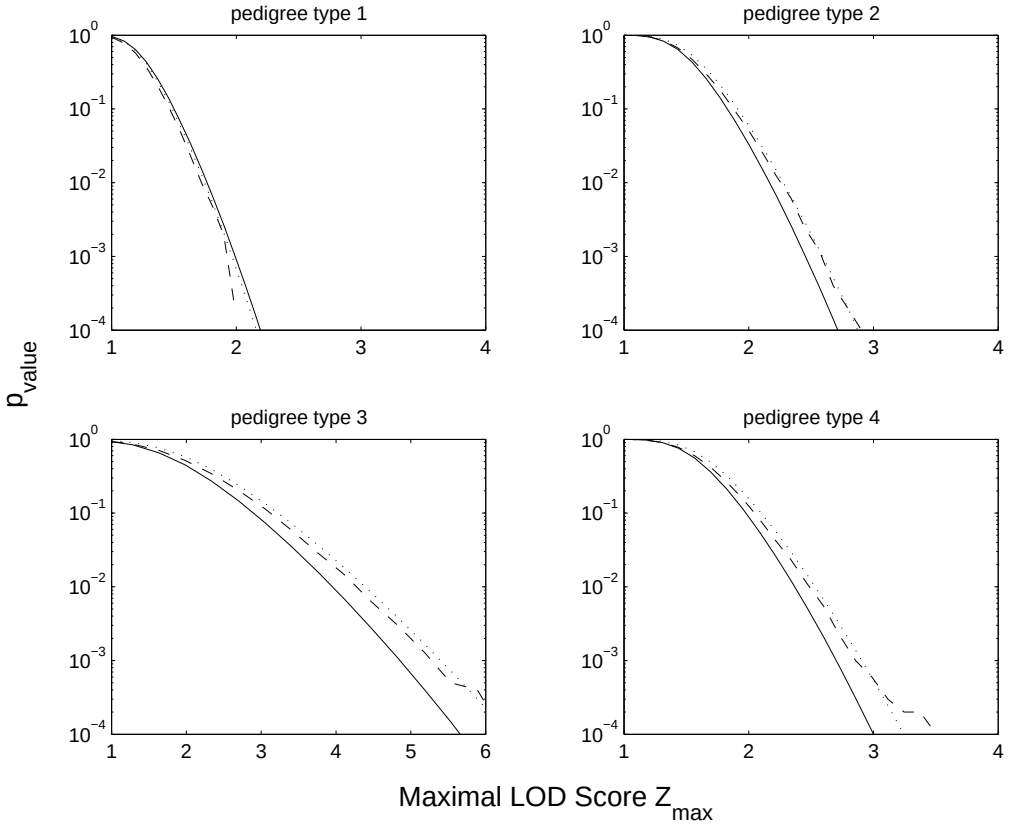


Figure 13: Comparisons between the genomwide p-values for the normal approximation ($-$), the simulation procedure (- -) given by (13) using 10000 replicates and the adjusted normal approximation ($\cdots$) for Model 2 and 180 families for each pedigree type. Marker data is perfect.

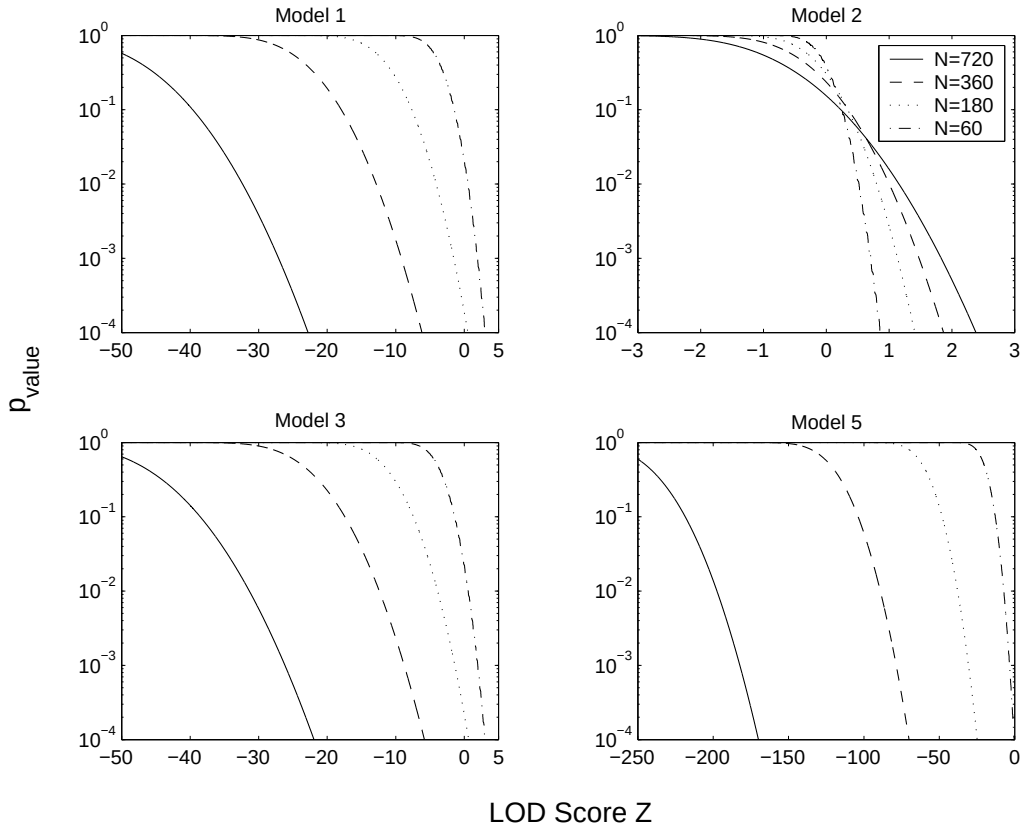


Figure 14: Comparisons between theoretical pointwise p-values for four different genetic models and four different sizes of data N of pedigrees type 1.

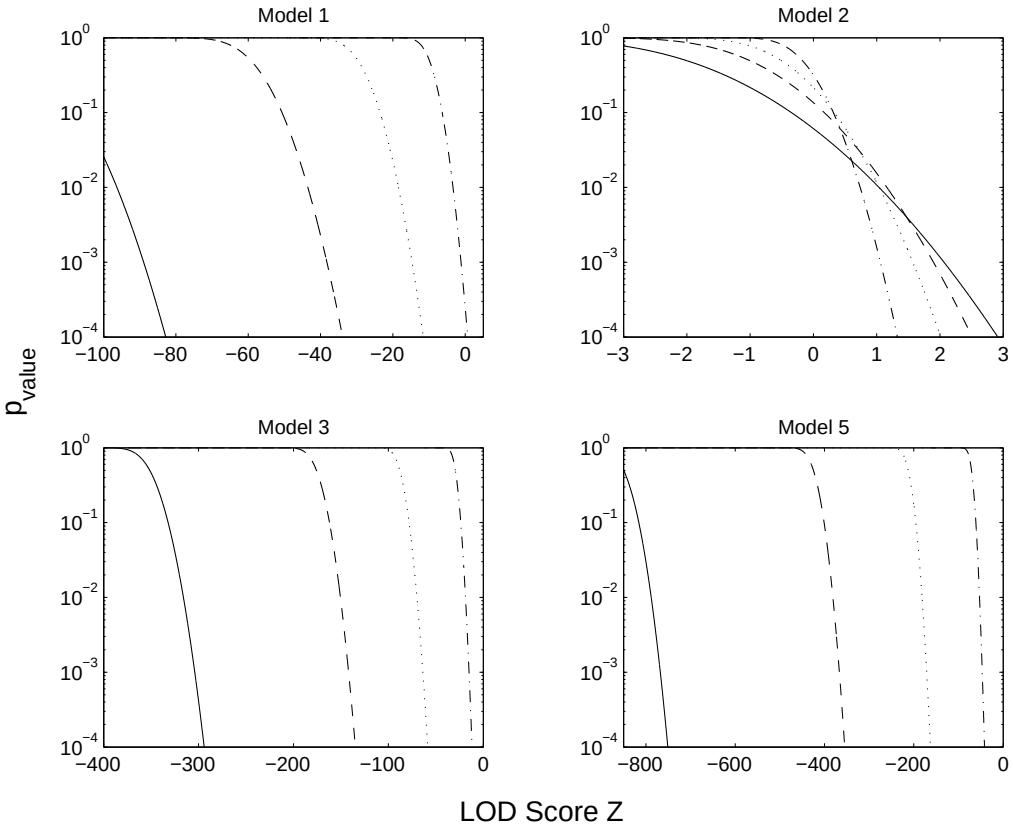


Figure 15: Comparisons between the theoretical pointwise p-values for four different genetic models and four different sizes (N) of data, 720 ($-$), 360 (- -), 180 ($\cdots$) and 60 (-$\cdot$ -), of peidgrees type 4.

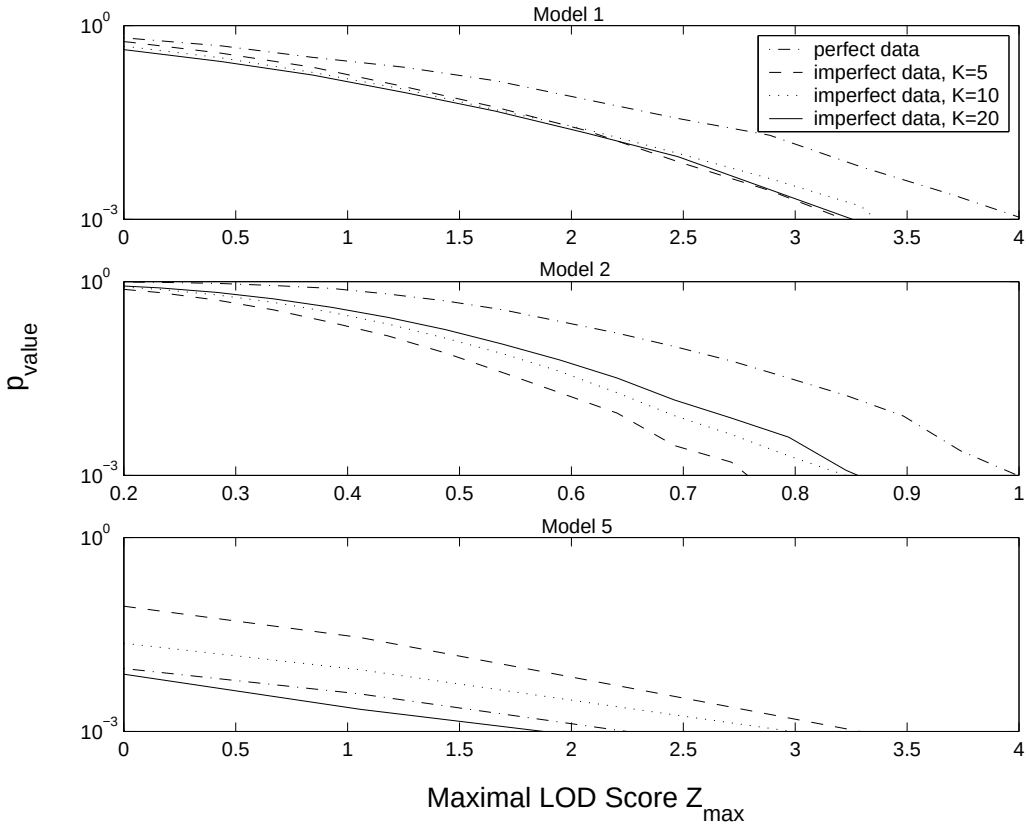


Figure 16: Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) p-values for perfect and imperfect data for 60 families of pedigree type 1 for the simulation procedure given by (13) using 10000 replicates. Simulations are done for three different models. All markers have K possible alleles with equal probability $1/K$. Distance between markers for imperfect marker data is 10 cM. Only nonfounders are genotyped when marker data incomplete.

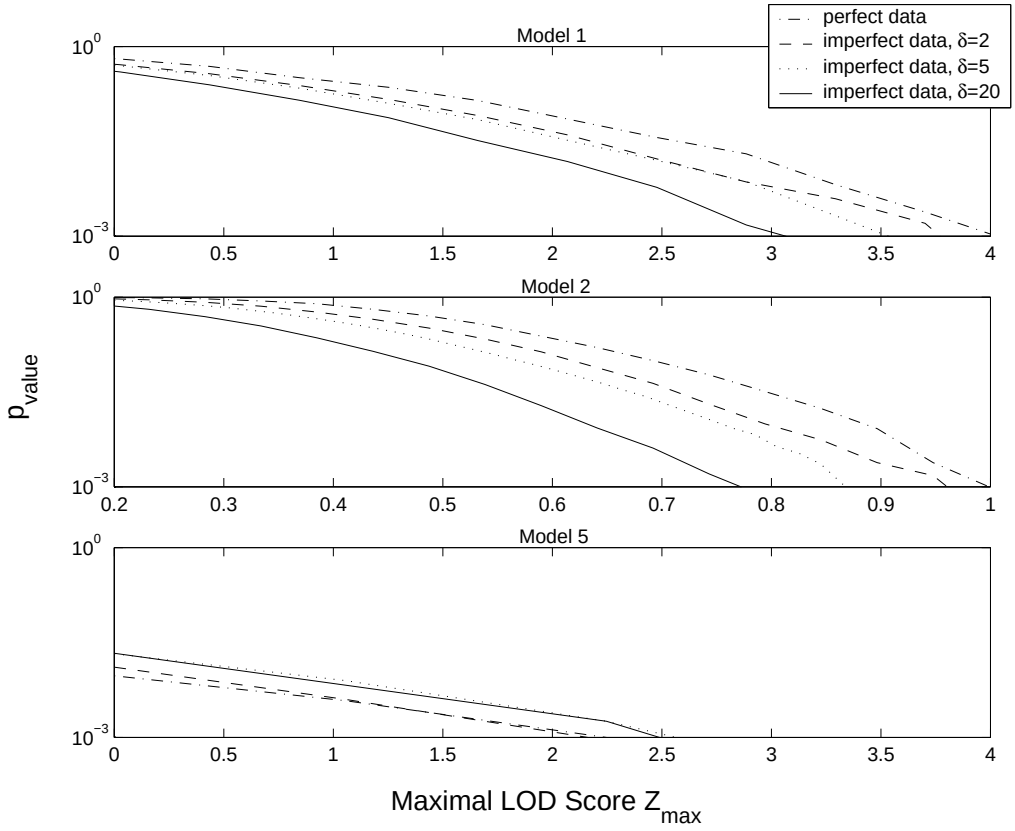


Figure 17: Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) p-values for perfect and imperfect data for 60 families of pedigree type 1 for the simulation procedure given by (13) using 10000 replicates. Simulations are done for three different models. For imperfect data, the markers are equally spaced with distance δ cM and only nonfounders are genotyped.

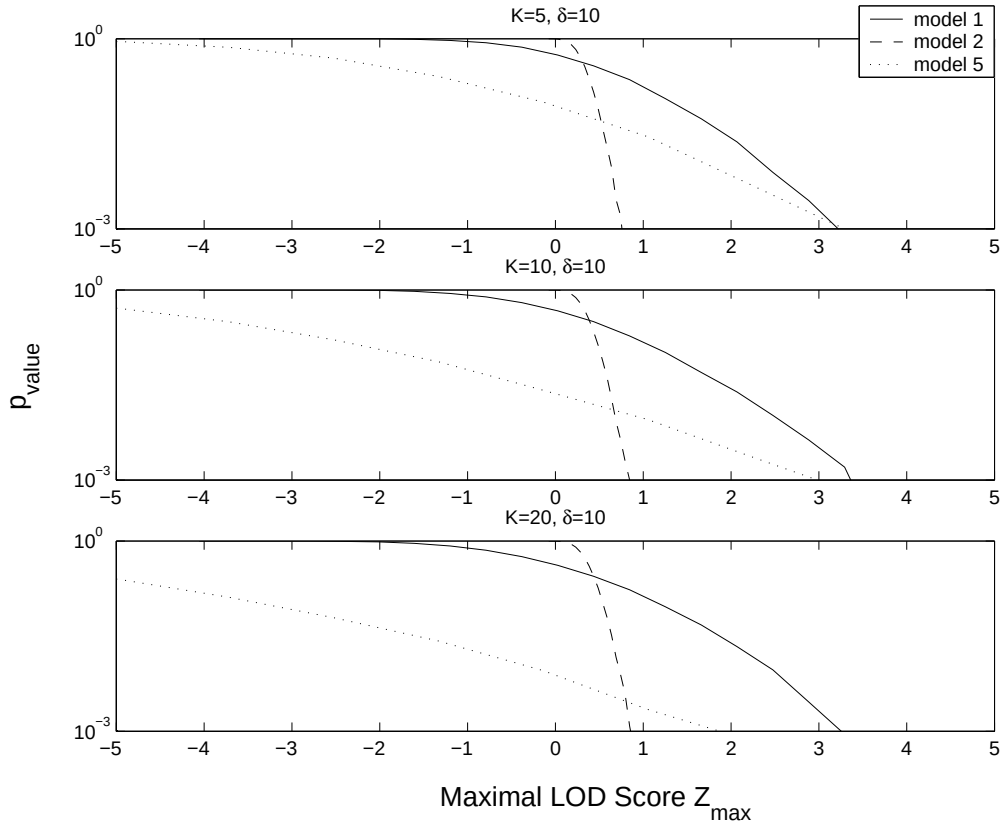


Figure 18: Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) p-values for incomplete data for 60 families of pedigree type 1 for the simulation procedure given by (13) using 10000 replicates. Simulations are done for three different models.All markers have K possible alleles with equal probability $1/K$. Distance between markers for imperfect marker data is 10 cM. Only nonfounders are genotyped when marker data incomplete.

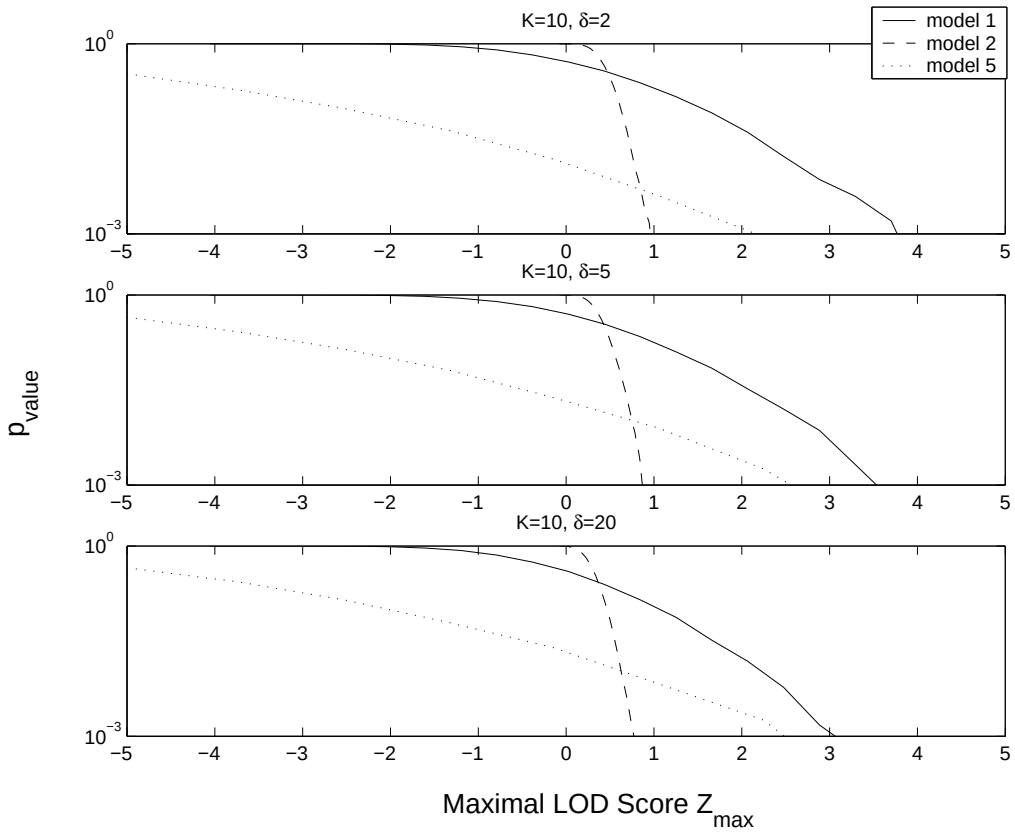


Figure 19: Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) p-values for incomplete data for 60 families of pedigree type 1 for the simulation procedure given by (13) using 10000 replicates. Simulations are done for three different models. For imperfect data, the markers are equally spaced with distance δ cM and only nonfounders are genotyped.